

**Contributor-Centric Analytics For OpenStreetMap:
Approaches to Full Stack, Metadata-Driven Analysis
Infrastructure for an Open Geospatial Data Platform**

by

Jennings Anderson

B.A., Carroll College, 2012

M.S., University of Colorado Boulder, 2017

A thesis submitted to the
Faculty of the Graduate School of the
University of Colorado in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Department of Computer Science

2019

This thesis entitled:
Contributor-Centric Analytics For OpenStreetMap: Approaches to Full Stack, Metadata-Driven
Analysis Infrastructure for an Open Geospatial Data Platform
written by Jennings Anderson
has been approved for the Department of Computer Science

Prof. Leysia Palen

Prof. Kenneth M. Anderson

Prof. Brian C. Keegan

Prof. João Porto De Albuquerque

Prof. Tom Yeh

Date _____

The final copy of this thesis has been examined by the signatories, and we find that both the content and the form meet acceptable presentation standards of scholarly work in the above mentioned discipline.

IRB protocol #17-0402

Anderson, Jennings (Ph.D., Computer Science)

Contributor-Centric Analytics For OpenStreetMap: Approaches to Full Stack, Metadata-Driven
Analysis Infrastructure for an Open Geospatial Data Platform

Thesis directed by Prof. Leysia Palen

OpenStreetMap (OSM), the free and editable map of the world—whose data is consumed by technology platforms, social media users, news media, global disaster responders, and many more—is much more than a simple digital map. With over 1M contributors, OSM is an active online community of hobbyists, humanitarians, professional geographers, and others who grow and curate a massive collection of spatial information. The map itself is a constantly evolving database of billions of points that describe our physical world, often being the most complete or even only source of geographic information for many parts of the world.

The data that can be analyzed is abundant, and yet conducting these analyses is difficult, especially for thorny questions about data quality. Contributor-centric analysis approaches reimagine OSM data analysis beginning with the bottom of the stack to prioritize the metadata about the individual edit which preserves data provenance and allows analysts to interrogate the history of the map’s evolution. These representations enable new scalable data processing workflows that drive improved data visualizations, allowing for more meaningful, contextualized interpretations of the evolution of the map.

This dissertation explores these analytical advantages by viewing OpenStreetMap not as a map, nor simply a geospatial database, but rather as the culmination of edits to hundreds of millions of objects that represent our physical world. I trace my development of OSM data analysis systems across three previous iterations and discuss the subsequent empirical research that each iteration supported. This culminates with the presentation of a fourth analytical framework and data schema capable of capturing the complete editing history and evolution of the map at a global scale.

Acknowledgements

First, and foremost, this dissertation would not be possible without my Advisor, Leysia Palen. Thank you for your continuous support, guidance, and mentorship over the past six years.

I'd also like to thank my committee who helped me think through and shape this work: Ken Anderson, Brian Keegan, João Porto De Albuquerque, and Tom Yeh.

My Project Epic colleagues, now close friends: Melissa Bica, Marina Kogan, and Robert Soden—who introduced me to OSM and then vouched for me. I have learned so much from you all and look forward to seeing where we go from here. Additional thanks to Jim Dykes and Gerard Casas Caez for keeping our computational infrastructure running.

Mikel Maron and my collaborators at Mapbox: Lucas Martinelli, Ramya Ragupathy, Sanjay Bhangar, Sajjad Anwar, Arun Ganesh, Rasagy Sharma, and Jinal Jofia. This would not have been possible without your support, expertise, and Mapbox tools.

My collaborators and coauthors, Seth Fitzsimmons and Dipto Sarkar: I look forward to continuing to work with you. And the research team at HeiGIT: I have learned a lot from your work, past and present, and am excited about our future collaborations.

Diane Fritz and the Maptime Boulder/MileHigh crew: Thanks for being another space to learn what we can do with maps and open data.

Finally, I'd like to thank my family, friends, roommates, past and present, for their encouragement and support over the years.

National Science Foundation (NSF) Grant IIS-1524806 funded this research, with additional computational resources from both Mapbox and the Chameleon testbed supported by the NSF.

Contents

Part I: OpenStreetMap	1
1 Introduction	2
1.1 OpenStreetMap & Volunteered Geographic Information	2
1.2 Contributor-Centric Research	3
1.3 Motivation	6
1.3.1 Data-Centric vs. Metadata-Centric: Making Visible the “Who”	7
1.4 Who Is Editing the Map: OSM Contributors	10
1.4.1 Many Communities	17
1.5 Global Scale	26
1.6 Related Work	27
1.6.1 Heidelberg Institute for GeoInformation Technology	29
1.6.2 OSMesa	30
1.7 Outline of the Dissertation	31
1.7.1 Inclusion of Published Work	32
2 Measuring OpenStreetMap	33
2.1 Defining “edits” to the Map	35
2.2 Minor Versioning of OSM Objects	37
2.2.1 Limitations and Pitfalls of Minor Versioning	40
2.3 Spatiotemporal Scaling: Volume and Velocity	40
Part II: First Ventures into an OpenStreetMap Analysis Infrastructure	43
3 Epic-OSM: A Software Framework for OpenStreetMap Data Analytics	44
3.1 Introduction	44
3.2 OpenStreetMap	48
3.3 epic-osm Framework	52
3.4 Implementation	59
3.5 Use of the Framework	61
3.6 Extensibility and Future Development	63
3.7 Conclusions	64
Epilogue: Epic-OSM Implemented	65

3.8	Time Series Analysis	65
3.9	Sharable Data Visualizations with osmdown	66
3.10	2015 Nepal Earthquake	67
Part III: Transition to Vector Tiles		70
4	Representations of OSM Data	71
4.1	Topologically Lossless OSM Data Formats	72
4.2	Topologically Lossy OSM Data Formats	73
4.2.1	Attribute Lossy Conversions	75
4.3	GeoJSON + OSM	75
4.3.1	Support for Object Histories	76
4.3.2	Creating GeoJSON from OSM	78
4.4	GeoJSON + Vector Tiles	79
5	OpenStreetMap Analysis Vector Tiles	81
5.1	OSM-QA-Tiles	81
5.2	Vector Tile Based OpenStreetMap Data Analysis	83
5.2.1	Implementation Example	84
5.3	Improving OSM-QA-Tiles	87
5.3.1	Simplifying OSM-QA-Tiles by Object	87
5.3.2	QA-Tiles-Plus: Turn-Restrictions	89
5.3.3	OSM-QA-Tiles: Summary	91
5.4	Historical Tile-Based Analysis 1: Annual Snapshots	93
5.5	Historical Tile-Based Analysis 2: Quarterly Snapshots	94
5.6	Tile-Based Analysis Approach 3: Full Historical Vector Tiles	96
5.6.1	Historical OSM Data Schema(s)	96
Part IV: OSM Historical Snapshot Analyses		100
6	Annual Historical Snapshots	102
6.1	Data Processing with Annual Snapshots	102
6.2	State of the Map US 2016 Presentation	105
6.3	Interactive Contributor-Centric Visualizations: First Generation	111
7	The Crowd is the Territory: Assessing Quality in Peer-Produced Spatial Data During Disasters	115
7.1	Introduction	115
7.2	Background	118
7.3	Dataset and Methods	123
7.4	Contributor-based Intrinsic Quality Metrics	128
7.5	Discussion	141
7.6	Conclusion	146
8	Quarterly Historical Snapshots	147
8.1	Improved Resolution with Annual Snapshots	147

8.1.1	Identifying Shadowed Edits Per Tile	149
8.1.2	Building Quarterly Historical Snapshots	151
8.2	State of the Map US 2017: OSM Analysis Dashboard	152
9	Corporate Editors in the Evolving Landscape of OpenStreetMap	158
9.1	Introduction	158
9.2	Materials and Methods	164
9.3	Results	166
9.4	Discussion	177
9.5	Conclusions and Future Research	178
 Part V: Full-History Tile-Based Analysis		 182
10	Full-History Tile-based OSM Data Analysis	183
10.1	Moving to Full-Historical OSM-QA-Tiles	183
10.1.1	Recreating Histories with OSM-Wayback	184
10.1.2	Other Concurrent Development	189
10.2	Full Historical Analysis To Date	192
10.2.1	State of the Map US 2018: OpenStreetMap Data Analysis Workshop	192
10.2.2	Analysis of the US	195
10.3	Full Historical Analysis: Future Work	196
10.3.1	Paid Editing Interactions and Map Seeding	196
10.3.2	Validation Behaviors	197
10.3.3	Scaling to Real-Time	198
10.4	Conclusion	199
10.4.1	Contributions	200
10.4.2	Final Remarks	201
 Bibliography		 203
 Appendix		
A	Glossary	221

Tables

Table

2.1	Classification of types of edits in OSM	36
3.1	Top Tags for OSM objects in New York City.	50
3.2	Differences in use of “school” tag	58
9.1	Known Corporate editing teams active in OSM	167
9.2	Corporate editing totals per year	173

Figures

Figure

1.1	Daily edit count and number of contributors mapping Nepal vs. the United States	5
1.2	Density of mapped objects in OSM	8
1.3	Density of editing activity broken down by number of active mappers in 2019	11
1.4	Daily number of contributors and edits to OSM	12
1.5	Date of first and most recent edit for all OSM Contributors	14
1.6	Lifespans of OSM Contributors who first mapped in 2015 (Nepal Earthquake)	16
1.7	Rise in humanitarian mapping efforts	20
1.8	Identifying Localized Editing in the Map	23
1.9	The rise of Corporate Editing in OSM	25
1.10	Google Scholar results for OSM Data Analysis	27
1.11	Missing Maps Leaderboard, powered by OSMesa	30
2.1	Building and intersection objects in OSM	34
2.2	Progressive changes in the geometry of a building in OSM	37
3.1	Port-Au-Prince, Haiti in OSM before/after	46
3.2	OSM Elements rendered on openstreetmap.org	50
3.3	The Domain Objects of Epic-OSM	54
3.4	The run-time objects of epic-osm.	55
3.5	Count of OSM Changesets and Users after Nepal Earthquake	61
3.6	Graphical Comparison of 2010 Haiti Earthquake and 2013 Typhoon Yolanda Mapping Response	65
3.7	Screenshot of the osmdown webpage for the 2015 Nepal Earthquake	67
3.8	Public engagement with our Nepal Earthquake Stats page	68
4.1	OSM XML Format	72
4.2	GeoJSON representation of The Taj Mahal OSM object	76
5.1	Creating OSM-QA-Tiles	81
5.2	Boulder, CO OSM Data (OSM-QA-Tile)	82
5.3	Tile-Reduce Spatial Data Analysis Workflow	83
5.4	Heatmap of Corporate Editing Activity	86
5.5	Duplication of metadata across tiles	88
5.6	Turn-Restrictions as Points in Panama City, Panama	91
5.7	OSM-QA-Tiles for London in 2007 and Today	93
5.8	Shadowed Edits in 2012	94
5.9	Quarterly-Snapshots of Boulder, CO	95

5.10	OSM @history attribute	97
6.1	Annual historical analysis workflow	102
6.2	Annual SQL query example	104
6.3	Active editors in the US per week	105
6.4	Cumulative growth of users active per US city	106
6.5	Cumulative count of buildings edited per city	107
6.6	Screenshot of road co-editing networks	108
6.7	Screenshot of building co-editing networks	110
6.8	Screenshot of per-editor annual editing summaries	111
6.9	Screenshot of Editors by country visualization (annually)	112
6.10	Screenshot of Edits/Contributors Graph for France	113
6.11	List of users active on a tile in France	114
7.1	Road Network in OSM	124
7.2	Editing OSM on openstreetmap.org	125
7.3	The 4 Study Tiles for Intrinsic Quality Analysis	127
7.4	Density of unique contributors by tile over time	130
7.5	Users active each year on the study tiles	131
7.6	Features tagged as building=collapsed in Port Au Prince.	132
7.7	Editing Preferences among OSM contributors	135
7.8	Building / Road Percentages vs Experience	136
7.9	Stages of Growth in OSM	139
8.1	Daily Edit Count Discrepancies in Snapshots	148
8.2	Screenshots of Shadowed edits over time	150
8.3	Tooltip of interactive map of shadowed edits	151
8.4	OSM Analysis Dashboard Comparisons	153
8.5	Tile-based analysis with enhanced resolution	154
8.6	Example of lower-zoom Aggregation	155
8.7	Edits per quarter in San Francisco	156
9.1	OSM Contributors Engagement	160
9.2	OSM and the 90-9-1 Rule	162
9.3	Where corporate editors are editing	172
9.4	Corporate editing percentages per object type	174
9.5	Characteristics of Corporate editors	176
10.1	osm-wayback Processing Pipeline	184
10.2	osm-wayback Processing Pipeline: Geometry	186
10.3	Amazon Aurora Example: Roads edited in Pokhara	191

Part I

OpenStreetMap

Chapter 1

Introduction

1.1 OpenStreetMap & Volunteered Geographic Information

This work studies the OpenStreetMap project (OSM), the largest volunteered geographic information (VGI) project in existence. Often called the "Wikipedia of Maps," OSM was created in 2004, before Google Maps was dominant, and amid the rising success of the Wikipedia project [24]. Though the term Volunteered Geographic Information (VGI) was officially coined after OSM was created, OSM is the most successful instance of a VGI project in terms of number of contributors and the amount of data produced [42]. OSM defined a new type of open data with the creation of the ODbL, a specific open license for databases that declares the data may be downloaded by anyone for any purpose, commercial included. The only requirement is that attribution be given to the original contributors, specifically seen as "© OpenStreetMap Contributors" in the corner of any derivative product involving OSM data, such as a rendered map.¹ OSM is both an open geographic database and an online community of millions of contributors. In the past five years, the number of registered users has grown from just under 2M to over 5M and the subset of these users who have actually edited the map has surpassed 1M [140]. These *mappers* are involved in the project for a variety of motivations ranging from open data enthusiasts to professionals [18]. I use the terms *mapper*, *editor*, and *contributor* interchangeably to describe a registered OSM user who has edited the map at least once, meaning they they show up in the recorded editing history. The term *registered user* may refer to any of the 5M+ users who have an account on openstreetmap.org,

¹ All statistics, figures, and maps rendered within this work are produced with Data © OpenStreetMap Contributors.

but may not have yet actually edited the map. Having a registered account is the only requirement to edit the map; there are no anonymous edits. A *consumer* is any person, service, or company that uses OSM data. Facebook, Instagram, Craigslist, Apple Maps, Snapchat, and by association all of the users of these platforms, are consumers to some degree.

Contributions to OSM are both difficult to quantify and highly unequal among all editors. As an online community, OSM is not immune to the participation inequality common in these groups [11]. The community adheres to the 90-9-1 rule where the majority of the registered users have made few to no contributions, some of the registered users have dabbled, and a tiny percentage of the users do the majority of the work [85]. *The crux of the work presented here is to view OSM not just as a map, but as the collaborative product of the more than 1M contributors producing billions of edits all over the globe.* Understanding the map then requires deeper understanding of the mechanisms of the data production and the interactions between the mappers and the platform as well as between the mappers themselves. Though this participation is greatly unequal, each of these individual contributions needs to be accounted for in our measurements of the activity to tell the complete story. To achieve this, I present new research approaches to prioritize a holistic understanding of evolution of the map, both a collaborative process and a cumulative product: *Contributor-Centric Research.*

1.2 Contributor-Centric Research

In the context of OSM, I define contributor-centric research approaches as data-driven analysis methods that prioritize the metadata about the edit to the map itself, not just the resulting object on the map. Technically speaking, the “map” is a geospatial database and an “edit” is then any change to an object in the database. However, due to the OSM data model, the relationship between edits to the database and the object on the map is not one-to-one. Between this, the large variation in the types of edits one can perform, and the drastic differences in the quantity of edits between users, measuring “edits to the map” is an overly-complicated task. For OSM, I find this is best done with a layer of abstraction above the raw OSM data that reveals the interaction between the

mappers and the database entries to define the change in a more measurable way. This requires first reconstructing the raw editing record (the history of the database) into an observable and computationally-efficient format. The work presented here documents and traces the evolution of this thinking through multiple iterations of analysis systems that have brought me to these conclusions.

In the context of crowd-sourcing and information quality, the term “contributor-centric” has been used to describe validating the resulting data against new value systems. Whereas a product is typically valued by its stakeholders (the data consumers), contributor-centric information quality metrics embrace the data provenance, valuing the mode of production and the values held by the data-producers [106]. Extrapolating beyond information quality metrics, contributor-centric approaches to data-analysis is similar: not looking at the end-product (the contribution), but rather seeking to understand the activity of the contribution itself, as performed by the contributor. As Chapter 7 will discuss in greater depth, knowing the larger context surrounding each contribution itself is paramount to understanding its impact to the map [5].

Furthermore, *who* is editing the map is becoming a more salient question as the number of active contributors grows. OSM has been described as a “community of communities,” a phrase that highlights this is not simply a peer-production platform, but instead that there exist a number of different communities with varying levels of influence, different value systems, and numerous goals associated with contributing to and supporting OSM [127]. As a global project, OSM must therefore cater to the needs of each of these communities, as it is no longer simply a collection of individuals seeking to create a free and open map of the world. In this way, OSM may have started as a crowd-sourcing project, but when examined closely, it has evolved past this label as the term “crowd” invokes a lack of unity and familiarity with one another. Today it appears the OSM community is much more connected than a crowd and self-identifies themselves as just that: Community. Chapter 9 will explore this further, showing that each of these communities imposes their own value systems, rendering any single assessment of the map, the community, and their evolution contested or flawed from all but that single perspective.

In relation to a data-centric architecture, the term contributor-centric is meant to more descriptively label the technical systems as data-centric architecture that prioritizes the metadata pertaining to the actions performed by OSM contributors. That is, this is not describing an infrastructure that is just data-centric, but rather *metadata-centric*. This involves designing around more complex questions such as “which users edited the buildings in this region, how many, and when?” Instead of “how many buildings are in this region?” This enables data-driven content analysis of the OSM editing record that specifically highlights the interaction between contributors and the map, not just how the data within the map evolved. This idea is a core conceptual contribution that drives the innovations described here. As an example, Figure 1.1 shows two dramatically different editing histories between the map of Nepal and the US in terms of number of contributors, the edits they made, when. These were calculated from the editing metadata that allows us to then identify and explain the high impact editing events for each region.

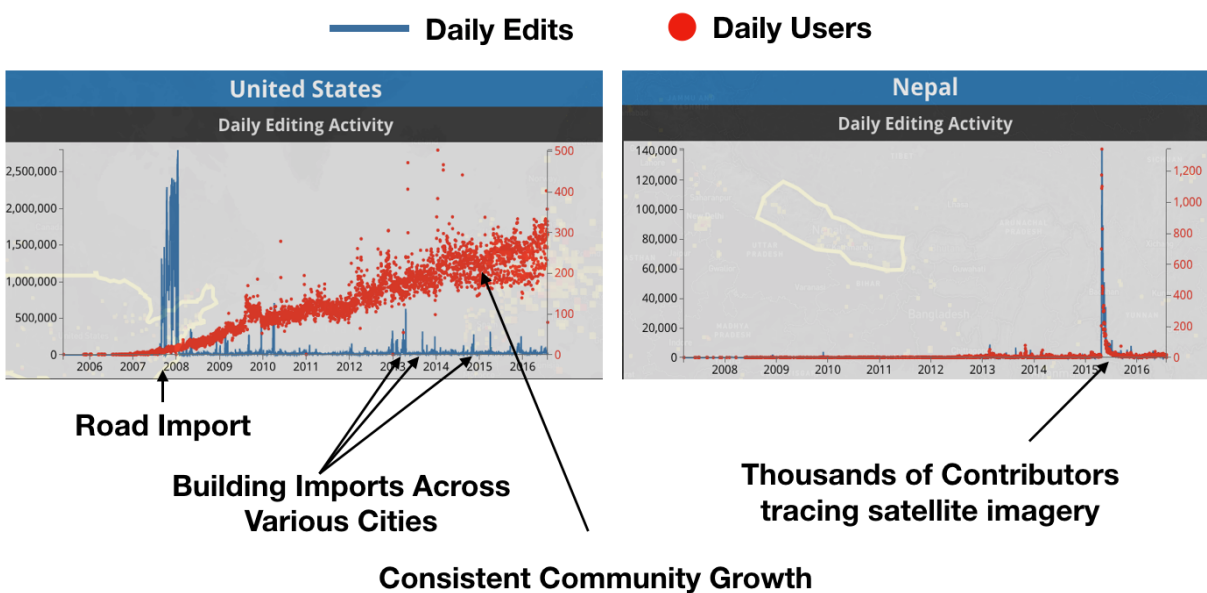


Figure 1.1: Number of daily edits versus mappers actively editing the maps of United States and Nepal. Annotations show two wholly different editing histories between the two Countries. The United States has seen a number of data imports and a consistently growing community of mappers. Nepal has a much smaller active community but saw over 1,200 contributors active daily following the 2015 earthquake. These are metadata-driven analyses of maps of these two Countries.

1.3 Motivation

As a participant observer in the OpenStreetMap community for the past five years, I have seen the number of registered users increase by more than 3M people and the number of active daily editors nearly double from 2,700 contributors to 4,800. As the community continues to grow, so does the desire to understand the nature of contributions and the evolution of the map itself. As described above, the data produced is being indirectly consumed by millions of people. Now being used by major companies, OSM is no longer a small, UK-based project: It is part of a major industry with its data underlying a great number of maps, digital and print. For many parts of the world, OSM is the most accessible, accurate, and sometimes only source of digital geographic information [125]. This makes OSM the primary base map for many humanitarian activities as well as a decision-making tool in time- and safety-critical situations, such as disaster response [105, 125].

This latter use-case is what originally brought me into the OSM research domain, looking to better understand the collaboration among the rapidly-converging volunteers in the wake of a disaster [105, 125]. At the time, it quickly became clear that existing tools for working with OSM data prioritized the spatial and physical attributes of the map, not the context and activity surrounding how the data was produced. While all of the editing history of the map is made available alongside a variety of low-level data-processing tools to work with the data, accurately and meaningfully measuring the editing activity from these datasets remained an unsolved problem [4].

More specifically, there did not exist a simple interface to ask questions such as, "how many users edited in this area at this time?" While others had asked these questions before, everyone had their own approaches to wrangling and processing the OSM historical editing record. While a trivially simple interface to ask this question still does not yet exist for the whole planet, there are now many tools and utilities that allow analysts to answer this question in a multitude of ways for a variety of regions around the globe. More importantly, this seemingly simple question can, and should, be broken down further: "How many users have ever edited this part of the map?" "How many users continue to edit?" "How many users have edited which types of objects?" "How

many of these *users* were really bots?” “How many of these users made more than one edit or edited more than one time?” and so on. The minute differences in each of these questions impacts our understanding of the creation of the map. As a constantly evolving project, these nuances are critical to capture in our data analyses.

While the one-size-fits-all analysis dashboard for OSM that I set out to build five years ago still has not been built in full, the infrastructure to support such an endeavour now does exist, and furthermore, we as a research community have a better understanding of why such one-size-fits-all approaches fail, and we know how to restructure our approaches to explore the more nuanced and impacting questions we should have been asking in the first place.

1.3.1 Data-Centric vs. Metadata-Centric: Making Visible the “Who”

Since the distribution of work among contributors is incredibly unequal, visualizing the number of edits across the map tells a different story than visualizing the number of editors actively editing the map. These two attributes are certainly related in many areas, as more active mappers often results in more edits. However, there are many parts of the map where these numbers are skewed by a few power mappers² performing a large number of edits or many mappers performing relatively fewer edits, such as the case with new mappers attending a mapathon or disaster mapping. These many mappers performing fewer edits are at risk of going unaccounted when looking at the number of edits alone; their work is just a drop in the proverbial bucket of millions of edits. Additionally, not all edits are equal in impact (societal or otherwise), and should not be compared as such. A road being digitized in a rural area for the first time by a new mapper who is simultaneously being exposed to the very concept of open-data and OSM is distinctly different in nature from an edit performed by a long-time contributor modifying an existing road in an urban area for which high quality geospatial data may have existed for a number of years. While an extreme example, the notion stands that not all edits are equal and rank differently within different value systems present

² Mappers who are responsible for the vast majority of the edits to the map. Exact quantity and percentages vary, but there are always a relative few who outperform all other editors locally, regionally, and globally.

in the OSM community.

Making even just one edit to the map means that a contributor has learned about OSM, taken the initiative to make an account, and learned the basics of editing the map. This person is now not only aware of OSM, but the larger concept of VGI. This is a core tenant behind projects like *YouthMappers* with the tagline “We don’t just build maps, we build mappers” [103]. With thousands of students engaged in over 150 chapters around the world, YouthMappers uses OpenStreetMap to introduce students to open data and to “define their world by mapping it” [103]. This group is just one example of the many smaller communities that make up OSM. Each of these communities has their own motivations and intents when it comes to participating in OSM, each of which comes with its own larger context that influences the evolution of the project [3]. While many mappers engage little with the data itself in terms of their editing footprint if measured by number of edits to the database, their existence alone in the database and larger impact within the community is non-trivial and therefore data analysis systems need to account for each of these activities to tell the complete story.

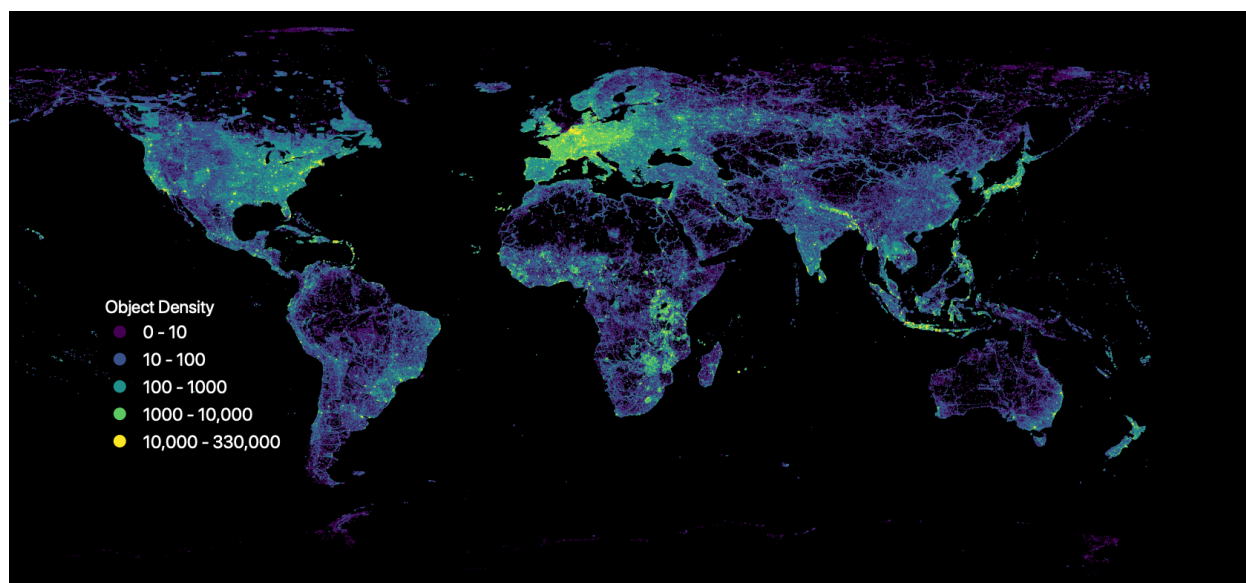


Figure 1.2: Density of objects in OSM: Objects per zoom level 12 tile. The Log10-scale highlights the incredible unequal distribution of map objects globally. Purple and blue sections of the map are 100 and 1,000 times less dense in terms of coverage than green and yellow regions. Data © OpenStreetMap Contributors 2017.

Figure 1.2 presents a data-centric answer to the question: “Where has the world been mapped in OSM?” While there are some obvious correlations with population density and urban/rural areas, this rendering highlights the Euro-centric nature of OSM. While a completely equitable map does not require the whole world be mapped as densely as Europe (there are certainly variations in urban densities around the world), this map is nonetheless much more purple and blue (i.e. less dense) in many areas than it should be if the entire planet were equitably mapped. This rendering is intended to highlight where it appears the world has been mapped more completely than others.³ In terms of completeness, the purple and blue parts of this serve to highlight areas that are *incomplete*. As Pascal Neis identified in 2016, there are many identifiable “unmapped” places in OSM.⁴ Areas on this map that appear purple and blue represent regions the size of a small city with less than 10 or 100 mapped objects. In many cases, these few objects are the names of towns or cities, perhaps with a road leading to it. These points really act as a placeholder claiming that there is something there, it just has not yet been mapped. This particular rendering is a double-edged sword in that way: At least there is *something* mapped in these areas and the map is not simply blank, but on further inspection, this area is still entirely incomplete on the map.

Breaking this down further, Figure 1.3 depicts the current state of the map with a more contributor-centric approach: showing the density of editing activity separated by the number of mappers actively doing this work. The incredible editing densities of Europe and other major cities around the globe match Figure 1.2 in terms of where the most editing continues to happen by the most users (at least 10 mappers active per zoom level 12 tile).⁵ Also highlighted in Figure 1.3a are the heavily edited areas of Southeast Asia and Sub-Saharan Africa by both corporate and humanitarian mappers [3]. In contrast, Figure 1.3b, shows that while there have been fewer than ten mappers actively editing across most of the world, there is truly global editing activity, albeit in smaller quantities. It should be noted that Figures 1.3a and 1.3b are mutually exclusive in the

³ Completeness is just one measure of spatial data quality, Chapter 7 explores measuring geospatial data quality.

⁴ Blog post available at neis-one.org/2016/06/unmapped-places-osm/

⁵ Zoom level 12 tiles are a common unit of analysis in my work. Chapter 5 will present them in greater detail, but on average they represent the area of a small city

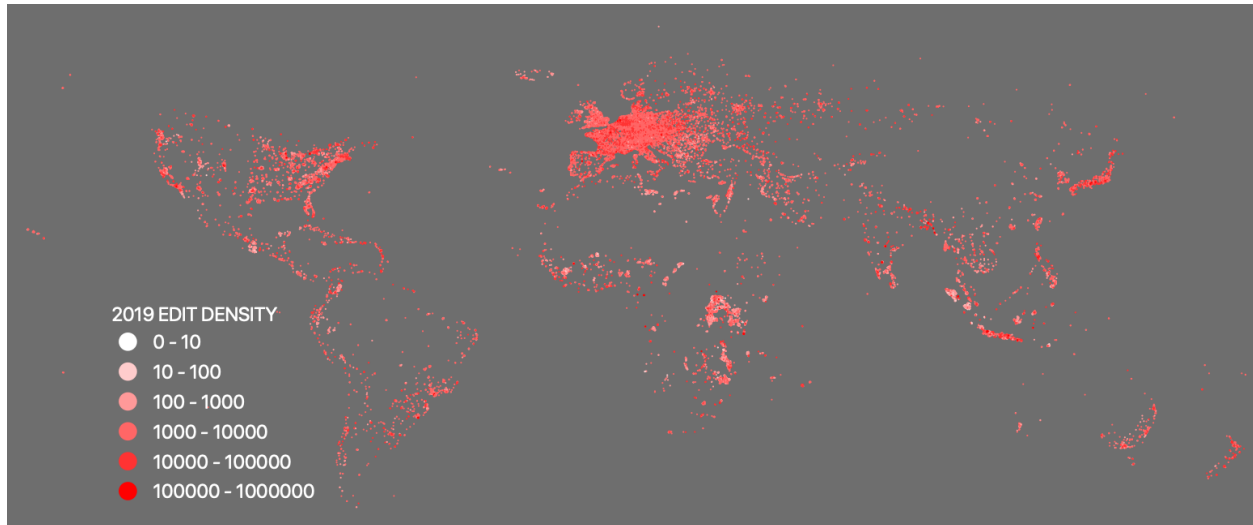
areas they represent in that each tile is only represented once between the two figures. The major takeaway is then: There is mapping happening across most of the world. The rate of mapping and the number of contributors varies drastically, but where there are not many active mappers, there are usually at least a few. The less-mapped regions are often drowned out in volume by the heavily mapped regions in terms of absolute activity, but so far in 2019, most of the planet has seen some level of mapping activity, even if it is just one mapper making a single edit in a large area: It is still progress towards completeness. Again, this is not to say the whole world has been mapped, just that there is mapping occurring at a variety of levels across much of the world.

In this manner, contributor-centric approaches to OSM data analysis can provide a more complete view into how, where, and when people are editing the map than other data-centric analysis that interrogate the more spatial qualities of the map such as how much data exists and where. These are the differences exposed between Figures 1.2 and 1.3, or more succinctly, the differences between “data-centric” and “contributor-centric” measures. Together, both types of inquiry tell the complete story about the development of the map. There are, however, different design choices that need to be made early in the analysis process to ensure the data being collected, organized, and analyzed can answer the appropriate questions in their entirety. *By prioritizing the editing metadata over the edit itself when designing and implementing data analysis systems, we can make visible the person and the activity that goes into building and maintaining the map.*

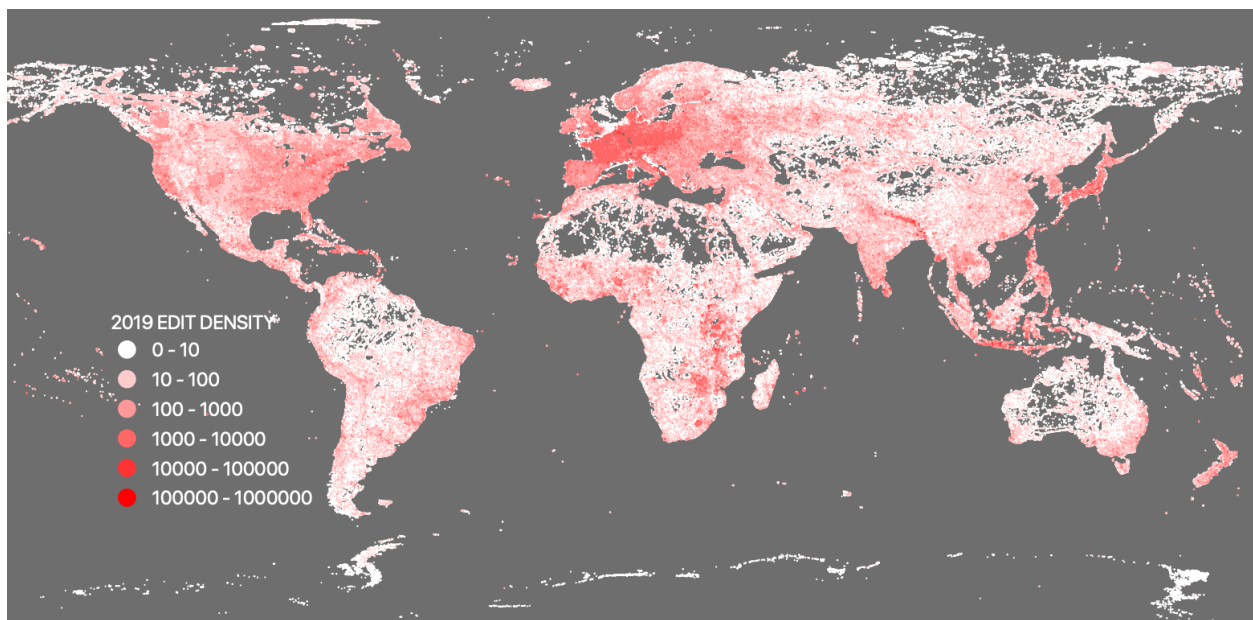
1.4 Who Is Editing the Map: OSM Contributors

Today, OSM has over 1M active contributors, each editing the map for a variety of motivations [18, 140]. These 1M+ contributors represent a small portion of the 5M+ registered users on the platform, with both of these numbers continuing to steadily grow [86]. In this section I present a series of figures and interpretations that highlight the nuances of OSM contributor activity including the growth, lifespans, and associations of the contributors.

Figure 1.4 shows the general editing activity in terms of both changes to the map and the number of contributors active each day. Both of these values continue to increase consistently,



(a) Edits performed in first half of 2019 where more than 10 mappers have been active



(b) Edits performed in first half of 2019 where 10 or fewer mappers were active.

Figure 1.3: Density of editing activity around the world for first half of 2019, separated by number of active mappers in an area. Number of mappers and editing density computed for zoom level 12 tiles (about size of small city).

with the number of contributors growing faster.⁶ The burst-like activities likely correspond to

⁶ These overview statistics are calculated from the record of all changesets that counts the number of number of OSM elements affected in each changeset. As Chapter 2 discusses, this is related, but not equivalent to the actual number of changes happening to objects visible on the map.

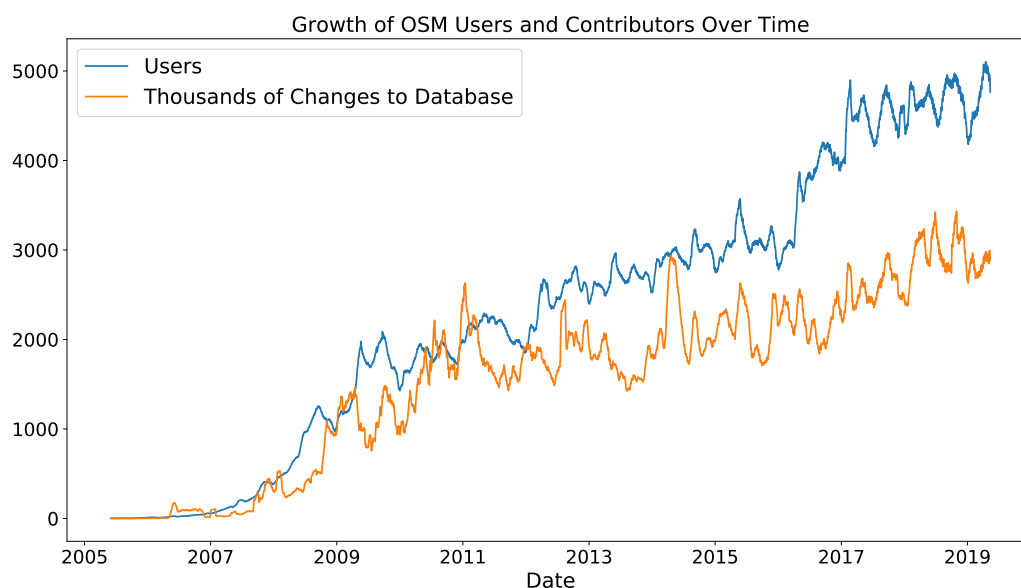


Figure 1.4: Number of active contributors and changes to the database happening daily, filtering out bot and import accounts. Results averaged over 30-day rolling window for readability at this scale.

specific organized mapping events such as humanitarian activations and mapathons. These create a pulse-like contribution pattern that appears to be becoming more intense in recent years with respect to the number of contributors active each day.

Figure 1.4 gives a more truthful account of the editing activity and community size than reporting on the database statistics alone: “5M+ registered users” is an impressive statistic in terms of community growth, but is misleading if discussing the number of contributors to the map. To this end, a more accurate report would be stating that less than 0.1% of the registered users edit the map daily.⁷ This is not to say that the community is not active or engaged with the map, it merely highlights the complexity of measuring engagement in an online community such as OSM.

Not accounted for here, however, are the other ways in which members participate in the community, such as organizing mapping events, introducing new people to OSM, being active on the wiki or mailing list, etc. Quantifying these engagements is not possible from data analysis of

⁷ Calculated from 2019 average of 4800 daily users and over 5M registered users

the database alone, but instead requires more qualitative investigation and observation of all of the community spaces. Calculating and reporting on these numbers is out-of-scope of this particular work, but these types of engagements are extremely important to the project and community as a whole and should be acknowledged as such. Another important nuance not represented in Figure 1.4 is the extreme level of inequality common in online platforms, in this case between power-mappers and other contributors: While more than 1.2M users have edited the map at least once, less than one-third of those users have made more than 100 changes. Furthermore, more than 900k previously active editors have not returned as of 2018. Figure 9.1 presents a visualization of this in Chapter 9.

One metric that can help elucidate the inequality of engagement is the time-span or life-span of a contributor. Bégin et al. first visualized this by plotting the date of each contributor's first edit to the map against their most recent edit [11]. This creates a visualization with time on both axis where the diagonal represents 1-time contributors: Where someone's first and last edits were on the same day. Figure 1.5 uses the approach proposed by Bégin et al. in [11], but with the additional visual channel of color to denote the total number of days a contributor has been active on OSM (not necessarily continuously). The addition of color allows us to distinguish between contributors who edit the map often and therefore have a continually growing lifespan and contributors who are active sporadically.

The majority of purple points across the top of Figure 1.5 indicate that users who have actively edited OSM on more than 21 days are still actively mapping regardless of when they started. This might suggest there is a saturation number of days that gets a mapper hooked that is somewhere around three weeks worth of mapping days.⁸

58% of all mappers exist along the diagonal in Figure 1.5; these users only edited the map one day. Furthermore, 84% of mappers fall into the <5 days of mapping activity category. Also observable in Figure 1.5 are the higher amounts of orange and red close to the diagonal. This represents a large number of users who were active for a number of days in a row, but then did

⁸ This is an oversimplification that does not take into account any larger contexts around the types of mapping one does during this time. Dittus et al. identify a number of more salient factors than number of active days when measuring contributor engagement during humanitarian mapping events [29].

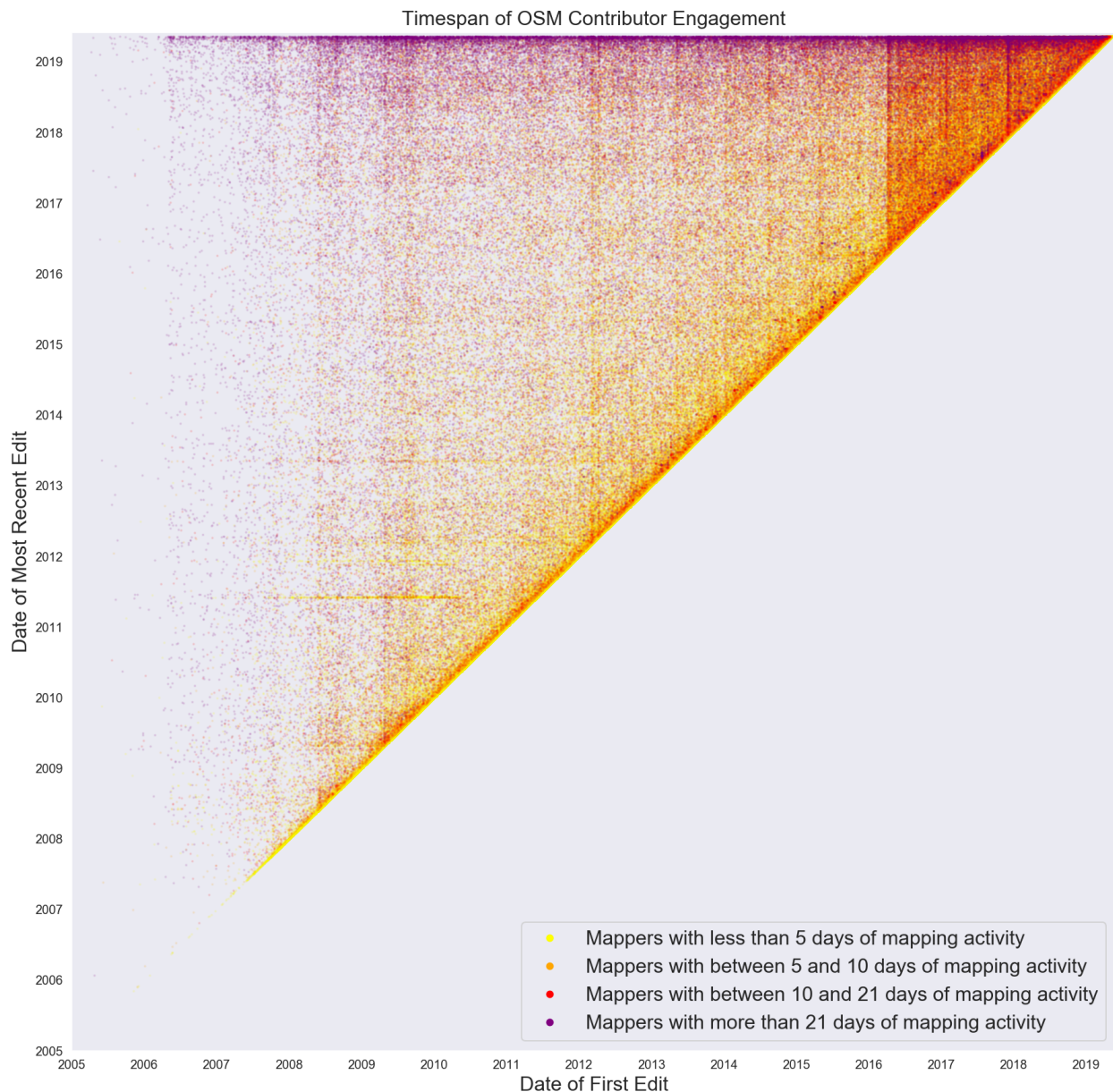


Figure 1.5: Date of first and most recent edit for all OSM contributors (1.3M as of May 2019). Color denotes number of days a mapper has been active between those two dates (inclusive). Created using approach from [11].

not return to the map. These are most likely new disaster mappers participating in humanitarian mapathons as described by [125, 105, 32].

The darker, denser triangle in the upper right-hand corner of Figure 1.5 represents a major increase in new contributors as of April 2016. Preliminary investigation into these contributions

shows they are likely associated with the addition of an in-map editor to the popular mobile mapping app, [maps.me](#)⁹. The addition of this in-app editor seemingly lowered the barrier to editing the map for many of the app users. Of particular interest is that this increase is not confined to the diagonal, but rather creates a vertical line, suggesting that many of these users continued to contribute for days, weeks, months, or years. The engagement and attrition rates of mobile mappers is therefore distinctly different from others. The subtle vertical and horizontal lines present in Figure 1.5 represent specific points in OSM's history such as the license change or the first mentions of OSM in mainstream media as identified by [Begin et al. \[11\]](#).

Figure 1.6 presents a slight variation on Figure 1.5 that highlights the response to the 2015 Nepal Earthquake. The larger blob of red on the lower diagonal immediately following the earthquake represents the large number of users who, while only ever active for the week or two following the event, contributed hundreds of changesets to the map. This slight variation of using color to represent total changesets instead of active days exposes a different pattern in contributor behavior, highlighting the short lived power mapping activity following a disaster event. There is a subtle, yet more dense section of this scatter plot that follows the vertical line representing the earthquake on April 25, 2015. These dots represent users who first mapped in response to the earthquake, and while the majority only mapped for a few days and then never returned, many users continued to map in the years since the event. [Dittus et al.](#) investigate this phenomena in much more detail, but many mappers are introduced to OSM through disaster mapping and continue to stay involved with the project in the years to come, as evidenced by this slightly denser portion of the scatter plot representing those users who first mapped after the 2015 Nepal Earthquake [125, 105, 29].

⁹ Based on the version release dates available at wiki.openstreetmap.org/wiki/MAPS.ME#History

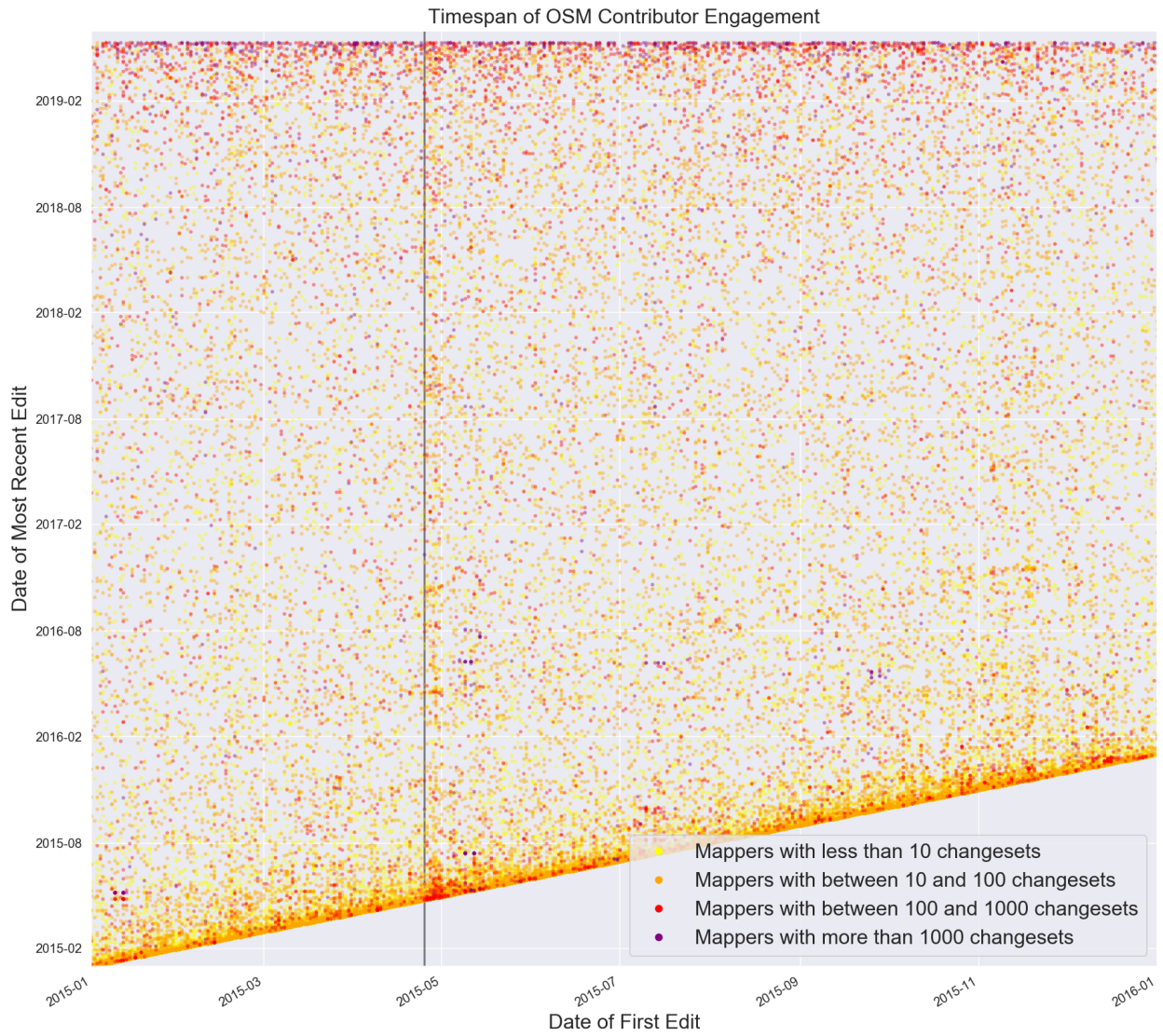


Figure 1.6: Lifespans of OSM Contributors who first mapped in 2015. Vertical line marks April 25, 2015, the day of the Nepal Earthquake.

1.4.1 Many Communities

Borrowing again from Solis’ presentation at the 2016 AAG annual meeting, I agree that OSM is best described today as a “community of communities” in that the project is made up of a number of smaller communities, each involved for their own reasons [127]. This distinction allows the map to be assessed from the context and perspective of any of those involved. While this may inevitably create tensions between these many active communities, it highlights that there are multiple perspectives, depending on the modes of data production and mapper involvement, as discussed more in Chapter 9.

An important caveat in this description, however, is that these communities are not mutually exclusive. Furthermore, defining rigid boundaries between these sub-communities likely creates more issues than it resolves. One distinction that has become more defined in recent years is that between “organized” and thereby the default “non-organized” editing efforts [3]. Broadly defined, organized editing refers to contributors engaged in a mapping effort that has a defined objective or goal that offers organization around said effort. Paid editing, humanitarian editing, and community import-efforts all fall into this category. As of Fall 2018, the OSM Foundation has released the *Organized Editing Guidelines* to provide a list of best-practices that organized mapping efforts should take to ensure openness, transparency, and engagement with the rest of the community [3].¹⁰

Since the boundaries of OSM communities could likely be defined ad infinitum, I will only identify a few of the larger communities within OSM who have particular, observable editing behaviors, while acknowledging that this is far from a comprehensive classification of all OSM contributors and that many mappers may participate in multiple communities, even moving between them over time. For example, a lot of mappers are introduced to OSM through humanitarian mapping efforts, thereby beginning as part of an organized editing effort, learning how to edit the map within a humanitarian context. If this mapper then continues to be an OSM contributor by mapping their hometown, then they have transitioned from a humanitarian mapper to becoming

¹⁰ wiki.osmfoundation.org/wiki/Organised_Editing_Guidelines

part of their local mapping community. Taking this further, perhaps it is solely a hobby, or perhaps it is done in conjunction with some personal gain, such as publishing and selling maps of important tourist locations? While a trivial example, the purpose of such speculations is to remind that OSM data can be and is used by anyone for any purpose. Defining these communities is helpful to understand the project at a whole, but imposing rigid definitions and binding individual contributors within each is ultimately unhelpful and simply divisive as the map and the communities continue to evolve [3].

Further, many of the top humanitarian mappers are also general power contributors to the map. Kogan et al. found that some contributors do not change their mapping workflows and practices when a disaster happens: simply where they choose to map, such as a German mapper who mapped every day after work [61]. When the 2010 Haiti earthquake happened, this reserved time for mapping became reserved humanitarian mapping time because he chose to shift his geographic focus to Haiti during his daily mapping efforts [61]. In this way, this mapper belonged to both his local mapping community and the humanitarian community. Any metrics that discredit this mapper from either of these communities have failed to adequately capture the full-story and context of the mapping activity. Next I will briefly describe a few of the major communities that can be found in OSM.

1.4.1.1 Humanitarian Mappers

Humanitarian mapping in OSM was first popularized in the days after the 2010 Haiti earthquake [125]. In the years since, an organization known as the Humanitarian OpenStreetMap Team (HOT) formalized this type of disaster response and continues grow as perhaps the largest community within OSM [105].¹¹ As highlighted in mailing list discussions during the December 2018 OSM foundation board election, HOT is a major community player with its own formal

¹¹ In terms of organized editing, this is the largest single community. However as a whole, local and/or craft mappers might have more by numbers, depending on where the (arbitrary) line is drawn. This simply highlights the pettiness of such measurements and my desire to stay away from such quantification and let the data speak for itself.

structure and a lot of joint membership between the OSM foundation and HOT leadership.¹² While being the largest active humanitarian mapping group, HOT could also be considered the largest recruiter of contributors to OSM. Driving these efforts are other groups that associate with HOT such as youth mappers (youthmappers.org) or missing maps (missingmaps.org). These organizations bring new mappers into the humanitarian mapping community and primarily organize their mapping activities in conjunction with HOT. In other words, they typically use the official HOT tasking manager (tasks.hotosm.org) to organize their mapping efforts and associate their edits with HOT by using the #hotosm hashtag in the changeset comments. Searching for this hashtag is a common way to identify edits done by humanitarian mapping efforts. This is not to say that all humanitarian mapping in OSM includes this hashtag, but certainly the majority of map edits performed in this context do. HOT also makes significant investment in training new mappers by sponsoring and directing users towards tools such as learnosm, a step-by-step guide to mapping in OSM (learnosm.org).

Figure 1.7 highlights the growth of the HOT community within OSM. This figure was calculated by identifying the changesets with “hotosm” in the comment text. Broadly, these would be considered part of an organized-editing effort. Of most notable importance in Figure 1.7 is the divergence in the two lines since 2014. While all editing activity continues to increase as a whole, humanitarian related mapping appears to be responsible for the majority of the growth in total editing activity.¹³ Because of the potential volatility of such statements, I am deliberately showing “changesets” and not “mappers” here. Classifying mappers as exclusively HOT-associated or not is a misleading calculation: Is there a minimum percentage of a user’s edits that need to be associated with HOT tasks in order to classify this user as a HOT contributor? More than 82k mappers were active only for one-day as a humanitarian mapper and never returned. On one hand, the work of

¹² Public archives available at lists.openstreetmap.org/pipermail/osmf-talk/. This has some of the community concerned about over-representation of HOT within the governing OSM foundation. It should also be noted that HOT is not the only group doing humanitarian mapping. While they are certainly the largest, they do not have an exclusive claim (nor make such a claim) on all humanitarian mapping. That said, HOT is certainly the largest active humanitarian mapping group in the OSM ecosystem [125, 105, 3].

¹³ Unfortunately when it comes to measuring these activities, it turns into: “HOT” vs. “non-HOT” or “Organized” vs. “Non-Organized” where the latter categories do not represent a unified group, but rather simply “other.” This leads to unfortunate “subgroup X” vs. “The Community” comparisons which does not seem right.

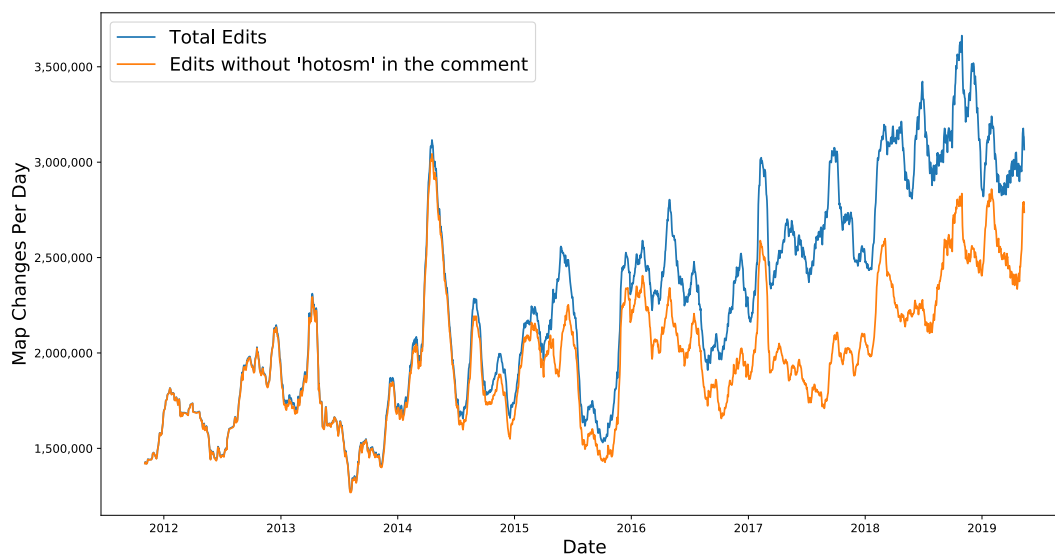


Figure 1.7: The rise in humanitarian mapping efforts over time: On average, the majority of the growth in the map is related to humanitarian (organized) editing efforts.

these mappers needs to be recognized as I argued earlier that all edits need to be counted. On the other hand, claiming this particular community has gained 82k mappers is overstated because these mappers were never active again. Most worrisome to me as an analyst is the propensity for these overly-simplistic measures to be easily “weaponized” on a volatile mailing list to further agendas.

While it is accurate to say that HOT-related mapping activity is a dominant force in both the map and community today, there is always more to be uncovered. Dittus et al. explored the engagement of HOT mappers and the retention of mappers across multiple tasks, finding among many other things, that prior or related experience with OSM had a significant increase on retention rates [29]. Findings like these add uncertainty to the trends in Figure 1.7. If the most prolific HOT mappers are also active in other communities, then it is unfair and biased to make claims that HOT-related mapping is specifically responsible for the growth of the OSM community as a whole. I thereby include Figure 1.7 here with caution and to prompt more critical thought about such measures. This graphic would wreak havoc on the volatile OSM mailing lists and are unfortunately

relatively simple to generate from the OSM-changeset database.¹⁴

One of HOT's largest contributions to OSM is the continued sponsorship of various analytical tools. With efforts in gamification, tracking, and analysis of edits performed within HOT tasks, the HOT developer community (anyone who contribute to projects in the hotosm ecosystem) builds and maintains a massive number of open source utilities. These tools will be discussed in greater detail in coming sections.

1.4.1.2 Craft / Hobby Mappers & Localized Mapping Communities

Though relatively easy to name, The community of craft and hobby mappers is exceedingly difficult to define and identify because mappers may be associated with this group for any variety of reasons, making identifying them in the data particularly hard. For the purposes of this classification, this group is perhaps best defined as anyone mapping non-organized, or perhaps, self-organized. Though, at what point does an editor or small group of self-organized editors become an organized group, editing for a specific purpose?¹⁵

Active local mapping communities take ownership of the map in a particular region and once the map is filled in, continue to edit and update the map as the world changes in real time. I use the term “localized” over “local” specifically to represent a particular location, not necessarily a mapper's home. There are many mappers who maintain and watch over a particular location on the map where they do not currently reside. For example, the user ‘chachafish’ is a self-described nomad who happens to maintain the map of Denver, Colorado even though does not currently live there. Whenever the map of Denver is edited, this user will inspect the edits and often comment on the changeset (alerting the contributor) if anything was mapped incorrectly.¹⁶

¹⁴ December 2018 (and seemingly every election period) sees these debates. As a researcher in this space, I tread lightly and err on the side of not producing overly-simplistic figures and measures that could be co-opted in these debates. I find these issues further inspire the search for better analytics to uncover the inherent difficulties and better tell the complete story.

¹⁵ Rhetorical question intended to remind reader that defining communities is helpful to better understand OSM as a community of communities, but rigid, strict, mutually exclusive definitions are ultimately counter-productive.

¹⁶ Self-described in his OSM profile: openstreetmap.org/user/chachafish. Recently, however, this user has been relying on external tools to cross-validate street names in the Denver area that are currently at-odds with the official city of Denver databases. This raises questions about the authority of localized editing and who should have the final word (osmcoloradoimport.info).

Since the actual home location of a contributor is not publicly available, there is no method to identify a user’s “local” mapping area from the public editing record. This has been tried, however, many times in various studies involving OSM. Popular approaches to determining the home location range from simply the country of a user’s first edit to more complex calculations such as the geometric median of a user’s edits [84]. For localized mapping communities, these approaches will likely be successful: these users tend to edit primarily in a region that is likely to be their home, or close to. The popularity of disaster mapping, however, makes these approaches more difficult and I have empirically observed fewer of these studies in recent years as many mappers are introduced to OSM through mapping regions thousands of miles from their home.

Tensions over the 2018 OSM Foundation board election expressed on the OSM mailing list suggest that some in the community see “craft mappers” as the original maintainers of the map and therefore today’s rightful owners and stewards. With this comes a value judgment that other forms of mapping are less important or meaningful. I mention this only to show that there is a perceived difference between mappers in this group and other groups I will describe. Attempting to measure or investigate these ideas further will only legitimize such divisive ideas. As a researcher in this space, I am especially cautious about appearing to validate or support these types of arguments through my work.

In general, these types of editors are largely difficult to quantify and summarize because the definition of these local communities can be so broad. For example, how many edits would one need to make in their “hometown” to be considered a local editor? I briefly explored defining this threshold as a percentage of a mapper’s annual edits, as shown in Figure 1.8. Empirically we do observe that many mappers have specific areas of interest (often even multiple regions) that receive a significant portion of their total mapping activity, but calling this a mappers’ home or area of localized knowledge is still purely speculative. Humanitarian mappers, for example, will still always appear to be local the largest event of the year with utilities like that shown in Figure 1.8.

Instead, I have observed that the best indicators of local mapping communities are not found in the map data itself but through other channels: A number of Facebook communities,

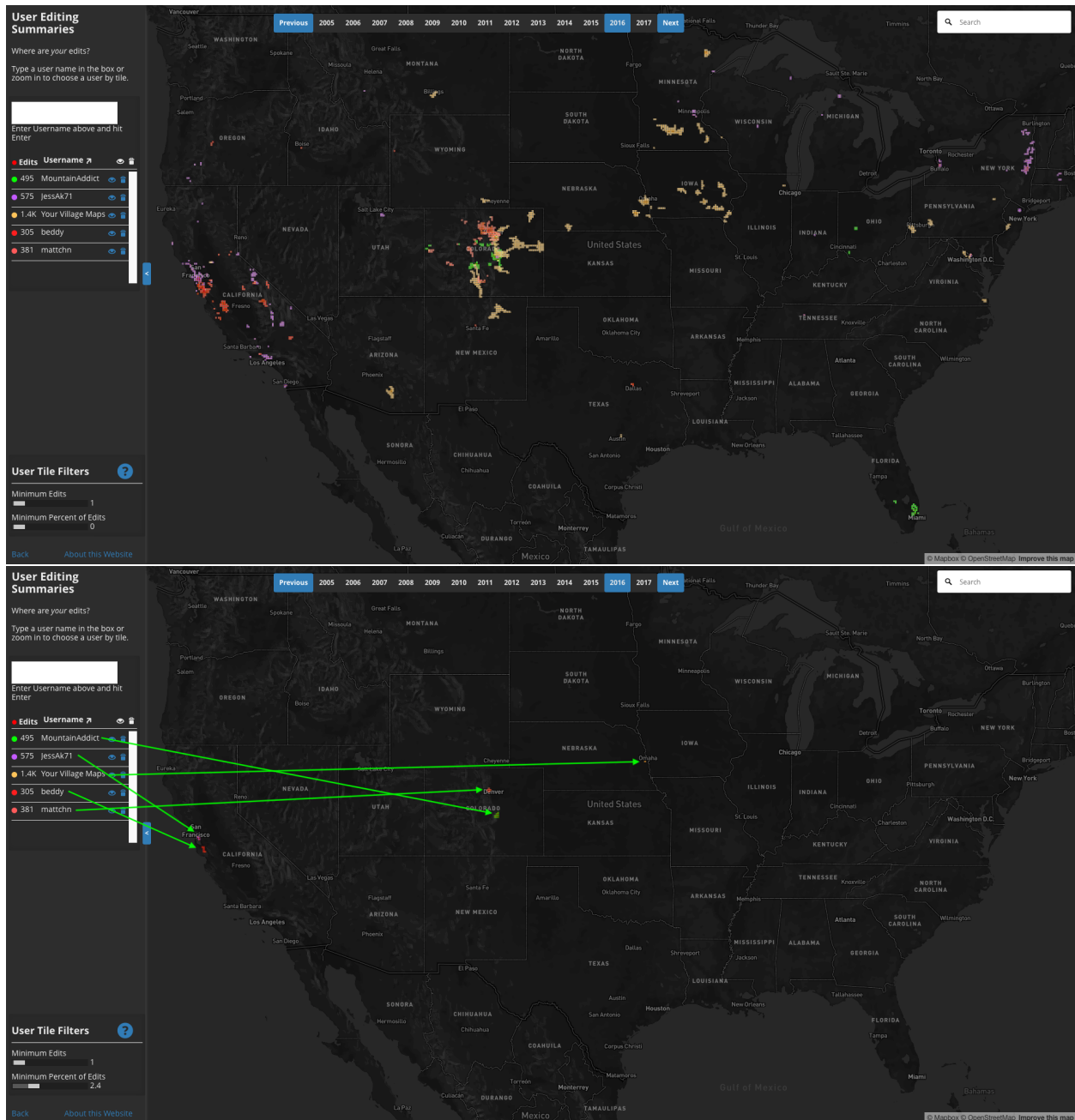


Figure 1.8: Top: The mapping footprint of five mappers in the United States in 2016. Bottom: The smaller regions where more than 2.4% of these mapper’s annual editing activity was isolated. Screenshots from mapbox.github.io/osm-analysis-collab, a research collaboration with Mapbox to be introduced in Chapter 6.

Telegram channels, Slack Workspaces, and Meetup groups, for example, are dedicated to specific regional mapping groups. OSM-Colorado, for example, is a regional Meetup group dedicated to

all things happening in Colorado related to OpenStreetMap. Many regional communities have dedicated mailing lists, such as the “talk-us” or “talk-ph” mailing lists for OpenStreetMap US and OpenStreetMap – Philippines, respectively. Some regional mapping groups are certainly more active, structured, and organized than others.

1.4.1.3 Corporate Editors

Corporate data teams have been active in OSM for years with growing transparency since at least 2014. These are teams of employees who are paid to contribute to the map. Presumably, the goal of corporate data-teams is to improve the map for a particular business use-case. An obvious example here is routing: many companies are using OSM Data as input to routing algorithms to provide directions in their maps. Amazon, Apple, Mapbox, Grab, Lyft, and Uber are all using OSM Data to improve their products or logistics in some manner. Perhaps not surprising for an open data community, there are mixed feelings about the presence of these corporate editing teams in OSM [3]. In the last 3 years, I have watched Facebook, Apple, and Lyft “come out” to the community in the form of presentations at annual conferences. These presentations are more warmly received at the US based annual conferences than the global conference, State of The Map. To this end, the OpenStreetMap US community is known to be more corporate-friendly than the global OSM community. Current increasing trends in the number of corporate edits on the map each month as shown in Figure 1.9 suggest that corporate editing is not going away anytime soon.

Currently, the best practice for a corporate data-team editing OSM is for the company to provide a list of usernames associated with the data-team on the company’s OSM wiki page. It is then largely assumed that edits associated with these usernames are associated with the company. The aforementioned organized editing guidelines apply to these types of editing activity. Even before these guidelines were published, however, many companies were already maintaining active Github repositories and OSM wiki pages describing the extent of their editing activities.

The presence of corporate editors in OSM raises new questions around geographic bias, corporate interest, and the potential pushing-out of local mapping communities. As such, I have

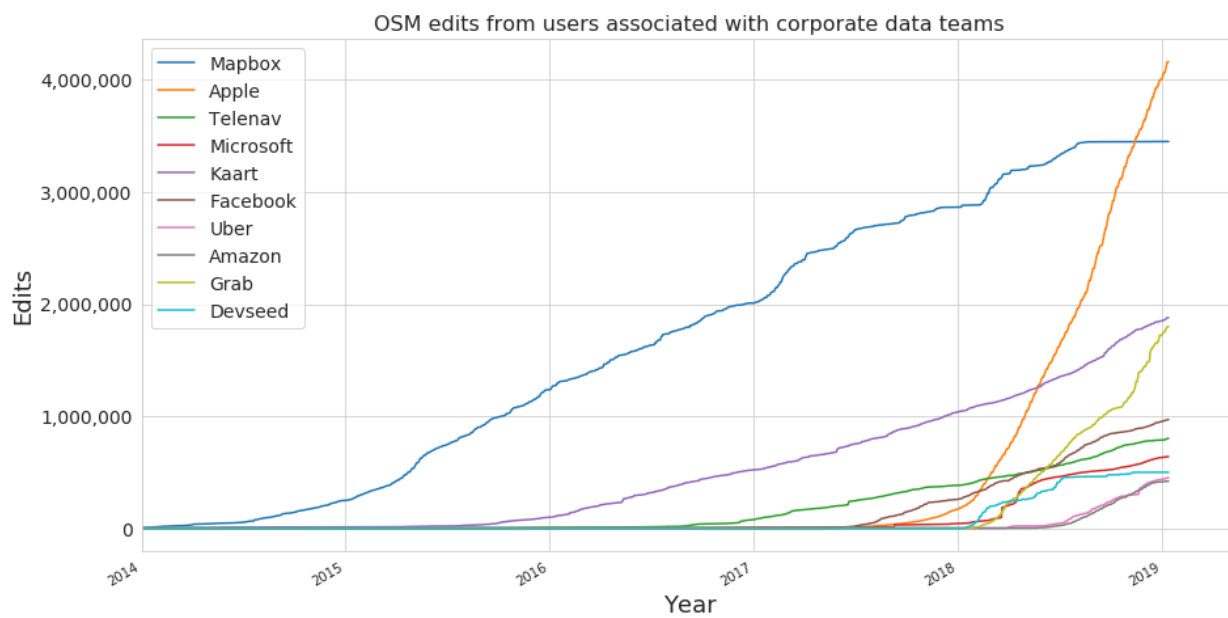


Figure 1.9: The rise of Corporate Editing in OSM

observed their presence to be a contentious topic. It is common for a presentation from a company talking about their mapping practices to be standing-room only (especially their first talk when they “come out” to the community). In the case of both a Facebook talk (SOTMUS 2017) and an Apple talk (SOTM 2018), the companies requested that their presentations not be recorded, forcing people to attend in person. Chapter 9 discusses the long-term involvement and influence that corporations had had in the history and evolution of OSM.

While there are studies that look at specific communities within OSM such as work by Dittus et al. that focus on contributors to HOT tasks [29, 30], it is (more) common for studies to classify OSM contributors as a whole based on their contribution activity: often classifying users in some level of hierarchy ranging from beginner mappers and experienced, power mappers [84, 11, 18]. While it is likely that the same distribution of edits per user exists within each of the communities identified here (hobby/ localized mappers, humanitarian mappers, corporate editors): The difference in editing patterns between these communities and their ultimate role in the development of the map is not as deeply explored, prompting the research presented in Chapter 9 [3].

1.5 Global Scale

This research addresses OSM data at a global, planet-wide scale. Because people typically focus on a specific area of interest, this global approach differs from how most users engage with the map. If, for example, one is only interested in simple analytics such as how many roads or buildings are in an area, then more simple, localized analytics can be done in geographic isolation. However, if we are to address, as this research attempts, ways in which we can derive more information about how the map was constructed and what that tells us about the quality of the map for example, we need to consider the map as an entire entity, because of the ways in which contributors themselves behave.

Figure 1.3 highlights that contributions to OSM, while uneven, are worldwide. Additionally, Figure 1.8 shows that even though many contributors are active everywhere, individual mappers tend to have a particular area of the map that receives more of their attention. Incorporating this background information is important for contributor-centric analysis because it reveals a mapper's previous mapping expertise. For example, a contributor introduced to OSM through participating in a disaster or humanitarian mapping project will have performed a significant amount of mapping—most likely not in their local region. If this mapper then maps a few features in their hometown, they will look relatively inexperienced in their local editing record, but they may have deep familiarity with the process of mapping and the norms of the community through many previous days of disaster mapping activity. Accounting for this expertise requires looking into a mapper's global history beyond the bounds of a particular area of analysis.

To this end, regardless of the actual area of interest, contributor-centric analysis needs to account for a mapper's previous activity. There are two ways to do this: First, a mapper's username can be looked up on openstreetmap.org (or through the public API) to learn when a contributor registered their account and how many changesets they have authored since. This offers two pieces of general information about a user, but does not yield understanding of where a mapper has been active or what type of edits they are performing. Second, all of the edits made globally need to be

included in any and all analyses of distinct sections of the map to account for the full context of an contributor’s lifetime activity. The first approach is implemented in Chapter 3, while the second approach is implemented in Chapters 6 and 7.

Additionally, if one hopes to perform analysis of a particular community of mappers within OSM, considering the entire globe is critical because many communities are active everywhere. Humanitarian mappers, for example, have contributed to mapping projects all over the world. As Figure 9.3 will show, corporate editing is a global phenomenon, so any questions around this community requires considering the entire planet to tell the full story.

1.6 Related Work

There are multiple categories of work related to what I am presenting here. Overall, Figure 1.10 shows the increase in Google Scholar results for articles with the terms *OpenStreetMap Data Analysis*. The increase in articles represents the growing awareness and use of OSM data in research, primarily as a source of spatial information, not necessarily research *about* OSM, but exposure for OSM nonetheless. While this exposure better validates OSM as a worthy source of geospatial information for research, it necessarily raises new concerns about how the data is processed and if these are the best approaches for the types of questions people are asking.

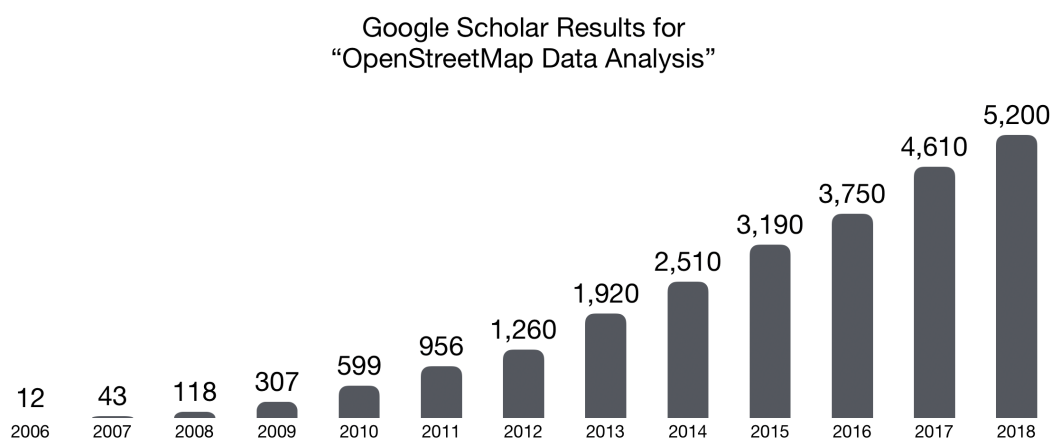


Figure 1.10: Approximate number of articles on Google Scholar published each year with references to OpenStreetMap Data Analysis

Empirically, I observe that these concerns do not yet raise alarm. Reading through the titles of the latest research confirms that the majority of this new research is either validating OSM data against another dataset to show its viability for a particular application (routing, planning, etc.), or using it as a primary (best, only, most accessible) source of geospatial data for a region of interest. Most of this work does not ask questions about the growth of the data or community itself and therefore is not losing anything by not embracing *contributor-centric approaches*.¹⁷

Working with OSM data requires first piecing together a data-processing pipeline that best fits one's analysis goals. For relatively small regions, a number of utilities exist to simply load the spatial data into traditional GIS environments such as ESRI tools or the open source alternative, QGIS.¹⁸ My observation is that much of the research in Figure 1.10 includes a variation on the following phrase: *We downloaded the regional extract for <region> and loaded it into QGIS/ESRI for comparison with <dataset>*. Depending on the software version and format/source of the extract, any number of the tools and frameworks just mentioned are used in this processing pipeline.

Closer to the work discussed here are processing workflows for *intrinsic quality analysis* built as toolboxes and plugins for QGIS. Such as one for analysis of road networks created by Graser et al. or a more extensible intrinsic quality assessment plugin as built by Sehra et al. [45, 120]. Intrinsic analyses approaches often involve the metadata and data provenance and can therefore benefit from contributor-centric approaches that prioritize this information. Chapter 7 includes a more comprehensive review of these and points to many other intrinsic quality analysis frameworks available for OSM.

With regards to full-stack development of OSM data analysis systems, there are a few other research groups and projects working in this space. Most notably, these projects are still very much in active development and are have been constantly evolving in parallel with the work conducted at the University of Colorado over the past five years.

¹⁷ As much as I would like to advocate my approach to OSM data analysis is the best, it is not the only valid approach to working with OSM data and is overkill for the majority of analyses that simply use the spatial attributes of OSM data.

¹⁸ See the OSM wiki for a comprehensive list of tools and frameworks that have been implemented for these purposes: wiki.openstreetmap.org/wiki/Category:OSM_processing

1.6.1 Heidelberg Institute for GeoInformation Technology

OSHDB, iOSMAnalyzer, and by association, many of the tools created by Pascal Neis all have their roots in the Geography Department—and now the Heidelberg Institute for GeoInformation Technology (HeiGIT)—at the University of Heidelberg, Germany [112, 10].¹⁹ An early intrinsic quality analysis tool, iOSMAnalyzer used a very similar data-processing workflow to our Epic-OSM utility presented in Chapter 3 to ingest an OSM history file and perform batch analysis on a particular region with a set of predefined quality measures and indicators, ultimately producing a PDF summary of the analysis [10].

OSM researcher Pascal Neis did his PhD work at the University of Heidelberg and has for years maintained a number of community-oriented tools available on his personal website. The most common of these tools is HDYC (“how do you contribute to OSM?”). This tool, available at hdyc.neis-one.org shows general summary statistics for any OSM user account. It is a common practice for a user to link to their “hdyc page” in their OSM profile to provide an overview of what type of mapper they are. While Neis’ webtools provide many useful metrics, they are dashboards that are not built to be customizable. The code is also predominantly closed source, so people cannot contribute to it or build from it. To this end, they are extremely useful dashboards, validated by their widespread use in the community, but different from an analysis infrastructure as presented here because they do not enable users to ask arbitrary questions of the data.

The most recent work to come out of HeiGIT is the OSHDB, the final iteration and implementation of a scalable, full-history spatiotemporal query engine that is intended to support OSM research [112]. Now with a public API in front of a global instance of OSHDB known as OHSOME, these tools are going to lower the barrier to entry for researchers looking to explore the history the map for any area. As Chapter 10 will discuss further, the OSHDB can function as a powerful back-end for future contributor-centric analyses.

¹⁹ heigit.org/, resultmaps.neis-one.org/

1.6.1.1 osm-analytics.org

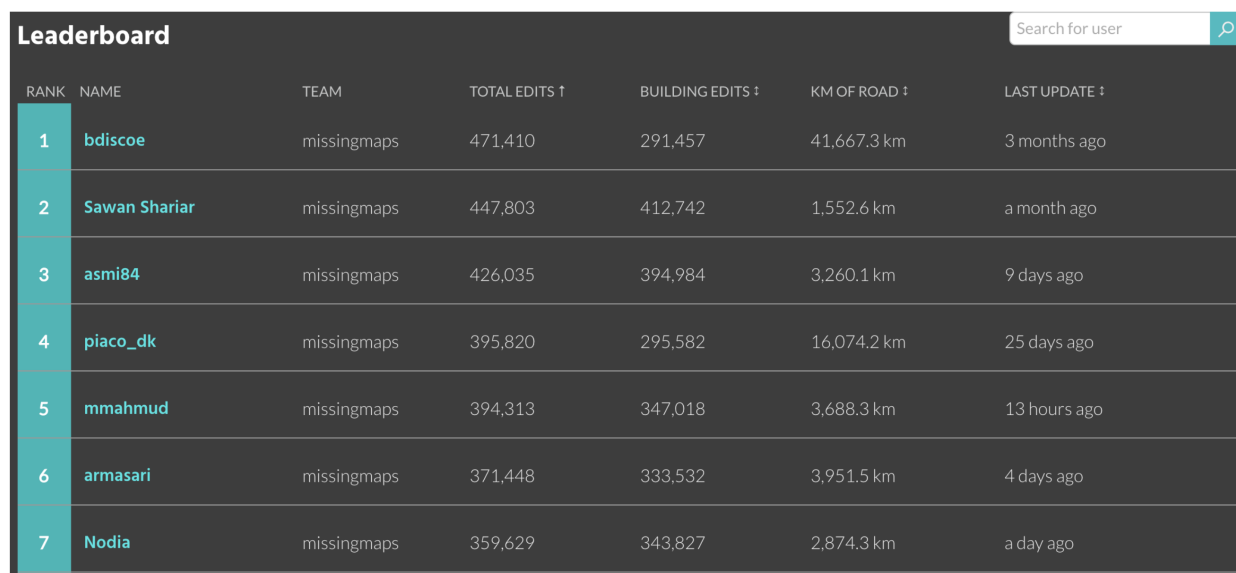
osm-analytics.org is an interactive dashboard for visualizing the global coverage of OSM data. The website is built with OSM-QA-Tiles (introduced in Chapter 5) and allows the user to compare the state of the map over time in terms of the number of buildings, roads, rivers, amenities, or hospitals. Added recently, a new feature of the dashboard is the comparison of OSM to global population data, identifying areas of the map that are likely incomplete. The processing workflow that powers this dashboard inspired the data processing pipeline presented in Section 5.2. This tool was built as part of a much larger collaboration involving HeiGIT.²⁰

1.6.2 OSMesa

OSMesa is another scalable suite of tools built on contemporary big-data technologies such as Apache Spark and Amazon Web Services.²¹ Also being actively developed in parallel to the work presented here, OSMesa is part of the processing pipeline behind most of the figures presented

²⁰ osm-analytics.org/about

²¹ github.com/azavea/osmesa



The image shows a screenshot of a leaderboard titled "Leaderboard" with a search bar for users. The table lists the top 7 contributors, including their rank, name, team, total edits, building edits, kilometers of road, and last update time.

RANK	NAME	TEAM	TOTAL EDITS ↑	BUILDING EDITS ↓	KM OF ROAD ↓	LAST UPDATE ↓
1	bdiscoe	missingmaps	471,410	291,457	41,667.3 km	3 months ago
2	Sawan Shariar	missingmaps	447,803	412,742	1,552.6 km	a month ago
3	asmi84	missingmaps	426,035	394,984	3,260.1 km	9 days ago
4	piaco_dk	missingmaps	395,820	295,582	16,074.2 km	25 days ago
5	mmahmud	missingmaps	394,313	347,018	3,688.3 km	13 hours ago
6	armasari	missingmaps	371,448	333,532	3,951.5 km	4 days ago
7	Nodia	missingmaps	359,629	343,827	2,874.3 km	a day ago

Figure 1.11: The Leaderboard available on missingmaps.org that shows the top contributors to the project. OSMesa powers the data analysis behind the dashboard which helps gamify humanitarian mapping to engage contributors.

in this chapter. With powerful planet-scale editing history reconstruction abilities, OSMesa is an optimal choice for the future back-end to contributor-centric analysis systems. This framework can also scale to perform near real-time analysis of OSM data, making it a powerful analytical background currently used by The Red Cross, Missing Maps, and a variety of other organizations to power a number of leaderboards and result maps based on editing statistics, such as Figure 1.11.

1.7 Outline of the Dissertation

This dissertation is organized into *Parts* that contain *Chapters*. Each part includes a preamble that briefly describes the chapters to come. Additionally, there is a glossary in the appendix with short explanations of various terminology. Terms with entries in the glossary are introduced with an asterisk (*).

Chapter 2 concludes **Part I** by introducing the technical and analytical challenges associated with measuring OpenStreetMap. **Part II** follows, and discusses the first OSM data analysis system built at the University of Colorado, Epic-OSM, and the multiple research projects that it supported. Within this Part, Chapter 3 is an exact reprint of [4], as listed in Section 1.7.1. The failures of the system to scale when implemented as a real-time analysis framework in the 2015 Nepal Earthquake inspire the adoption, implementation, and extension of vector-tile based analysis approaches, as introduced in Chapter 5.

Part III presents the technical contributions of this work and discusses the iterations on and innovations to existing systems that I have performed to produce this work. Specifically, the chapters included discuss the pros and cons of vector-tile based analysis of OSM data and how it can support contributor-centric research.

Part IV consists of four chapters, each discussing a completed research project demonstrating the analytical capabilities of vector-tile based analysis. Within this, Chapters 7, and 9 are exact reprints of [5] and [3], as listed in Section 1.7.1.

Part V discusses the current state of full history vector-tile based OSM data analysis. This includes work completed to date and lays out concrete plans for future implementations and work

currently underway.

1.7.1 Inclusion of Published Work

Chapters 3, 7, and 9 are reprints of already published work, included here with the permission of my coauthors:

Chapter 3: Jennings Anderson, Robert Soden, Kenneth M. Anderson, Marina Kogan, Leysia Palen (2016). *EPIC-OSM: A Software Framework for OpenStreetMap Data Analytics*. In Proceedings of the 49th Hawaii International Conference on System Sciences. 5467-5477.

Chapter 7: Jennings Anderson, Robert Soden, Brian Keegan, Leysia Palen, and Kenneth M. Anderson (2018). *The Crowd is the Territory: Assessing Quality in Peer-Produced Spatial Data During Disasters*. International Journal of Human-Computer Interaction. doi:10.1080/10447318.2018.1427828

Chapter 9: Jennings Anderson, Dipto Sarkar, and Leysia Palen (2019). *Corporate Editors in the Evolving Landscape of OpenStreetMap*. ISPRS Int. J. Geo-Inf. 2019, 8, 232. doi:10.3390/ijgi8050232

Chapter 2

Measuring OpenStreetMap

Differentiating itself from other geospatial data, objects in the OSM database are internally referenced and related to one another. This means that geometries like lines and polygons do not actually contain any geographic information themselves, but instead reference other objects that contain the coordinates. In contrast, it is common for representations of LineString or Polygon objects in other formats to contain a list of coordinates that define the vertices. In practice, working with these types of objects is computationally simpler: a single entry or row in a file or database contains all of the object's information. In OSM, however, the coordinates need to be looked up in a location cache to resolve the geometries.

This creates a topological structure in OSM where the *node element* represents the smallest building block of objects on the map. *Way* and *relation elements* then reference these nodes. In this manner, duplicate points on the map are minimized: One node at a particular geographic location can be referenced by any number of objects. An intersection, for example, can be represented as a single node that is referenced by both roads that cross to create the intersection. This node can then have additional attributes such as the identification of a stoplight (traffic signal) at the particular intersection. In this case, the topological structure of the OSM data model is convenient and efficiently represents all of the necessary information. Furthermore, this design is particularly well-suited to construct a routable road network. A standalone building, however, likely does not share any corners with another object. Because of this, way elements created to represent buildings often reference multiple nodes in which most or all of them have no other attributes than their

geographic location. It is unnecessary for these nodes to exist independent of the building object. Similarly, the vertices of a winding road, coastline, or border are typically only referenced by one parent object. These represent the vast majority of nodes in the OSM database and their existence as standalone objects, independent of their parent element is unnecessary. Today, however, this is simply an artifact of the initial data model and is unlikely to change.

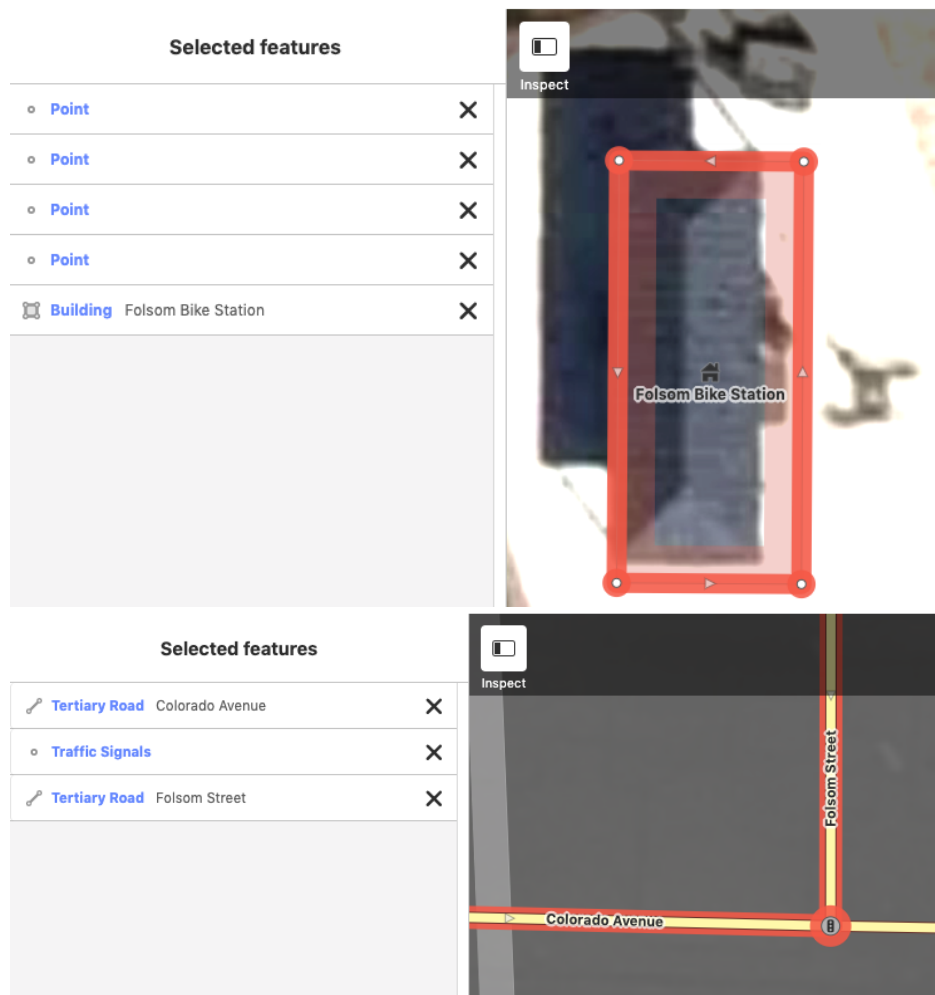


Figure 2.1: Top: A square building in OSM is represented by five elements, four nodes (Points) that mark the corners and a way element tagged as “building” that references the nodes. Bottom: An intersection with a traffic signal in OSM references the same node element multiple times: The traffic signal itself as the point and then both roads joining at that point. Screenshot from iD editor on openstreetmap.org; Data © OpenStreetMap Contributors.

While there are storage, computational efficiency, and cross-compatibility arguments to be— and have been—made [133] against the topological data model, I am most concerned with the

analytical challenges that it inadvertently introduces. Since map objects may reference one or more other elements, the number of edits to objects in the database does not equate to the number of edits made to the map in a logical sense. For example, there are over 5B elements in the OSM database across the nodes, ways, and relation tables. However, when the map is rendered, there are fewer than 1B objects on the map. An edit to any one of these objects may then have cascading effects within the topology. This combined with the sheer volume and global scale of OSM data makes measuring the contributions to OSM a difficult, multi-faceted problem.

2.1 Defining “edits” to the Map

A contributor-centric approach means first abstracting this topological relationship between OSM elements away, leaving just OSM objects and the edits that affect them. This change produces output more similar to standard geospatial data structures where each object has its own geometry and set of descriptive attributes. With this abstraction, the OSM database transforms from a collection of billions of nodes, hundreds of millions of ways, and millions of relations to a single collection of hundreds of millions of geographic objects representing the physical reality of our world. Though still a massive and messy dataset, it is now easier to quantify what an edit to one of these objects might look like. For consistency, I use the term “element” when referring to a node, way, or relation in the OSM sense and the term “object” when referring to the abstraction of an entire object (as it might be rendered on the map).

For example, consider the common rectangular building in OSM—of which there are hundreds of millions in the database [87]. Creating this building involves a minimum of five edits to the database: the creation of four node elements to define the building’s corners and a single way element that references these node elements. Any subsequent edit to this building object could change any or all of these elements in the database. Editing the building’s name, for example, would be an edit to only the way element. Moving one of the building corners would be an edit to only a single node element. In the database record, these two edits are unaware of each other, even though they are really both edits to the same OSM object. For this work, I define the following five distinct

types of edits in OpenStreetMap that can occur to an OSM object:

Change	Type of change at the object level (and the OSM elements changed)	Primary Element Edits	Ref. Element Edits	Version
1	Creating a new object	1	0 or more	Major version = 1
2	Slightly modifying an existing object's geometry (moving existing node elements)	0	1 or more	Minor version += 1
3	Deleting or adding references to other elements (major geometry change)	1	1 or more	Major version += 1
4	Editing an existing object's attributes (tag changes)	1 or more	0	Major version +=1
5	Deleting an existing object	1	1 or more	Major version += 1

Table 2.1: Classification of types of edits in OSM

Referenced objects in Table 1.1 refer primarily to nodes referenced by ways, but this classification of editing extends to OSM relation elements as well. With this abstraction of OSM element(s) to a single OSM object, we can then identify, classify, and assign all of the edits to the many OSM elements that may be involved instead as edits to OSM objects. We can then record the username, timestamp, and specific type of edit. Only then can we reconstruct the complete history of an object: Crediting and accounting for each editor that has contributed to the current state of the object as it exists on the map.

This taxonomy of edit types may at first seem excessive, but it is complete in terms of the types of edits that need to be accounted for when examining the potential history of an OSM object. Of particular importance is the notion of a “minor version” introduced here and further described next. For reference, and as a major sign of progress in the realm of OSM data science, the concept and terminology of a minor version is becoming more familiar today and is being actively implemented and used by other OSM researchers.

Chapter 4 explains the nuances of this particular abstraction to the “OSM object” in more detail, but in general, the flattening of the OSM data model to single geometric objects is a fairly common data conversion required for working with OSM data in more traditional GIS environments (such as converting OSM data to the more common ESRI shapefile). This lossy conversion (Chapter 4) often discards various attributes of the editing information. The metadata or less-common tags, are then lost, and/or the topological relationship is lost entirely. For most data-use purposes (like rendering a map), these are completely acceptable losses. For analysis, however, these types of lossy conversions discard valuable information, particularly metadata.

2.2 Minor Versioning of OSM Objects

The minor version of an object increments when the child elements of an OSM element are updated independently from the object itself. The most common example of a minor version occurs when a mapper moves existing nodes (edit type 2 in Table 2.1). In practice, this edit could be to better align the object to newer, more clear imagery, or to straighten the corner angles of a structure. If the mapper only adjusts existing nodes and does not change any other attributes of the OSM element, then the parent element is not aware that anything has changed.



Figure 2.2: Progressive changes in the geometry of a building in OSM over multiple years. The four nodes that make up the building have been moved five times, creating five distinct historical geometries that are independent from the the versioning history of the way element.

In practice, this only creates a problem for data analysis. Though there are multiple versions of the nodes, only the most recent version should be used to define the geometry of an object. The current *planet file*, a downloadable database dump of the current map, then only includes one version of any OSM element: The most recent. Further, it is relatively easy (time-intensive, but not

complex) to view the map at any point in time by creating a historical “snapshot” of OSM. To do this, one only has to truncate the database to exclude any edits after a specific point in time, and then keep only the latest version of an object. When referenced, only the geometry at that point in time exists to be returned. This creates a replica of the *planet file* as it existed at that point in time. Chapter 6 further describes how these can be employed as a tool for historical analysis.

Most OSM data-processing utilities are equipped to handle OSM data where only one version of an element exists, such as the planet file or other data extracts. OSM extracts that contain more than one version of any OSM element (the ID is not unique) are referred to as *history* files. A history file of all of the OSM database is made available to download weekly. It is known as the *planet-history file* and contains over 9B OSM elements (including deleted objects).

While all of the information required to recreate an object’s history exists in the history file, most software cannot handle the nuances. For example, there might only be one version (version=1) of a way representing a building in the history file, but multiple versions of the nodes that it references (thereby creating minor versions). This is a common situation in which the first edit created the building and a secondary edit (not always a different user), moved these nodes, typically to square-up the corners of the building. From an edits-to-the-map perspective, there are 2 distinct edits: the creation of the building and then changing its shape to match the community norm of having square building objects on the map. Looking only at the history of the way element, however, only shows the first edit. Worse, counting the edits to node objects could show up to eight edits: The creation of four nodes to represent the building’s corners and then the subsequent edit to any of these four nodes as they were moved to create version 2.

For accurate data analysis, we need a representation of this object that reflects two versions and shows the proper metadata for both versions. For example, the second version of the building with the updated shape needs to include the metadata from the edits to the referenced node elements, because even though this change was only to the nodes, the relevant change is not to the node objects, it was to the shape of the building.

To accurately reflect this editing history, then, we introduce the minor version. This building

would currently exist as version 1, minor version 1. Its history includes version 1, minor version 0, representing the object as initially created. The minor version allows us to maintain the primary version attribute so that it matches the OSM database, while adding the granularity of a minor version. Over 120M objects have minor versions, which is about 20% of all of the map objects on the planet. Here is an example schema of how this fictional object and its history can be represented to accurately count the users involved in the evolution of this building on the map:

```
{
  id: <ID of way element>,
  history : [
    { version: 1,
      minorVersion: 0,
      user: <The mapper who created the building on the map>
      timestamp: <When the mapper created the building on the map>
      geometry: <Location of version 1 of the nodes, as they were created>
      changeset: <ID of changeset in which the building was created>
    },
    { version: 1,
      minorVersion: 1,
      user: <Mapper who moved the nodes>
      timestamp: <When the user moved the nodes>
      changeset: <ID of changeset in which the nodes were moved>
    }
  ]
}
```

Here, the **version** attribute remains unchanged and therefore still matches the version number of the way element as it exists in the main OSM database. Furthermore, there can be any number of minor versions associated with each version. Minor versions can be reconstructed from the planet-history file, but they are far more complex and computationally challenging to produce than historical snapshots. Section 5.6 goes into further detail about how to represent these changes in a new data schema and the tools that currently exist to create it, while Chapter 10 presents my current implementation of said schema and further discusses my development of a utility to reconstruct these OSM objects with full-histories from the planet-history files.

2.2.1 Limitations and Pitfalls of Minor Versioning

For reasons of practicality, minor versions should be limited to geometry changes only. They can be thought of as *geometry versions*, but will retain the name “minor version” for consistency with current implementations. Consider, again, the case of two intersecting roads and the later addition of the traffic signal to the node that represents the intersection (edit type 4 from Table 2.1). By the definition described above, two minor versions would be created: One for each of the intersecting roads because a node element referenced by both ways was changed. This change, however, really does not affect the way elements and furthermore, inflates the number of edits, propagating this single change to the parent way elements as minor versions. This can be safe-guarded against by filtering for a location change, discussed in Chapter 10.

Computing minor versions can be extraordinarily resource intensive because in the worst-case scenario, the creation of accurate minor versions requires comparing a massive cross-product: The sequences of all referenced nodes, including previously referenced nodes that have since been removed, with each version of each of these nodes. In practice, this is rarely the case because it would be an odd editing pattern, but some objects that are heavily edited will require thousands or even millions of node location sequences to be checked. Chapter 10 shows techniques to avoid these when calculating minor versions.

2.3 Spatiotemporal Scaling: Volume and Velocity

As a spatial database containing over 800 million distinct objects and more than one-billion previous iterations of these objects, OSM data analysis becomes a big data problem.¹ As such, I borrow the terms *volume* and *velocity* from big data discussions to further describe issues of scalability associated with OSM data processing. As Section 1.5 discussed, to capture the full context of editing activity, analysis systems need to be capable of global-scale data processing. This significantly increases the volume of data that needs to be considered beyond just the map data present in a particular area of interest.

¹ Computed with Amazon Athena from OSM Full history objects as built by OSMesa (github.com/azavea/osmesa)

Additionally, the time range of the data being analyzed introduces two complications, first at the data level with regards to volume, and second at the application level when it comes to the implementation and use of OSM analysis systems. With previous versions of OSM objects, analysis of larger time-ranges significantly increases the volume of data to be processed (more than two-fold if doing the full history). An increase in volume causes an increase in the processing time required (impacting velocity). For analysis of mapping that has already happened, concerns of processing time are primarily about convenience for the researcher, but do not affect the usefulness of the results. However, if attempting to produce real-time analytics that capture the editing activity as it is happening, an increase in volume and velocity will negatively impact the usefulness of the system. Consider analysis of a disaster mapping event: If exploring a previous mapping activation, whether it takes one or two hours for the analysis process to run has little consequence for the results. For real-time analysis of a disaster mapping activation, however, the time it takes to consume, parse, and process the data creates the delay between the activity itself and producing the analysis. Being one hour or two hours behind is a much more significant difference in the temporal accuracy of the results. Section 3.5.1 will present a specific example of this real-time scenario and a patched solution.

The work presented here ultimately chooses to scale spatiotemporally in terms of full-global and historically-complete data to tell the story of the evolution of the map, but leaves real-time analysis to other platforms, such as OSMesa which is capable of continually ingesting the latest changes to OSM and producing a number of summary statistics that can power analytics like the mapper leaderboard shown in Figure 1.11. Making this distinction between historical and real-time analysis clarifies the goals of the systems presented here, which aim to tackle the full history of the planet, addressing the spatial and historical scaling complications, while acknowledging the added complexity of real-time systems that will remain beyond the scope for the time being.

Now that I have introduced the difficulties in measuring OSM and the importance of new analytical systems that can help us understand the evolution of the map and the data, the rest of this document will look at the iterative development and evolution of OSM data analysis systems through a series of projects and publications. While the design requirements of each system have remained similar, the approaches have differed greatly, optimizing first for smaller-scale analysis of particular regions, and then for global scalability. Throughout this process, there have been consistent requirements with regards to rapid-iteration and consistent output formats so as not to lock data analysts into single visualization and analysis environments.

It often appears that the number of tools and techniques in Data Science are growing and evolving faster than the systems they are built to research. While the approaches to OSM data analysis presented here differ in how they handle and process the historical editing record of OSM, consistent output formats and schemas allow for the least amount of change in down-stream analysis work as the input changes. For example, interactive maps built to visualize the output from data-processing steps can be initially developed around a specific data-schema and future output should require the smallest amount of modifications to be compatible. The compatibility extends to analysis environments such as Jupyter or Zeppelin Notebooks. Section 10.2.1 describes the advantages of this further.

JSON (and GeoJSON) is a standard data schema, specifically for web-based tools. The human-readability of JSON allows analysts to manually investigate the output without the need for additional translation and identify errors quickly, which supports rapid iteration. For these reasons, all of my workflows work almost exclusively in this format.

Part II

First Ventures into an OpenStreetMap

Analysis Infrastructure

Chapter 3

Epic-OSM: A Software Framework for OpenStreetMap Data Analytics

This chapter is comprised of an article published in the proceedings of the Hawaii International Conference on System Sciences, reprinted here with permission from my coauthors.¹ This paper presents the first infrastructure we developed at the University of Colorado Boulder to ingest and analyze historical OSM data, specifically to support crisis informatics research. It should be noted that this work predated and ultimately lead to the identification of the challenges just articulated in Chapter 2. Those challenges are therefore not yet addressed in this chapter. Following this article, the Chapter 3 Epilogue reviews how the infrastructure was implemented, other work it supported, and discusses its ultimate shortcomings to scale, which prompted the switch to vector tile based analysis, as introduced in Chapter 5.

3.1 Introduction

We live at a time when organizations of all kinds increasingly have the means to generate, collect, and analyze large volumes of data via software systems. These systems—collectively known as data-intensive software systems or “big data” systems—are challenging to design, develop, and deploy [6]. One application area that requires the development of these systems is crisis informatics [104], which investigates how social computing can impact the practice of emergency management. Of particular interest is the use of digital maps to support disaster response, an activity known as

¹ Jennings Anderson, Robert Soden, Kenneth M. Anderson, Marina Kogan, Leysia Palen (2016). EPIC-OSM: A Software Framework for OpenStreetMap Data Analytics. In Proceedings of the 49th Hawaii International Conference on System Sciences. 5467-5477.

crisis mapping.

However, the analytics of geospatial data are especially challenging to resolve. This is because a) map datasets tend to be extremely large—often consuming terabytes or petabytes of information—and b) map datasets are not good at conveying how they were created. That is, for any given version of a map, all one sees is the final aggregate map, not the individual edits that were performed to create it. This is what separates collaboratively-edited geospatial data from collaboratively-edited text documents—such as articles on Wikipedia—which can much more easily display editing history across users.

In the new world of crowd-sourced data generation where information can be produced quickly for open use, understanding the collaboration that went into the construction of the map can be as important as the map itself. This is especially true for action-oriented communities, like the crisis mapping community, that are trying to understand their evolving work practices while they work to produce maps that can be used to aid crisis response. These communities seek to understand their work in situ to improve upon it. Social computing researchers desire the same understanding to both document what digital crowds can achieve and with an eye towards designing better tools to support that work in the future. For the big data community, this type of research is important, as it requires the novel use of data analysis techniques both for the batch processing of existing data sets as well as the real-time analysis of edits that stream in during a crisis event.

In this paper, we report on the design and development of a big data software framework that can be used to analyze the edit history of OpenStreetMap (OSM), making it possible to study the cooperative work that occurs there, including but not limited to the intensely collaborative periods of crisis mapping where much is at stake for humanitarian groups using these maps on the ground. At the time of this writing, there are no other frameworks that perform this type of analysis for OSM data; indeed, use of our software framework has been steadily increasing since its initial deployment for studying and monitoring the mapping activity surrounding the 2015 Nepal Earthquake event. This increased use is the direct result of the unique analysis capabilities our framework provides on top of OSM data.

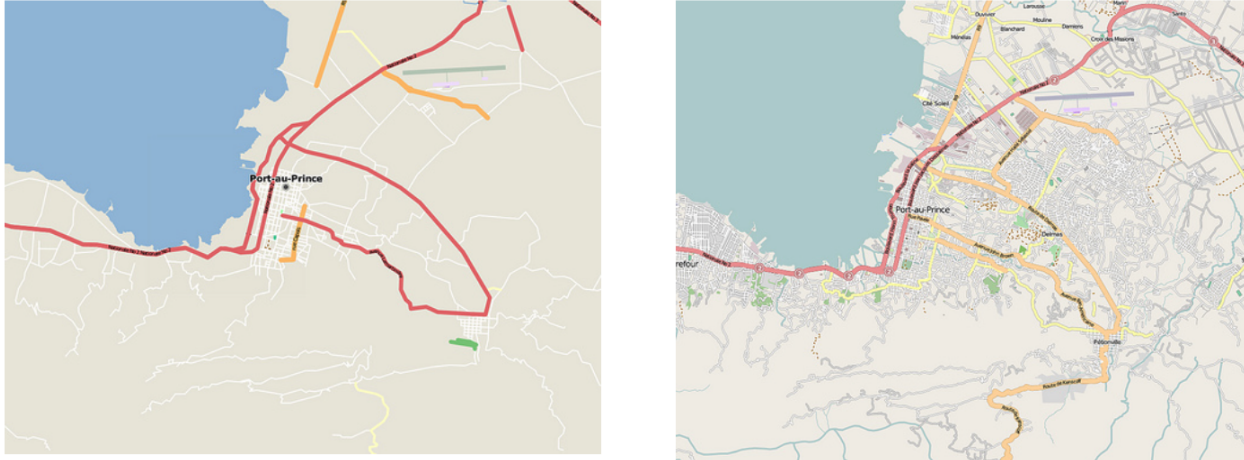


Figure 3.1: Port-Au-Prince, Haiti in OSM, before the 2010 earthquake (left) and 4 days after (right). Remotely-located volunteer mappers added all features by tracing aerial imagery [70].

OpenStreetMap is an open geographic data initiative that provides a map, and its associated geospatial data, that anyone can contribute to and access. Our software framework, known as epic-osm, can scale to process gigabytes of OSM data by employing a variety of techniques to both analyze data for desired metrics and visualize the results in ways that are meaningful to mappers themselves and the larger OSM organization and community. It also makes details of this enormous data-producing organization [105] available to researchers in the way that Wikipedia has been studied extensively for years as a notable site of collaborative data production. Our framework is more than just a design; our code is available on GitHub and the software tools that have been built on top of this framework are in active use. Our experiences designing and implementing this framework can be of use to others. We demonstrate how to address the challenging data modeling issues that arise in the design of data-intensive software systems [118], as well as issues of extensibility, scalability, and interoperability.

3.1.1 Studying the OSM Community

With some notable exceptions [74], the majority of research on the social organization of the OSM community has been based upon qualitative research methods such as participant observation,

interviews, and surveys. These studies have provided insights into participant motivation [18] and demographics [117]. Our team saw that examination of the OSM database itself—which contains a complete record of every edit ever made—is critical to the advancement of the 2.1M OSM member organization, which needs to better understand its production functions to manage its growth [105], as well as for social computing researchers to characterize the nature of cooperative crisis mapping. Understanding the social processes governing the creation of OSM data is especially important for crisis informatics, since these behavioral phenomena can affect the quality of the geographic data produced. This can have real human consequences as OSM is frequently used as the primary base map in humanitarian response [125]. One likely reason that so little analytical research of socio-behavioral phenomena in OSM has been conducted (in comparison to the vastly-studied Wikipedia organization) is the challenges of manipulating OSM data. A complete download of the OSM history database is over a terabyte in size and is continuously growing as new edits are made. This difficulty affects not only scholars, but also the OSM community itself, which struggles to track its own activity, and hence its growth and impact [105]. To address this knowledge gap, we have identified a number of OSM members who have been willing to contribute to the development of the epic-osm framework as well as deploy and test it for a range of purposes. As will be discussed, this engagement has helped push the development of our framework and its surrounding toolset in new directions. Furthermore, this has catalyzed discussion within the OSM community about the need for new tools, as the existing community toolset, prior to the creation of our framework, is sparse and does not provide in-depth analytical capabilities.

3.1.2 Crisis Informatics and OpenStreetMap

When a major disaster occurs, a subset of the OSM community rapidly converges on the map around the impacted geographical area. The first well-documented case of this was after the 2010 Haiti Earthquake, where what few mapping products did exist were lost to the destruction of the office buildings of the national mapping agency. The international humanitarian responders converging onto the scene needed accurate maps to perform their work [125]. As depicted in Figure

3.1, hundreds of remote mappers from all over the world dramatically improved the digital map coverage of the affected areas in a matter of days by digitally tracing aerial imagery to build the map. This map then became the primary resource used in relief efforts [125]. Known as high-tempo events, these activations are of interest to the OSM community as a way to understand and communicate its impact. It is also of specific interest to crisis informatics researchers because of the rapid, large-scale convergence of “digital volunteers” from around the world, which demonstrates new forms of collective behavior [58, 105]. However, to begin asking questions of how this collaboration occurs, we must first create new tools to access and explore the “site of work”—the database supporting the map itself. This is the motivation behind the development of epic-osm—to create the first open framework for easily analyzing the large OSM dataset. Initially developed to support crisis informatics research, the use cases we will discuss are abundant and the framework provides great flexibility for all types of OSM research.

3.2 OpenStreetMap

Created in 2004 by students in the UK in response to restrictive licensing on geographic data [21], OSM has become the most widely used platform for “volunteered geographic information” [35, 42]. OSM is supported by a worldwide network of developers and volunteers committed to the open data values of the platform. Today, OSM has over 2.1M registered users, a small subset of whom are active editors [84], and 2.9B individual geographic points [88]. The website itself is a Ruby on Rails application on top of a PostgreSQL database. OSM incorporates an in-browser map editor and provides an API to interface with external tools.

3.2.1 OSM Data Structure

Six domain-level data types are found in the OSM database. Three of these primary objects construct the map itself: nodes, ways, and relations. Nodes are the most basic building blocks of the database and represent single geographic points. A way is composed of an ordered series of nodes, representing a line or polygon. A relation is a collection of nodes and/or ways, such as a

country border or a noncontiguous set of polygons. When an object is first created, its version is set to “1.” Any subsequent edit to that object will increment the version number; such edits also track the user who performed them and the changeset (discussed below) to which this edit belongs. Representations of nodes, ways, and relations are shown in Figure 3.2.

Beyond the primary map objects, the OSM database contains changesets, users, and notes. A changeset is the digital receipt associated with every edit to the map. Each time a user commits their edits to the database, a changeset is generated with information about the editing session. The changeset id is recorded with every map object it contains, allowing a user to view a complete grouping of all the objects edited within a single changeset.

A note object is a geographically-located comment that a user adds to the map. These notes are marked as either open or resolved and may contain a comment thread as users discuss the note. Notes document a discussion between users on how to represent a feature on the map, which can be another important element for understanding map creation.

The OSM user database contains the user display name, a unique user id, and the date on which the user created an account on openstreetmap.org. `epic-osm` makes use of the date when a user creates an account to determine their experience level with OSM. This facilitates comparison of behavioral differences between novice and experienced editors.

3.2.2 Tags

The descriptive, non-spatial characteristic of each map object within OSM is a set of tags. These are unrestricted key-value pairs that can be added to any map object. An active wiki supports discussion about best tagging practices for consistency within the map, and editing tools offer default tag suggestions, but there are no database rules to enforce tagging schema or structure. For instance, Table 3.1 shows some of the top keys and common values for OSM objects in the map for New York City at the time of writing. From this table we can observe that information regarding the building footprints and heights for NYC is of major interest to the subset of the OSM community mapping in NYC, and is therefore not representative of all cities within OSM. This

highlights the non-uniform characteristics of OSM contributions, calling for analysis tools that are capable of handling this dynamic nature.

Table 3.1: Top Tags for OSM objects in New York City.

Objects with tag	Key	Most-common values
66%	building	garage, house, school
64%	height	8.2, 8.0
13%	highway	residential
11%	name	(various)
2%	amenity	parking, bicycle parking



Figure 3.2: OSM Elements as Rendered on openstreetmap.org. Each element shows various aspects of possible metadata (truncated) associated with OSM elements. Data © OpenStreetMap contributors.

Map rendering software then uses these tags to properly display an object. For example, a way tagged with “highway”: “pedestrian” represents a path, while a way tagged as “building”: “yes” represents a building. Examples can be seen in 3.2

The importance of tags in OSM analysis cannot be overstated. However, given the open and dynamic nature of tags and tagging practices as the map evolves, an analysis tool must be robust to handle filtering by tags. For example, it is common for current OSM analyses to report summary statistics of OSM data by reporting on the number of new nodes added to the database. However, reporting that 956,725 nodes were added to the map in the month after the 2010 Haiti earthquake reveals very little about the manner in which the collaborative mapping was achieved. Filtering and sorting intelligently with tags instead can achieve results like this: *“308 users added 40,067 roads to the map and 162 users added 20,696 buildings to the map. 148 of these users were the same, adding buildings and roads.”*

Even this first-step expansion is a much richer summary of user contributions. The requirement, therefore, to develop a framework that is tag-aware is critical in understanding the creation of the map. As a result, epic-osm has advanced support for tags, and a mechanism for incorporating knowledge about the types of tags that the OSM community uses to create its maps (see Section 3.5). It can use this mechanism to find “all buildings” in a region even though different users tag buildings in different ways.

3.2.3 Planet Files

OSM provides its data in a common XML format via a RESTful API. Unfortunately for our analysis, this data represents the current state of the map, or the most recent version of the map objects, which, as we discussed above, is not of primary interest to those who study crisis mapping and the creation of the map itself. More useful are the “full-history planet files” that OSM strives to make available for download on a weekly basis. These files are bulk exports of the complete OSM database containing every edit to every object. Available in the Google protocol buffer format (PBF), these files are about 60gb in size, whereas the uncompressed history database in the OSM XML format is over a terabyte in size. While the PBF exports make obtaining the full history easier, working with the files requires specific knowledge of the file format and structure, and is computationally intensive to manipulate. This creates a requirement for an analysis framework: any

OSM analytical framework must be able to handle the processing of full-history PBF files, which will continually grow in size as the OSM community continues to work.

3.3 epic-osm Framework

This section describes the current implementation of the framework and its features. epic-osm has supported crisis informatics research throughout its development. This iterative, domain-driven approach to development has been shown to be useful when creating data-intensive systems [9]. As we refined our OSM research questions, the framework was adapted and refactored to support the processing of those questions. This agile development process has enhanced the usability and capabilities of the framework, thus supporting a main design goal which was to encourage the adoption and use of the framework among the many different communities interested in better understanding OSM data and mapping practices.

3.3.1 Features

The central object in our software framework is called an analysis window (**aw**). This is a spatio-temporal bounding box for a researcher's given geographic area and time frame of interest. All data analyses operate within the scope of an analysis window. An analysis window is thus defined by specific start and end times and a set of polygonal geographic bounding boxes; in addition, an analysis window includes the queries to be performed on that subset of the database and other metadata such as the the contact person and associated data directories. The framework does not limit the size or timeframe of an analysis window. However, we recommend working with a bounded analysis, especially during initial research. Since OSM is home to many different types of mappers with a great deal of variance around mapping practices, careful boundedness in space and time will yield results that are easier to interpret; one can then build on those results with progressively larger bounds, if desired.

3.3.2 Queries

Queries are associated with a specific analysis window and a specific temporal unit of analysis. Since every OSM object has a date and time associated with its creation, all queries return data sorted by these common features. A specific time unit for analysis can currently be set to hour, day, month, and year. These increments are then used to create time buckets for sorting the data returned. All queries return arrays of the form:

```

{start_of_aw, bucket_end, results},
{bucket_start, bucket_end, results},
...
{bucket_start, end_of_aw, results}

```

The first bucket will always start at the beginning of the analysis window and will end on the first unit of analysis after that. For example, if the unit were specified as “month” and the analysis window started at 2014/06/15, then the first bucket would include results from this date up to 2014/07/01. The second bucket would include all data for the range 2014/07/01 to 2014/08/01. This design decision ensures that the colloquial units of analysis make sense. If a user is looking to perform an analysis on months, then their results are returned in time buckets of the common month, not a grouping of 28 days starting from the beginning of the analysis window. In the event no unit of analysis is specified, then a query will return an array with one item:

```

{bucket_start, end_of_aw, results}

```

The framework is therefore designed to treat time as the default structure for analysis. This design decision supports the current practices in crisis informatics research and other observers of time- and safety-critical events. This makes our framework unique in comparison to other OSM data services that return the map data as it exists in real-time such as the official OSM API. These services are designed to deliver up-to-date geospatial data and map rendering, while epic-osm is designed for analysis of user contributions within a given period of time. Furthermore, this ensures

the results that are returned by queries represent individual edits, not necessarily distinct map objects. In other words, the same map objects with different versions may appear across multiple buckets of returned results. This allows users to explore the creation of the map by tracking changes to individual objects through time.

3.3.3 Conceptual Framework

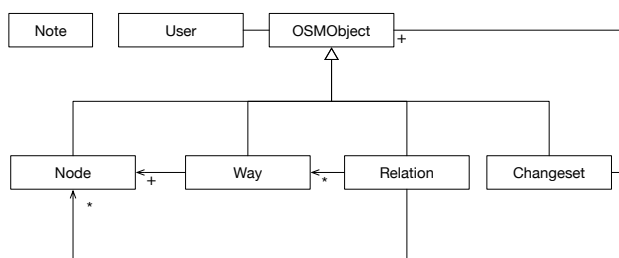


Figure 3.3: The Domain Objects of Epic-OSM

In Figure 3.3, we show the semantic relationships between the various data objects in our framework. The root class is `OSMObject`; it has attributes such as geometry, date created, user id, object id, and version number. Each `OSMObject` has an associated user who edited that particular version of that particular object. Nodes, ways, relations, and changesets are all subclasses of `OSMObject`.

The UML diagram shows that ways consist of one or more nodes and relations consist of some number of nodes and ways. While in practice this is true, our analysis framework performs extra work during import to ensure that each of these objects stands on its own. In particular, when importing a way, we traverse all of its associated nodes and embed the geographic information of those nodes in the way itself. We do the same thing for a relation, accessing all of its associated nodes and/or ways and embedding these objects into the relation itself. Therefore, when `epic-osm` performs a query on ways or relations, the query only has to access way or relation objects in `epic-osm`'s persistence layer.

The decision to perform this extra work during import was twofold: a) improving run-time

performance and b) reducing complexity during analysis. With respect to the former, we did not want to incur a run-time penalty during an analysis workflow spending time accessing a way or relation’s constituent parts. With respect to the latter, users may edit attributes of either the way or relation itself, or the nodes and/or ways associated with it. In such cases, the associated objects may not be aware of these changes. To properly reconstruct the object requires resolving the geometries based on dates and changeset ids and “burning-in” the geometry as it existed in that specific version of a way or relation. We determined it was best to absorb this computational cost just once during import. This type of tradeoff is common in the design of big data software frameworks.

Changesets contain information about the editing session such as a geographical bounding box of the extents of the user’s edits and the length of the editing session. Changesets themselves are unaware of the objects contained within the editing session, but the edited objects contain the changeset id of the changeset in which they were edited, allowing these relationships to be established after the fact. Note: although the semantics of our UML diagram allow changesets to include other changesets, this does not happen in practice: each changeset stands on its own and does not reference other changesets. Finally, our notes class contains attributes that allow OSM notes to be retrieved from the database and analyzed.

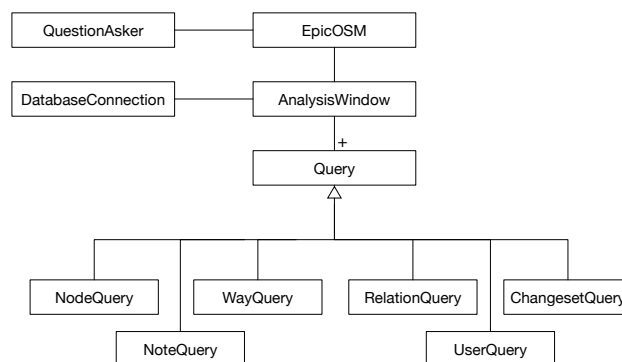


Figure 3.4: The run-time objects of epic-osm.

Figure 3.4 presents the framework classes that are used to perform an analysis at run-time.

An instance of EpicOSM acts as a controller for the analysis session, creating the requested analysis window, asking it to connect to the database, and invoking its associated queries. The QuestionAsker acts as a proxy for the user who invoked epic-osm, and can influence where the results of the analysis are stored, provide other metadata about the invoking user, or further process the results of the invoked queries. The classes in Figures 3.3 and 3.4 are connected because query objects return instances of the domain objects. Thus, node queries will return instances of nodes that can then be further analyzed.

3.3.4 Current Technology Stack

In keeping with OSM’s mission of open geospatial data, our framework is built on open source technologies. The logic of the framework is currently written in Ruby and is supported by a variety of open source libraries, developed by the greater OSM community and available on GitHub, for processing and importing OSM planet files. Given the importance of OSM object tags and their key-value structure, we chose to use a NoSQL document database, MongoDB, with inherent key-value support for persistence. Mongo stores each domain-level OSM object in namesake collections (i.e., nodes, ways, relations, etc.). Common fields such as date created, user id, changeset id, and geometry are indexed by MongoDB to speed up most queries; specific tags such as “highway” or “building” are indexed as well to support queries against these objects of interest.

3.3.5 Flexible Query Language

To support the goal of extensibility, our framework makes use of metaprogramming techniques [107] to avoid binding clients of the framework to a particular set of metrics and query methods. Metaprogramming facilities have been a part of programming languages for many years and include techniques such as “monkey patching” in Ruby, Python, and Javascript and key-value observing in Objective C. In epic-osm, we make use of a feature provided by the Ruby run-time system known as “method missing.” This feature is invoked whenever a client calls a method on an object that does not have an implementation of that method either within itself, its included modules,

or its superclasses. Though normally this situation would generate an exception that can crash a running program, Ruby's runtime instead calls the object again this time on a method called `method_missing`. It passes to this method a description of the method the client was trying to invoke. If that object has an implementation of `method_missing` and it can handle the processing of the failed call, the call will instead succeed. If it cannot handle the invocation, then, finally, an exception will be raised. In `epic-osm`, almost all querying-related methods are handled by `method_missing`. This convention allows us to handle a wide range of possible queries that can be expressed using a domain-specific language that our method parses at run-time and allows for new queries to be added in an incremental fashion. For instance, a call to the method `nodes_x_year` will be interpreted by an analysis window as a request to return all edited nodes that fall within its constraints, grouped by year. That same functionality (retrieving all nodes) can be invoked but have the data grouped in a different way by simply calling the method with a different argument after the 'x', i.e. `nodes_x_month` or `nodes_x_day`. Since the desired structure of the results is defined by the name of the function, arguments passed to the queries are for further filtering of the results and are passed through `epic-osm` to MongoDB unaltered. This allows users to take advantage of MongoDB query capabilities in their own `epic-osm` queries. For instance, the query: `ways_x_month(constraints: { "tags.highway" => "pedestrian" })` will return every version of a way which represents a pedestrian footpath which was edited or created within the analysis window, grouped into months. In this example, `epic-osm` handles grouping the results of the query into months while MongoDB finds all of the relevant ways while ensuring that all returned ways have a tag called "highway" with the value "pedestrian." For improved performance, users can externally index the underlying MongoDB collections to support common queries.

3.3.6 Question Modules

As shown in Figure 3.4, query objects target a specific type of domain object: Node queries return nodes while note queries return notes. This modular design allows analysts to focus their queries on just the domain objects they need. However, many questions require querying multiple

types of objects. `epic-osm` provides this type of query via the use of Ruby’s support for modules. A specific module is created that contains all of the code that is needed to query across multiple types of domain objects; this module exports a single method that can then be invoked on an analysis window to execute the query at run-time. As an example, consider the need to ask an analysis window about the number of schools that were edited within its geo-temporal bounding box. For this particular query, it is important to check both nodes and ways to find all possible schools “hiding” in the map. According to OSM’s community guidelines, the best practice for marking a school on the map is with the tag: `{\amenity": \school"}`. However, the actual OSM object that should contain this tag is not strictly defined. Mappers are encouraged to use an area (a polygon comprised of a closed way) that outlines the school’s geographic footprint; however, the Wiki also states that mappers can “place a node in the middle of the site if [he or she is] in a hurry” (wiki.openstreetmap.org/wiki/Tag:amenity=school). As a result, the question of “how many schools were mapped during the analysis window” becomes far more complicated than a simple query for objects with the school amenity tag. Instead, one must query both the ways and the nodes collection, identify distinct versions of interest and then resolve any geographic overlap in which both a node and a way mark the same school. To illustrate this, Table 3.2 shows the results of this query for the 2010 Haiti Earthquake across different types of `OSMObjects` and shows how the numbers change when accounting for geographic overlaps:

Table 3.2: Differences in use of “school” tag

Query: “amenity”: “school”	Nodes	Ways	Geo-Unique
Added	145	41	166
Edited	32	27	57
Unique Sum	146	52	173

Ultimately, one may conclude that 173 schools were edited in Haiti within OSM in the month following the 2010 Haiti Earthquake. As mentioned, these more complex queries are isolated into Ruby modules—that `epic-osm` calls question modules since they contain all the code needed to ask

a particular, complex question—that are then accessed via a single method with all support code cleanly hidden away from the main classes of the framework. If OSM community guidelines change for a particular tag, just the code in the relevant module has to change in response. If one analyst has a broader (or more narrow) definition of what constitutes a particular entity, they can create their own module for finding instances of that entity. These modules can then be easily shared and plugged into any instance of the framework. This is important because defining questions such as “how many schools were edited” as shown above are not immediately straightforward, so turning that question into a single method within a reusable module ensures that all users abide by the same rules when querying the data. This modular design has also affected the development process by encouraging developers to write many questions in separate modules and then refactor common helper functions into the analysis window to make them available to all other question modules, thereby making the functionality provided by the core objects more powerful over time.

3.4 Implementation

Above, we presented the concepts and capabilities contained in the epic-osm framework. Here, we discuss how we have created a set of tools that use the framework and some of their implementation-related concerns. The advantage of creating a framework that can be incorporated into a wide range of tools is the large number of analysis use cases that can then be supported. Our initial set of tools handles the processing of a large amount of OSM data via the use of batch processing. First, command line tools are used to download and import OSM history data into MongoDB. Second, an input file is used to specify the parameters of a desired analysis window along with the desired queries. Third, a command line tool was created to read the input file, create instances of the objects shown in Figure 4, and kick off the processing of the specified queries. The output of that process is a directory of easily read JSON files. This straightforward set of tools and components can be used to process gigabytes of map data, ensuring scalability. It is important to note that this same framework can be incorporated into a web application and be used to dynamically query MongoDB in response to user commands; indeed, we plan to develop

such tools and, as we discuss later, we have already made changes to the framework to allow for more real-time processing of OSM data by analysis windows. Next, we discuss a few additional implementation-related concerns in more detail.

3.4.1 Persistence Layer

As mentioned above, MongoDB is used to store OSM history data and to perform the bulk of the work with respect to the queries that users specify. Storing the history data in this way allows users to have the flexibility to easily track changes to their queries over time. For example, a user may define an analysis window for their hometown over the past month. With each new month, they can create a new analysis window with the same geographical bounds, but with new start and end dates. As the user learns more about their data through defining new questions, persistence of previous analysis windows allows them to rerun those questions without having to re-import the underlying data. Furthermore, using a database ensures that the size of objects referenced by an analysis window can scale beyond the physical memory constraints of a user's machine. While MongoDB was selected for its ease of use and deployment, any key-value store or document store could be used as the persistence layer for epic-osm.

3.4.2 Output

In an effort to support interoperability via many types of analysis and by not forcing OSM researchers to use a single tool, epic-osm writes output to a pre-defined file structure: a series of JSON files. These files can then be easily parsed and visualized by a variety of libraries and analysis tools, leaving the visual inspection and analysis environment open to a user's preference. Currently, we build a static website from these JSON files that can be used to view and easily share the results of the analysis but many other options for how to make use of these files from more interactive web-based dashboards to network analysis toolsets are being pursued, both by our group and the OSM community. These multiple pursuits validate our design decision to create a common output directory of single JSON files.

3.5 Use of the Framework

At the time of this writing, our framework has supported academic research by our group as well as OSM community members. The initial release was in support of our post hoc research on the growth of the OSM organization between 2010 and 2014 in response to two distinct humanitarian events [105]. This required the processing of a month’s worth of historical OSM data for each event, consisting of edits by nearly 500 users and 1500 users, respectively. Since then, the framework has been available on GitHub and has been forked, contributed to, and adapted to support real time analysis and statistics of specific OSM mapping events.

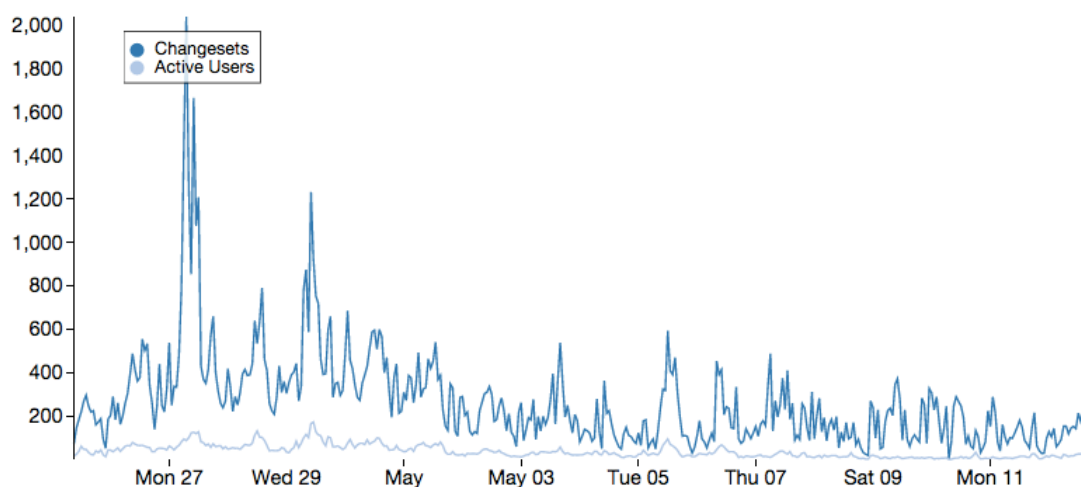


Figure 3.5: Count of OSM Changesets and Users. Graph shows the by-hour contributions to the map of Nepal after the April 25, 2015 earthquake.

For example, MapGive, a mapping initiative sponsored by the U.S. State Department, used epic-osm to visualize results of a competition between two universities to see which could create more data (mapgive.state.gov/events/mapoff). Additionally, it was deployed to monitor the first-ever mapping event at the White House (mapgive.state.gov/whmapathon). Another project, moabi.org, is also running an instance of the framework to monitor the mapping of logging roads in the Congo (loggingroads.org). The statistics are used to populate a “leaderboard” showing the highest-contributing users.

3.5.1 Nepal Earthquake Deployment & Improvements for Real Time Analysis

On April 25, 2015 a 7.8 magnitude earthquake struck central Nepal, killing over 8,500 people and destroying over 500,000 homes. Due to previous OSM work in the country [123], the city of Kathmandu was already mapped in detail. Yet many of the affected rural areas outside of Kathmandu were not well covered on the map. In what is believed to be the largest convergence of OSM mapping activity to date, over 7,000 contributors from all over the world mapped roads, buildings, and other features.

Our team deployed an instance of epic-osm immediately following the earthquake, which proved to be a valuable test case. A real-time import module developed by an epic-osm contributor that interfaces with a newly available OSM changeset streaming service (github.com/osmlab/osm-meta-util) supported this instance. Figure 5 illustrates this impressive convergence as tracked by epic-osm, showing the number of users editing and the number of changesets created per hour for the weeks following.

However, tracking this huge mapping activity in real-time exposed a problem. Designed to be a static snapshot in time that reads historical edits from a database, the analysis window could mimic near-real time results by running new queries every 10 minutes with bounds that spanned the time from the event to the current time. This solution worked well until the second day when the database had grown so large that the time it took to run the queries was longer than 10 minutes, creating a backlog.

To resolve this problem, we added a new feature: a rolling analysis window that would update the analysis window's constraints at each run to start at the top of the hour and end at the current time, thus never querying more than an hour's worth of data. These results were then output to separate directories, which could be iterated over to create the new totals. As a result, the framework was able to support a website providing visualizations of edits over the past hour. This site received over 1,700 unique visitors from 79 countries in the first week and was the OSM community's primary tool during the response for tracking its activity. This ad hoc solution worked in this particular

use-case, but more importantly, exposed the weaknesses in the framework for similar use cases, which have since generated great interest in the OSM community.

3.6 Extensibility and Future Development

The desire to support both historical and real-time analysis of user contributions to OSM is strong across both industry and academia. At a June 2015 OSM conference (The State of the Map US) held in New York City, OSM users from the Red Cross, the US State Department, and three digital cartography-oriented start-up companies held a Birds-of-a-Feather discussion on the need for developing and supporting analysis tools such as epic-osm.

3.6.1 Stream Processing

The real time tracking of mapping activity in response to the Nepal earthquake identified a very powerful use-case for epic-osm that will significantly influence the next development iteration, specifically the ability to process the edits to the map as an incoming stream directly, instead of first importing to a database and extracting distinct time chunks. We will use contemporary big data solutions such as Apache Spark and its streaming capabilities to achieve better real time performance.

3.6.2 Database Improvements

With an emphasis on stream processing, the role of the persistence layer will also change in the next iteration. New user-level models will need to be developed to track mapping behavior, while the persistence of the individual object edits should also be preserved for later analysis, should users desire to perform new queries post-event. Alternative geo-spatial database technologies will be explored as well, which may improve query performance for geographic oriented analysis, such as “how many kilometers of a road did a particular user map?”

3.7 Conclusions

We have presented and discussed the design of epic-osm, the first full software framework to support the analysis of volunteered geographic information contributed to OSM. The framework was initially developed to support crisis informatics research surrounding the production of map data in two major crisis events, and has continued to grow and gain exposure to a larger community of developers and mappers alike, with hopes of allowing the entire OSM community to better reflect on its production of open geographical data. Our framework makes use of a number of techniques to efficiently handle large volumes of OSM data and serves as an example of how to design frameworks for data-intensive software systems. We believe that our framework, our lessons learned from initial deployments, and our iterative development approach, which is deeply grounded in empirical knowledge of a target domain—in this case, crisis mapping—will be of use to other designers and researchers of data-intensive software systems.

Acknowledgments

This material is based upon work sponsored by NSF Grants IIS-0910586 and IIS-1524806. We thank the OSM community for their involvement, particularly Mikel Maron of MapGive, Humanitarian OpenStreetMap Team and Kathmandu Living Labs.

Chapter 3 Epilogue: Epic-OSM Implemented

As mentioned in Section 3.5, Epic-OSM was used for the data analysis that visualized the improved coordination among humanitarian mappers in OSM through the implementation of the Tasking Manager from the Humanitarian OpenStreetMap Team [105]. This is shown in Figure 3.6 below:



Figure 3.6: Differentiating individual user contributions to OSM by color shows where each mapper was active. This graphic shows that during the mapping response to the 2010 Haiti Earthquake (left), contributors were all mapping in the same region, often right on top of each other. Three years later, with the introduction of the Tasking Manager, the mapping in the Philippines in response to Typhoon Yolanda (right) has more separation between mappers, as noted by the more "chunky" blobs of color. This is especially notable because there were about three times as many mappers active in this event. Each image represents 116k nodes for consistency. Creating this visualization was part of my contribution to [105], Data ©OpenStreetMap Contributors.

3.8 Time Series Analysis

The spatiotemporal bounding and binning capabilities of epic-osm described in Section 3.3.2 allow for Time-Series analysis of OSM contributions. This feature was used by Kogan et al. to construct snapshots of contributor-interaction networks for 8-hour bins of activity in the mapping response to the 2010 Haiti Earthquake [61]. Analyzing the structures of these networks further uncovered inter-mapper interactions that were qualitatively investigated. In this way, epic-osm

supported a mixed-methods study of coordination and collaboration during disaster mapping activities. The new OSHDB and OHSOME API built on top of it now offer similar capabilities for spatiotemporal queries that can work with larger regions [112].

3.9 Sharable Data Visualizations with osmdown

As described in Section 3.4.2, Epic-OSM produces JSON files with a specific schema based on the duration of the timebins as defined in the analysis window. A visualization tool called osmdown was developed to ingest a custom form of markdown, denoted with the suffix `.osmdown`.² An osmdown file has three primary attributes: First, YAML front-matter may define global variables such as the location of the JSON files that were produced by Epic-OSM. Second, the main body consists of markdown formatted text. This allows analysts to easily insert their interpretations of the data in plain text. Third, any code in between the standard markdown denotation for code (````), is evaluated at compile time and the results are added to the output file. This yields a static HTML document that references the JSON output from Epic-OSM.

Since the output is a single HTML file, it can be easily shared on the web with any static-website host. The page can be rendered once to share results from historical analysis, or can be continually rendered as needed to support more real-time tracking. Additionally, the JSON output from Epic-OSM can be read in two ways. First, the files can be loaded and parsed at compile time to inject the values directly into the body of the web page. Alternatively, interactive graphics driven by d3.js can load the JSON output from any accessible directory. If Epic-OSM is continually running and producing output to a publicly accessible directory, then the statistics page is able to reference these continually updating files directly and users will see the latest data reflected in the interactive charts embedded in the osmdown output. Figure 3.7 presents a screenshot of the osmdown page we built to track the mapping response to the 2015 Nepal Earthquake.

² osmdown (github.com/project-epic/osmdown) is a portmanteau of "OSM" and "Markdown" and built on top of Tom Yeh's VizDown (github.com/doubleshow/visdown) engine.

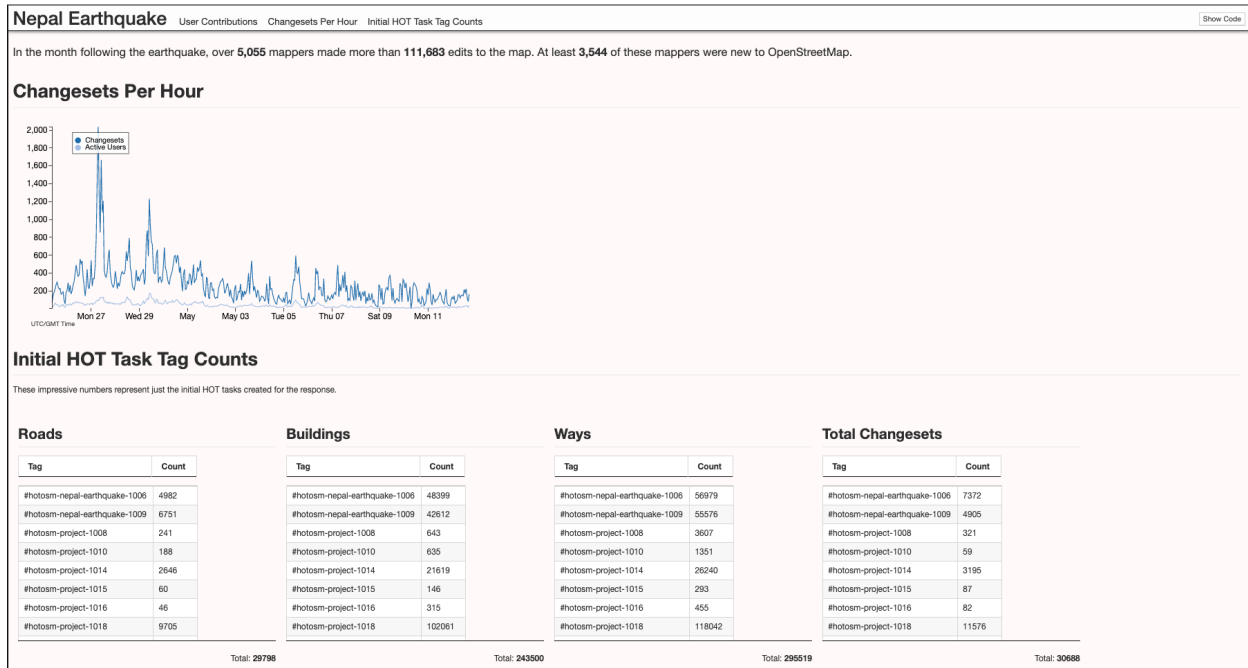


Figure 3.7: Screenshot of the osmdown webpage for the 2015 Nepal Earthquake

At the upper-right corner of Figure 3.7 is a button labeled **Show Code**. This button will toggle the code blocks that are embedded in the output. This helps increase transparency and reproducibility, allowing viewers to see how the figures are generated. This particular page includes an embedded graph of contributors and changesets that is populated by referencing the JSON output from epic-osm and built by d3.js. As Epic-OSM continues to run in the background, this graph will load the newest data when the page is refreshed. Additionally, the summary statistics embedded in the text and the tables showing road and building counts by HOT hashtags are static HTML tables that are generated from the JSON output when osmdown runs. Therefore, to update these tables, both Epic-OSM and osmdown need to continually run.

3.10 2015 Nepal Earthquake

As reported in Section 3.5.1, Epic-OSM ultimately failed to keep up with the number of contributors active during the Nepal Earthquake when implemented as a real-time analysis tool. As Figure 3.8 shows, the webpage was viewed by thousands of people following the earthquake,

especially during the mapping response.³ In order to keep the data fresh, Epic-OSM and osmdown were set via a `cronjob` to run periodically. At first, these tasks were running every five minutes. By the end of the first 72 hours, however, there was too much data to complete the analysis tasks in that time window. Keeping the tracking page running became an arduous task as I continually patched the infrastructure to keep it running. This involved creating a rolling analysis window as introduced in section 3.5.1. At first, this solved the problem by not recalculating the statistics for the initial high volume days, but it also failed to scale as we were left with a massive number of output files to be read into osmdown for the stats page regeneration. The webpage continued to update hourly until June 2, 2015 when the majority of the editing activity slowed and we turned off the Epic-OSM server. I then moved the page to a static host where it could continue to be viewed, but no longer actively update. As Figure 3.8 shows, the page continued to get a handful of visitors in the months and years to follow.

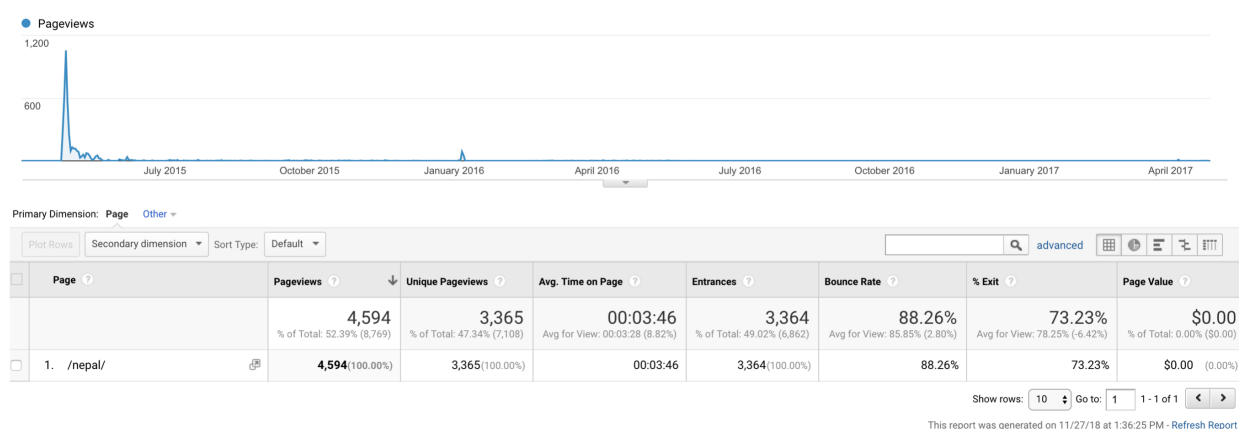


Figure 3.8: Public engagement with our Nepal Earthquake Stats page

Ultimately, Epic-OSM proved useful for spatiotemporal analysis of discrete events, specifically for Time-Series investigation of past-editing activity. The failures and complications to scale during the 2015 Nepal Earthquake response taught me two things: First, Epic-OSM cannot gracefully handle real-time analysis (not an initial requirement in the first place, but a desire nonetheless).

³ At one point a WIRED article reported the same number of buildings to date that our page showed (presumably using our page as their source).

Second, the nearly 16x increase in number of mappers from the 2010 Haiti Earthquake response showed that these analysis systems need to be designed with scalability in mind from the beginning because the next major disaster-mapping activity will likely have even more contributors.

While this event showed the desire for real-time analysis, I limited future scalability concerns to the volume of OSM data, both in spatial volume (the entire planet) and in temporal volume (the full history of the map), not about the speed in which we can consume, parse, and process the data. Providing real-time analysis of OSM editing activity is clearly important, but out-of-scope of the particular systems presented here which were built in support of historical OSM data analysis. In implementation, a continually evolving analysis system with no guarantees of uptime is unsuitable for the global disaster response community to rely on for disaster mapping analytics. A system to support this community needs to be stable, always available, and maintained throughout its deployment. This added complexities and dependencies beyond what I could support while designing a system for historical, planet-scale research. This is not, however, out of scope for future work, especially in collaboration with the development of other systems like `OSMesa` that are far more optimized for real-time analysis.

Furthermore, Epic-OSM did not address many of the smaller concerns put forth in Chapter 2. This was not a deliberate oversight at the time, but rather a complete lack of enlightenment to the true complications of measuring OSM contributions that I was slowly discovering. With these developments, I went back to the drawing board on how to approach OSM data analysis. This started with re-imagining how OSM data is represented in the first place by embracing new analysis technology that Mapbox had just released: Vector-tile based OSM data analysis.

Part III

Transition to Vector Tiles

Chapter 4

Representations of OSM Data

Most users interact with OSM data only as a map. A map, however, is just one way to (visually) represent the OSM database; it is a data visualization of OSM objects rendered with respect to cartographic rules. While this representation of OSM data is the most seen, used, and important, there are other ways to represent the objects contained in the OSM database that are more optimized for contributor-centric data analysis. This chapter discusses these different formats and the benefits that certain formats can provide for data analysis, addressing the complications introduced in Chapter 2. Of most value to contributor-centric analysis is the ability to accurately account for the change and evolution of the map at the individual object level. This enables us to identify and classify the individual edits according to Table 2.1 presented in Chapter 2.

In its most raw form, OSM is a relational database of nodes, ways, and relation elements. The OSM XML format represents the data exactly as it exists in the database. Figure 4.1a shows OSM XML for the node with id of 1. This iconic, very first node to exist in OSM, has been modified 20 times and is currently located in the UK in Greenwich Park as part of a larger way that represents the Prime Meridian of the World, depicted in Figure 4.1b.¹

Next, I will describe at a high level the trade-offs of converting OSM data into other formats and the types of common conversions. I borrow the *lossy* and *lossless* terminology from data compression to describe these conversions.

¹ People have tried to move this iconic first node to position [0,0] on the map to represent a weather Buoy that exists at this location. At one point the node was moved to represent a restaurant. It has been deleted and restored seven times.

```

<node id="1"
  visible="true"
  version="20"
  changeset="64040630"
  timestamp="2018-10-31T10:20:19Z"
  user="SomeoneElse\_Revert"
  uid="1778799"
  lat="51.4779481"
  lon="-0.0014863"
/>

```

(a) OSM XML of Node #1. Only node elements contain geographic information (highlighted)

```

<way id="268533450"
  version="18"
  timestamp="2019-06-06T01:52:33Z"
  user="bjankuloski" >
  <nd ref="1"/>
  ...
  <tag k="loc_name"
    v="Prime Meridian of
      the World" />
</way>

```

(b) OSM XML of Prime Meridian (Way) referencing Node #1 (highlighted). Full attributes truncated for space.

Figure 4.1: OSM XML representation of the Node element 1 and the Way element that references it. Each of these objects has its own metadata and editing history.

4.1 Topologically Lossless OSM Data Formats

A lossless schema is one in which every attribute of the initial database is maintained, along with the topology. This requires maintaining the identity of every element (node, way, relation) as well as the topological relations between them. The OSM XML format just shown in Figure 4.1, typically denoted by the `.osm` suffix does this. The protocol buffer format (PBF), often denoted as `.osm.pbf` is a binary version of this format. Much more compact than the XML, the PBF format is optimized for downloading or transferring OSM data. Multiple versions of OSM elements (historical versions) are valid in these formats, but there is no concept of a minor version (Section 2.2); these still need to be computed from the raw data present in the file. Objects in these files are unique based on type, id, and version number.

Because these formats preserve the topological relationships between OSM elements, there are multiple steps required to reconstruct the OSM objects themselves, preventing them from being read line-by-line. This is something that all OSM data-processing tools are equipped to do, but can be resource intensive. To handle the geometries associated with ways and relations, OSM data parsing utilities need to first construct a node location cache (typically in memory) that can be

subsequently referenced for point locations when the ways and relations are read. For city-sized and small-country extracts of OSM data, this can be done with modest personal computers, but for larger amounts of OSM data, this requires non-trivial computing infrastructure. The `osmium`* library and command-line tool is a robust utility maintained by the OSM developer community to perform these tasks. Written in C++ with bindings for Python, Node, and Java, this open-source library is found in many OSM data processing tool chains.

Lossless database exports are made available at regular intervals and include access to the full history of the every object on (or deleted from) the map. These exports of OSM data are often the primary data sources for OSM data analysis. Therefore, ingesting these files is the first step in an analysis workflow, especially if one needs any editing metadata or historical information. The next step is likely to convert the OSM data into something more workable and familiar, typically a more common spatial data format. I consider these to be *topologically lossy* with respect to the OSM node, way, and element data model.

4.2 Topologically Lossy OSM Data Formats

Converting OSM data into another more traditional geospatial format comes with some level of loss, especially with respect to the OSM topology. The largest change that happens is removing billions of stand-alone nodes from the database that exist only to mark the vertices other elements. This is done by embedding the locations of these nodes as vertices in the geometry of the parent objects. Then, if converted back to a node/way/relation topology, these points would become nodes, but not have the same attribute information such as the ID and last edit metadata (version, changeset, user, etc). While this information could be maintained in the metadata of the way, this would bloat the object—neutralizing many of the advantages of the more compact, efficient representation.

Consider, for example, the intersection example from Chapter 2 again. This node representing the traffic signal was a part of at least 2 ways that intersect. When this intersection is converted to a more standard format such as GeoJSON, three objects will be made: 2 `LineStrings`, each of which

will contain the coordinates of this intersecting point, and a single `Point` that marks the traffic signal. While no map-level data has been lost (all of the data required to render this intersection remains), reconstructing the exact OSM topology would require tracking node references and/or more expensive geographic calculations to rebuild a topological relationship. While these are both technically possible, they are unnecessary because converting OSM data into a more common format for analysis is typically a one-way conversion with a specific purpose. In practice, I have found tracking just OSM way IDs in order to reference the original element if needed is enough. To this end, this conversion is *lossy* in that all of the original information is not always maintained, but the new object is still a complete, accurate representation of the original with editing metadata left intact.²

If, however, the OSM data schema were to ever evolve—a topic that is brought up now and again, such as a recent conference talk [133]—a likely change would be including geographic coordinates in certain types of ways, effectively performing the same topological flattening described here to conform to the more common geographic objects: `Points`, `LineStrings`, `Polygons`. In terms of OSM data analysis, this change would allow for more accurate versioning and solve the issues of tracking minor versions introduced in section 2.2. I bring up the concept of topological lossy conversions here only to highlight the caveats, not to argue for a change in the OSM data model. Today a variety of tools exists to accurately and efficiently perform these conversions, tools that did not exist five years ago when I started investigating these problems. For now, these tools (`osmium`, `OSMesa`, `osm-wayback`, `ohsome`)* are critical first steps in data analysis that alleviate the need to perform these conversions upstream. A main feature of these tools is their ability to keep all attributes of an OSM object intact. Object attributes are critical to analysis and can be affected by many types of conversions between geospatial data formats.

² In the same way that a JPEG image is a lossy conversion from a RAW image. For the purpose of the digital image, the JPEG is complete and not compromised. It does not, however, contain all of the information from the camera's sensor like the RAW file.

4.2.1 Attribute Lossy Conversions

The limitations of some traditional file formats (such as Shapefiles) may result in the loss of attribute information when converted. First, there may be character limits on the length of column names that will truncate a number of OSM keys. Second, the open key-value tagging schema cannot be converted to a table format with a finite number of columns and capture all-possible key-value attribute pairs.³ In application, the namespace of commonly used tags is relatively small, making this a minor concern for the majority of use cases. For example, if someone wanted to calculate road coverage, they would need only the `highway` attribute from the OSM way. Likewise for analysis of buildings, knowing if the `building` tag exists is enough. Assumptions like these, however, will likely introduce problems in downstream analysis if all attribute information is not maintained.

Editing metadata is also often discarded in lossy data in the name of efficiency. Similar to the loss of topology, this is typically not a problem for the majority of analysis, especially extrinsic data analysis as discussed in Chapter 7. The loss of these metadata in contributor-centric data analysis, however, is crippling. Fortunately, there are formats that alleviate these concerns. To safely avoid truncating the attribute space, it is best to use a format without a predefined schema. As discussed in the Epic-OSM framework (Chapter 3), JSON is a powerful, flexible data format choice. More specifically, GeoJSON, the standard for representing geospatial data in JSON is the ideal, commonly understood spatial format for representing OSM data.

4.3 GeoJSON + OSM

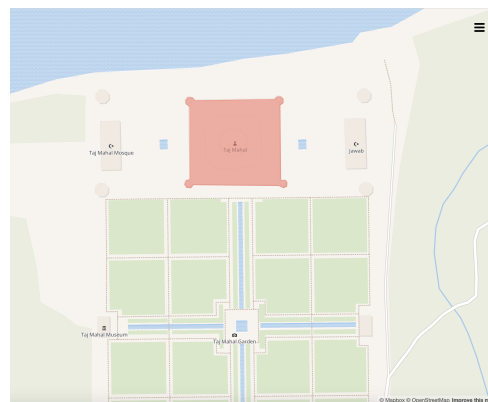
The GeoJSON format allows for human-readable, attribute-complete representation of OSM objects that can be easily parsed and understood by a variety of data processing and analysis systems. Not topological and internally referencing like the OSM data model, the file can be read line-by-line without the additional overhead of a node location cache because geometries are embedded into the objects themselves. These files can then be processed in smaller sections at a time, requiring less

³ This was a primary reason for choosing MongoDB when developing EpicOSM (Chapter 3). Today, however, PostGIS has support for key-value stores.

```

{ "type": "Feature",
  "properties": {
    "@id": "way/375257537",
    "addr:city": "Agra",
    "addr:country": "IN",
    "addr:state": "Uttar Pradesh",
    "building": "yes",
    "height": "73",
    "historic": "monument",
    "name": "Taj Mahal",
    "start_date": "1632",
    "website": "https://www.tajmahal.gov.in/",
    "@timestamp": "2019-03-27T02:10:34Z",
    "@version": 23,
    "@user": "pizzaiolo",
    "@uid": 1772368
  },
  "geometry": \{"type": "Polygon", "coordinates": [[
    [78.042627, 27.1753829], [78.0426626, 27.1754064],
    ... (21 more points) ...
    [78.0426359, 27.1746263],[78.0426270, 27.1753829]]] }

```



(a) The Taj Mahal in GeoJSON. Retrieved from Overpass Turbo, all properties truncated for space (overpass-turbo.eu).

(b) The rendered GeoJSON representation of the Taj Mahal.

Figure 4.2: As GeoJSON, The OSM object that represents the Taj Mahal is a Polygon with 24 distinct points held in the coordinates array within the highlighted `geometry` attribute. In the standard OSM schema, these 24 points would be individual nodes. Data ©OpenStreetMap Contributors

computational resources, or broken into smaller chunks for more efficient, distributed processing. Today, GeoJSON is a common, standard geospatial format, especially on the web. The majority of geospatial data processing and visualization tools support it. While it might not be the most space-efficient representation, it can be easily encoded into more compact formats. Ultimately, the open-tagging schema makes GeoJSON an obvious choice to represent attribute-complete OSM data.

Figure 4.2a shows a *GeoJSON Polygon Feature* object representation of the OSM way element that represents the Taj Mahal in Agra, India (rendered in Figure 4.2b). The metadata for this version includes the details about the most recent edit to the object on March 27, 2019.

4.3.1 Support for Object Histories

As a data format, GeoJSON does not offer any more support for object histories than the standard OSM data model. It is possible to have more than one version of an object, but these versions remain unaware of each other. GeoJSON does, however, greatly simplify tracking versions by removing inter-element references and can easily support the concept of a minor version with

the addition of an attribute to a standalone object. As Figure 4.2 shows, in the OSM data model, the way element references 24 distinct nodes. Each of these nodes then has its own editing history. Understanding how these histories relate to the history of the way element creates the *minor versioning* problem of Section 2.2. Embedding the temporally appropriate coordinates of these node elements into GeoJSON representations of the object produces 25 distinct versions of the Taj Mahal: Each with a distinct version and minor version property. These 25 distinct versions are comprised of the 23 versions of the way element along with two minor versions resulting from node changes.

In contrast, the complete history of the Taj Mahal in OSM consists of 23 way elements and 29 node elements. Three of the 24 nodes have previous versions, making for 29 node elements in total. Furthermore, the metadata (or at least `timestamp`) for each of these 29 node objects must be maintained in order to know how (when) they relate to which of the 23 versions of the way element. Important to contributor-centric analysis: These minor versions were created by two mappers that do not show up in the other 23 versions of the way element. Those 23 versions were created by a different group of 15 mappers (some edited more than once and other edits were from bots). If we do not count minor versions, these additional two editors are discredited from participating in the mapping of this iconic object.

In sum, the the complete history of the Taj Mahal can be represented either as 25 GeoJSON features or 52 OSM elements that internally reference each other, requiring additional processing to reconstruct. One criticism of representing each version as a distinct GeoJSON feature is that there are really only three unique geometries associated with the Taj Mahal. 25 distinct GeoJSON features then creates a lot of repetition. To this end, I encode the geometric history as `TopoJSON*` to create a different type of topology from nodes/ways, but to shared arcs between objects. Section 5.6.1.1.⁴ discusses this further.

⁴ Successfully implementing this encoding was a pivotal moment in the successes of my analytical approaches that was informed by conversations with University of Colorado Boulder Geography Professor Carson Farmer.

4.3.2 Creating GeoJSON from OSM

Conversion to GeoJSON requires a significant number of rules to best determine how an object should be converted. Node elements that have none of their own attributes, for example, should be excluded from the output. A node without any of its own attributes exists solely to define a vertex of another element. Its coordinates will be embedded into the parent element's geometry. However, many nodes have relatively meaningless tags, such as a `source` tag that declares where the data originally came from or various import related artifacts that have yet to be deleted.⁵ Likewise, these elements exist only to define vertices within other elements and should not be converted into their own Point features. Handling this nuance is not computationally difficult, but it requires a well-curated and comprehensive list of exempt-tags to be compared against.⁶ Well-curated means that this list must continually grow and adapt as more of these types of nodes are added to the map. Similarly, other well-curated lists must exist to define the target geometry types for specific OSM elements.

While common map objects like highways and rectangular-shaped buildings may convert directly to LineStrings and Polygons, these must be defined by a set of rules. Utilities that handle these conversions like `osmium-tool` and `OSMesa` have large lists of presets to define these conversions.⁷ These get more complicated with relation elements like administrative boundaries, coastlines, and intricate buildings which get converted into MultiPolygon features. Though these conversions can be difficult and computationally demanding, utilities like `ohsome`, `OSMesa`, and `osmium-tool` are capable of performing them at the full-planet scale when provided with enough resources. The ability to handle historical versions, however, only came about within the last two years and is unique to `ohsome`, `OSMesa`, and `osm-wayback`. Ultimately, converting OSM data to GeoJSON is made possible today by a number of tools and while it can be resource-intensive, bulky, and depends on the community to maintain a set of evolving tag-based rules, there are analytical advantages in

⁵ Not that attribution is not important, it can be just as effectively declared once on the main object. It does not need to exist on every vertex.

⁶ One such list includes over 50 keys (github.com/geotrellis/vectorpipe) that catches nearly 300M points.

⁷ A list of way tags that should be represented as areas (Polygons) used by the iD editor is here: github.com/osmlab/id-area-keys. `OSMesa` uses this list as well.

the ability to accurately represent any version of an OSM object as a standalone GeoJSON feature. The largest disadvantage of GeoJSON is a current inability to represent some OSM elements that describe abstract relations between other OSM elements.

4.3.2.1 Abstract Representations

A primary feature of the OSM data model is the ability to represent abstract relationships between two elements, such as a turn restriction. There are over 1M of these relations in the OSM database.⁸ Turn restrictions reference multiple geometries and describe a relationship between them, such as *there are no right turns from this road onto that road*. When these roads are turned into GeoJSON, however, there is no obvious way to represent this turn restriction: which road does it go on? How does routing software need to relate these objects? These remain open questions in terms of implementation, so default behavior is currently to ignore these types of relations. GeoJSON conversions are then always incomplete because these abstract objects are not retained on account of not having a representable geometry. Section 5.3.2, however, discusses an implementable workaround for contributor-centric analysis.

The rest of the work presented here relies on first converting OSM data to GeoJSON. I primarily use the `osmium-tool` for this because it is a robust command line utility that can be easily invoked as a processing step in any workflow.

4.4 GeoJSON + Vector Tiles

Vector tiles are stores of geospatial data that are typically used for delivering data across the web, and therefore optimized as such. In contrast to raster-tile servers which will return a pre-rendered images that the browser will tile together to make a webmap (known as a slippy map), vector tile servers return vector data, tiled in the same manner. Rendering the vector data into a map is then done by the client browser. The primary advantages of vector tiles include smaller tile sizes (vector tiles are often lighter than rendered images of a region) and flexibility in how the vector

⁸ taginfo.openstreetmap.org/restrictions reports 1.1M relations with a `restriction` key on 2019-7-3

data is rendered. The browser has control of how the map object gets displayed and can therefore be instructed to render a map of any cartographic style. While vector tiles are primarily used as the back-end for interactive maps, they can be extended to hold an arbitrary amount of geospatial information in a spatially-aware manner. The analysis techniques described here exploit this ability.

All of the work discussed here uses vector tiles of the Mapbox vector-tile-spec format.⁹ Based on the Google Protocol Buffer format, Mapbox vector tiles (typically denoted with the suffix `.mvt`) are supported by Esri software, QGIS, and the majority of open source mapping software frameworks. Further use of the term vector tile refers to geospatial data encoded in the Mapbox vector tile format. These vector tiles are capable of representing global quantities of GeoJSON data efficiently and there is a well-supported and documented open source command-line-utility: `tippecanoe`* that can convert massive quantities of GeoJSON into vector tiles, specifically SQLite files containing vector tiles known as `.mbtiles` files. This work heavily relies on this utility.

The primary purpose of encoding GeoJSON into vector tiles for the work presented here is to create a persistent, spatially indexed data store of geographic features with arbitrary amounts of metadata. As single files on disk, these tilesets do not require any major infrastructure to persist (no running database) and are easy to share. The next chapter will explain these benefits of these files in a data analysis workflow in terms of reproducibility, rapid iteration, and data sharing.

⁹ docs.mapbox.com/vector-tiles/specification/

Chapter 5

OpenStreetMap Analysis Vector Tiles

While not the only possible pipeline to produce vector tiles from OSM data, the pipeline consisting of the open-source utilities just introduced in Chapter 4 allows for a planet's worth of geographic features to be converted into vector tiles. Converting the full planet file from the OSM data model into vector tiles creates a spatially-indexed collection of OSM objects that can be easily decoded into GeoJSON. This opens new possibilities for data analysis that were first realized and implemented through the creation and subsequent analysis of the *OSM Quality Assurance Tiles* produced by Mapbox.

5.1 OSM-QA-Tiles

Initially designed for efficient, parallelized data analysis of OSM data, OSM-QA-Tiles are vector tiles containing all of the OSM data that can be effectively represented as GeoJSON. These tiles are not optimized for serving over the web as they contain all of the OSM attributes for an object as well as the metadata for the most recent edit. Since they are not built for rendering on a slippy map, they are only rendered at one zoom-level, currently 12. These tiles can then be processed in parallel with the open source `tile-reduce`* framework, discussed next.

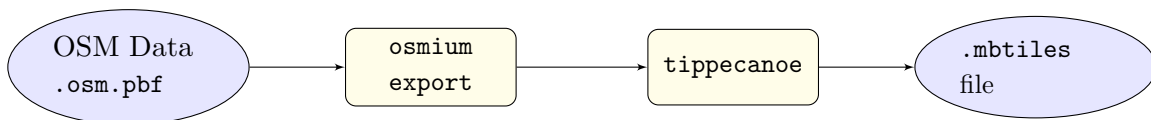


Figure 5.1: Creating OSM-QA-Tiles

Figure 5.1 shows the process of creating OSM-QA-Tiles using the open-source tools discussed in the previous chapter. The `osmium export` command converts OSM data into GeoJSON, so only features which can be accurately represented as GeoJSON exist in the resulting tileset. The `tippecanoe` utility then encodes the GeoJSON into vector tiles to create a single `mbtiles` tileset.¹ Section 5.3 will discuss how this workflow can—and has been—improved to handle the consistently growing map. In an OSM-QA-Tile, all metadata attributes are prefixed with the `@` symbol, such as `@id`, `@user`, `@timestamp`, `@changeset`. This is useful in separating these values from the rest of the object attributes and analysts can depend on these attributes to always exist. As of 2019, OSM-QA-Tiles have always been and continue to be maintained by Mapbox and are released for public use at osmlab.github.io/osm-qa-tiles.

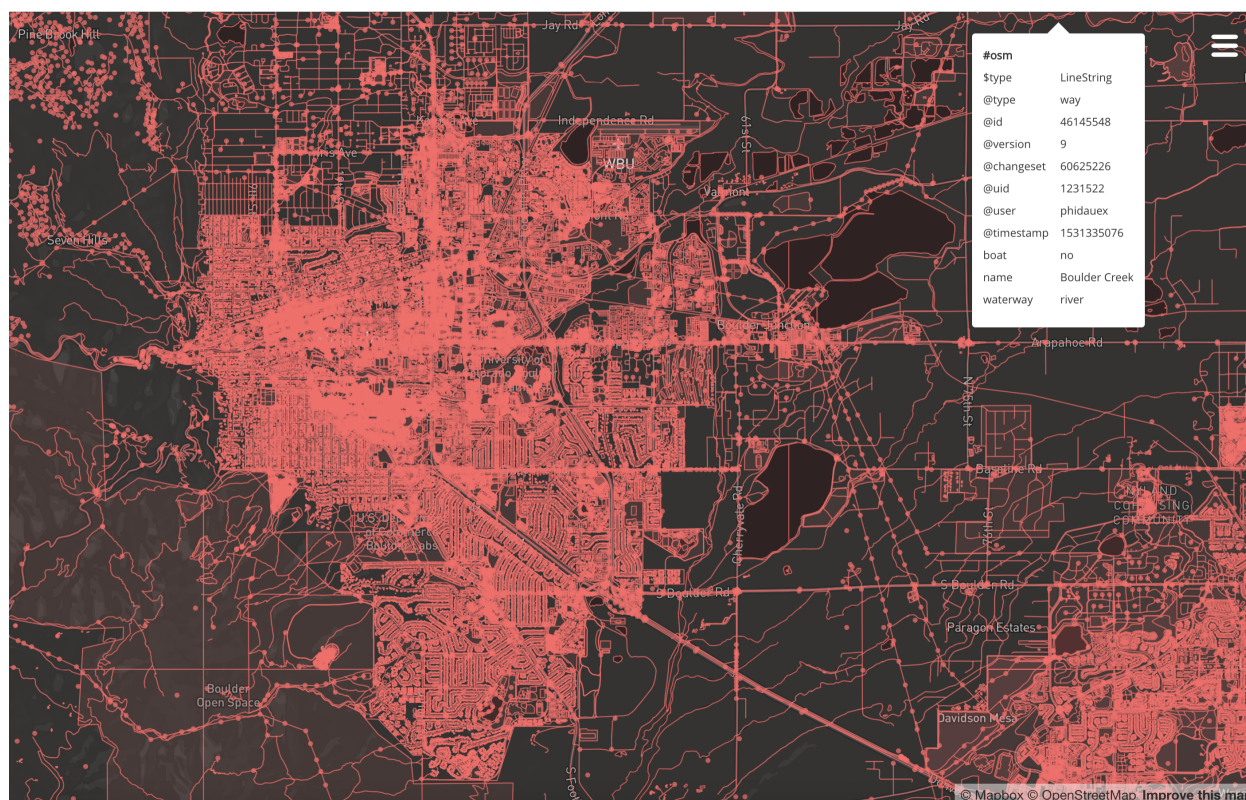


Figure 5.2: OSM-QA-Tile showing all of the map data around Boulder, CO. Pop-up shows example metadata and attributes that are associated with an object.

¹ For specific configuration options, see github.com/osmlab/osm-qa-tiles

5.2 Vector Tile Based OpenStreetMap Data Analysis

Tile-based analysis allows for parallel processing of data as each tile can be passed to a different thread. In this manner, the map-reduce big-data processing techniques can be used [28]. The `tile-reduce` Javascript library was created to allow analysts to write map-reduce jobs against vector tiles in the `.mbtiles` format. This library reads the database and passes each tile (optionally limited by a geographic area) to a worker thread which returns a single result to the reduce job. This allows for spatially-aware, parallel processing of OSM data in a vector tile. This analysis workflow was created by Mapbox for efficient planet-scale quality assessment of OSM data. For my analysis purposes, it is important that worker threads can be used for more than just passing data back to the reduce job. This enables us to perform analysis at both the global and tile level and provide the most flexibility for downstream analysis.

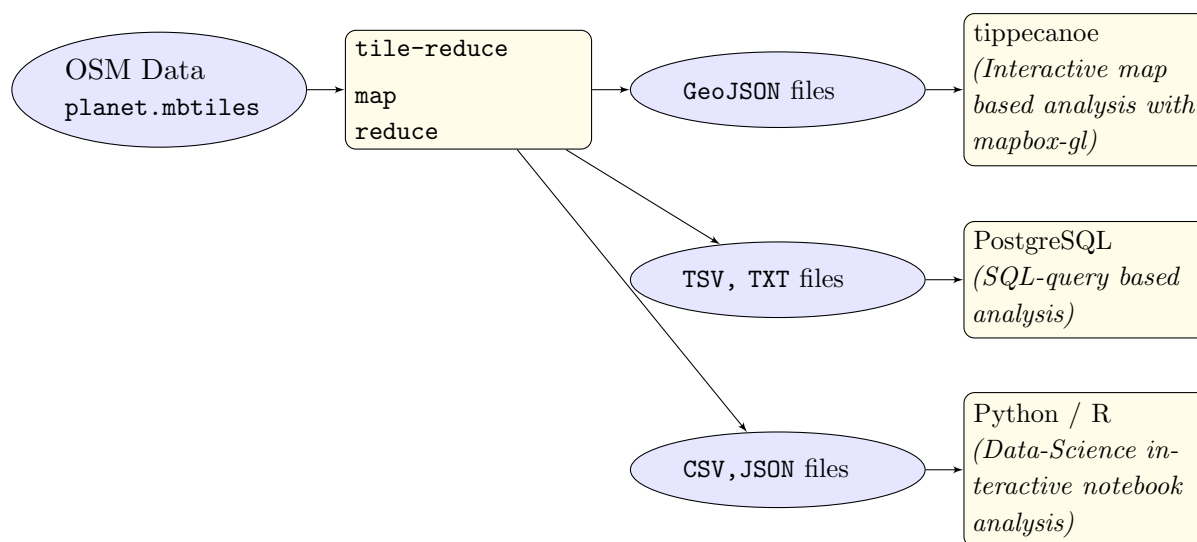


Figure 5.3: Tile-Reduce Spatial Data Analysis Workflow. While still benefiting from the computational efficiency of parallel processing with `tile-reduce`, I configure the process to produce multiple pieces of output that will each be used in a different data analysis environment. This is where my approaches push the boundary of what this workflow was designed to do.

5.2.1 Implementation Example

Next, I will walk through a recent implementation of the tile-based analysis workflow as an example. This particular analysis was extended to perform the historical analysis for Chapter 9 [3]. The question is: *Where are corporate data-teams mapping, how much work are they doing, and how much has this increased in recent years?* As depicted in Figure 5.3, we first choose the appropriate inputs and then define the two functions of the tile-reduce script: `map` and `reduce`.

`input`: OSM-QA-Tiles for the planet. Specifically, I will use a custom extension that includes turn-restrictions to be described in Section 5.3.2. Since OSM-QA-Tiles are made available daily, this entire workflow could be run daily or weekly to be kept up to date. The second piece of information we need to include is the list of usernames associated with each data-team [3].

`map`: This function will run on every tile. Since the primary question involves knowing who the editors are, the first thing we do is group all of the features present on this tile by username. Next, we compare this list of usernames to the list of known corporate editors. If there is overlap, then we build lists of features edited by each user within each corporation. Next, we want to better understand how the editing volume has changed (grown) over time. For this, we further segment these lists of edits by day. Ultimately, we create a contributor-centric breakdown of the features on the tile that consists of the following hierarchy:

`corporation` → `date` → [list of features edited by data team members]

This is a *contributor-centric* breakdown of features because it prioritizes the editing metadata of *who* and *when*. Then, we calculate the total kilometers of road, number of buildings, number of points-of-interest, and the total number of edits (catch-all) for each list of features. This creates per-corporation, per-day, per-tile statistics. These daily editing summaries are then written to disk as single GeoJSON point objects representing the centroid of any corporation’s daily editing activity with these statistics as properties. These points will answer the “where” question when later rendered on a map. Finally, these per-day, per-corporation summaries are returned to the reduce script.

reduce: As the per-day, per-corporate summaries of editing activity are returned from each tile, they are aggregated into per-corporate editing summaries. When all of the tiles have been processed, this global summary of daily activity by corporation is written to disk.

This tile-reduce job will yield two results. The first is a file of line-delimited GeoJSON point objects with per-day, per-tile, per-company statistics. Currently, this file is less than 1M lines, but this workflow could scale to produce a file of any number of points. As depicted in Figure 5.3, this GeoJSON file can then be ingested by `tippecanoe` to produce the vector tileset that drives the interactive map seen in Figure 5.4. The second output is a CSV file with the following per-corporation, per-day attributes: `date`, `corporation`, `km of roads`, `number of buildings`, `number of POIs`, `total number of edits`.² As depicted in the Figure 5.3, this CSV file is read into a Jupyter Notebook environment, aggregated by company and date, and produces the graphic seen in Figure 1.9 in Section 1.4.1.3 about corporate editing. This answers the growth over time question.

Figure 5.4 shows the screenshot of an interactive map built to answer the *where* and *how much* portion of the initial question about corporate editing. The map itself is built with `Mapbox-GL*` and the timeline is powered by the `D3.js` visualization library [16]. This particular visualization was built from a boiler-plate interactive map framework I created in mid 2018 for faster iterations of interactive maps as data-analysis tools.³ The dropdown menu in the top-left lets an analyst choose to visualize the editing activity of one of 10 corporations. The brushable timeline at the top allows the analyst to filter editing activity to a specific time-window. Animation controls allow the viewer to step through time automatically to see how the editing focus changes overtime.

The map is powered by the individual editing summary points produced by the `map script` and converted to vector-tiles for efficiency and performance. Each point is rendered on the map and the various calculated properties allows styling of the map appropriately. First, as individual points,

² In the case of more in-depth investigations such as Chapter 9, these are broken down further into the creation of new objects and the editing of existing features.

³ This example analysis was first conducted in September 2018 after a conversation with Mikel Maron about the feasibility of measuring corporate data-team editing. To speak to the power and efficiency of tile-based analysis: It took about 4 hours to conceive, iterate, and ultimately produce and run the analysis scripts and the first interactive map. In 2015 this would have taken weeks.

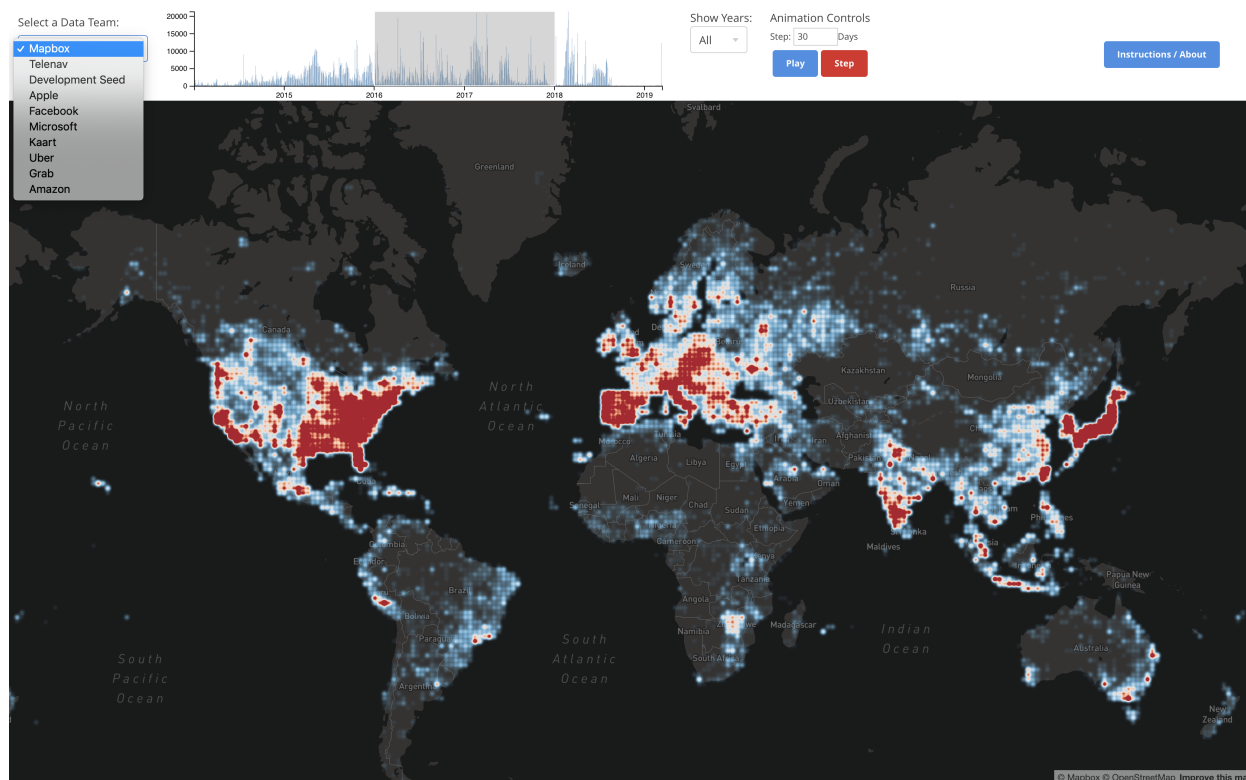


Figure 5.4: Screenshot of an interactive heatmap built from the output of a tile-reduce analysis workflow. Showing relative volume of edits from Mapbox data-team between 2016-2018

the heatmap style creates the aggregated visualization to show the relative volumes of editing: Where there are more points, the map appears to glow red-hot. Second, the time-slider at the top of the screen filters against the `date` for each point, representing the total edits performed by a corporation on a given day. As points are filtered on or off by date, the map will re-render appropriately to reflect the presence of the edits. Third, the number of edits associated with each point influences the weight for the heatmap, meaning that the visualization accurately reflects the *volume*, *location*, and *date* of edits performed by each team.

Not described in this example is the second type of output shown in Figure 5.3: `TSV/TXT` for ingesting into PostgreSQL. This particular analysis does not benefit from this from step, however, Chapter 6 will describe how this tile-based analysis workflow can act as an effective conversion between the OSM editing history and a relational database. This provides a powerful interface for

writing SQL queries against the entire editing record without having to recreate a replica of the entire OSM database.

5.3 Improving OSM-QA-Tiles

Though it is unknown how many people are actively use OSM-QA-Tiles, we do know that these tilesets are an important data source in a number of projects, such as *osm-analytics.org*. So far, I have shown just one example of a complex question that can efficiently be answered with OSM-QA-Tiles. Part IV will present more of my explorations of OSM data with this tile-based analysis approach. However, OSM-QA-Tiles are not always the most efficient or accurate dataset for every type of question about OSM. Slicing and distributing analytical processing along the boundaries of a tile can lead to complications and misleading results. Furthermore, as the amount of data in OSM continues to grow alongside improved GeoJSON representation of objects, efficiently generating useful tilesets for analysis will grow more difficult (it has already). If a tileset of every object in the world grows too burdensome for simple analyses, then we will need to rethink our approaches to tile-based analysis. Here I will discuss some of these challenges along with solutions, both implemented and proposed, for continued success with tile-based analysis.

5.3.1 Simplifying OSM-QA-Tiles by Object

When an OSM object extends past the bounds of a single tile, the object may be represented in one of two ways: By default, the geometry is clipped at the tile-boundary (or a buffer thereof), and the feature is effectively duplicated across multiple tiles, each tile containing the relevant geometry for that section. This creates the most efficient individual tiles as they only contain information about objects that exist on that tile when rendered. At the tile boundaries, these objects are seamlessly stitched back together, making them appear as one object (such as a coastline or Country border) when viewed across multiple tiles.

However, this can yield misleading metadata when the attributes of the object are copied to all of the sections as it is split across tile boundaries. Figure 5.5 exhibits this issue. Though the

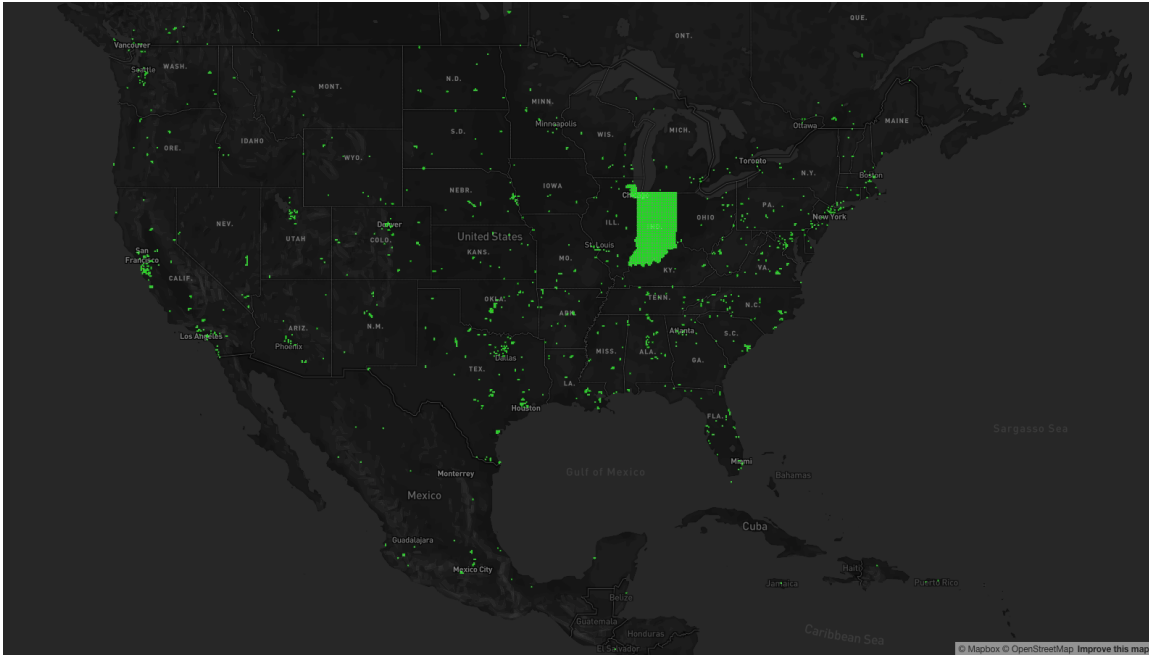


Figure 5.5: A map highlighting tiles where a single mapper was active in 2011. The mapper edited the OSM object which represents the administrative boundary of the State of Indiana. The metadata of this object is propagated to every tile that includes part of this geometry. This mapper’s edit count therefore includes every tile in Indiana, though the original edit was only to one object.

mapper made an edit to a single object, that object is split across all of the tiles within the state of Indiana because each of these tiles contains part of this geometry. This mapper’s username then appears in the metadata for every part of the object, distributed over the entire State. If we count this mapper’s total number of edits across all tiles, it will be greatly exaggerated by these objects on all of these tiles.

The second method is to avoid the duplication of features and instead only write an object to a single tile. While this creates a more accurate representation of editing activity (one edit per distinct object), it might bloat the single tile that contains the object since that tile contains geometry information beyond its own geographic extent. Also, tiles which should contain parts of this geometry are unaware that it exists. This creates the potential for imbalanced tile-reduce jobs as these larger objects are not distributed across multiple workers. While there are pros and cons of both approaches, they depend entirely on the question being asked. This highlights the complexities

and trade-offs of an all-purpose analysis vector tileset, currently an open discussion within the OSM data community.

A new complication in the generation of OSM-QA-Tiles comes with technical improvements to GeoJSON conversions of OSM elements in the past year: Country Boundaries and Coastline objects. At Zoom level 12, the majority of the objects on the map is under 2.5M tiles. This creates a tileset of about 39GB. Including administrative boundaries, however, creates large Polygons representing Territory and Maritime border objects all over the world. As a result, the full tileset now contains nearly 9M tiles, most of which are in the ocean, representing the coverage of a political boundary. These tilesets are now closer to 90GB and take much longer to process with the addition of these 6M tiles.

The immediate solution was to produce filtered tilesets: One with administrative boundaries and one without. Today, these two distinct OSM-QA-Tiles files are generated daily: One that includes all available OSM data and another that excludes coastlines and administrative boundaries. This additional “compact” tileset has been generated since early 2019 after a major update to the processing workflow, resulting from many discussions between Mapbox engineers and myself.

This solution, however, is a first step in producing more streamlined tilesets. Object-specific tilesets could dramatically improve efficiency as the total number of map objects continues to grow. Generating building or road-only OSM-QA-Tiles will allow more targeted, efficient analysis for those interested in analyzing a specific type of object on the map. In terms of efficiency, it would likely be the difference between downloading a 9 or 90GB tileset for analysis, a non-trivial advantage if automatically processing these tilesets daily or weekly. However, the daily generation of multiple tilesets increases overhead for whomever is producing and hosting the tilesets. This remains an open issue in the OSM data community. Section 10.1.2.2 presents current work addressing this issue.

5.3.2 QA-Tiles-Plus: Turn-Restrictions

As previously mentioned, abstract relations in OSM such as turn-restrictions cannot be represented in GeoJSON, and therefore are left out of OSM-QA-Tiles that are built from GeoJSON

representations of OSM data. For contributor-centric analysis, the lack of turn-restrictions in these tilesets means the exclusion of over 1M edits. Further, turn-restrictions are relatively complex features that are assumed to be edited by a mapper informed with “ground-truthed” information. Accounting for these edits then has major implications for intrinsic data quality analysis as discussed in Chapter 7. Therefore, a *geometric representation* of the turn-restriction is not as important as simply knowing that a turn-restriction exists on or near an object. To account for turn-restrictions, I created an extension of OSM-QA-Tiles called OSM-QA-Tiles-Plus that includes the metadata of a turn-restriction object with basic geographic information about its relative location. This name comes from the term *contextual-stream-plus* as introduced by Bica et al to describe a more contextually complete Twitter dataset [13].

OSM-QA-Tiles-Plus are created by first extracting all of the turn-restrictions in the current OSM planet file with the `osmium tags-filter` utility to identify all relations with the `type = restriction` attribute. Next, I extract *any* geographic information associated with that restriction. Typically, turn restrictions have three attributes: `to`, `from`, and `via`. In many cases, `via` is the node representing the intersection of the two other elements. The coordinates of this node then become the geometry of the new object. We now know there is a turn-restriction edited by a specific mapper at a specific time at a specific location. An example of all the turn-restrictions in Panama City represented as points is shown in Figure 5.6.

In cases where the `via` attribute does not exist or is not a single node, the script continues looking for any geographic information on the `to` or `from` elements. These restrictions are then encoded as single points and merged into the current OSM-QA-Tiles, creating the new OSM-QA-Tiles-Plus tileset. When these new tiles are analyzed, these points will show up as single features on a tile and be counted as an edit associated with a time and a location. For extra accuracy, I add a boolean attribute `@tr` so that analysts can filter for these turn-restrictions if desired, as seen in Figure 5.6.

For the majority of contributor-centric questions that I seek to answer, this is a passable solution to representing the abstract relation of a turn-restriction on the map: It simply acknowledges

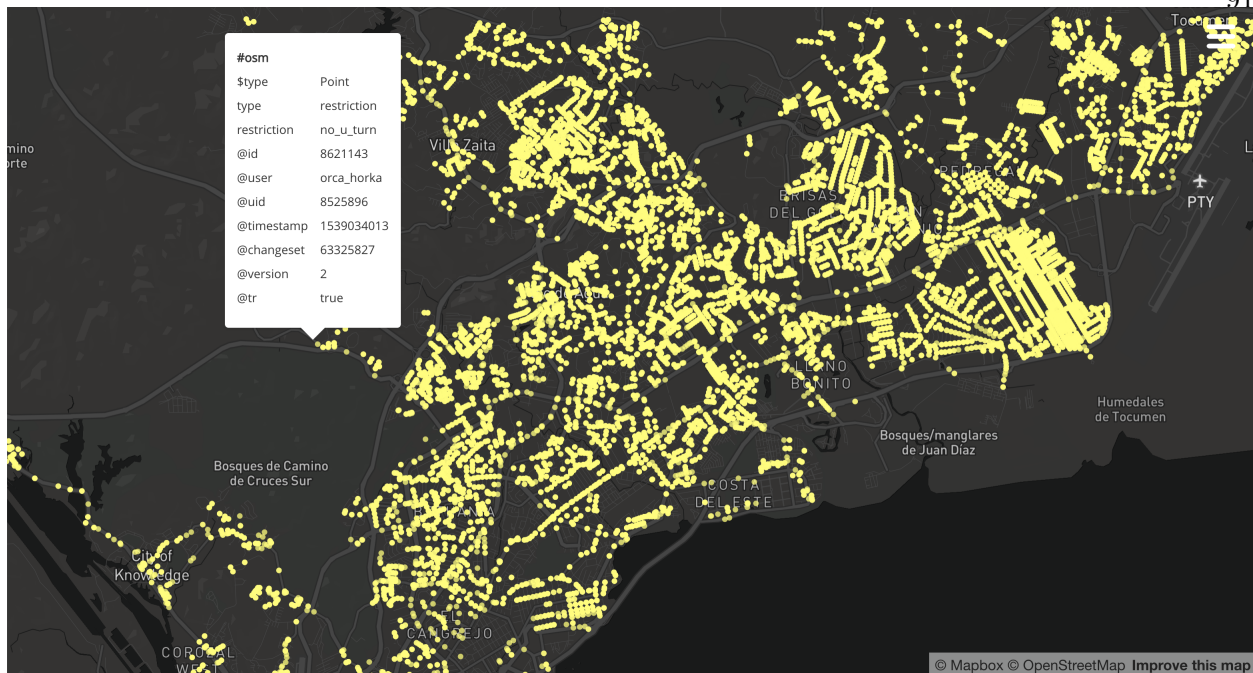


Figure 5.6: Turn-Restrictions as Points in Panama City, Panama. The popup shows the metadata associated with one of the points representing a restriction on a u-turn

the relative location of a turn-restriction. For any routing or applied purpose, this is not a viable solution because it preserves none of the actual actionable information and will therefore likely never be implemented in any general tileset.⁴ Additionally, this approach could be extended as a fall-through processing method for GeoJSON representations of OSM elements. That is, any object in OSM that cannot be converted into valid GeoJSON could still be created as an object with the appropriate attributes and metadata, but the geometry would be a *placeholder Point* feature. This would recognize the existence of the feature, and the mapper who last edited the object is given credit for their contribution.

5.3.3 OSM-QA-Tiles: Summary

Tile-based analysis workflows utilizing the `tile-reduce` framework and OSM-QA-Tiles allows for fast, robust, and scalable analysis of OSM data. Vector tiles allow us to maintain all of the

⁴ When I shared this code with a maintainer of the `osmium` utility, he helped identify a performance improvement but did not respond to my proposing this approach as a potential solution to the representation of turn-restrictions in GeoJSON.

OSM metadata, enabling contributor-centric analysis of the map. Another major benefit of this analysis workflow is that it relies only on a single file for input. It is not dependent on the OSM or Overpass* API, nor requires configuring and running a database. My hope is that this will continue to lower the barrier to entry for other analysts to do global-scale OSM data analysis and promote reproducibility. Additionally, with the only computational dependency being Node.js, entire analyses can be packaged and distributed. For example, the corporate editing analysis shown here is packaged and hosted on GitHub. It can be cloned and then run against the latest OSM-QA-Tiles at any point to identify the current geographic footprint and volume of corporate editing in OSM.⁵

The largest drawback of the current OSM-QA-Tiles is that they only represent the most recent version of the map. This means they do not include the editing history of an object, a critical component of contributor-centric analysis. For example, the corporate editing example analysis of Section 5.2.1 certainly captures the majority of edits and the growing trends, but it likely undercounts the number of edits. If a corporate editor created a road or building in the last five years that was subsequently edited by another user at any point, then that initial edit is not counted because the OSM object in the OSM-QA-Tile only reflects metadata of the most recent edit. In contrast, the similar analysis in Chapter 9 uses an improved approach to historical tile-based analysis, presented next.

Since the tile-reduce framework can be run with any tileset with the OSM-QA-Tiles schema, it is possible to run the same analysis against multiple OSM-QA-Tiles produced at different times, allowing us to compare the data at two different points in time. This is a different approach to Time Series Analysis than that of Chapter 3, but it is more scalable. Until now, this was the only way to do tile-based historical analysis. This is the foundation of two of the three approaches to tile-based historical analysis that I will discuss next.

⁵ I invited a number of corporate data-team managers to this repository and it is my assumption that they have cloned this repository and continue to run this analysis against the latest OSM-QA-Tiles.

5.4 Historical Tile-Based Analysis 1: Annual Snapshots

Historical snapshots allow analysts to see the map as it existed at a certain point in time. When OSM-QA-Tiles were first introduced, so were additional “*annual snapshots*” that contain OSM objects as they existed on January 1 from 2005 to 2016.

Because of the complexities of measuring changes over time in OSM as discussed in section 2, historical snapshots represent efficient collections of OSM data at specific points in time to allow for something closer to time series analysis. Tile-based analysis of annual snapshots involves running the same `tile-reduce` job over multiple snapshot files, obtaining results for the map data as it existed at any point in a year. If the analysis involves counting edits over time, annual snapshots provide the guarantee of at least annual resolution of editing activity. This means an object created in 2014 and edited by other users in 2015 and 2016 will be first counted as an edit that occurred in 2014 when the script is run against that year, and also count the edits in 2015 and 2016 when those years are processed. In contrast, only the 2016 edit will appear in the current OSM-QA-Tiles. While this increases the processing time about 10-fold to run the job over and over for each year, there is no added complexity because the data schema is consistent between each year. Analysis of pre-2009 tilesets is much faster because as Figure 5.7 shows, there was much less data on the map.



Figure 5.7: OSM-QA-Tiles for London in 2007 (left) and Today (right).

5.5 Historical Tile-Based Analysis 2: Quarterly Snapshots

Annual snapshot osm-qa-tiles allow us to ask detailed questions of the map data as it existed at the beginning of each year, but as the rate of editing continues to increase, we are blind to a significant number of edits that happen each year to existing objects. I call these “shadowed edits” because they are effectively hidden by the next edit to the object, or in the case of annual snapshots, they are masked by the edit that happened closest to midnight on December 31. Figure 5.8 shows the number of shadowed edits by location in for 2012. Section 8.1.1 will further explain how I identify and quantify these shadowed edits.

The large number of shadowed edits, however, should not entirely discount the power, abilities,

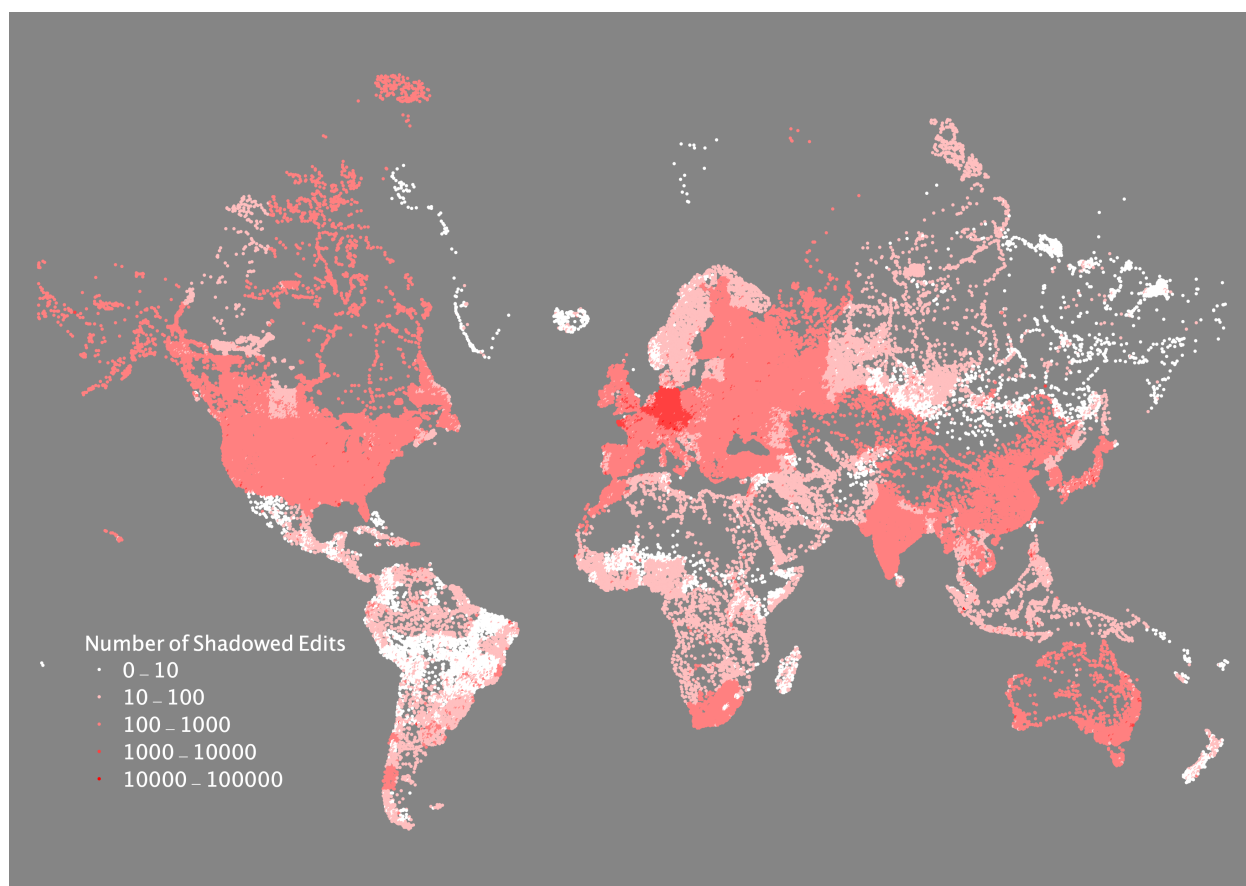


Figure 5.8: Global footprint of invisible edits from annual snapshots. With annual resolution, these “shadowed” edits go discounted between each year. This map represents edits that were essentially hidden in 2012.

and conveniences of using historical snapshots of the map for historical analysis. Data analysts just need to be honest about the limitations of the dataset: these are not complete records of the editing history, but they are still accurate to a specific resolution. Furthermore, if the goal is to compare the map between two points in time from a completeness or density perspective that does not require knowing the individual object histories, and snapshots are the ideal datasets for comparison.

To continue using snapshots for tile-based historical analysis, I increased the resolution to every three months, or *quarters* to reduce the number of shadowed edits. This “annual quarters” unit seemed appropriate because an analysis goal at the time was to show improvements in terms of data density. Likening these to the more traditional “quarterly report” of the business world to measure growth, I worked with Mapbox to create quarterly snapshots of OSM data in the OSM-QA-Tile format. Generating these at any higher resolution becomes burdensome as the collection of quarterly snapshot tilesets from 2006 thru 2019 is already 759 GB. For processing, these tilesets need to all be accessible on a single, large machine: Remaining under 1TB seemed like a reasonable compromise. Quarterly snapshots offer four-times the resolution of annual snapshots and therefore take four-times as long to process. However, the processing complexity is not increased as the same tile-reduce job is run across each of the quarters. Chapters 8 and 9 present analyses using quarterly-historical

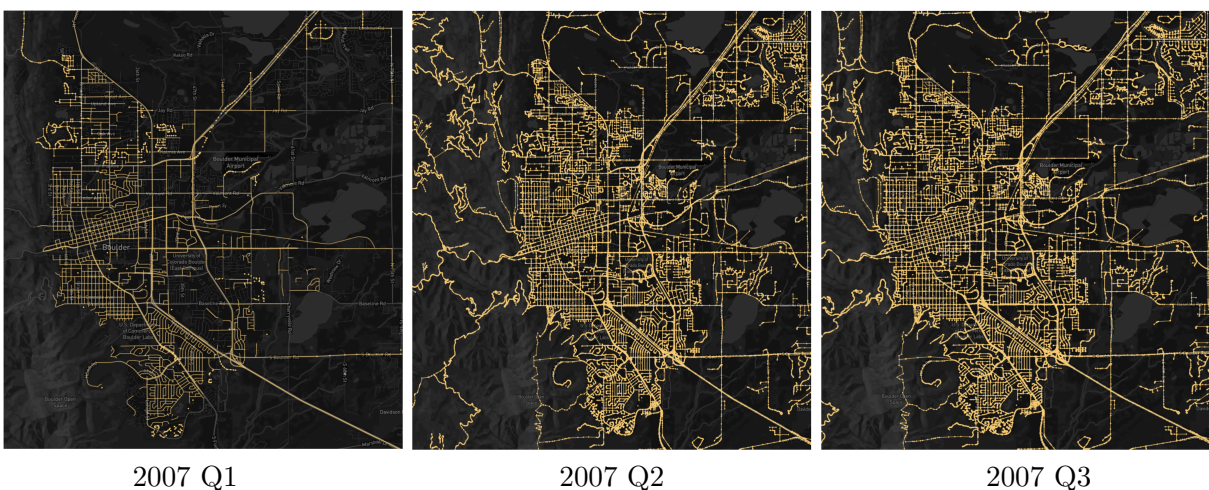


Figure 5.9: The evolution of the map of Boulder, CO at quarterly-snapshot resolution, capturing the bulk import in the 2nd quarter of 2007.

snapshots.

5.6 Tile-Based Analysis Approach 3: Full Historical Vector Tiles

Ultimately, the complete history of all objects on the map cannot be represented by snapshots of any resolution, especially with tile-based analysis. Since processing is distributed among individual tiles, different versions of the same object between snapshots are unaware of one another. This makes it impossible to observe interaction between users such as validation, vandalism, or tagging disputes. If, however, all of the versions of a single object were on one tile, then historical tile-based analysis of the evolution of the map at the object level is not only feasible, but can also embrace the processing efficiencies of tile-based analyses. All of the previous work leads to this single objective: *Full-history vector tiles*.

5.6.1 Historical OSM Data Schema(s)

In this section I identify two data schemas that are extensions of the GeoJSON representation of OSM objects in the OSM-QA-Tile format. These schemas allow us to represent the full history of an object in two forms: First, a historically accurate, complete account of every edit; second, a format optimized for rendering so that analysts can easily interact with the data on a map to observe the changes over time. These two schemas are *Embedded Object Histories* and *Individual Historical Versions*. The first step to converting OSM historical data into these schemas is to compute the full history of an object over time, including minor versions, from the full planet history. Currently, the only tools capable of doing this include `osm-wayback`, `ohsome`, and `OSMesa`. `osm-wayback` is a tool initially developed between Mapbox and me, and is currently maintained by me. I used and further developed this utility to develop these schemas. Section 10.1.1 will explore this particular tool in further detail.

As discussed in Chapter 2, an object's complete history involves any number of five different types of edits. Changes can occur to either the geometry of an object or its attributes. Geometry changes can either produce a new version of the object or a minor version. Changes to attributes

can be one of three forms: addition of a new tag, modification of an existing tag, or the deletion of an existing tag. Utilities that reconstruct an object's complete history should identify these attribute changes and categorize them accordingly.

Figure 5.10 introduces the extension, showing how the addition of a `@history` attribute to an OSM object's GeoJSON representation can capture an object's entire history while accounting for all contributors, including minor versions, thereby addressing all of the challenges put forth in Chapter 2. These history objects look very similar to an OSM object in an OSM-QA-Tile with a few exceptions. First, there are now two timestamps for an object: `@validSince` and `@validUntil`. `@validSince` represents when this version (or minor version) was created while `@validUntil` is added to an object when a newer version (or minor version) is created. For the current version of an object, this value is `null`. The difference between these two timestamps defines the lifespan of this historical object. Instead of storing the complete attributes for every version of the object,

```

@history : [
  {
    @version:      <number>,
    @minorVersion: <number>,
    @user:         <string>,
    @uid:          <number>,
    @changeset:   <number>,
    @geometry:    <GeoJSON geometry>,
    @validSince:  <timestamp>,
    @validUntil:  <timestamp>,
    @tags_added:  {
      'new_key': 'new_value '
    },
    @tags_modified: {
      'existing_key': [ 'old_value ', 'new_value ' ]
    },
    @tags_deleted: {
      'old_key': 'old_value '
    },
  },
  <all previous versions>
]

```

Figure 5.10: The addition of the `@history` attribute to OSM objects will capture an object's entire evolution and allow analysts to quickly see what changed.

historical objects can track which attributes were changed (Edit type 4 in Table 2.1). This allows an analyst to look through an object's history and quickly identify a specific edit type. The addition of a 'name' attribute, for example, is an interesting edit to track as it likely implies the addition of local or ground-truthed knowledge to the map. When the history object is a minor version, the editing metadata is copied from the changeset that produced the minor version so that the mapper responsible for the minor-version appears in the editing record of this object. In minor versions, there will be no changes to attributes, but the rest of the metadata will be updated. This `@history` object can be integrated into current OSM-QA-Tiles in the following two ways: Embedding the history into each object as an attribute, or creating distinct map objects for each version.

5.6.1.1 Embedded Object Histories

As an additional top level attribute, the `@history` object can exist among the current properties of an OSM object in a current OSM-QA-Tile. This approach is optimized for tile-based per-object history analysis. When the object is decoded into GeoJSON, the `@history` object can be easily parsed by the tile-reduce framework's `map` function to give analysts access to the complete history of the object. Iterating over the object's history lets analysts see which mappers have made which changes. Since the distinct geometries of each version are embedded, geospatial processing can be employed to compare geometry changes between versions as well.

Storing individual geometries for every version, however, is not space-efficient. Especially because these geometries are likely to be similar between versions. The TopoJSON* format is a topologically organized extension of GeoJSON that minimizes duplication by storing line segments as *arcs* and then defines geometries by referencing these arcs. Performing this conversion adds slight overhead to the production of the tiles, but it greatly reduces the size of the `@history` object. To compare the geometries between any two versions, the TopoJSON can be decoded for each version to return two distinct GeoJSON geometries. This can be helpful if an analyst wants to find buildings that were modified to have square corners, a popular minor version.

The only drawback of this representation is that it requires parsing the entire object's history

to know what the map looked like at a certain time. To solve this, objects may be encoded as individual historical versions.

5.6.1.2 Individual Historical Versions

Another approach which is optimized for rendering is to represent each version of an OSM element as its own object, as described in the Taj Mahal example earlier in Section 4.3.1. This will inevitably be less space-efficient because every version has its own geometry, but they are historically accurate. Therefore, in this schema, the most important attributes are the `@validSince` and `@validUntil` timestamps. This allows an interactive map framework like `mapbox-gl` to filter by time to render the map exactly as it existed at any point in time, down to the second. This allows us to build an exploratory map that can show the exact changes over time.

I have implemented and iterated on these three tile-based historical analysis approaches across multiple research projects which I will now share in Parts IV (snapshot) and V (full-history).

Part IV

OSM Historical Snapshot Analyses

This part contains four chapters, each of which correspond to a specific project, presentation, or publication that utilized and advanced *tile-based analysis of historical snapshots of OSM data*.

Chapter 6 reviews and discusses work I completed as a Mapbox Research Fellow in 2016.⁶ This was my first exposure to the annual-snapshot historical OSM-QA-Tiles. I will review the data-processing pipeline I developed, present the visualization utilities I created, and summarize some of our findings as they were co-presented with Mikel Maron at the State of the Map US 2016 conference in Seattle, Washington.

Chapter 7 is an exact reprint of a paper that used the annual-snapshot historical OSM-QA-Tiles for analysis. The paper is reprinted here with the permission of my coauthors, for which the full reference is:

Jennings Anderson, Robert Soden, Brian Keegan, Leysia Palen, and Kenneth M. Anderson (2018). The Crowd is the Territory: Assessing Quality in Peer-Produced Spatial Data During Disasters. International Journal of Human-Computer Interaction. doi:10.1080/10447318.2018.1427828

Chapter 8 reviews and discusses my innovation of moving from *annual* historical snapshots to *quarterly* historical snapshots. Much of this work was completed in partnership with Mapbox as a Research Fellow in 2017. I will review the need for and creation of these quarterly historical snapshots as well as an interactive analysis dashboard collaboratively developed with Mapbox to visualize the growth of the map over time. This analysis dashboard was presented at State of the Map US 2017 in Boulder, CO.

Chapter 9 is an exact reprint of a paper that used the quarterly-snapshot historical OSM-QA-Tiles for quantitative analysis. The paper is reprinted here with the permission of my coauthors, for which the full reference is:

Jennings Anderson, Dipto Sarkar, and Leysia Palen (2019). Corporate Editors in the Evolving Landscape of OpenStreetMap ISPRS Int. J. Geo-Inf. 2019, 8, 232. doi:10.3390/ijgi8050232

⁶ Chapters 6 and 8 include work done in collaboration with Mapbox as a Research Fellow during the summers of 2016 and 2017, respectively. These chapters include summaries of research projects and conference presentations. They are included here with the permission of my collaborators.

Chapter 6

Annual Historical Snapshots

This chapter describes the implementation of a data-processing pipeline developed to better understand the growth of the map on an annual basis, using the annual-resolution historical-snapshot OSM-QA-Tiles that were originally developed and released publicly by Mapbox. As part of a 2016 Research Fellowship with Mapbox, I extended the analytical workflow to create a suite of interactive visualizations that can better explore the editing history of OSM. First, I will describe this data processing pipeline.

6.1 Data Processing with Annual Snapshots

As originally presented in Section 5.4, annual historical snapshots promise at least annual resolution of the OSM editing history. These analyses are achieved by running the same tile-reduce

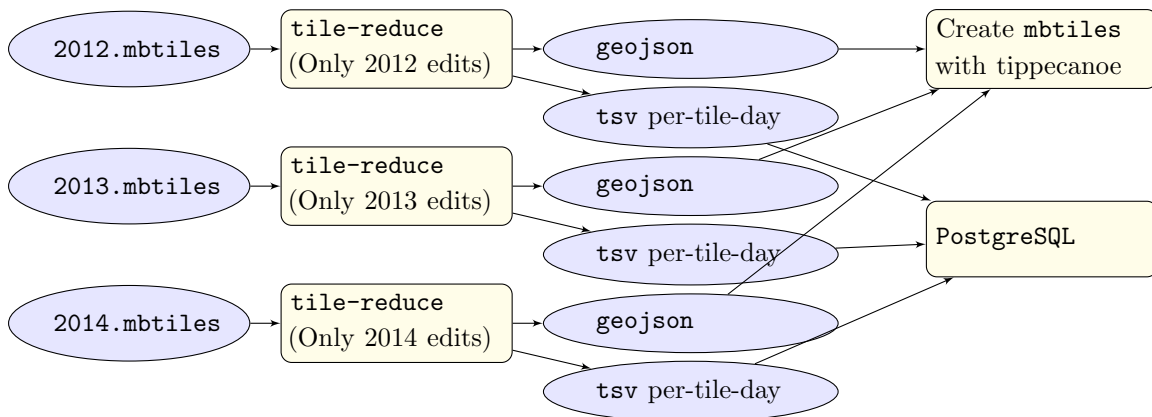


Figure 6.1: Tile-Reduce workflow for annual analysis (example for 2012, 2013, and 2014). This approach ensures at least annual accuracy of historical OSM data analysis.

job over consecutive years with annual time-filtering to capture editing activity within a given year. Figure 6.1 shows this workflow for just three years. Annual OSM-QA-Tiles are used as input to the tile-reduce script which then filters for just the edits that happened during that year, grouping them by day, and then computing summary statistics: per-tile, per-day, per-year. These summaries are then aggregated per year as GeoJSON which can be turned into vector tiles to power interactive visualizations.

To recombine the annual editing activity per tile into a queryable format, a TSV file is produced as output from the `map` function. Each line of this file contains per-day, per-tile editing statistics such as the total kilometers of roads or buildings edited. As a TSV, this file can be efficiently transformed into a PostgreSQL database with the `\copy` command. This results in a database containing tables of annual editing statistics per day for each tile. Additionally, tiles can be indexed by quadkey to allow for spatially bounded queries without the need for any geospatial extensions or indexes.¹ Figure 6.2 shows an example query to get the total number of buildings created on January 13, 2010 (the day after the 2010 Earthquake) around Port Au Prince, Haiti.

In 2016, running this analysis for the entire planet over the previous 11 years of annual OSM-QA-Tiles, and then creating the annual databases required about a day of processing time on a single (modestly large) machine.² Creating per-day tile summaries meant that the same tile might exist in 365 different rows. These were found to be rare however, and the total number of tiles could be adequately managed by a PostgreSQL database. Because this workflow and the resulting databases were a successful step towards allowing us to recreate the evolution of the map, I next turned to the editing metadata to make these databases more *contributor-centric*. This meant separating per tile-statistics not by day, but by contributor. In this way, per-tile, per-editor, per-year editing statistics were generated. These databases were larger, with up to 15M rows in 2014 and 2015, though still manageable by PostgreSQL. This meant it was now possible to ask questions such

¹ Quadkeys are numerical ids of map tiles in base-4. They are the basis of the Bing Maps Tile system. docs.microsoft.com/en-us/bingmaps/articles/bing-maps-tile-system. A key feature of a quadkey is prefix matching: All higher zoom tiles that fall within tile 0123 start with the quadkey 0123.

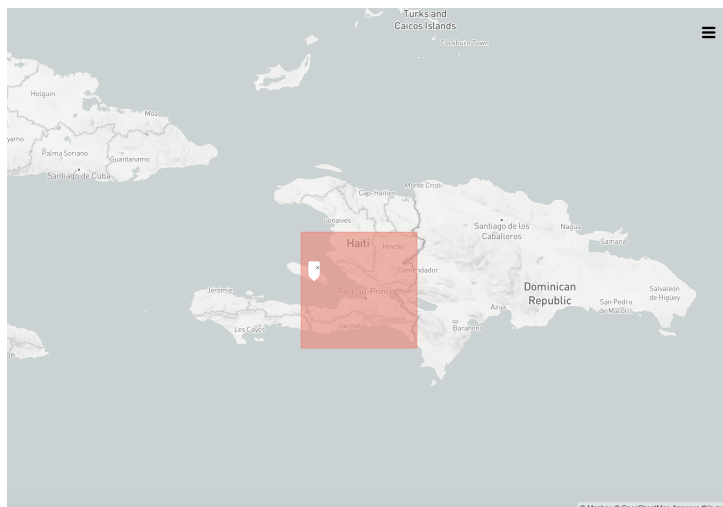
² These were created on a single Amazon EC2 c4.8xlarge instance. This is a one-time job of about \$30 to generate the output needed to populate the databases.

```

SELECT
    sum(new_buildings)
FROM
    per_tile_day_2010
WHERE
    day = 13 AND
    quadkey LIKE '03221120%';

```

(a) SQL query to count the number of buildings edited on the 13th day of 2010 on any tile with a quadkey starting with 03221120.



(b) The bounds of the zoom level 8 tile (quadkey 03221120)

Figure 6.2: Tables containing daily editing summaries indexed by quadkey allow us to perform queries based on quadkey prefixes. Figure 6.2b represents a zoom level 8 tile that covers the region of interest. There 256 zoom level 12 tiles that fall within its bounds. The quadkey for each one of these tiles starts with 03221120. This query is then restricted to these 256 tiles via quadkey prefix matching. The day constraint only returns the rows representing the subset of these tiles that were edited on January 13: The 13th day of the year 2010.

as, “how many kilometers of road did a particular user edit in 2015, and where?” Previous methods of enabling such questions required establishing significant infrastructure involving a mirror of the full OSM database and additional logic to reconstruct the state of an element at a specific time. By leveraging `OSM-QA-Tiles`, `tile-reduce`, and `PostgreSQL`, I was able to establish a much simpler infrastructure that could answer a multitude of questions about contributor activity in OSM in a more straightforward manner, relying on specific precomputed metrics such as the number of buildings or roads added or edited.

6.2 State of the Map US 2016 Presentation

The daily-editing global summary databases just introduced allow for daily-resolution editing statistics by quadkey. I used these databases to generate statistics for a presentation at the 2016 State of the Map US conference in Seattle, Washington.³ This section will share some of the main takeaways of that presentation: Stories of the US map’s evolution as reconstructed from annual-historical snapshots. First, Figure 6.3 gives an overview of the number of mappers editing the map of the US per week. These values were computed by the workflow presented in Section 6.1 which allowed me to count these users at the individual map-object level, per tile, per year. As this was my first iteration in scalability from Epic-OSM (Chapter 3), these object level, Nation-scale results validated tile-based OSM data analysis approaches for me.

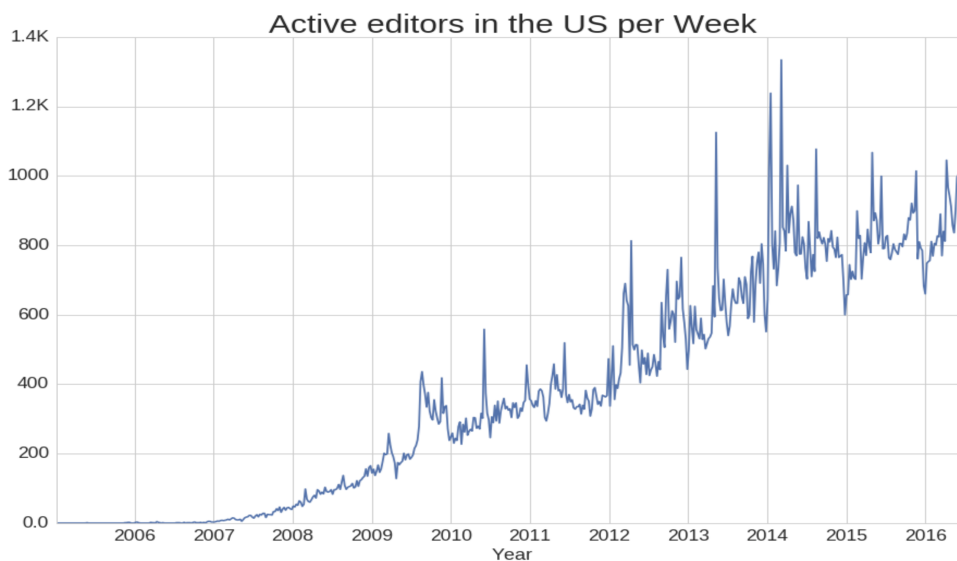


Figure 6.3: Active Editors in the US per week, calculated from annual historical snapshots. There is an observable linear rate of growth in the number of active US editors.

Filtering by quadkey allows us to group results by specific location, as shown in Figure 6.2. Here, I used quadkeys with bounds that cover various cities in the US with relatively active OSM communities to compare the evolution of these communities between the cities. Figure 6.4 shows

³ Mikel Maron and Jennings Anderson (2016). OpenStreetMap US by the Numbers, for the Community. Seattle, WA. July 23, 2016. Recording of the original talk available at 2016.stateofthemap.us/osm-us-by-the-numbers-for-the-community/

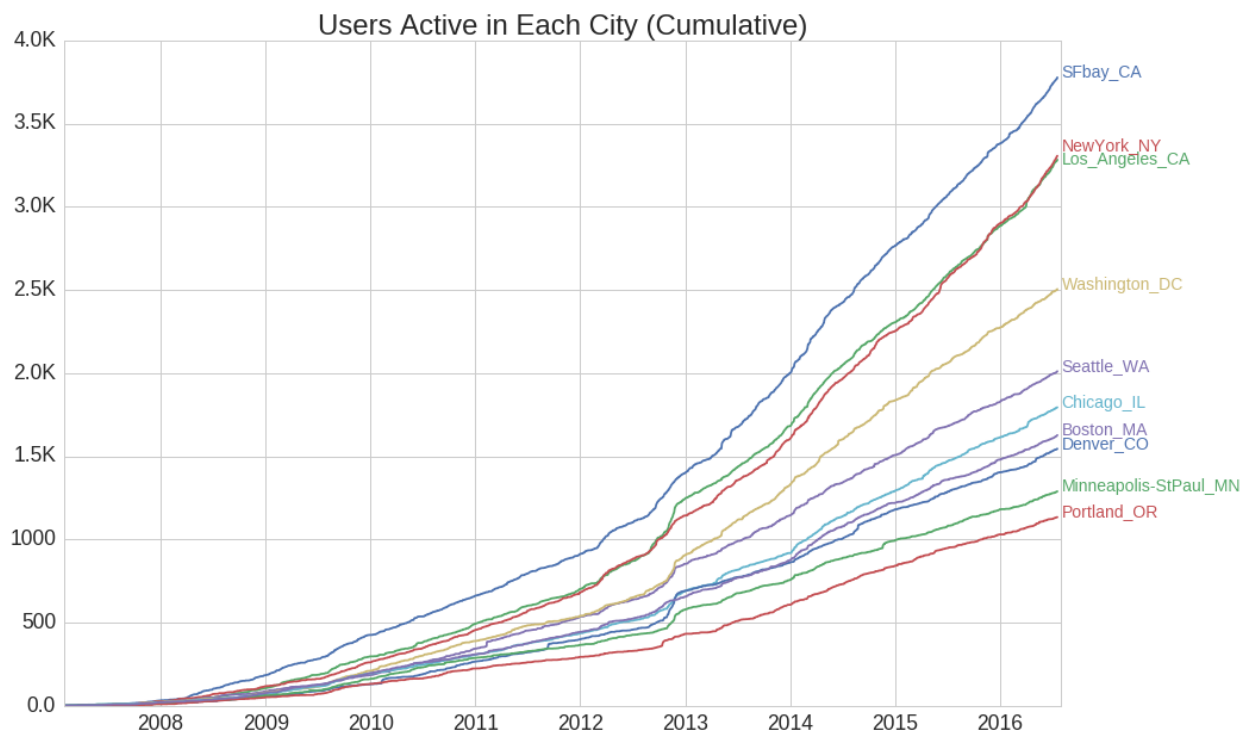


Figure 6.4: Cumulative count of editors in major US cities.

the cumulative growth in the number of mappers who have contributed to the map of the top 10 US edited cities in OSM. This figure shows relatively consistent growth in the editing communities across all of these cities at slightly different rates. There is a significant bump in contributors active in most of cities at the end of 2012. Investigating this finds a massive number of edits to the road network in the US, likely an organized data-cleaning effort that affected all of these cities. This is corroborated by a sustained spike in weekly activity shown in Figure 6.3. Of more interest, however, is that this event appears to be an inflection point to an increasing rate of local community growth in most of these cities.⁴

Similarly, we can compare the number of buildings edited over time across these cities, as shown in Figure 6.5. For these cities, the majority of the buildings have been added to OSM through community-led data imports that can be seen here as large steps in the graph. A building import involves acquiring building geometry data (footprints/roofprints) that have been made public (or at

⁴ “If you built it/fix it, they will come maintain/grow it?” - This is something I plan to investigate in the future.

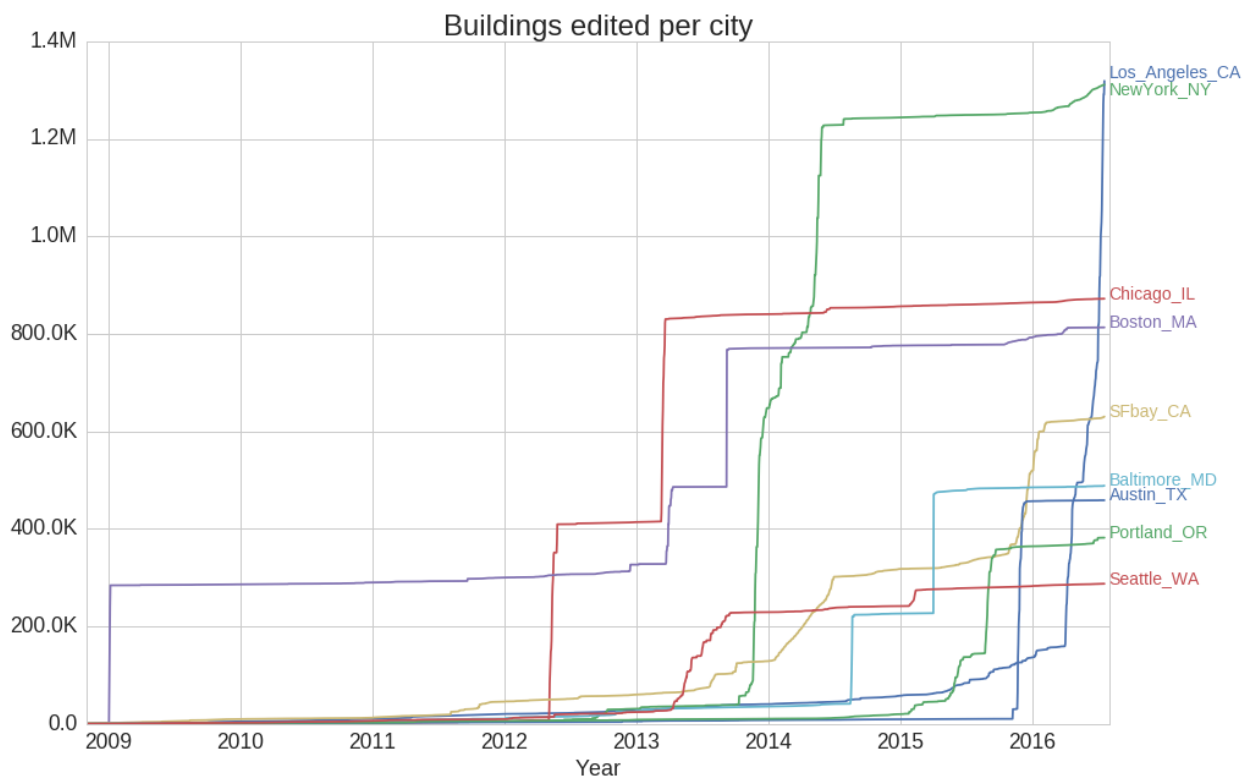


Figure 6.5: Cumulative count of buildings edited per major US city.

least released with an OSM-compatible license), and then systematically adding them to the map. It is best practice for mappers to create a different import-specific user account to separate the activity.⁵ Some imports have been done in multiple steps, such as Chicago which was paused due to licensing issues and then resumed.⁶ Other notable patterns in Figure 6.5 include the slow but sustained growth of building edits in Los Angeles before the import in mid-2016, and the three year head-start that Boston has on all of the other cities from the *MassGIS* import: An early addition of Massachusetts public data to the map in the US.

Taking a contributor-centric approach, I began querying the per-editor, per-tile, per-year database created via the workflow shown in Figure 6.1 to identify mappers who were active on the

⁵ It is common for users to simply add `.imports` to the end of their account to maintain an association, such as my import account name: `jenningsanderson.imports`.

⁶ Part of best practice for a building import is documenting the process and giving space for community feedback through the wiki. See the wiki for Chicago's import here: wiki.openstreetmap.org/wiki/Chicago,_Illinois/Buildings.Import

same tile in a given year. Earlier work of ours explored co-editing patterns where mappers were actively editing near each other [61]. Extending this idea with zoom level 12 tile-boundaries as the unit of analysis, I constructed co-editing networks based on the quantity of edits to specific object types. In these networks, nodes represent mappers and there exists a link between two nodes if both mappers edited a given quantity of the same object type (building or roads) on at least one tile together in a given year. For example, two mappers who edited at least 30 buildings and 50km of road on the same map tile would be connected in the 30-building/50km network. Using steps of 10 as quantity limits, I generated all of the possible co-editing networks for mappers editing in North America.

Figure 6.6 shows the co-editing network for mappers who edited at least 100km of roads in 2008. The various connected components on the right represent the clusters of mappers in 2008

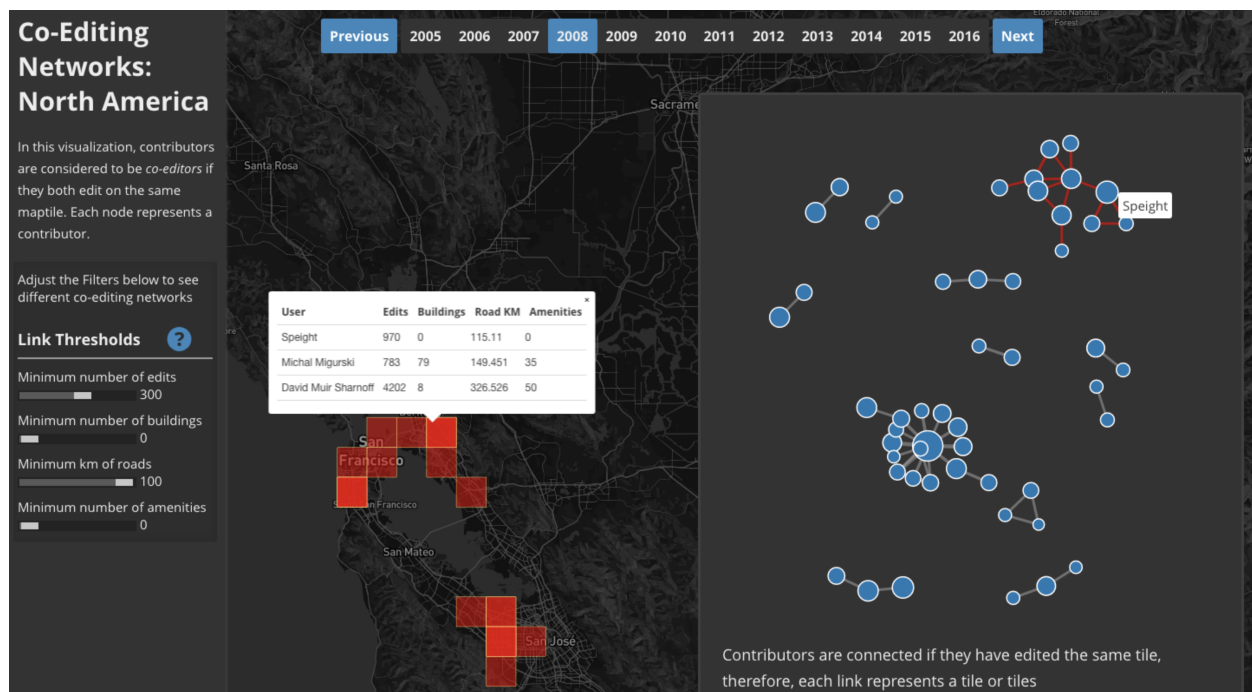


Figure 6.6: Screenshot of a visualization showing tiles where mappers edited at least 300 objects, affecting at least 100km of roads. The highlighted component in the top-right (with red edges between the nodes) represents all of the tiles that are shared between the users in this component. These tiles (in San Francisco) are shown on the left with per-editor statistics highlighted. Adjusting the link thresholds on the far left will further subset the network in the right-hand panel to match the constraints.

that satisfy these conditions. The large hub-and-spoke component in the center of the network is the TIGER road import in the US that year. The large node in the center represents the import account and the smaller nodes connected to it represent those editors who have edited over 100km of roads on tiles where there were imported data, likely integrating the imported data.

Created in the same way, Figure 6.7 shows the co-editing network for mappers with over 40 edits to buildings on the same tile in 2010. This Screenshot tells part of the story of the 2010 mapping response to the Haiti Earthquake [105, 125]. The highlighted cluster of nodes represents mappers who edited at least 40 buildings on the same tile in 2010. The map on the left shows that these tiles were all in Haiti. Put another way, most of the mappers who edited more than 40 buildings on a single tile in all of North or Central America in 2010, did so in Haiti: Presumably in response to the earthquake.⁷

Next, I used this same database of per-tile, per-editor, per-year editing summaries to calculate *every mapper's* global editing footprint: What they edited on which tiles over the entire world. Once these annual editing summaries for a mapper are computed, they are saved as single GeoJSON files and stored in an Amazon S3 bucket so they can be easily retrieved over HTTP. In sum, this is a few million GeoJSON files, each of which are under 1MB. Figure 6.8 shows a screenshot of the visualization tool I built to further explore these global editing footprints. The viewer first selects a year and then enters an OSM username. The tool will then request the mapper's editing summary from Amazon and render it on the map, showing how many edits happened on each tile, and moreover, what percentage of their total annual editing activity each tile comprises. Filters can be set to show only those tiles with with at least a specified number of edits or percentage of annual activity. Editing footprints for additional mappers can be added and will show up on the map in a different color so that multiple editors can be compared.

Figure 6.8 shows some of the same users from the 2010 building-editing network shown in Figure 6.7. Here we see that some of these mappers were also active all over the globe in

⁷ Comparing to Figure 6.3, the total number of active mappers in Haiti the week after the earthquake (about 500) is nearly twice the average number of weekly mappers in the US at that time.

2010. However, the selected user, *dbusse*, was most active in Haiti with 14% of their total 2010 editing activity on the selected tile, and over 50% of their 2010 editing activity when including the surrounding tiles. This implies more than half of this mapper’s 2010 editing activity was as part of the disaster mapping response to the 2010 Haiti Earthquake.⁸

This presentation of historical tile-based analysis of the map using annual-snapshots demonstrated to me the ability to effectively analyze the growth of the map at scale using zoom level 12 tile-boundaries as units of analysis. The ability to process years of planet-scale data was a major step in scaling from the previous Epic-OSM infrastructure.⁹

⁸ Also visible in Figure 6.8 are the artifacts of splitting up large Polygon geometries as explained by Figure 5.5.

⁹ Additionally, both the talk and the interactive visualizations were well received by the audience, which further validated this analysis approach, for me anyway.

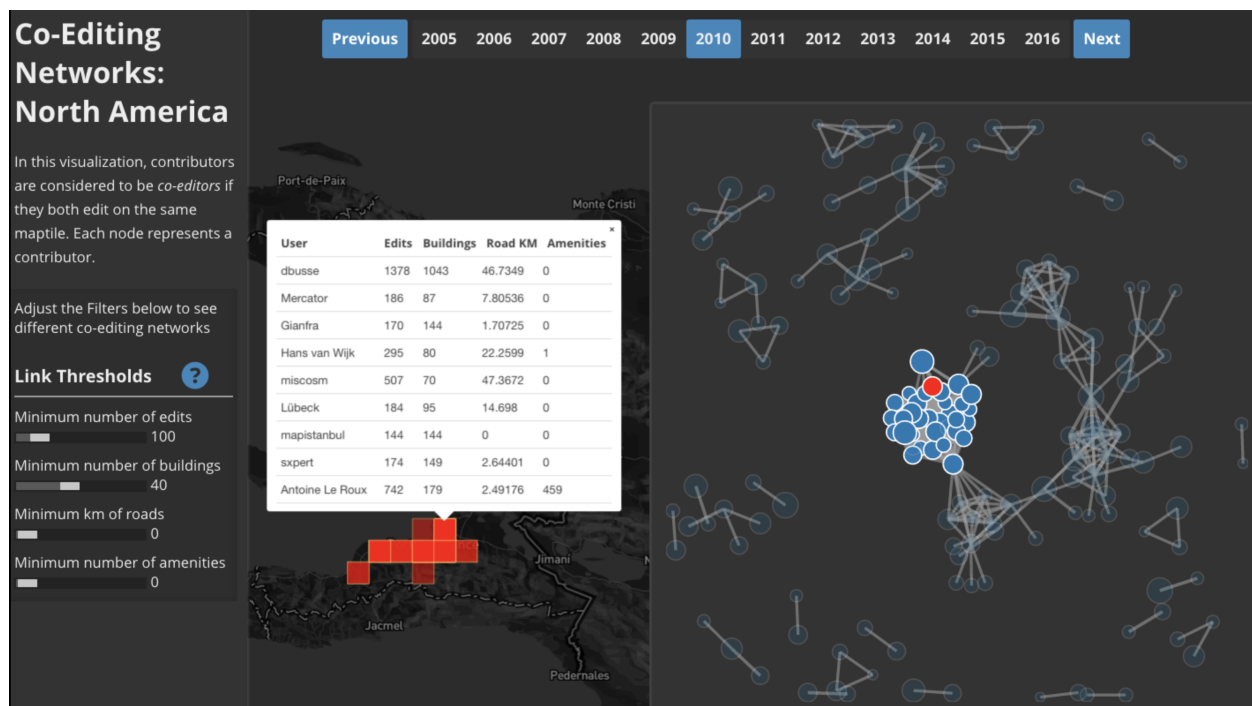


Figure 6.7: Screenshot of an interactive visualization showing North-American co-editing networks for users who edited over 40 buildings on the same tile. The selected cluster shows those who mapped more than 40 buildings in 2010 on the same tile in North America. The map on the left shows that this cluster represents mappers responding to the 2010 Earthquake [125].

6.3 Interactive Contributor-Centric Visualizations: First Generation

I published all of the visualization tools just presented along with some other more basic interactive maps: *editing density*, *recency of edits*, and *types of objects* at mapbox.github.io/osm-analysis-collab: A summary of my 2016 Mapbox Research Fellowship.¹⁰ I now refer to these as the first generation of contributor-centric visualizations because they are powered by datasets produced through my first iteration of contributor-centric approaches to OSM data analysis—reconstructing the evolution of the map through annual-snapshot OSM-QA-Tiles. These particular tools are *contributor-centric* because they rely on the metadata about the edit to the object more than the object itself. The final visualization tool I built this way does per-country aggregation of roads, buildings, and contributors, calculated annually.

¹⁰ These tools are still available at this address for the years 2005 through 2016.

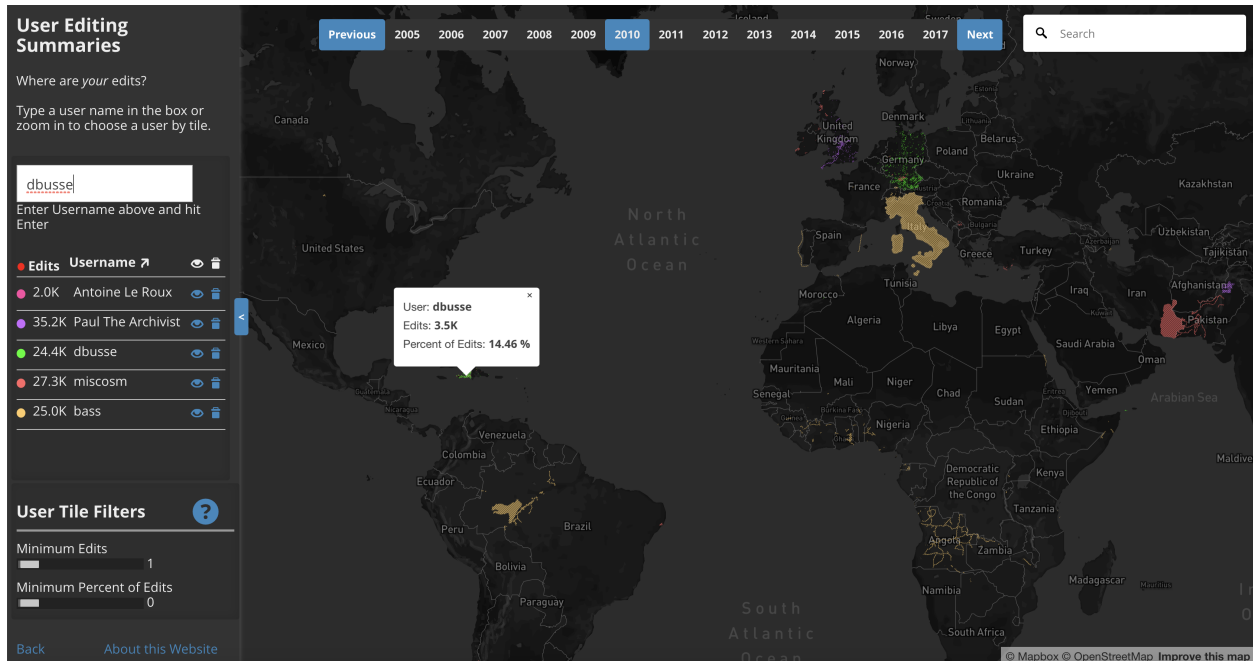


Figure 6.8: per-editor editing summaries show all of the tiles where a user was active in a given year. This screenshot shows all of the tiles that the users who co-edited on the tile selected tile in Figure 6.7 were also active on, individually. For the selected user, 14% of all of their edits in 2010 were on one tile in Haiti.

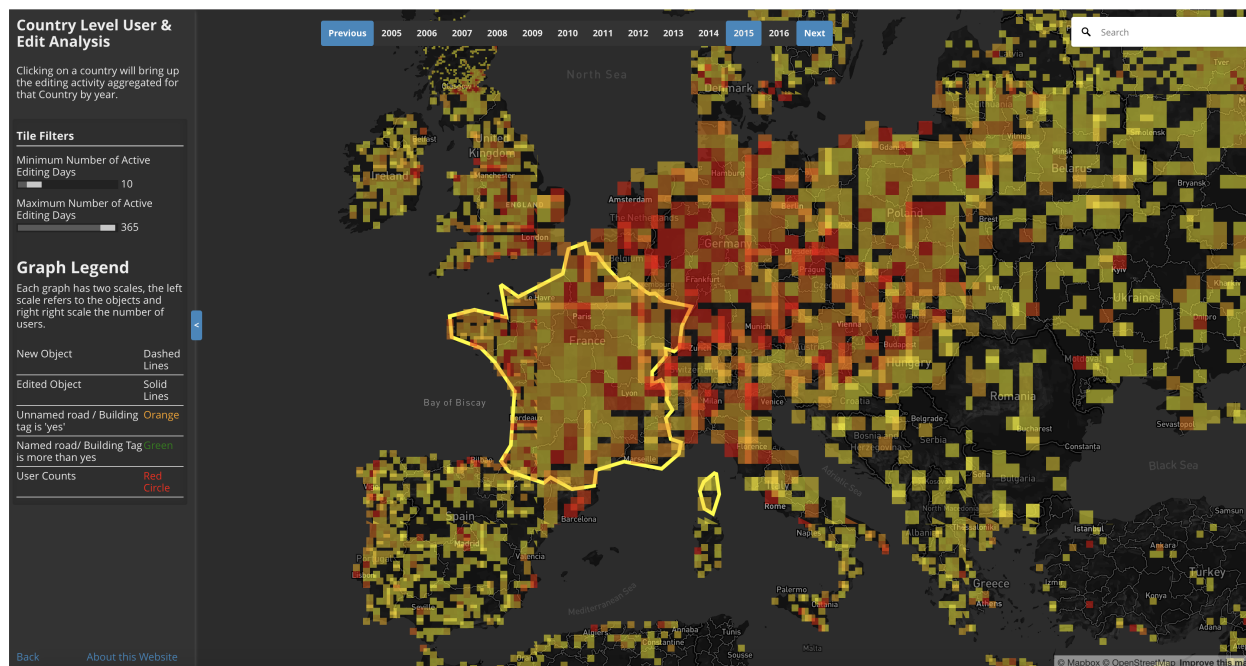


Figure 6.9: Screenshot of the *editors per country* interface. The user can set thresholds on the left to show only areas where editors were active for a specific range of days.

Figure 6.9 shows this visualization tool where an analyst has first selected 2015 (top) and the map is highlighting regions where editors were active for more than 10 days (non-consecutive) in 2015. The user has selected France by hovering their mouse over the geographic bounds of the country, which highlights it in yellow. When they click, A window with multiple graphs of editing activity over time pop up, as shown in Figure 6.10. These graphs compare the total editing activity over time to the number of editors active each day. For European Countries, Figure 6.10 is a fairly representative example: The number of daily editors grows steadily in the early years and then reaches some level of mapper saturation. In France’s case, 200 people edited the map of France each day between 2012 and 2016, on average. Since the number of daily editors (red) does not spike at the same time as the prominent spikes in daily editing activity (blue), these spikes are likely the result of automated edits. Using this utility, I found that the self-proclaimed bot-account, *PierenBot*, did 9.1M edits in France in 2011, which was likely responsible for the larest spike in daily editing in late 2011.¹¹

¹¹ If I were to recreate this analysis, I would attempt to identify automated edits and remove them. Unfortunately there is no foolproof method of doing this. I currently search for ‘bot’ in the user name, such as ‘xybot’ or ‘PierenBot’.

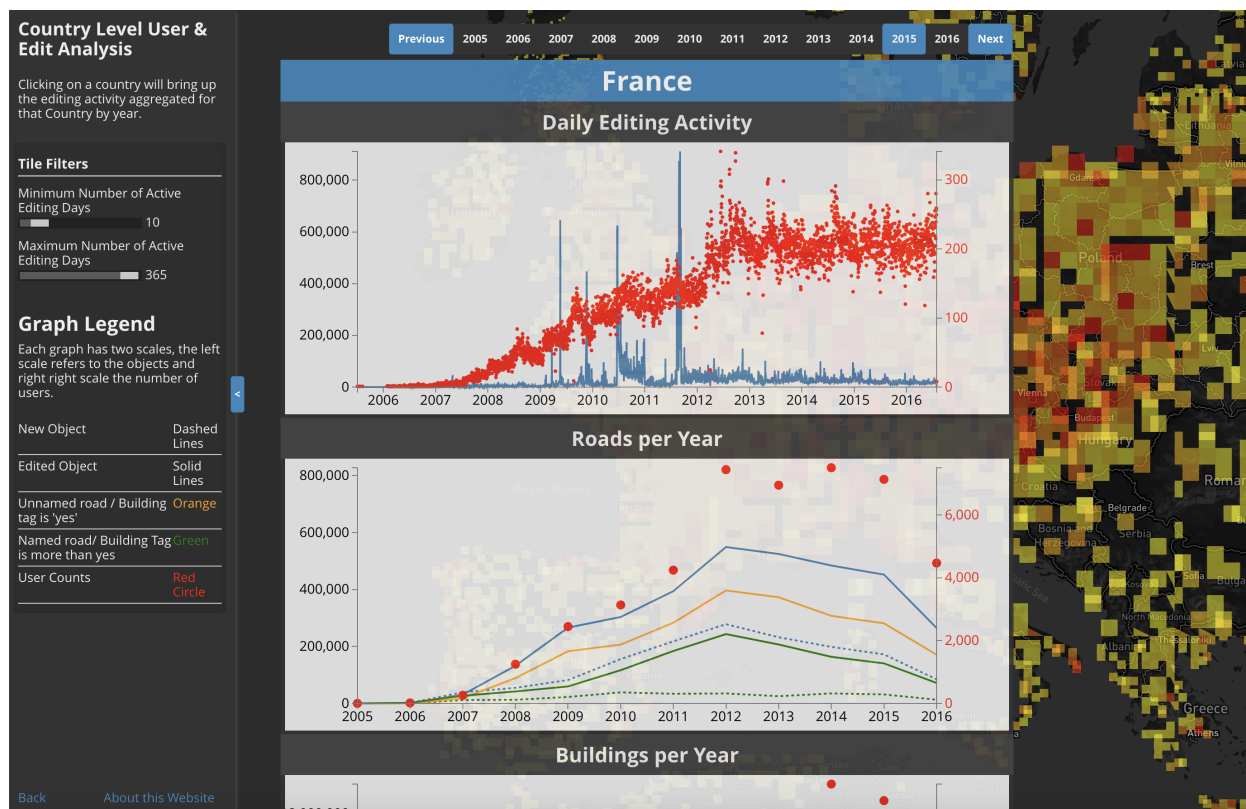


Figure 6.10: When a country is selected, the editing history is shown in a graph where the number of users active each year is depicted by a red dot (corresponding to the right-hand y-axis) and the total number of edits per object per year are represented by solid lines.

By clicking on a tile, the user can see general annual statistics about the roads and buildings as well as a list of users who edited on that tile that year as shown in Figure 6.11. Clicking *See all edits* will open this mapper's annual editing footprint, the visualization in Figure 6.8.

At the time, the primary motivation behind these analysis was a focus on *intrinsic quality analysis* of OSM data, driven by the contributor. These are meant to dive deeper into the quality metrics put forth in Haklay's early OSM data quality article [48]: Identifying areas with more recent or more contributors in general has major quality implications.¹² Building from this, I next used the data-processing workflow presented here for the data analysis supporting the research in the next chapter.

¹² See mapbox.github.io/osm-analysis-collab/osm-quality.html for a casual discussion of OSM data quality that I wrote on this topic regarding these visualizations.



Figure 6.11: Clicking on a tile will show the list of users active that year and a quick overview of created/edited roads and buildings. Clicking on *See all edits* will open that mapper's Global Footprint for that year (Figure 6.8)

Chapter 7

The Crowd is the Territory: Assessing Quality in Peer-Produced Spatial Data During Disasters

The following section is an exact reprint with the permission of my coauthors of an article published in the International Journal of Human Computer Interaction: Special Issue on Social Media in Crisis in Early 2018.¹ This work explores quality assessment of Volunteered Geographic Information, specifically assessment of OSM data produced in response to disasters and humanitarian mapping efforts. The data analysis here uses annual snapshot historical OSM-QA-Tiles processed via the workflow presented in Figure 6.1.

7.1 Introduction

Here we examine methods for assessing the quality of peer-produced spatial data, or Volunteered Geographic Information (VGI), for use in crisis response. For many parts of the world, VGI is the primary geospatial data source because it is the most accessible and complete source of data for the area [105, 134]. As such, crisis responders often use these datasets during disasters. For example, the 2010 Haiti Earthquake destroyed much of the country's government buildings, and with them, access to official mapping resources [125, 105]. In just a few days, organizing online, hundreds of contributors to OpenStreetMap (OSM) created the most complete map of Haiti in existence. This map became the de-facto basemap for subsequent rescue and relief operations [125]. This early

¹ Jennings Anderson, Robert Soden, Brian Keegan, Leysia Palen, and Kenneth M. Anderson (2018). The Crowd is the Territory: Assessing Quality in Peer-Produced Spatial Data During Disasters. International Journal of Human-Computer Interaction. doi:10.1080/10447318.2018.1427828

instance of “disaster mapping” was a catalyst in creating a new form of volunteer disaster response work [105]. Today, thousands of online volunteers mobilize before, during, and through the recovery phase of a disaster to answer the data needs of the community and responding organizations.

Crisis informatics research seeks to understand how new technologies enable volunteers to mobilize, create, and process information about a disaster event. Therefore, of specific importance to both the OpenStreetMap and the crisis response communities is disaster mapping (or crisis mapping). Disaster mapping is the practice of volunteer contributors converging online to improve the map for a region experiencing a disaster or crisis ([32, 108]. In the case of OpenStreetMap, the Humanitarian OpenStreetMap Team (HOT) is an active community organizer in coordinating these tasks all over the world [105, 108]. These activities leave a very distinct contribution pattern on the map: specific regions with considerably more coverage of certain objects (typically roads and buildings) than the surrounding area. These improved maps are used for emergency response, planning, routing, and more [125, 126]. With the widespread use of OSM data in disaster response, developing and validating measures of information quality for it is essential.

Studies of online peer production systems like Wikipedia have demonstrated that high-quality, open source content can be generated by integrating contributions from non-experts [12]. VGI systems like OpenStreetMap emulate many of the features of these peer production systems, but the spatial—rather than textual—knowledge they encode requires alternative methods for measuring and validating the quality of their user-generated content. Developing methods for assessing the quality of spatial information is a fundamental issue within the study of geographic information science (GIScience). Any representation of spatial information necessarily involves some loss of detail and thus quality. The challenge that this presents is illustrated by Borges’ famous parable that imagines a civilization so obsessed with precision that they constructed a 1:1 scale map of their territory. The result of their labor was perfectly accurate but totally unusable (Borges, as quoted in [33]). This problem is faced not just by the designers of maps, but of all information systems: maps are abstract, incomplete, and imperfect portrayals of the phenomena they are created to represent. This condition, illustrated by Korzybski’s famous maxim that “the map is not the territory,” renders

questions of information quality more complex than they might initially appear [62].

A central motivation for this work is the lack of authoritative reference geographic data in many parts of the world, making more traditional quality analysis by comparing to reference data sources—referred to here as extrinsic methods—impossible. Our research expands upon existing methods within GIScience for assessing map quality, which rely on attributes such as completeness, consistency, and accuracy, to take advantage of the behavioral meta-data of VGI contributors' activities that are unavailable to traditional data sources. We draw on previous research measuring the information quality of Wikipedia articles based on the intrinsic processes generating them, such as the number of editors or how recently the data has been updated. We identify analogous generative features in OSM data and evaluate three metrics drawing on contributors' histories and temporal contexts to examine their relationship with alternative intrinsic information quality metrics. Both intrinsic and extrinsic quality assessments of VGI have been explored in GIScience. Our metrics are distinct in that they rely primarily on the metadata of the individual contributions and contributors: the details and context of how the digital volunteers converged, not just the geographic features that were contributed. This distinction connects this work from the more traditional approaches of GIScience to the fields of social computing and human computer interaction.

Using a quantitative case study method, we identify four different areas of the global map that have been the geographic focus of disaster mapping activities in the past. For each of these areas, we apply our three proposed intrinsic quality metrics, which expose varying histories of contributions, each telling a different story, consistent with its associated crisis event. We then apply these metrics to areas of the map known and agreed to be of very high-quality for comparison. The differences—exposed by these metrics—suggest we are capturing substantively different mechanisms by which VGI information is contributed, which, in turn, has implications for quality assessment. The following sections will, first, discuss the background of information quality and quality assessment in peer production; second, discuss the OSM project and describe our dataset; then next introduce three approaches to intrinsic quality analysis based on contribution metadata; and then, finally, evaluate our methods applied to various parts of the map that have been the sites of disaster

mapping in the past. We conclude with a discussion of how these metrics fit within the larger domain of geospatial data quality assessment and offer suggestions for future work.

7.2 Background

How to measure information quality has been the subject of a substantial body of research across information science, management-related fields, and geography. We begin by reviewing work on measuring information quality using extrinsic and intrinsic data sources in the context of peer production and spatial information. The majority of prior literature assessing the quality of OpenStreetMap—with a few notable exceptions (Barron et al. [10, 82])—has typically focused on assessing quality relative to authoritative data sources; it, therefore, overlooked the potential offered by specific intrinsic features unique to VGI to measure the quality of peer-produced data. This gap between the value of intrinsic features for measuring information quality and the underutilization of these features unique to VGI for assessing quality in OSM motivates our subsequent analysis to employ contributors’ histories and temporal contexts as *intrinsic* sources of VGI quality.

7.2.1 Data and information quality

In this section, we (1) identify commonly-used dimensions for measuring information quality through extrinsic and intrinsic dimensions; (2) examine how the quality of spatial information is traditionally assessed; and (3) discuss the importance of spatial information quality for safety-critical operations such as disaster response.

7.2.1.1 Information quality frameworks

There are many sources of variance in information quality. Information quality problems arise because of incomplete, ambiguous, inaccurate, inconsistent, or redundant mappings between real world properties and their representation in an information system [137, 65]. We employ a taxonomy that differentiates information quality based on their use of *extrinsic* or *intrinsic* metrics.

Extrinsic information quality metrics focus on the accuracy, completeness, or consistency of

the object-based measures by referencing external data sources. Questions about the syntactics (conformity to other collected data; e.g., consistency) or semantics (correspondence to external or authoritative phenomena; e.g., accuracy) are paramount. In contrast, *intrinsic information quality metrics* use features of the target dataset itself to assess quality by examining contexts, reputations, and processes for generating information. Questions about pragmatics (use and interpretation of information; e.g., timeliness or authority) are paramount. This dichotomy, while simplistic, is useful for identifying gaps in existing approaches for measuring information quality, especially in the context of online peer production communities like Wikipedia and OSM.

7.2.1.2 Quality assessment of spatial information

Though there are many approaches for assessing map quality, those offered by the International Standards Organization (ISO), codified as ISO Standard 19113, are widely accepted. The standard has five primary approaches to assessing quality [10], summarized here:

- (1) *Completeness* - Is the dataset complete?
- (2) *Consistency* - Are the spatial and thematic attributes of the data in a uniform fashion?
- (3) *Positional Accuracy* - How accurate are the coordinates of the map objects?
- (4) *Temporal Accuracy* - If the data has a temporal element, is it accurate?
- (5) *Thematic Accuracy* - Are the quantitative/qualitative attributes of the data accurate?

As we discuss below, each of these dimensions, apart from consistency, implements quality assessment as an extrinsic information quality metric by referencing similarity to an authoritative dataset. In some cases, extrinsic measures have used proxy datasets, such as kilometers of road in relation to population density [31], in an attempt to assess completeness when a suitable source of reference data is not available. Consistency, on the other hand, is the sole example of a quality metric in this ISO Standard that uses features intrinsic to the target dataset to assess quality.

7.2.2 Information quality metrics for peer production

Wikipedia’s radical “anyone can edit” model integrating user-generated contributions into an authoritative encyclopedia justifiably raised concerns about the quality of the resulting information.

Evaluations of Wikipedia quality emphasize that features such as the quantity of information or the number of links in an article are the most important determinants of end users' trust in Wikipedia content [60, 134, 141]. Despite the major differences in the substantive content of contributors' edits, the technical designs of both the Wikipedia and OSM systems implement analogous methods for merging user contributions into a single canonical version as well as capturing similar kinds of meta-data in revision event logs about user IDs, timestamps, and content versions. This opens the possibility for translating information quality metrics from a well-validated domain like Wikipedia to a less studied domain like OSM. We compare the extrinsic and intrinsic information quality metrics used in prior research on both Wikipedia and OSM below.

7.2.2.1 Extrinsic information quality metrics

We define extrinsic information quality metrics as object-based measures focusing on syntactic or semantic “correctness” that reference external authoritative data sources. Online peer production systems like Wikipedia and OSM were created to replace authoritative incumbent products like *Encyclopedia Britannica* and government land surveys (respectively) created by expert organizations. Thus, assessing the quality of user-generated information by comparing it to expert-generated counterparts is a natural validation step. Extrinsic metrics for assessing the accuracy of Wikipedia articles have used experts to compare the number of errors in Wikipedia against other works of reference, finding that error rates were similar to or lower than authoritative sources [40, 113]. Other studies have explored the completeness of Wikipedia's coverage by measuring the representation or overlaps in topical coverage across sources [17, 50, 115, 116].

The first scholarly investigation of OSM's extrinsic information quality assessed the completeness and positional accuracy of the OSM road network for the United Kingdom as compared to the authoritative Ordnance Survey [46]. Although it was inconsistent, the OSM data compared favorably to the government's dataset and judged to be of good quality. Such findings are consistent with work that examined other geographic locations and employed a wider range of quality measures [144].

7.2.2.2 Intrinsic information quality metrics

We define intrinsic information quality metrics as process-based measures focusing on pragmatic or contextual “authority” by examining the processes generating information. Most Wikipedia studies employ intrinsic measures to assess information quality and validate against community-generated labels of article quality [56, 130, 139]. Behavioral features like the number of revisions, the number of revisions from administrative, registered, or anonymous editors, the number of unique editors, number of reverts, and time since last revision are intrinsic characteristics that are easily computed from revision event logs [129]. Content features such as word count [15]), number of references [69], images, and tables [1] also provide popular metrics.

Intrinsic measures of data quality are growing increasingly important to assess the quality of OSM data due to the lack of authoritative reference datasets. For many parts of the world, OSM is the most complete geographic dataset. This situation can arise because of a lack of good, official data—as is the case in some developing countries—or simply because contributions from an active local mapping community outpace official survey work. Whatever the reason, the lack of high-quality reference data limits the utility of extrinsic measures of quality in these situations. Barron, et al. acknowledge that “the quality of OSM data also depends on the project’s contributors” [10]. Preliminary frameworks exist for evaluating intrinsic quality [10], evaluating the consistency of tagging schemes [4, 135], and investigating “user-centric” quality metrics based on contribution meta-data [46, 75, 61]. GIScience has recently seen many new intrinsic quality metrics introduced with respect to OSM. Barron et al. introduce a framework discussing 25 measures requiring no external reference sets [10] with a comprehensive review of current approaches and explanations of intrinsic quality metrics as applied to VGI. More recently, Sehra et al. created an extension for the QGIS open-source software project to allow for easier OSM data analysis, analogous to tools in commercial GIS software [120].

Barron et al.’s intrinsic quality assessment framework highlights the importance of “fitness for purpose” in quality analysis, and separates the 25 metrics into six distinct categories to connect

metrics and indicators (assessments) to specific purposes. For example, road network completeness is an assessment relevant to the use case of routing and navigation [10]. One of the areas Barron et al. consider is “user and information behavior” to include contributor activity as an indicator for quality assessment. They find that the distributions of edits per user are heavily skewed—with a few contributors doing most of the work—a finding that is common among all peer production systems. Importantly, they note while it may be expected that contributors with high edit counts create higher quality data, this thesis remains untested [10]. Their call motivates the work presented here; we further explore user-based metrics with respect to the number of contributors and their respective expertise. Furthermore, as Eckle and Albuquerque point out, OSM data contributed during disaster mapping events often contains just the raw geometry (that is, an outline of a building or the path of a road) without the contextual information of attributes describing it, making existing intrinsic quality assessment techniques which rely on the object’s attributes alone (such as name or type) difficult or impossible [32].

Given this, intrinsic quality assessment based on contributor metadata becomes the most feasible type of quality assessment available for many areas of the map and this observation motivated our work in developing the metrics presented below; our metrics can be applied to any part of the map, independent of reference datasets or detailed object attributes.

At a high-level, our three metrics are straightforward to understand. Our first metric is a variation on a simple contributor-based metric: the absolute number of contributors that have been active in a region of the map. Our second metric looks at the types of objects that different contributors prefer to edit and the amount of that object type they have edited before. Our third metric looks at the overall editing evolution of a region in terms of what objects are being collectively edited by the contributors. Each of these metrics relies solely on the basic object type and the contribution metadata. This provides the *who*, *what*, *when*, and *where* attributes of each edit, and enables investigation of how the map developed in any given region. Our metrics are not replacements for other quality metrics, but rather provide richer context from which to understand the resulting OSM data.

7.3 Dataset and Methods

7.3.1 OpenStreetMap

Started in 2004, OpenStreetMap is an open geospatial database released under the Open Database License. The main rendering of the database can be viewed as an interactive map on www.openstreetmap.org. The map is also available as a set of tiles through a web service. As a result, OSM is used as a basemap for many interactive web-based maps. The OSM website currently has over 4 million registered users; though less than 1 million users have ever edited the map data. The database has over 4 billion unique geographic points that make up the objects on the map. To illustrate the degree of completeness of the global map, Figure 1 shows just the road network in OSM.

Objects in OSM are defined by a set of *tags*: key-value pairs that identify a country boundary from a park or a bike path from a major street. The objects we focus on are roads and buildings. These are the most common objects in OSM as well as the most edited objects during disaster mapping. They are tagged as *highways* and *buildings*.

7.3.1.1 Roads (Highways)

In OSM, a road is a geometry known internally as a ‘way’ which is semantically tagged with the key ‘highway’ and an associated value describing its relative prominence such as primary, tertiary, walkway, etc. When roads are traced from remote imagery—as is common in disaster mapping—they are rarely tagged with a ‘name’ attribute. Indeed, a road with a ‘name’ attribute can be considered to contain some level of local, ground-truth knowledge, likely implying higher quality.

7.3.1.2 Buildings

A building is denoted by a tag describing its purpose, such as {‘building’: ‘residential’} or, in many cases, simply {‘building’: ‘yes’}. In OSM, buildings are typically represented by

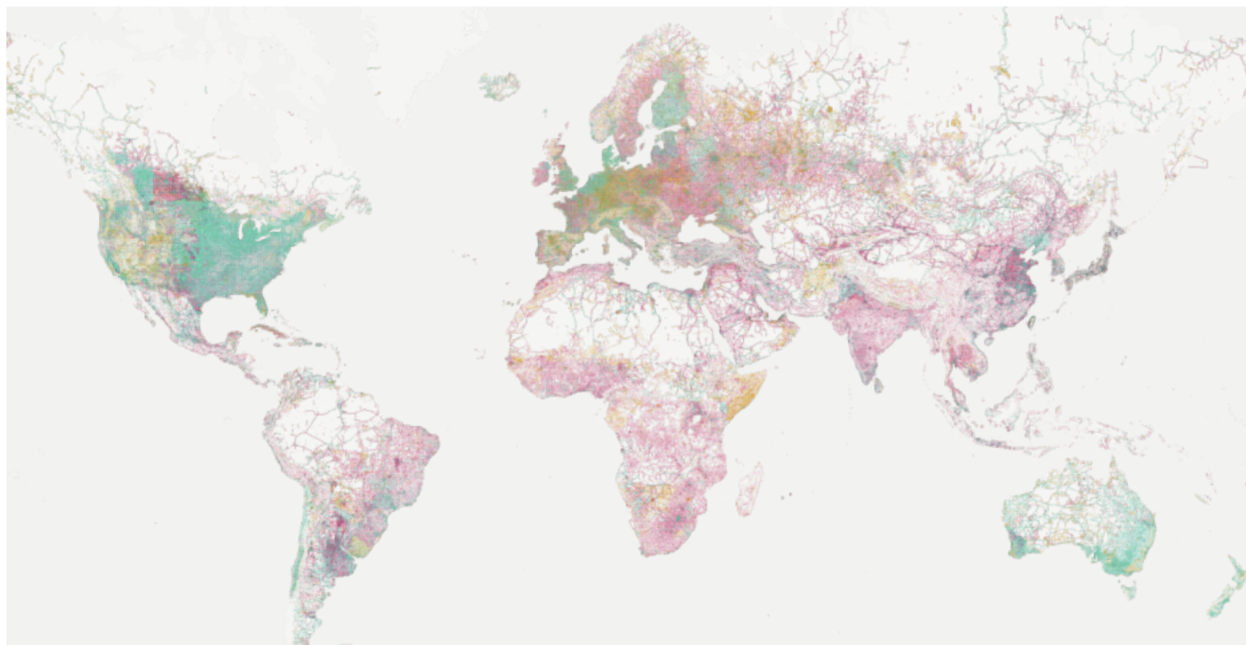


Figure 7.1: The Road network in OpenStreetMap, showing global coverage and colored by existence of the name attribute. Cyan roads include a name; magenta or orange roads and paths do not. Map Data ©OpenStreetMap Contributors.

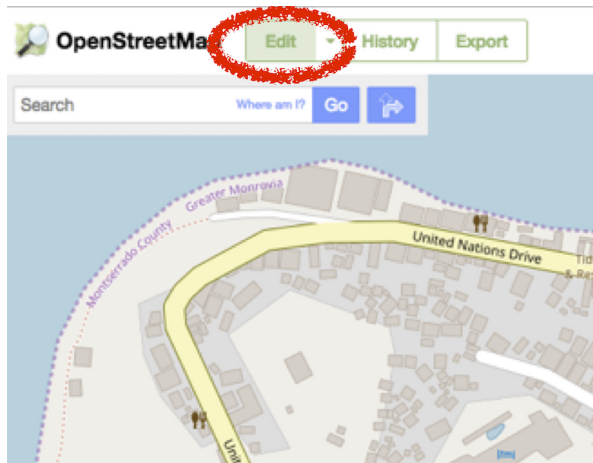
closed ways, i.e., line geometries that have the same start and end points. As of October 2017, building is the most common tag in OSM, with over 5.5% of all objects having this tag.²

Contributors can edit OSM through an in-browser editor on openstreetmap.org or through stand-alone map-editing tools that communicate directly with the database through the API. Editing OSM is depicted in Figure 7.2.

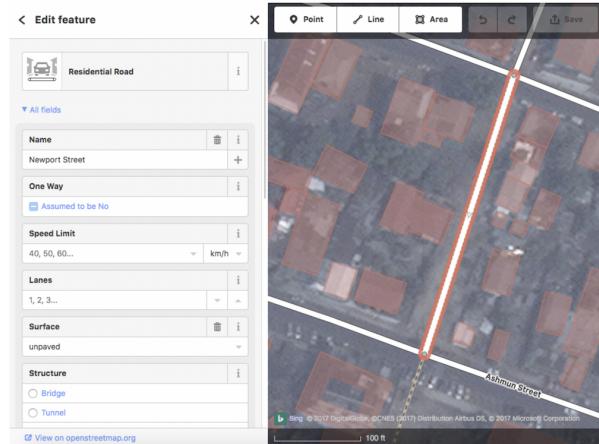
7.3.2 Obtaining OSM Data

Obtaining and manipulating OSM editing data is possible through a set of public APIs and downloadable database files. A number of open source tools are available for converting the data between popular geospatial data formats. A format made popular by the web for efficient storage and serving of map data is the vector tile. A vector tile stores the geometry, attributes, and metadata for every map feature organized by geographic location. Tiles can be created at various zoom levels,

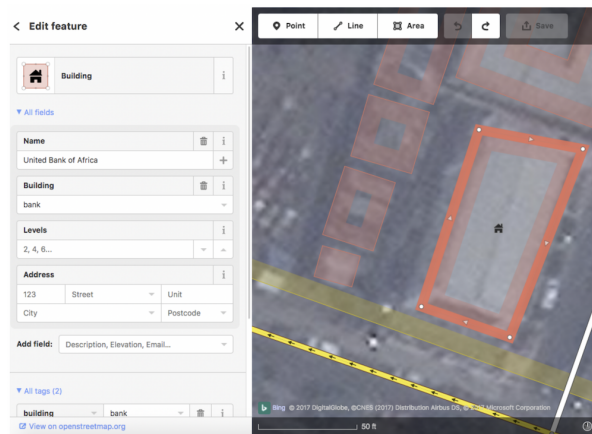
² taginfo.openstreetmap.org



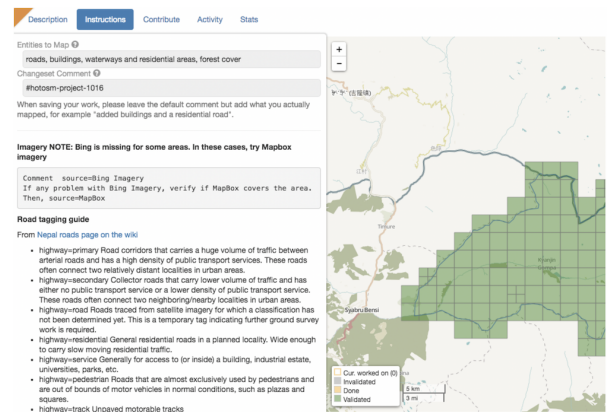
(a) The edit button on openstreetmap.org above the map



(b) Editing a road object with the in-browser editing interface, the iD editor.



(c) Editing a building feature with iD. Suggestions of attributes to add such as number of levels or the specific address are presented on the left.



(d) The Tasking Manager from Humanitarian OpenStreetMap Team (HOT) gives instructions on what to map for a specific disaster event and helps mappers coordinate their efforts. (tasks.hotosm.org)

Figure 7.2: Editing OSM on openstreetmap.org

each with a different resolution of data. The tiles used for our analysis are generated at zoom level 12. At this level, the inhabited part of the earth is comprised of about 2.5 million tiles, and these tiles have an area of roughly 100 square-kilometers at the equator.

For each of our analyses below, snapshots of the map at annual intervals from January 1, 2006 to January 1, 2017 are used to achieve annual granularity of the history of the database. We also note that in some cases where the same object was edited multiple times in one year, only the last

edit of that year is counted. Some of our reported numbers are therefore an under-representation of the total editing activity in OSM. We use an open-source Javascript framework called tile-reduce to process these vector tiles in parallel.³ We perform all spatial analysis with open-source GIS tools, and our full data processing pipeline includes a combination of javascript, postgresql, and python.

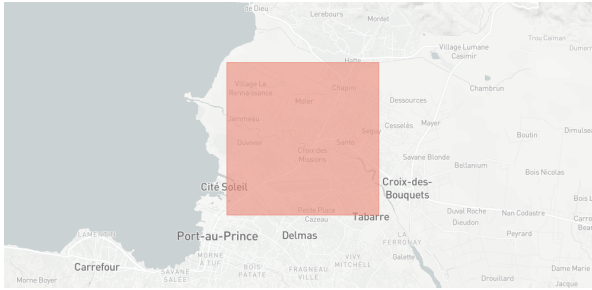
7.3.3 Our Dataset

To evaluate our three intrinsic quality metrics with respect to the insight they can provide into the practice of disaster mapping, we selected four distinct tiles on the map that have been the geographic focus of disaster mapping events following different kinds of events (see Figure 7.3). These areas are: (1) Port Au Prince, Haiti, the scene of one of the first instances of major coordinated disaster mapping following the 2010 Earthquake; (2) Tacloban, Philippines, where disaster mappers digitally converged before, during, and after Typhoon Yolanda in 2013; (3) Monrovia, Liberia, a region that was part of a year-long humanitarian-focused mapping project to help relief and prevention efforts during the 2014 Ebola outbreak; and (4) Trisuli Bazar, Nepal, a region heavily impacted by the 2015 Earthquake.

At the time of these events, each of the associated disaster-mapping activations was the largest to date in terms of number of contributors. Study Tile 3 (Monrovia, Liberia) is different from the rest because the activation in response to the ebola outbreak was not a single, rapid convergence of contributors, but rather a long-term project that saw thousands of volunteers over a period of months. In comparison, the other events saw a period of rapid mobilization as contributors converged on OSM in immediate response to the natural hazard. We expect to see distinct differences then in our results between these regions. The mapping tasks are similar across all the events: for disaster mapping, tasks focus on performing detailed mapping of buildings and roads in specific regions.

For quality comparison, we have chosen two well-validated areas of the map: London, UK and Heidelberg, Germany. Previous extrinsic quality research found that these tiles are of high quality when compared to external reference datasets [7, 48]. We compare the study tiles with

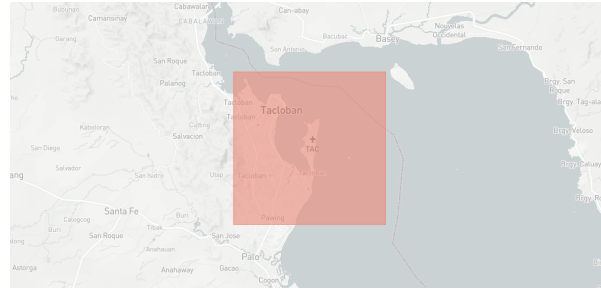
³ github.com/mapbox/tile-reduce



(a) Study Tile 1: Port Au Prince, Haiti

km of road	1,006 km (54% with names)
# of buildings	12,141 (7% labeled)
Contributors	494

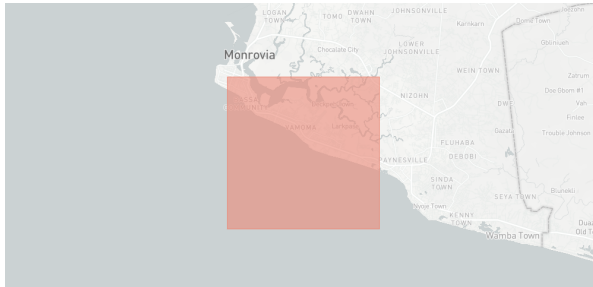
In response to the January 2010 Earthquake, hundreds of users contributed tens of thousands of features to the map to aid disaster relief, creating the most comprehensive map of Haiti to date [125, 145]



(b) Study Tile 2: Tacloban, Philippines

km of road	257 km (35% with names)
# of buildings	29,573 (71% labeled)
Contributors	371

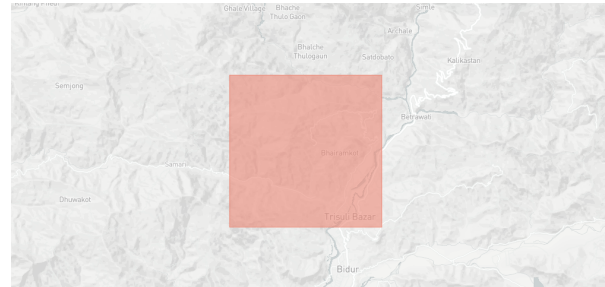
Striking the Philippines in 2013, Typhoon Haiyan (Yolanda) prompted contributors to improve the map in the Tacloban region, specifically updating buildings on the map for damage assessment [105]



(c) Study Tile 3: Monrovia, Liberia

km of road	174 km (32% with names)
# of buildings	19,193 (6% labeled)
Contributors	202

In 2014, HOT coordinated disaster mapping efforts in West Africa in response to the Ebola Outbreak. Lasting many months, this was distinctly different from the rapid convergence of contributors on the other three tiles.



(d) Study Tile 4: Trisuli Bazar, Nepal

km of road	324 km (3% with names)
# of buildings	7,596 (16% labeled)
Contributors	257

The largest disaster mapping event to its time, thousands of contributors were mapped Nepal in response to the April 2015 Nepal Earthquake. Though contributors mapped all over the country, this tile represents one area with a lot of activity. [108]

Figure 7.3: Details of the four study tiles selected for contribution-based intrinsic quality analysis. Data retrieved at the beginning of 2017. See the OpenStreetMap wiki for more information on these activations:

wiki.openstreetmap.org/wiki/Typhoon_Haiyan, [2014_West_Africa_Ebola_Response](http://wiki.openstreetmap.org/wiki/2014_West_Africa_Ebola_Response)

wiki.openstreetmap.org/wiki/2015_Nepal_earthquake

these high-quality tiles for each metric. The differences suggest that our metrics are capturing contribution patterns unique to disaster mapping.

Though today these tiles may appear complete, our metrics aim to expose the differences in the histories of how the data were contributed. For each metric, we discuss the specific implications the findings may have for measuring intrinsic information quality in VGI.

7.4 Contributor-based Intrinsic Quality Metrics

We extend one existing intrinsic quality metric and propose two new intrinsic information quality metrics for VGI. These metrics apply to vector tiles of OSM data. Our metrics explore attributes of the data beyond geometries and visible properties; instead, they examine features specific to peer-produced spatial data. This includes information about a contributor’s previous experience with the platform for each individual contributor and the time when an object was last edited. Specifically, our metrics are:

(1) **Contributor Density Over Time**

How many users have been active on a given part of the map? Denser maps should have higher quality as more people have been active in the area. This is a straightforward measure that was first explored by Haklay et al. [48]. Our extension to this measure focuses on temporality, looking at the cumulative density over time and marking when the bulk of contributors were active.

(2) **Contributor Experience**

How long has a contributor been active in the OSM community? What types of objects have they mapped before? We expect that areas with experienced contributors should have higher quality. This metric works to supplement the straightforward measure of contributor density by further inspecting who the contributors are. The need for such a metric becomes especially important when considering mapping events that attract newcomers. This measure asks: “Who does a majority of the work: many new contributors, or fewer experienced power users?” Depending on this distribution, the cumulative number of contributors per square kilometer may not be as important.

(3) **Tile Maturity**

How is the composition of objects changing over time? Areas where contributions are focused on maintaining existing features instead of adding new features may have achieved some level of completeness, itself a quality measure. Instead of examining qualities of individual contributors, this measure instead considers collective editing activity by looking at the bulk of types of edits in a region over time.

7.4.1 Metric 1: Contributor Density

In one of the first intrinsic quality studies of data quality in OSM, Haklay et al. found that after 15 mappers have been active in a given square-kilometer, the positional accuracy below 6-meter resolution is “very good” in comparison to government data [46]. This study also revealed that the first five mappers to an area make the greatest impact to the positional accuracy of the data. This contributor-density method draws inspiration from Linus’s Law of open source software development: “given enough eyeballs, all bugs (in software), are shallow.” For OSM, the contributor-density method assumes that more mappers contributing to an area provides a greater chance that some level of data validation and quality assurance has been achieved [48].

Globally, less than 1% of zoom-12 tiles reach Haklay’s threshold of 15 contributors per square kilometer. When we examined our tiles, we found that both Port Au Prince and Trisuli Bazar reached this threshold during their respective disaster mapping events (see Figure 7.4). This initially suggests that the quality of these tiles became “very good” as contributors mobilized in response to the event. The spikes in contributor activity at the time of the event for Tacloban and Monrovia are significant and represent the most activity ever to occur on these tiles, but still do not reach this particular threshold of 15 contributors per square kilometer. Figure 7.4 also shows the density of contributors in London and Heidelberg, which surpassed 15 users/km² in 2008 with steady growth of an active OSM community since.

Figure 7.5 looks beyond cumulative density to contributor count over time to reveal the rate of growth for the number of distinct contributors. As expected, the time of the disaster-mapping event creates the most significant spike in contributors for each tile. This spike shows many contributors active during a relatively short amount of time and then never returning to edit in this area. This may lead to the staleness of the map data (discussed next).

In contrast, London and Heidelberg show the sustained growth of a contributor community. These communities grow steadily from the beginning and seem to level off in recent years, perhaps suggesting a level of saturation of contributors in the region. Knowing these tiles are of high-quality

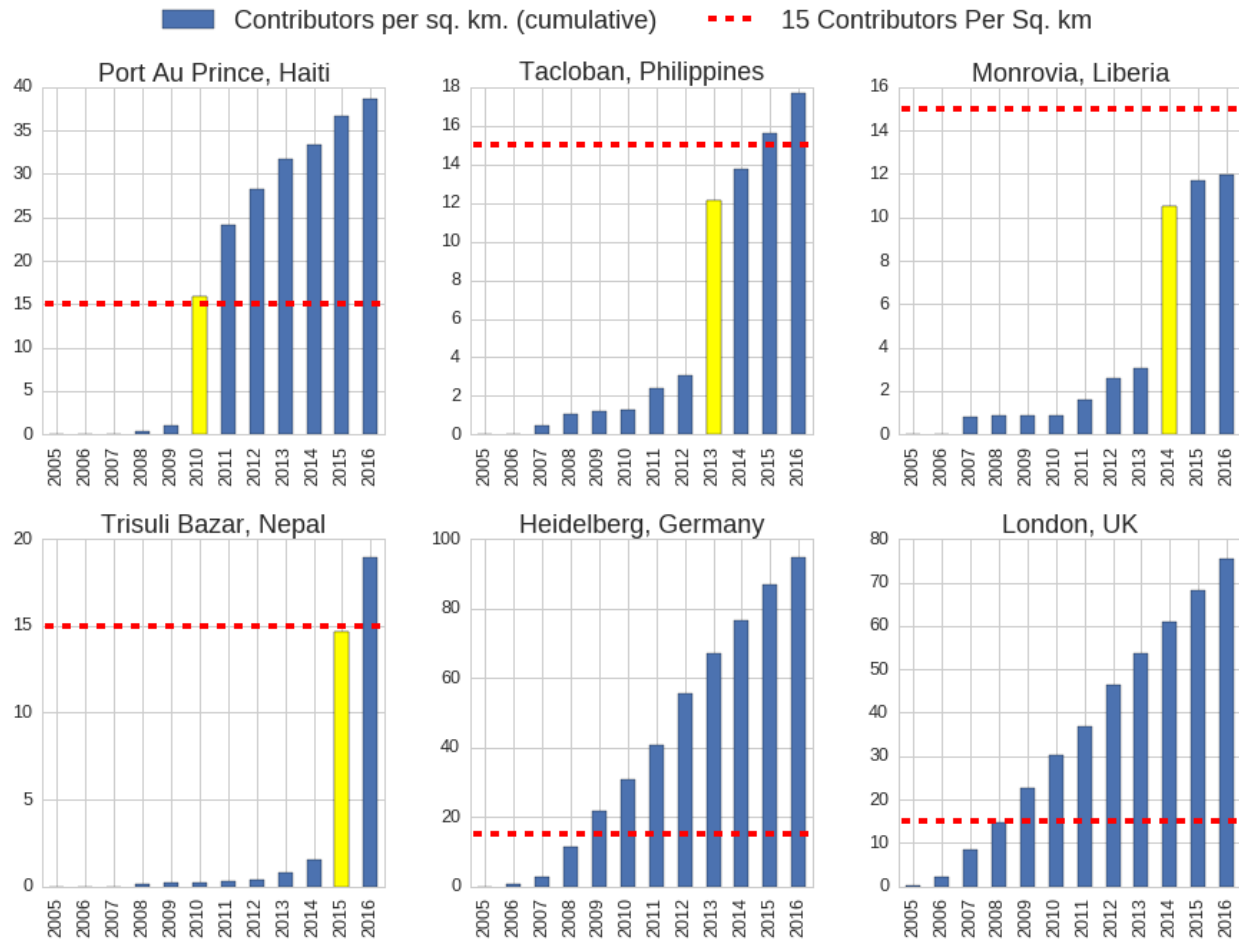


Figure 7.4: Density of unique contributors by tile over time (cumulative - in users/km²). Event year is highlighted in yellow.

suggests that a sustained, growing community of contributors is a positive quality indicator for the map overall.

Port Au Prince continues to maintain an active community in the years following the earthquake, significantly larger than the contributor activities of the other tiles; this is likely the result of the work of a local mapping community group, COSMHA (Comunite OpenStreetMap de Haiti), which formalized and incorporated as part of the response to the earthquake [125]. This sustained community of contributors has positive quality implications for the resulting map.

In contrast, an indicator of potential lower quality as a long-term result of these rapid, single mobilizations of contributors is staleness of the data. Figure 7.6 shows that six years after the

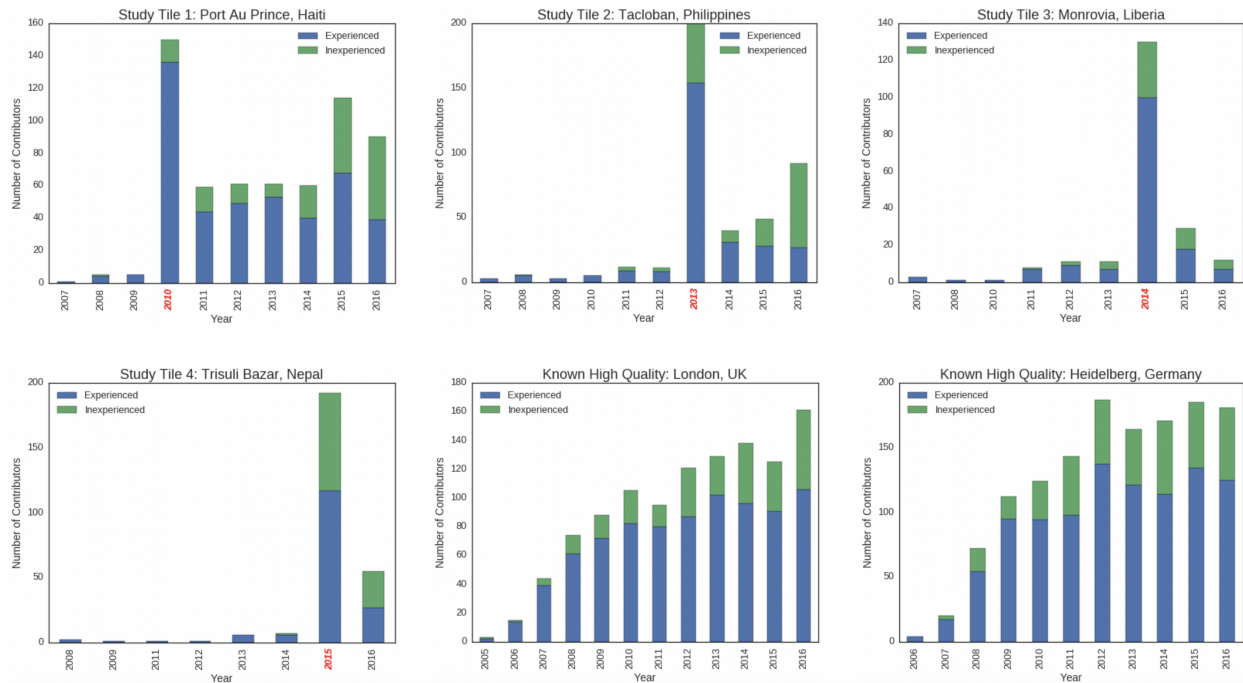


Figure 7.5: Users active each year on the study tiles and two known high-quality tiles for comparison. The years of disaster events for our four study tiles are labeled in red. Inexperienced and Experienced users are denoted by color. Metric 2 explores these differences.

event, the Port Au Prince tile has many features still tagged as `building=collapsed` which have not been edited since the earthquake. While these buildings may have not been rebuilt and are indeed represented accurately in the database, we cannot know for sure without more recent timestamps in these edits.

7.4.1.1 Implications for assessing information quality

This metric shows that areas that have experienced the rapid mobilization of contributors during disaster mapping events may superficially satisfy quality measures based on density of contributors with one-time contribution activity. Quality evaluations need to take into account the previous editing context and consider the amount of sustained editing activity, which requires new contributor-density measurements over time. In this vein, [48, 10] warn that OSM quality evaluations should be localized and performed with “fitness for purpose.” In the cases under study

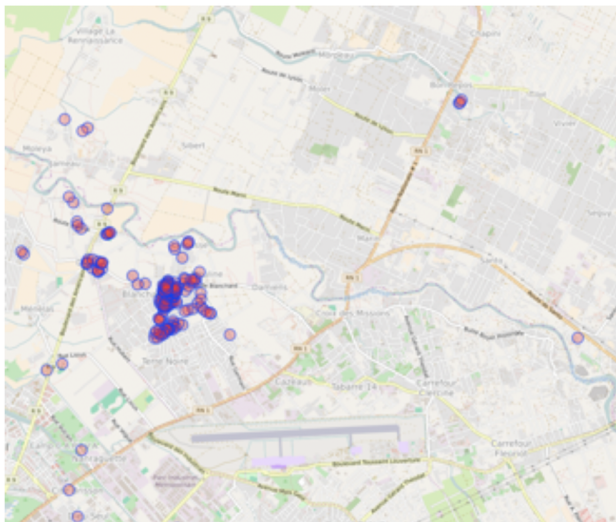


Figure 7.6: Features tagged as building=collapsed in Port Au Prince.

here, the purpose was to create roads and buildings data where there previously was none, and for immediate use. This is a different type of mapping activity than a local community performing sustained, detail-oriented mapping. Quality evaluations of these data need to then be aware of these generative differences in the map so as to evaluate the data within context.

Furthermore, the timing of these contributions raises the question of staleness as well. Our first metric expands on previous work by considering the age of the contribution [10]. Overall, this metric is simple, yet powerful, because the results seem intuitive and can locate areas of the map where high numbers of contributors (relative to others) have been active, and moreover, how long they were active.

7.4.2 Metric 2: Contributor Experience

Our second metric expands quality investigation to the amount of editing experience a contributor has with the objects they are editing. Note: our use of the term “experience” refers to a user’s familiarity and expertise with the OSM platform. The relationship of contribution experience to map quality has been explored by a variety of methods, but most commonly it is defined by the *number of edits that a mapper has made* [84]. We explore a new notion of experience in terms

of *days active on the platform*. Barron et al. remind us that while it seems plausible that editors with more contributions create higher quality data, this has not been formally evaluated [10]. For the purposes of this metric, we take a slightly modified approach to the notion of experience by classifying users with seven or more days of activity as “experienced,” and users with less than seven days as “inexperienced.” Because over half of all contributors have only made one edit while other contributors have made millions, the distribution of editing days per user is highly unequal and non-uniform. We empirically selected seven days because it retains an approximate log-normal distribution of edits per user, consistent with other online communities. This threshold retains 97.7% of the total edits but only 13% of the users for global OSM editing. We find this definition of experience more illuminating than previous definitions because it takes into account sustained interest and activity in OSM. A contributor active for only a weekend mapping event may create a lot of data, but has less overall experience with the platform and community norms than a contributor who has been active for more days. For this research, then, we take the equivalent of a week-long experience with the platform to be a useful minimum for understanding a range of basics about the platform and the OSM community, based on our experience with training others on OSM. However, this is a flexible variable that can be chosen at different thresholds for other purposes; we chose seven days for the models here. How definitive a line between new and experienced is drawn at this threshold is an area of active research.

Referring back to Figure 7.5, we see a difference between experienced and inexperienced contributors per year. For each study tile, activity spikes consistently have more experienced contributors than inexperienced. This suggests that more experienced contributors participate in disaster mapping activations than inexperienced. This has important quality implications for the data contributed during these events: specifically that these data are likely of good quality because the contributors have previous editing experience. However, the ratio of inexperienced contributors increases with each event from nearly 10 experienced users for every inexperienced user active in 2010 in Port Au Prince, to 1.6 experienced users for every inexperienced user active in Nepal in 2015. This suggests that more new mappers are becoming involved in the disaster mapping

community. While this is encouraging for the overall growth of the larger OSM community [29], it comes with the potential that recent and future events may include more and more data from first time contributors not yet aware of specific editing or community norms. Observations of the OSM mailing list during the Nepal earthquake confirm that experienced mappers were frustrated that new mappers were not following community norms and creating square buildings.⁴ To combat this, OSM editing tutorials are constantly being developed, updated, and customized for different disaster events, such as learnosm.org.⁵

We next look at what types of objects contributors have edited to explore a richer notion of experience with the OSM project. This measure assumes that, with time, a contributor's proficiency in editing specific object types improves. We look specifically at contributor preferences for mapping buildings and roads. Figure 7.7 shows editing habits of all OSM editors by object type. The number of buildings and road kilometers edited is calculated for all contributors and then plotted against one another. The color represents the number of contributors having edited $\langle x \rangle$ kilometers of road and $\langle y \rangle$ number of buildings. The legend on the right matches color to number of users.

The majority of the activity lies along the x- and y-axes near the origin, indicating that most users edit 1) very little, and 2) only one type of object or the other, not both. The lighter trend down the diagonal indicates that, as contributors edit more (and therefore become more experienced), their preferences for one object over the other may fade and they map both types of objects, though the majority of contributors do not exhibit this behavior. This distribution is consistent with power contributors in peer production systems like Wikipedia [64]. This prompts the question for the quality of our study tiles: are the ratios of buildings and roads edited by power contributors versus others higher or lower than other regions of known high-quality?

Figure 7.8 shows the differences in object editing experience among contributors and their respective number of edits, an indicator of their experience with this object type. For each study tile, we also show the distribution for London and Heidelberg for comparison (the faint red and

⁴ May 2015 HOT mailing list archive lists.openstreetmap.org/pipermail/hot/2015-May.txt.gz

⁵ learnosm.org is an open source project maintained by the HOT and OpenStreetMap Communities.

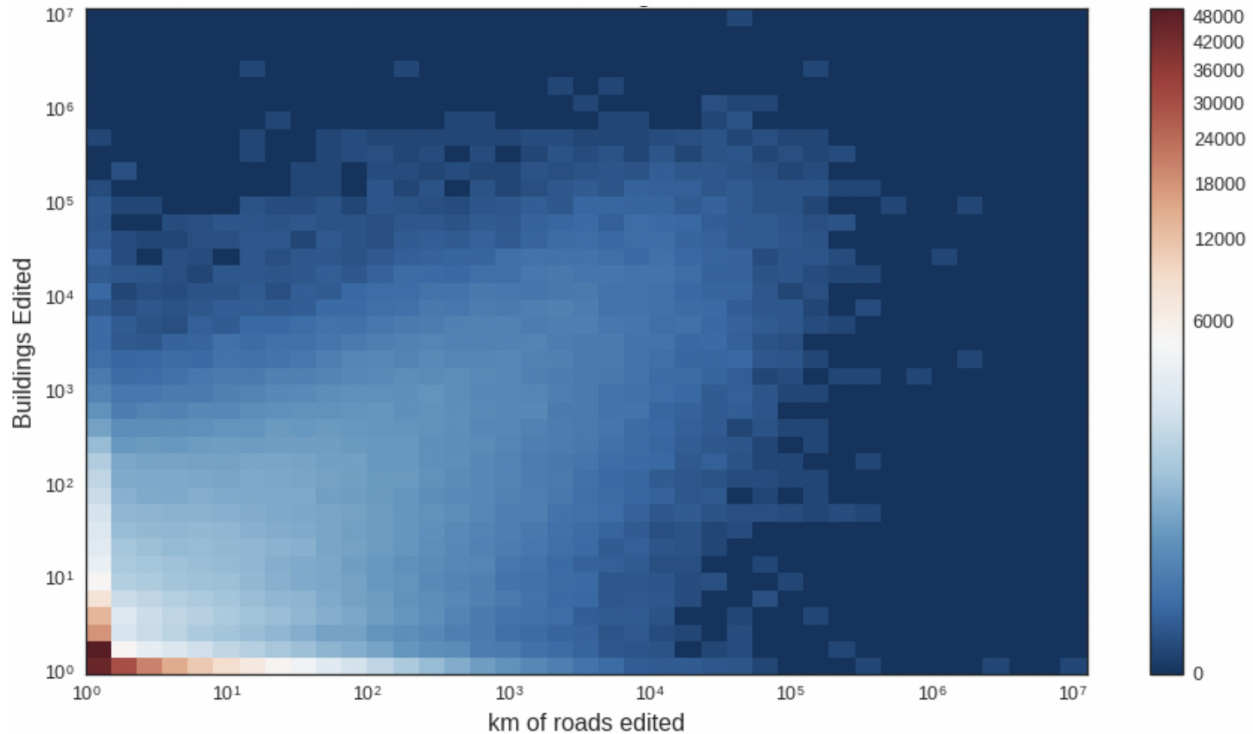


Figure 7.7: Editing Preferences among OSM contributors

green dotted lines). The similarities between the distributions for London and Heidelberg suggest that this shape of distribution may yield good quality. On both tiles, we see that contributors with experience editing over 1,000 buildings map over half of all the buildings for each region. There are both positive and negative quality implications here. Fewer more-experienced users doing the bulk of the editing suggests specific expertise, but limits the amount of crowd validation that may occur (referring back to our first metric).

In both Port Au Prince and Trisuli Bazar (Tiles 1 and 4), the distribution for buildings differs significantly from the known high-quality tiles. In these cases, less-experienced contributors edit a higher percentage of the total buildings. In Tacloban, however, this trend is the opposite, with more-experienced building mappers performing the bulk of the building edits.

Study tile 3, Monrovia, has the most similar distributions to the known high-quality tiles. This is fitting because the particular disaster mapping event consisted of a sustained mapping activity by

an engaged mapping community over a longer period of time. This mirrors the engagement of an active local mapping community, as seen in Heidelberg and London. Across all of the study tiles, there is no notable difference in the distribution of road mapping experience and the amount of roads mapped. Further analysis is required to identify the differences here.

7.4.2.1 Implications for assessing information quality

If most of the buildings or roads in an area were created by contributors without any prior experience creating those kinds of objects, then one may be suspicious of the quality of that section of the map compared to other areas where the majority of an object type is edited by contributors with prior experience working with that object type. On the other hand, if just a few power contributors have edited most of the objects, fewer eyes have seen this part of the map, lowering the potential for more validation opportunities.

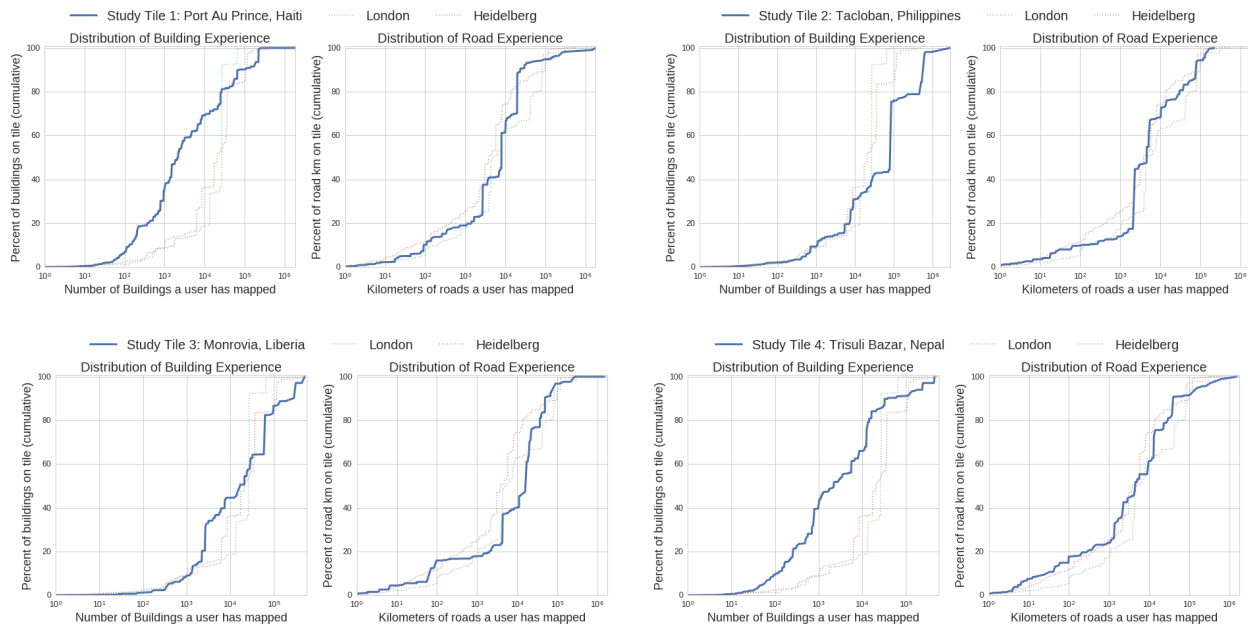


Figure 7.8: Percent of buildings and roads edited on each tile versus the number of buildings or kilometers of roads a user has mapped (experience). Thick blue lines represent object-level experience (X-axis) per cumulative amount of total edits to that object on the study tiles (Y-axis). The faint lines represent the same values for our High quality comparison tiles. Differences between these distributions highlight the differences in the amount of editing experience among contributors and their contributions.

Ultimately, the differences in these distributions cannot definitively say that one tile is of higher quality than another. However, the similarities in the distributions for our two high-quality tiles may suggest a target distribution of experience versus amount of objects mapped that yields a good quality map. Departure from this distribution would then have implications for the quality of the final map, though comparing to only two high-quality tiles is not sufficiently representative to make this claim definitively. Future research should expand this study of high-quality regions to achieve a statistically significant target distribution from a larger sample of known high-quality tiles. For now, however, there is no denying that the distributions of experience with mapping buildings to the amount of buildings mapped during a disaster is distinctly different in regions that have been the subject of disaster mapping activities with a rapid convergence of contributors, for better or worse.

7.4.3 Metric 3: Tile Maturity (Stages of Growth)

The current version of the OSM database is the aggregate product of hundreds of millions of edits from hundreds of thousands of users. Our third metric, what we call a tile maturity measure, breaks down the types of edits that occur in an area over time to identify distinct stages of growth. By looking at both the object type and the timestamp, we can identify several distinct stages of editing behavior that the map progresses through. These stages include the creation of new roads, the addition of new buildings, and, finally, a maintenance phase, where less new data is added and the bulk of contributions are edits to existing objects. In general, we know that the map grows from the road network outward [22]. The maintenance period has been called “map gardening” [72], in which continued editing of existing map objects, versus the creation of new ones, becomes the characteristic pattern of editing.

For comparison at a macro scale, we computed these stages of growth for the United States in OSM: While the number of edits continues to grow, the map does not fill in proportionally by object. In the US, 40% of the total road editing activity done to date was completed by 2009 (largely the product of a massive import of road data conducted in 2007/2008). However, it was not until five

years later that buildings caught up and 40% of the total building editing activity was complete. In the last two years, only 10% of the total road editing activity has occurred, but more than 50% of the edits to buildings have taken place. There is a clear trend of roads being added first, and while these roads continue to be maintained, contributors in the US are currently in a building phase. We, among others, find this pattern to hold in general for OSM globally [22].

Figure 7.9 shows the breakdown of new roads and buildings in comparison to editing of existing objects for each of our study tiles through the years. Across every tile, we see agreement that the first stage is the creation of roads. As new road activity subsides, there is a rise in the amount of new buildings. Port Au Prince (study tile 1) appears to currently be in a building phase, where the majority of edits in the past couple years have been the creation of new buildings. However, the years after the earthquake show a majority of maintenance activity, likely editing and maintaining data produced during the event. This creates a false sense of completeness where one may expect the building phase to be over. As evidenced by the new building activity occurring in the last two years, however, the region is not actually in a maintenance phase, but instead back in a building phase.

Similarly, Study Tiles 3 and 4 (Monrovia and Trisuli Bazar) both appear to be in a maintenance phase. With their respective disaster mapping activities occurring more recently, it is unknown whether this current maintenance phase is the product of editing the features created during the event (similar to Port Au Prince), or if the region indeed has reached some level of building completeness and has naturally entered a maintenance phase. In both cases, the types of edits occurring during the event nicely match the type of tasks outlined by HOT, which was to add buildings to the map. And in the case of Trisuli Bazar, also perform “detailed mapping” of the area. These maintenance phases likely represent the annotation of descriptive tags to features created by remote mappers during the event. It is still unclear, however, whether the tiles will enter another building phase in the future, as we have seen with Port Au Prince.

Study Tile 2, Tacloban, on the other hand, saw many new buildings, but mostly editing of existing features during the year of the event, prompting further questions about the exact

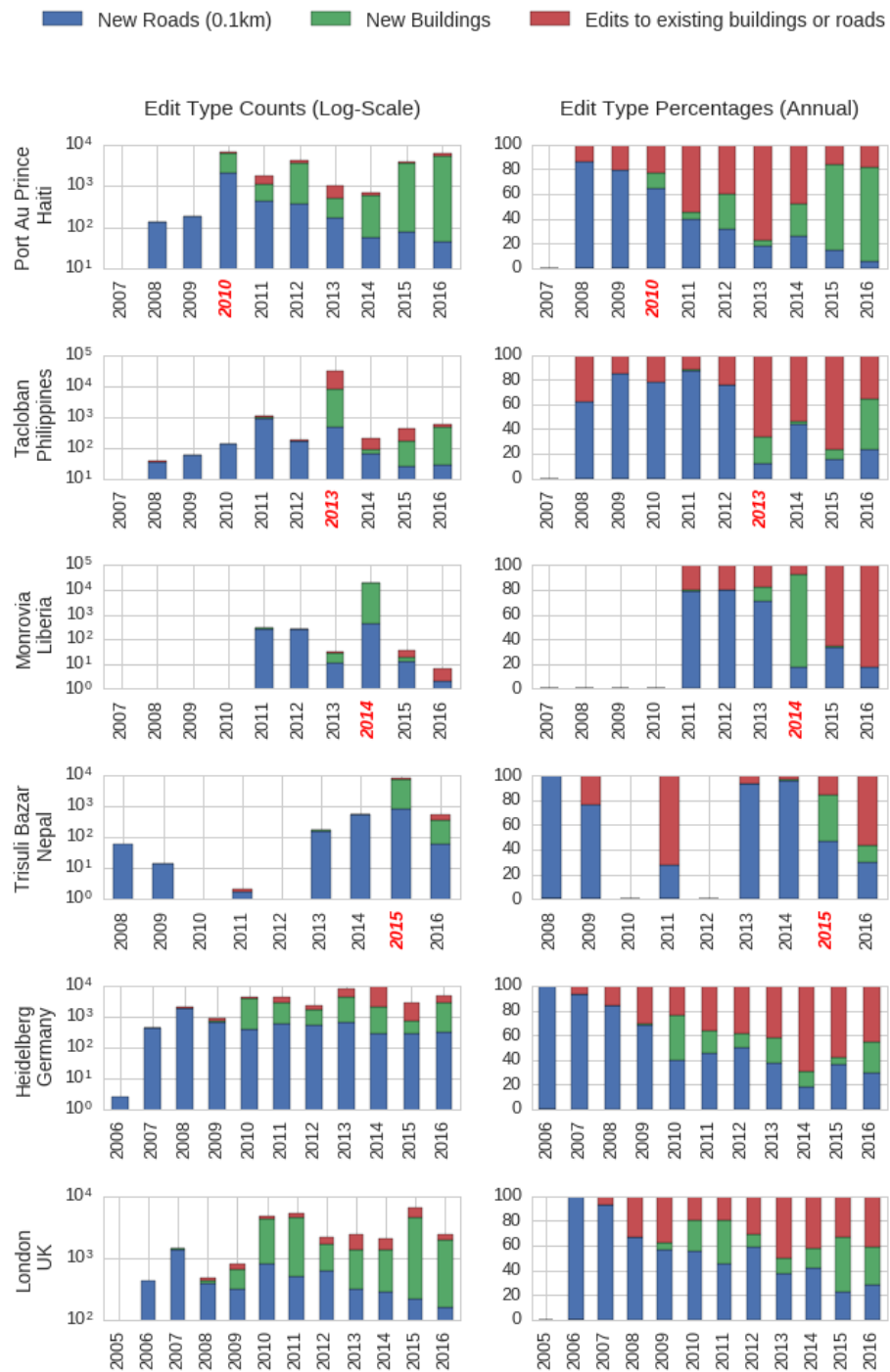


Figure 7.9: Stages of Growth as shown by edits of each type each year. The Y-Axis representing edit-counts are log-scaled to allow non-disaster event years to show. The percentages are shown on the right to better express the relative amount of activity. For study tiles, event year is denoted with red, italic label

disaster-mapping activity. Furthermore, the tile parallels Study Tile 1 by appearing to enter a maintenance phase after the event (though with a surprising number of new roads), and is currently going through another building phase.

This potentially premature maintenance phase is common across all these regions, making the tiles appear more complete than they are, relative to other parts of the map that appear to progress through the phases of growth in an orderly fashion. However, these regions are still significantly better mapped now than they were, having been the target of disaster mapping. For comparison, the stages of growth are shown for Heidelberg and London. Heidelberg clearly follows the standard trend with maintenance behavior increasing in recent years as both building and road creation slows down. London has seen increased building activity in recent years, but still follows the general trend of maintenance behavior being more common in recent years than new roads.

7.4.3.1 Implications for assessing information quality

Given the specific order in which the map grows and matures, knowing which phase of growth a given part of the map is in gives an indication of its level of completeness (a standard quality measure). Determining these phases strictly on percentages of edit types requires neither external reference data nor specific object attributes, merely the geometry type and version number. This makes analysis of any region possible since these are basic attributes present in every map object.

Disaster mapping activity, however, interrupts this natural sequence, making the map appear to be in a different stage than it likely is. Our metric is good for showing relative tile maturity between different regions, but the context of a region is important to consider. Comparing the apparent stage of growth with the specific tasks outlined for a disaster mapping activity can provide this context. Ultimately, these stages of tile maturity are relatively easy to compute for any region of the map and offer a measure of object-level completeness, a metric that is typically only possible with extrinsic quality analysis relying on an external reference dataset.

7.5 Discussion

Information quality is an important concern for online peer production systems like Wikipedia and OpenStreetMap, especially in safety-critical situations. Despite the similarities in the systems' affordances, the well-validated contributor-based intrinsic metrics for assessing information quality in Wikipedia have not been translated into OSM. While other intrinsic quality assessment techniques relying mainly on the spatial attributes of the target dataset have been explored for OSM within the field of GIScience, we presented three metrics using VGI meta-data about who made spatial contributions and when to develop alternative perspectives for intrinsic information quality than what is found in related work [10]. We find this shift in emphasis from the spatial attributes of VGI data to contributor information in turn establishes a bridge from the GISciences into the fields of social computing and human computer interaction.

These new metrics are especially important in understanding the quality of map data produced from a large mobilization of contributors during disaster mapping. Because these data are created for use for disaster preparedness, response and recovery, having ways to assess map quality becomes a safety-critical task. The intrinsic quality metrics offered here rely on metadata about contributor activity that, as opposed to other approaches, are likely to be available in disaster mapping scenarios. They have been tested against four distinct areas of the map that were the sites of large mobilizations of volunteer mappers. These metrics exposed differences in the contributor activity between each of these areas and areas of the map known to be of very high-quality and not impacted by disaster events. The variation in the results suggests these metrics capture distinct generative processes that have implications for assessing the quality of the final product.

Metric 1 showed that while the number of contributors active in a region may indicate the size of the OSM community with direct correlation to the quality, events that draw many remote contributors to the area artificially inflate this density with one-time activity. While contributor density has been shown to be a useful intrinsic measure of quality [48], we show that it is important to also include the temporalities of these contributions in quality assessment. Metric 2 reveals

that mapping done by power contributors looks different in areas with sustained and active OSM communities than in areas experiencing the rapid convergence of digital volunteers. In terms of buildings, power contributors had less influence over the total edits in Port Au Prince and Trisuli Bazar than they have in regions with more continually active contributors. It should be noted that both these events were earthquakes—that is, sudden onset events—prompting a rapid convergence of contributors. Metric 3 reveals that disaster mapping activity may disrupt the natural evolution of the map away from the distinct phases of editing, creation, and maintenance.

Given the fundamental difficulty of extrinsic quality assessments of spatial information, intrinsic quality metrics used with other features help identify nuances in the different processes for generating peer-produced spatial information. Ultimately, each of the regions we investigated become better mapped than they were before as a result of the volunteer contributions, but as discussed above, this process played out in unique ways across each site. By combining these metrics, users of the map data can develop a richer understanding of exactly how the map came to be, such as understanding how stale the data may be due to a one-time very active community or learning about the specific expertise breakdown of the contributors. As we have shown, and as with traditional metrics of data quality, none of these metrics convey uncontested assessments of data quality. Rather, they are intended to be used in combination with other measures to provide historical context of the editing in the region to help better understand the evolution of the map. Additionally, these analyses must be performed with a consideration of how the data will be used [10]. This is further complicated when considering time- and safety-critical applications of the data such as emergency response. Ludwig et al. suggest that in emergency situations, notions of general information quality assessment are less important than the specific fit and purpose (emergency use) of the information itself [68]. Referring back to Figure 7.6, the “staleness” of the data today and therefore its potential to lower overall information quality for the area seems a worthy tradeoff for the value that data held during the specific emergency task for which it was contributed in 2010.

7.5.1 Implications for practice and design in disaster response and beyond

Authoritative data sources that can support extrinsic approaches to assessing VGI quality are often difficult to obtain outside of advanced industrialized countries. In the absence of objective ground truth, examining how user behavior and temporal context interact to generate data can identify gaps. Because these metrics only rely on the OSM database and not external sources, they can be used immediately to help disaster mapping efforts better understand the contribution patterns. Who is editing the buildings? How much experience do they have? These represent real concerns; discussion occurring on the Humanitarian OpenStreetMap Team’s mailing list during the Nepal earthquake response highlighted frustrations of experienced mappers over the non-square buildings being mapped by new users that cost valuable volunteer time. Our metrics could help organizers of disaster mapping activities more quickly inform their volunteers as to what is happening and/or prompt intervention where it may be most helpful.

Intrinsic methods also allow for identification of stale data in the map, requiring only the date of the most recent edit. This type of analysis could inform contributors where they should focus validation efforts. As we have shown, even in places where the map appears relatively complete, there may be stale artifacts that degrade map quality. The scale and complexity of these data coupled with the fundamental difficulty of establishing extrinsic quality for spatial information also suggests that developing and validating intrinsic quality metrics will also be essential for filtering out vandalism and attacks. Consider a map tile rapidly accumulating edits from novice or non-local contributors: Is this an instance of coordinated vandalism or disaster response? Automatic, algorithmic approaches to vandalism detection have yet to be perfected and similar approaches on Wikipedia have distorted behavior in the community and discouraged new contributors [38, 51].

7.5.2 Limitations and Future Work

Our methods are currently limited to the resolution of the specific OSM vector tiles as they are generated, both in temporality (annual snapshots only count the latest edit to an object per

year) and in size (zoom level 12 may be too big to identify more spatially nuanced editing activities). Computationally, however, this approach utilizes advanced methods for parallel processing of the massive OpenStreetMap database, making analysis faster and more scalable than previous methods. Because these techniques use a contributor's editing history, having entire histories instead of annual snapshots will be more accurate in the future, though this is currently an unsolved problem at scale for this domain. Furthermore, there are currently no scalable methods of tracking over-written geometry changes. For example, if an editor squares up all the buildings in a region or slightly moves the path of a road to better match updated satellite imagery without changing other attributes of the building or road—a common type of edit—the database remains unaware of the change at the object level. That is, if only the spatial geometry of a complex feature like a road or building are changed, the change does not propagate to the object itself. Due to the data structure, identifying and tracking these activities is non-trivial and no solution exists yet for performing this at scale. These types of edits represent validation and correction and their existence has major implications for the quality of the map in that region. Incorporating such features in future research is paramount to better intrinsic quality assessments.

As indicated by a growing number of contributors with each subsequent event, data contributed to OSM in disaster mapping situations will become more prevalent. In general, this will improve the overall completeness of the map. These mapping activities help attract new members to the OSM community, create large amounts of open geographic data, and most importantly, help to satisfy the informational needs of emergency responders. As data contributed in these events become more common in the OSM database, future work could explore more longitudinal questions of community engagement and maintenance of the affected regions. For example, Dittus et al. present a study of 26 disaster mapping campaigns that sheds light on contributor engagement (and retention) across different types of disaster mapping events; of specific relevance to this work, they propose quality metrics based on data persistence and quantify user expertise and engagement using methods proposed in [38] around the concept of an editing session, not simply number of edits or editing days [30]. Knowing that the percentage of newcomers is increasing with each disaster mapping event,

more and more of the map will be the product of novice editing. Future map data quality research could further examine the correlation between new mappers and data quality across more events. Dittus et al. find that the success of these events is not dependent on the large number of novice mappers because novice mappers work slower and produce less data on average [30]. At the same time, a novice mapper that joins for a disaster event and remains part of the community inevitably becomes a more experienced contributor. While the actual amount of data contributed per mapper will vary, future work could investigate if the level of experience (and volume of contributions) per returning contributor is increasing at a rate greater than novice contributors are producing data. This would lead to a population of disaster mappers with community mapping characteristics of a non-disaster contexts like those of London or Heidelberg discussed here.

Thus far, this work is rooted in exploring metadata of VGI contributions to expand more traditional VGI quality assessment methods. Another direction is to build from quality assessment techniques in other forms of user-generated content independent of geospatial data such as social media posts. Reuter et al. discuss the implementation and usefulness of a social media API that incorporates post-specific metadata to perform quality-assessment of the data based on a variety of data use-cases [114]. Future work along this vein could incorporate more social media research techniques: network analysis, content analysis, sentiment analysis, etc. that are independent of the geospatial information. Moreover, new technological solutions to improve coordination of these disaster-related crowd-sourced and peer-production activities were not discussed in depth here, current work in this domain such as Ludwig et al. present novel methods to ensure coordination among volunteer responders to disaster events, even in the presence of network outages [67]. Such systems are invaluable to communities of disaster volunteers with many quality implications for the data produced.

Future work may also provide valuable insight to the fields of Crisis Informatics and VGI by exploring potential theoretical and methodological consequences of these types of comparisons to community behaviors (and the metrics) to peer-production in non-disaster contexts.

7.6 Conclusion

The openness and availability of VGI presents new opportunities to use spatial data for applications including in essential humanitarian and safety-critical situations where rapid availability of high-quality data is paramount. We draw from the peer-produced OpenStreetMap database to propose and evaluate three intrinsic quality metrics for spatial data based on the provenance of these data that build upon user behavior and temporal context. These metrics are not introduced in opposition to or replacement of existing quality assessment methods that respond to traditional concepts of quality such as positional accuracy, map completeness, or the other ISO 19113 standards. The intrinsic measures presented here can instead expose specific aspects of the map's history that can provide context—especially useful when assessment by comparison is not possible. Moreover, these metrics are especially suited for identifying small and sometimes hard-to-detect changes to the map in regions that are affected by rapid disaster mapping. For safety-critical situations of disaster, where humanitarian decisions are based on maps being read by outsiders converging on an area to help, a suite of intrinsic measures that strive to communicate peer-produced map quality from the inside out, perhaps in real-time, is essential. If we anticipate that peer production platforms will continue to populate our future information environments, and certainly in times and places like disaster when convergence of information is a natural and age-old socio-behavioral phenomenon, then attention to developing rapid metrics of quality for digital data generated under socially distributed conditions will ascertain how much risk is assumed when life-and-limb decisions must be made upon them.

Acknowledgements This work received made possible with funding from US NSF Grant IIS-1524806 and a research fellowship with Mapbox. We thank our reviewers for their feedback.

Chapter 8

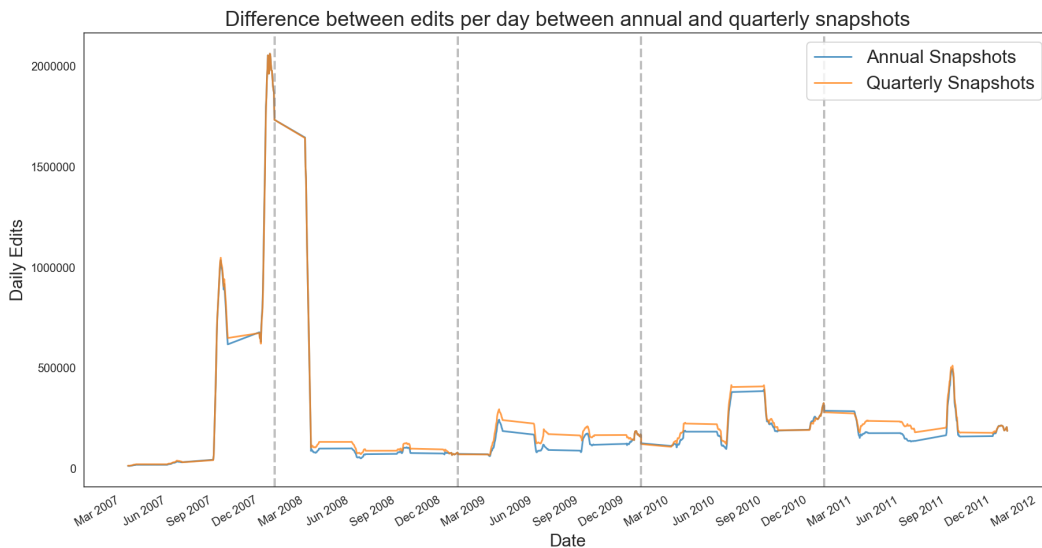
Quarterly Historical Snapshots

This chapter reviews and discusses innovations completed during my second Research Fellowship with Mapbox (2017). This involved first identifying shortcomings of the annual-resolution historic snapshots and then creating *quarterly resolution* historic snapshot OSM-QA-Tiles. These new tiles became the analytical backbone of an improved visualization tool to compare the state of the map between two quarters. I presented this tool at State of the Map US 2017 in Boulder, CO.

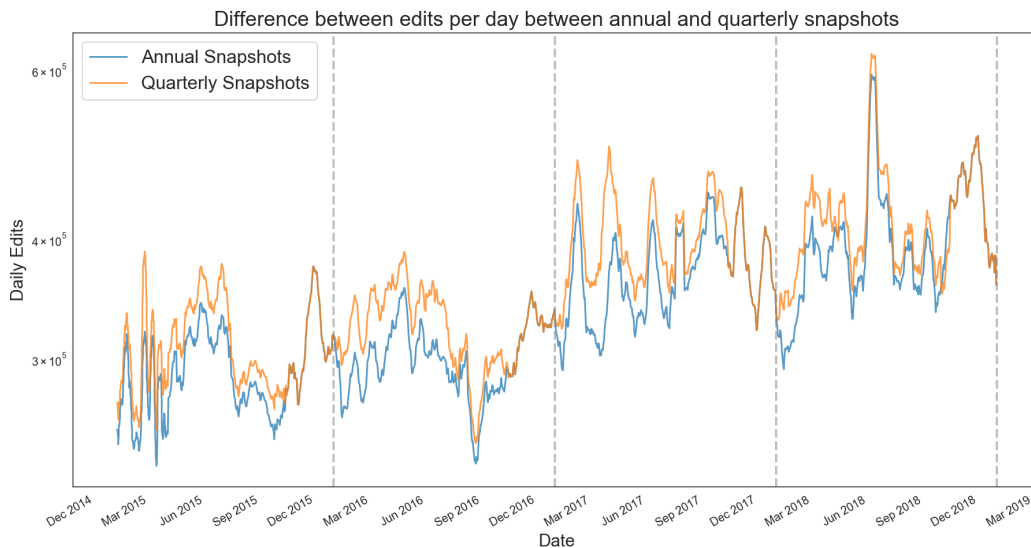
8.1 Improved Resolution with Annual Snapshots

Historical tile-based analysis with *annual snapshots* proved to be a powerful, scalable approach to global analysis of the evolution of the map using tile-boundaries as the units of analysis. However, with the growing number of contributors, I developed an increasing concern over the amount of editing activity that we were missing in recent years by using *annual* snapshots. The term “missing” here refers to the literal loss of editing metadata in an annual-snapshot when quantifying edits at the contributor level. I call these shadowed edits, as first presented in Section 5.4 and Figure 5.8. Additionally, the map is evolving so quickly that annual time-steps between snapshot comparisons cannot adequately capture the full story behind the map. To address this, we look into *quarterly resolution* for three month time-steps instead of annual.

Figure 8.1 shows the discrepancy between annual and quarterly snapshots when using them to count the number of edits per day. Higher edit counts in the earlier days of a year identified by quarterly snapshots indicate edits that get masked at the annual resolution.



(a) Daily edit count differences: 2007 - 2012 (~12M missing in total, ~2.5M / year on average)



(b) Daily edit count differences: 2015 - 2019 (~34M missing in total, ~8M / year on average)

Figure 8.1: Discrepancies between the number of daily edits as counted with annual snapshots and quarterly snapshots.

While obtaining the *guaranteed true* count of daily edits to the map with OSM-QA-Tiles like this would technically require daily snapshots, Figure 8.1 shows that the quarterly snapshots were able to identify an average of 8M edits more per year in recent years. In earlier years, with less editing activity, the difference between using annual or quarterly snapshots is about 2.5M edits per year. To get a better idea of where these shadowed edits are occurring, I developed the following formula to compare the difference between each tile over the years.

8.1.1 Identifying Shadowed Edits Per Tile

Since historical snapshots only contain one version of an object, it is impossible to know who has edited that object before. The version number of an object, however, can at least alert us to the presence of previous edits.¹ Learning more about these edits requires comparing versions of the object across snapshots. However, performing this comparison at the individual object level can really only tell us how many edits are missing, there will not be any metadata for the missing edits. Additionally, it would be computationally expensive. Instead, I developed the following formula to count the number of edits per tile that were shadowed between two consecutive years:

$$\begin{array}{r}
 \text{SUM OF ALL VERSIONS OF OBJECTS ON TILE} \\
 - \text{TILE VERSION SUM FROM YEAR PREVIOUS Year} \\
 - \text{TOTAL NUMBER OF EDITS THIS YEAR} \\
 \hline
 \text{SHADOWED EDITS}
 \end{array}$$

The sum of the version number for every visible object in OSM is equal to the total number of edits to the map.² At the tile level, then, the sum of the version numbers of all the objects on a tile at any given time represents the total number of edits ever performed on that tile, up to that point in time. While we still do not know who is responsible for the previous edits or when they occurred, those objects with a version number >1 can tell us *how many* times they have been edited.

The only situation in which no edits get shadowed would be for objects to only be edited once

¹ The version number of a converted way-element in an OSM-QA-Tile does not account for minor versions, so the total number of shadowed edits is likely higher.

² This is also going to be lower than the true number of edits because there are no representations of deleted objects in OSM-QA-Tiles

during the year, this could either be an edit to an existing object or the creation of a new object. In this case, the difference in the sums of all of the versions between two consecutive years would be equal to the number of edits that happened that year, and most importantly, the metadata for each of these objects would then reflect the (only) edit that happened to that object during that year. This is the logic behind this formula: Identifying the difference in these sums after accounting for the edits that we know happened in a given year. We can learn more by then separating these edits by object type, tracking the object-specific version sums per tile each year. Figure 8.2 shows how the number of shadowed edits has increased globally over the years. Figure 8.3 shows an example of calculating the total number of shadowed edits in 2015. The popup represents a tile in Kathmandu, Nepal, where by tracking the object type, we can identify that 804 of unaccounted edits were to roads and 1116 were to segments of roads.

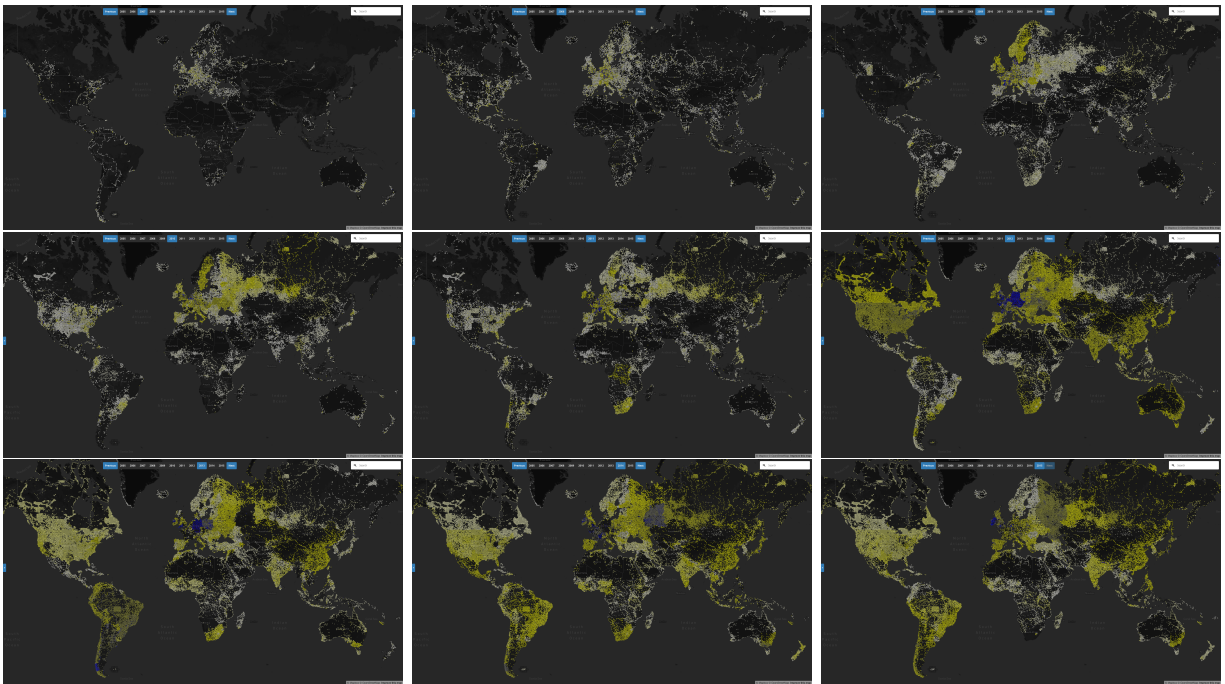


Figure 8.2: Screenshots of the interactive shadowed-edit map for years 2007 (top left) - 2015 (bottom right). Zooming in will show tile-level shadowed edit counts.

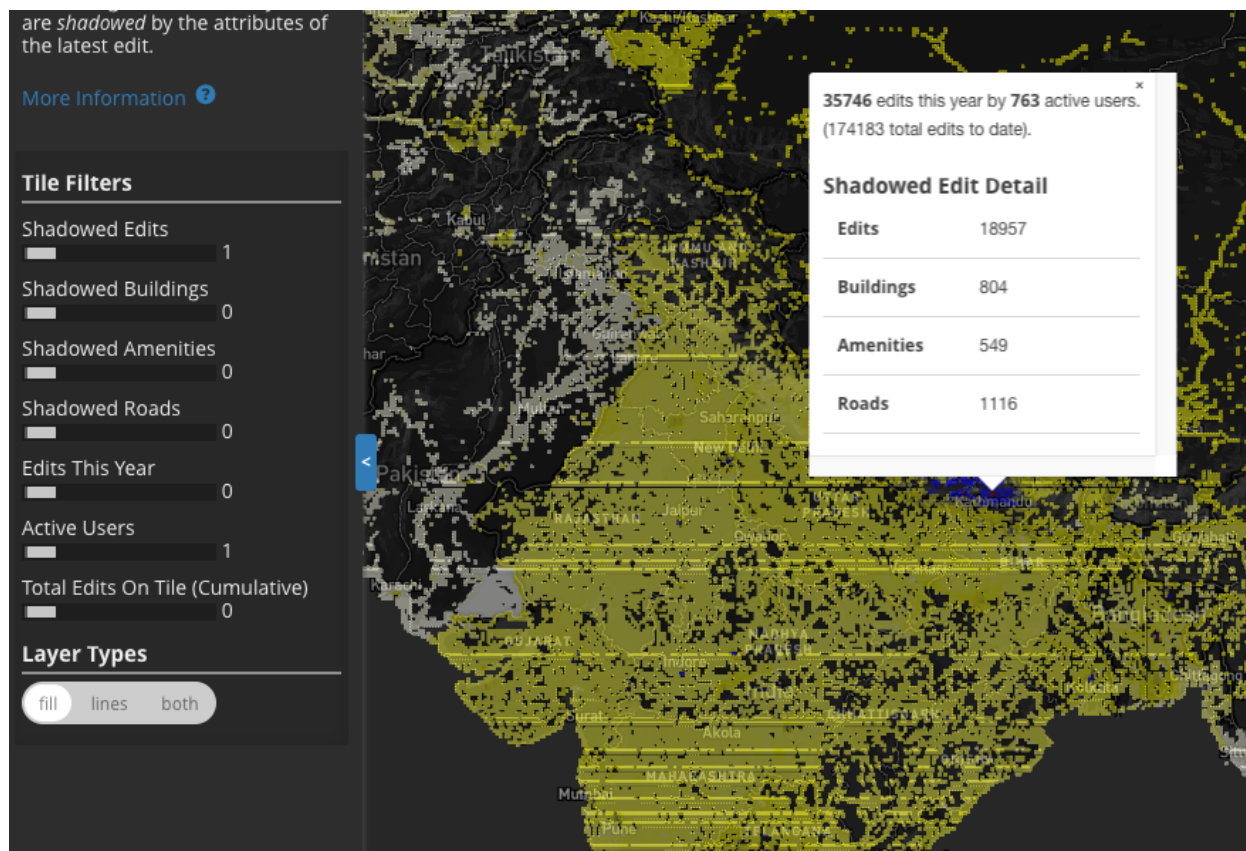


Figure 8.3: Closeup of the interactive tool-tip that shows the breakdown of what types of edits are being shadowed. The region selected here represents the city of Kathmandu, which had many shadowed edits in 2015 as mappers continued to clean up and edit the map after the earthquake in April.

8.1.2 Building Quarterly Historical Snapshots

As Section 5.5 described, to improve the tile-based analysis of historical data, I transitioned the tile-based historical analysis workflow to rely on quarterly snapshots. Quarterly-snapshots can be processed with the same workflow as Figure 6.1, just with four-times as many input files. Collaborating with Mapbox, I created the quarterly-snapshots in a consistent manner with the existing OSM-QA-Tiles so that existing processing tools could be simply pointed to these new files.³ Today, any OSM data analyst can download the historical quarterly-snapshots from 2005 through

³ An added bonus of re-generating historic OSM-QA-Tiles datasets included a new version of the GeoJSON conversion utility which cleaned up many complications such as duplicated multi-polygons. For exact configuration details, see: osmlab.github.io/osm-qa-tiles/historic.

the end of 2018 at the OSM-QA-Tile website, osmlab.github.io/historic.html. This page also includes detailed information on the limitations of these files and how they were created in an effort to better educate other analysts about the pros and cons of doing snapshot-based historical analysis of OSM data. I continued to maintain this page and produce these tilesets through the last quarter of 2018, but have since shifted my focus and preference to the full-history schemas as proposed in Section 5.6. The quarterly-snapshot tilesets, however, are still powerful analytical datasets if looking to quantify changes to the map between two points in time, and give an adequately representative account of editing activity over time, at least to 3-month resolution. The first project to implement the quarterly snapshots for analysis was the *OSM-Analysis-Dashboard*.

8.2 State of the Map US 2017: OSM Analysis Dashboard

During my 2017 Research Fellowship with Mapbox, we used the quarterly-snapshot historical OSM-QA-Tiles to create the *OSM Analysis Dashboard*. This interactive map and analysis dashboard contains snapshot statistics for North America from 2005 through 2017 along any of the dimensions shown in Figure 8.4. I presented the dashboard and the innovative analysis approach behind it at the 2017 State of the Map US conference in Boulder, CO. This section will summarize the innovations to the previous tile-based analysis workflows and share some of the key takeaways from the presentation.⁴

Distinguishing this particular analysis workflow from that shown in Figure 6.1 is additional resolution of analysis, both temporally and spatially. Though OSM-QA-Tiles are generated only at zoom level 12, they still contain *all* of the OSM data for that tile—as opposed to a vector tile optimized for rendering a map would only contain features visible at that zoom level, like major roads and city names. Instead of using the zoom level 12 (z12) tile-boundaries for analysis, we cut the tiles up further, into the 64 zoom level 15 tiles (z15) that fit within a single zoom level 12 tile. Whereas z12 tiles represent the area similar to a small city, z15 tiles cover an area closer to 1

⁴ Jennings Anderson and Ramya Ragupathy (2017). Watching The Map Grow. State of the Map US. Boulder, CO. October 21, 2017. Findings summarized here with permission from my collaborators. Video of original presentation available at 2017.stateofthemap.us/program/how-we-know-the-map-is-ready.

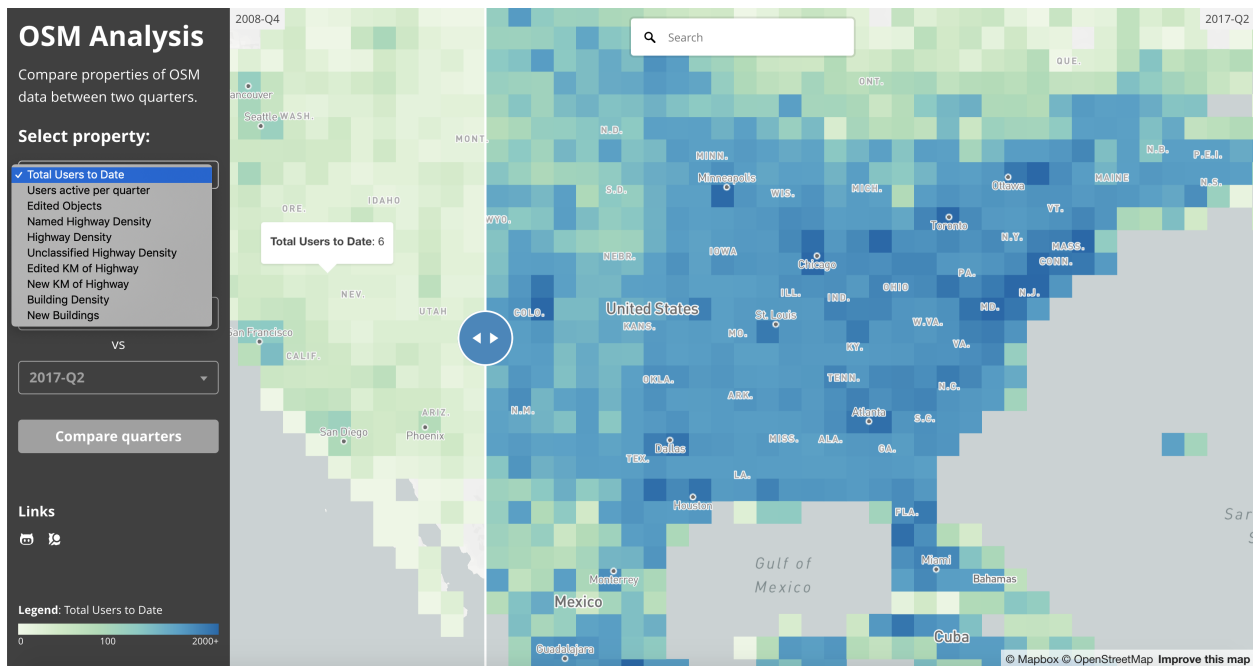


Figure 8.4: Screenshot of the OSM-Analysis-Dashboard visualization tool showing the available properties that have been precomputed and can therefore be rendered. Currently displaying the total number of users ever active in an area at the end of 2008 (left) and the middle of 2017 (right). The user can move the slider between the two for comparison.

square-kilometer. Furthermore, though more complex, this tile-reduce workflow does not depend on any post-processing of tile-summaries for aggregation or visualization as the previous workflows did. Instead, Figure 8.5 shows how the `map` and `reduce` functions can be leveraged to perform all of the analysis at once, aggregating as it goes and saving the results.

To achieve this, I modified the order in which individual tiles are processed by the *tile-reduce* job. By default, tiles are processed by column (longitudinally). Since the planet is represented as a grid of tiles, each tile has an x,y address (at zoom level 12, this grid is 4096×4096). Tiles are typically distributed to worker threads in the following order: $(0,1), (0,2), (0,3) \dots$. Instead, I segment tiles into larger blocks so that the grid is not read by column, but by larger blocks of varying size, such as 4×4 , which in z_{12} tiles, represents the area covered by a single zoom-level 10 tile: These 16 tiles are then all passed out to workers in parallel and the reduce job continues to check if all 16 of these tiles have been processed. Once all 16 tile summaries returned, these results

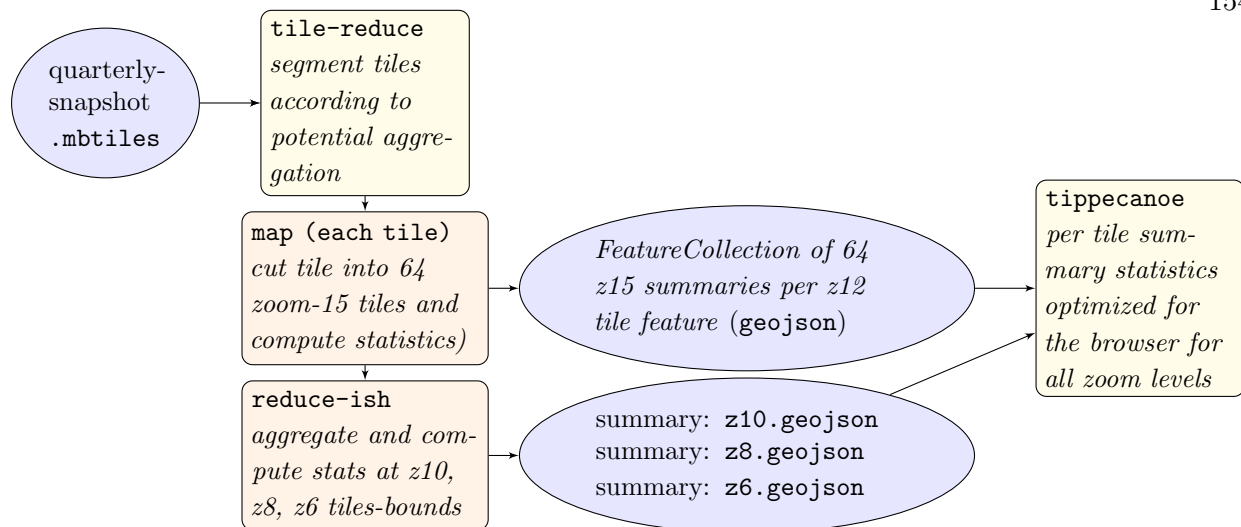


Figure 8.5: Tile-reduce workflow used to generate the datasets behind the OSM-Analysis-Dashboard visualization (mapbox.github.io/osm-analysis-dashboard). This workflow performs analysis at square-kilometer resolution and produces results aggregated at lower resolutions to be fed into interactive tools.

are again aggregated into a zoom-level 10 summary and a geographic representation of the z10 statistics are saved. In implementation, z12 tiles can be processed in any block size, allowing for arbitrary levels of aggregation, including multiple levels of aggregation as Figure 8.6 will show.

Figure 8.5 shows this workflow in which I label it `reduce-ish` because the information returned from the each tile as processed by the `map` function is not immediately reduce-able, but requires additional aggregation. This workflow is optimized, however, to hold these results in memory for the shortest amount of time because of the order in which the tiles are distributed. Once these aggregated statics are computed and saved to disk, memory is freed for the next block to be processed. In this way we can efficiently process the entire planet in parallel while continuously aggregating the results.⁵

Ultimately, this tile-based workflow ingests every quarterly-snapshot OSM-QA-Tile and computes all of the statistics seen in the dropdown in Figure 8.4, aggregating these statistics at larger and larger geographic areas. These statistics are then exported as GeoJSON Polygon

⁵ Hierarchical organization of spatial data by tile for efficient processing and serving is an active area of development: medium.com/@mojodna/tapalcatl-cloud-optimized-tile-archives-1db8d4577d92.

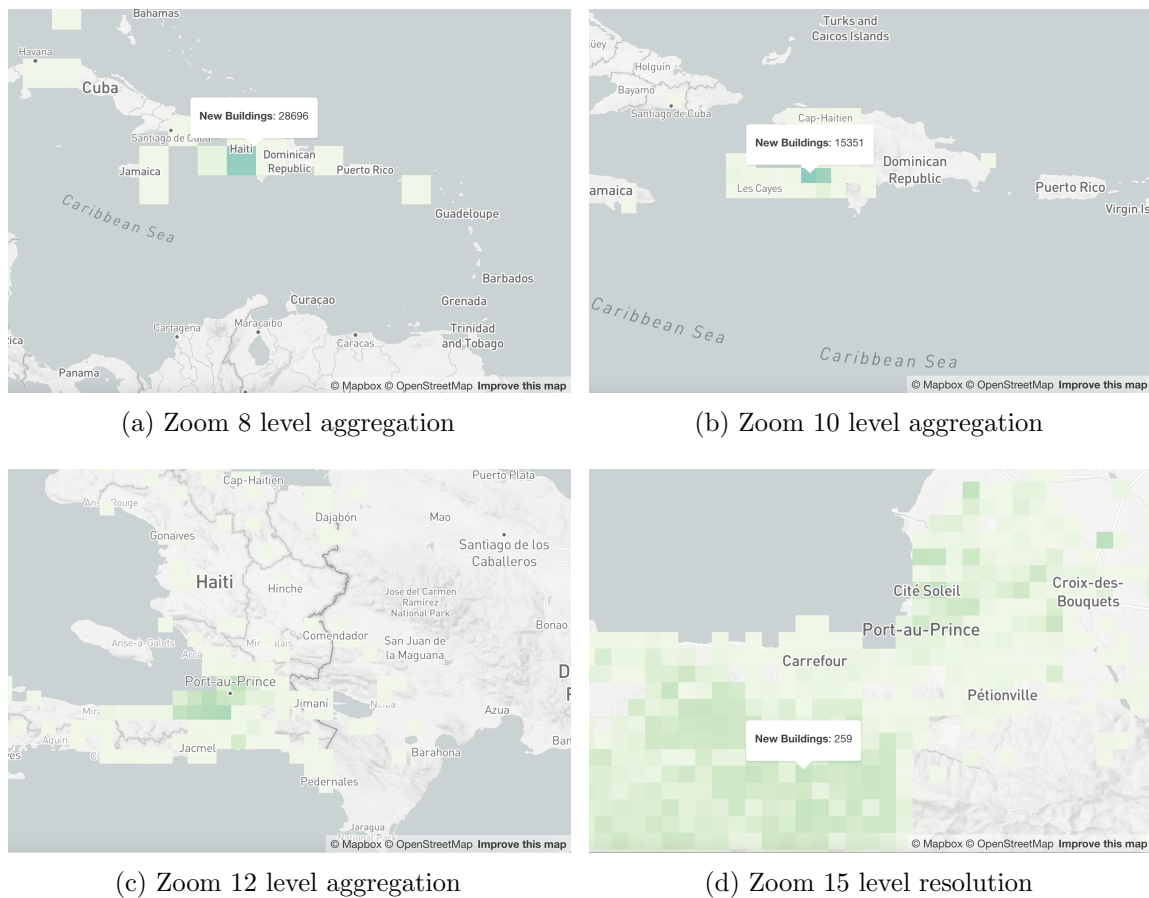


Figure 8.6: As the user zooms in and out, the OSM-Analysis-Dashboard will automatically load a different layer with statistics computed at a higher or lower resolution.

features and turned into vector tiles that power an interactive map that accurately displays these quarterly statistics at various zoom levels: The browser loads the appropriate aggregation based on which zoom-level an analyst is currently viewing. Figure 8.6 shows the four levels at which the editing statistics are aggregated. Statistics like *new buildings* as shown here are simply summations, however statistics like *density of roads or buildings* requires more careful calculation to not under- or over-weight at lower zoom levels (avoiding taking the average of averages). This adds complexity to the reduce job, but ensures higher accuracy in results.

The OSM-Analysis Dashboard lets analysts *compare the map* between any two quarters. This allows us to visualize growth over time (or at least at quarterly resolution). It also innovated (my)

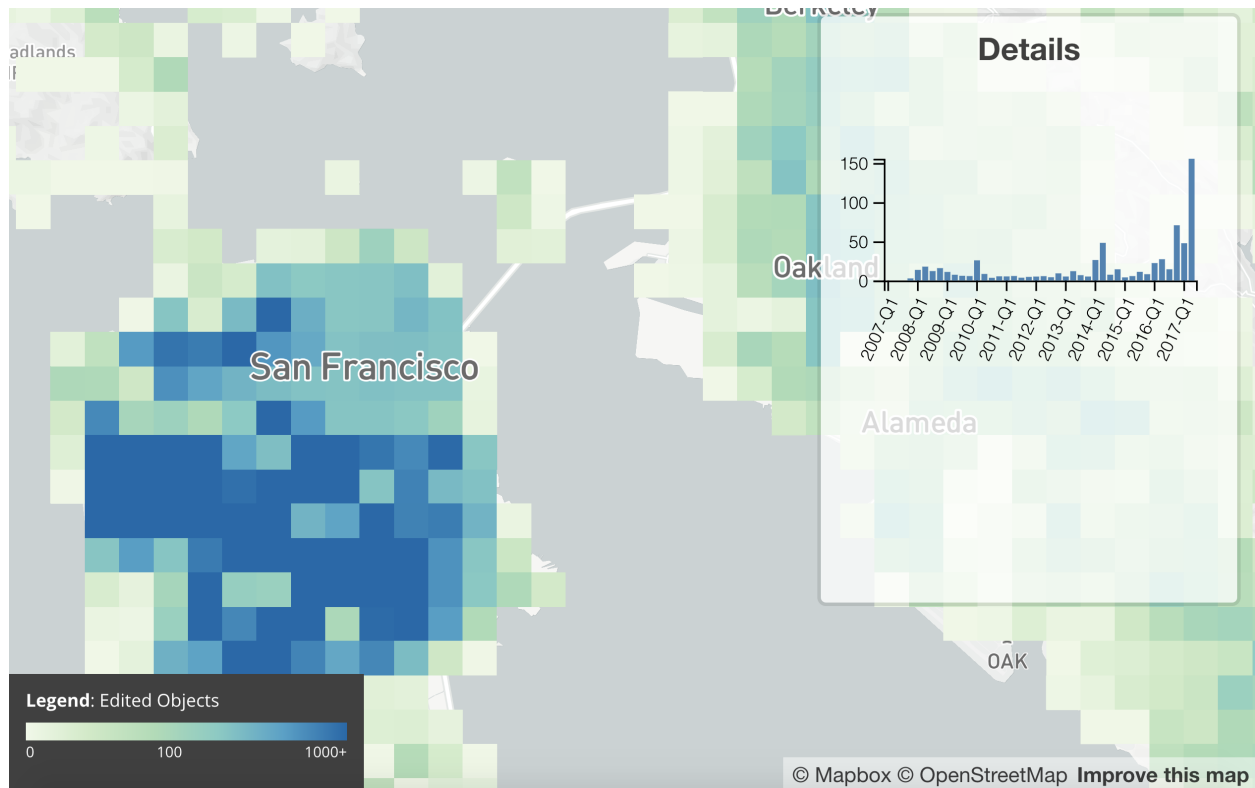


Figure 8.7: The OSM Analysis Dashboard is capable of querying all of the quarters to generate a graph (right) of how the given attribute—Number of edited objects per quarter in this case—has changed in the region over time.

tile-based analysis workflow by introducing aggregation on-the-fly as the tiles were processed.⁶ Additionally, since all of the editing statistics for each quarter have been precomputed, it is possible for the tool to query every quarter and generate a graph for any area the analyst is currently viewing to see how the given attribute has changed over time, as shown in Figure 8.7. In this case, we see the first significant activity in San Francisco in 2008. This corresponds to the TIGER import. Then 2014 sees a spike in the first two quarters. This is corroborated by an increase in the number of buildings edited per day seen during the same time period in Figure 6.5.

Ultimately, creating the quarterly-snapshots improved the resolution of analysis from the annual snapshots and improved the accuracy of contributor-centric measures by decreasing the number of shadowed edits in a given year. An enhanced analysis workflow was able to leverage

⁶ It also simply looks much better than the previous suite of analysis tools because professional UI and UX designers helped with the front-end interface.

this increased resolution to build datasets that can drive more advanced and accurate interactive visualizations to explore and compare the map between previous quarters. Used in this way, these tilesets are valuable in providing planet-scale information about the evolution of the map, at least 3-month resolution. These tilesets and the workflow presented here supported the data analysis in the next Chapter.

However, truly contributor-centric, individual-edit information is still not available for these tile-based analysis work-flows. Having observed the scalability and extensibility of a tile-reduce workflow, I became determined to produce a different form of OSM-QA-Tiles that can support the most contributor-centric analysis: Showing exactly who changed which objects, how, where, and when across the entire globe, at any time. To achieve this, I directed all further development to creating the Full-history OSM-QA-Tile schema, presented in Part V.

Chapter 9

Corporate Editors in the Evolving Landscape of OpenStreetMap

With the permission of my coauthors, the following chapter is an exact reprint of an article published in the International Journal of Geo-Information in May 2019.¹

9.1 Introduction

OpenStreetMap (OSM) is a freely available and openly editable map of the world founded in 2004 by Steve Coast in response to the prohibitively expensive geographic data owned by the Ordnance Survey [24]. Since this time, OSM has grown into the world’s largest Volunteered Geographic Information (VGI) platform. OSM is comprised of the consumable product—the mapped, geographic data produced by millions of people around the world—and the massive community that maintains it. At its technical core, OSM is a geospatial database with billions of entries that denote hundreds of millions of physical objects in the real world. Several researchers have commented on the growth in the volume and the evolution of this geographic content in terms of accuracy and completeness [41, 10, 46, 109, 76]. The constantly-evolving map is supported by a growing community of mappers with a variety of motivations [18]. In addition to individual mappers, various groups formed around OSM also provide clues about the diversity of interests in the OSM community. These include for-profit organizations that use the map-data, organizations such as the Humanitarian OpenStreetMap Team (HOT), which creates geospatial data both in preparation of and response to humanitarian crises around the world, or the many formal and informal local

¹ Jennings Anderson, Dipto Sarkar, and Leysia Palen (2019). Corporate Editors in the Evolving Landscape of OpenStreetMap ISPRS Int. J. Geo-Inf. 2019, 8, 232. doi:10.3390/ijgi8050232

OSM communities that organize mapping parties and other events to encourage participation and data contribution. As such, OSM can be described as a “community of communities” that curate and edit map data on a single platform, compelled by a range of individual and shared motivations, but with the over-arching objective of creating a freely accessible, open, and editable map of the world [127]. The continued growth of OSM is a testament to the idea that maps are never fully formed, and are thus an ever-evolving product of embodied, social, and technical processes [59]. Maps represent snapshots of the moment, reflecting the values and priorities of their creators. The various communities within OSM edit the map with different goals and motivations with the hope that the common platform results in a uniform product useful for all. The ongoing efforts of this “community of communities” make OSM a constantly evolving map-of-the-moment adapted to the requirements of the day.

The last two years have seen major growth of a particular type of community: corporate editors. These are paid editors that curate the map professionally. While numerous for-profit corporations have always been involved in OSM—typically through using OSM data in their services and products—the rapidly increasing number of paid-editors on the platform is new and has become a contentious issue for some in the community. Presumably, the corporations employing these editors are investing in OSM in relation to their product. For example, some core Mapbox products rely on maps built on OSM data. As such they were one of the first companies to engage in this activity, beginning as early as 2014. Other companies, such as Amazon Logistics, claim to use some OSM data in their internal routing algorithms. In turn, they contribute back information from their drivers to improve the vehicle routing abilities of OSM data [89]. In this article, we identify ten corporations that transparently employ teams of professional editors. We explore the editing activity of each team to better understand the impact on the map and community. Though some editing mishaps have made the OSM community suspicious of corporate editing, guidelines around transparency and community engagement are now in place that these corporations attend to—and in so doing, make the usernames of their editors available. To the best of our knowledge, this is the first article exploring the role and contributions of corporate entities editing OSM at scale. We

consider the discourse about corporate involvement in OSM to inform and contextualize quantitative analyses of the OSM database to measure the global footprint of the ten companies.

9.1.1 OSM Contributors

OSM relies on volunteer contributions to build and curate the map: specifically, this means that OSM does not offer financial incentives to mappers. Currently, there are more than 5 million registered users, over 1 million of whom have edited the map. The growth of the entire OSM community is shown in Figure 9.1. Researchers have noted the motivations for contribution to OSM as ranging from altruistic to vandalistic as a result of intrinsic self-motivations and external societal, economic, or political drivers [89, 25, 81, 8]. The legal entity behind the OpenStreetMap project is the OSM Foundation (OSMF). OSMF is a U.K.-registered non-profit that supports OSM by fund-raising, managing servers, organizing and sponsoring conferences, and supporting working

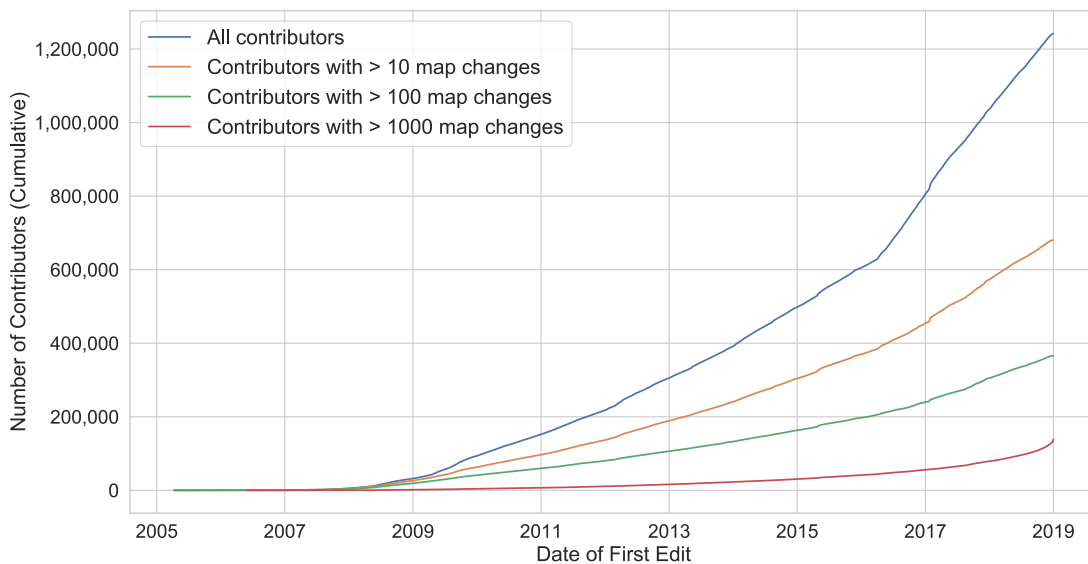


Figure 9.1: OpenStreetMap Contributors: Over 1 million users have made at least 1 change to the map. Far fewer contributors have contributed more than 10, 100, or 1000 times. Results calculated from an OSM changeset database, created from the OSM changeset files by the open source tool: github.com/toebee/changesetmd.

groups that attend to various business functions such as licensing, operations, or communications. OSMF is run by a board which is elected by due-paying members [102]. Membership with the OSMF is separate from having a user account on openstreetmap.org, which is required for mapping. There is no requirement to join the OSMF to be part of the OSM community (that is, as a mapper, data consumer, etc.). Though there may be overlap in personnel, projects, and donors, but there are no formal governing links between OSM subcommunities—such as HOT, local OSM groups, or companies—and the OSMF.

The response of the OSM community has been notable in the wake humanitarian crises [125, 5]. In particular, HOT mobilizes and coordinates global mapping events in response to disasters, including Typhoon Yolanda (2013), The Ebola Crisis (2014), and the Nepal Earthquake (2015), to name just a few. Additionally, local OSM communities organize mapping parties to recruit and support new participants as well as to map previously unmapped areas [36, 53]. Regional and global State of the Map conferences are also organized by active OSM groups, typically with support from the OSMF and regional OSM organizations. In addition to the map itself, there are active mailing lists and a wiki which also serve as venues for user contributions and discussion.

Not all users contribute equally to the map. OSM is no exception to the 90-9-1 rule found in online communities where only a small number of active contributors account for most of the contributions [85]. By our calculations for OSM, the top 1.4% of editors are responsible for 90% of all the map changes (Figure 9.2). On a monthly basis, approximately 1 to 13 percent of users actively contribute data [86]. Figure 9.1 shows that though over 1 million contributors have edited the map, less than 700,000 have made more than ten changes to the map.

Like other online platforms, OSM also reproduces offline inequalities. Several groups of people are underrepresented, including women, people in the Global South, people of color, and non-urbanites [128, 43, 27, 84, 39, 110]. The skewed participation in OSM produces several artifacts in the data [138, 131, 19]. For example, the predominance of male participation in OSM has created an apparent over-representation of features that are correlated to male interests [128]. Availability and access to the internet, technical knowledge, barriers created by the gatekeepers of the platforms,

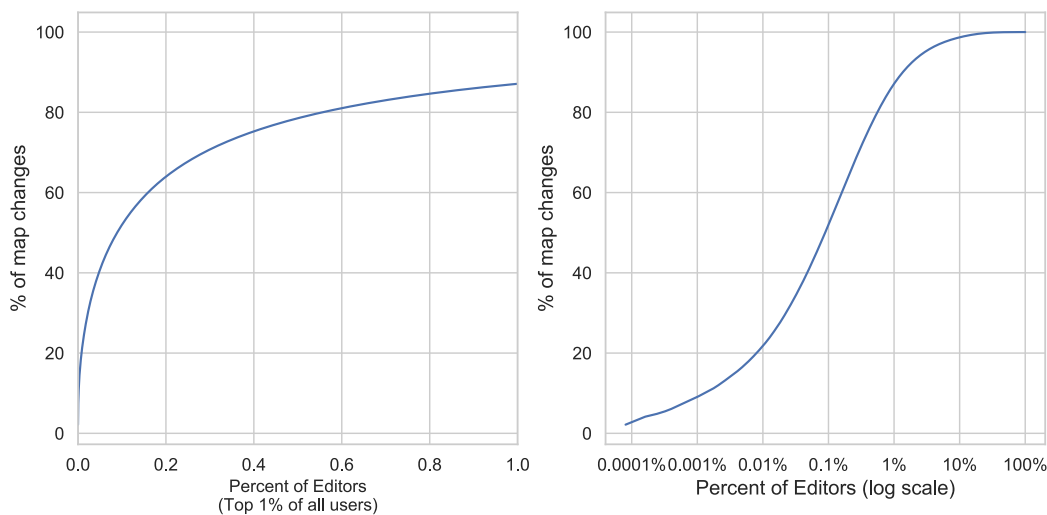


Figure 9.2: Left (a): The top 1% of users are responsible for 87% of all the changes to the map; Right (b): OSM adheres to the 1% rule: a very small percentage of the editing community contributes the majority of the data. Results calculated from the OSM changeset database described in Figure 9.1

and lack of free time and opportunity to contribute have been recognized as some of the hindrances to equal participation [128, 39, 19, 49, 117]. In addition to systemic barriers, researchers have also highlighted that the global political landscape has significant impact on contributors and consequently, on the data produced [117, 111, 44, 14, 20, 66].

9.1.2 Landmark Corporate and Government Contributions to OSM

While the rise of corporate editing teams is a new phenomenon in OSM, corporate presence is not new to OSM. For over a decade, corporations, governments, and other organizations have been heavily involved in shaping OSM as it exists today. These involvements are documented through the OSM wiki, mailing lists, and blogs, and cannot be traced through the scientific literature alone. As one example of this, the OSM founder, Steve Coast, also founded Cloudmade, a company that provided geo-services based on OSM data [90]. In this we see that special-interest groups are not new to the OSM community; corporate editorship is not simply a case of capitalist appropriation of an open data project, but rather the latest stage in an evolving project comprised of a wide-array of

stakeholders, each coming from a different value system.

Next, we highlight a few key involvements of external groups that have had significant impact on shaping the community and the map since its inception. First, the ability to trace features from Yahoo! aerial images as of December 2006 removed the barrier of requiring GPS devices for contributing to OSM [91]. This enabled “armchair mappers” to create and edit data for remote locations. However, armchair mapping comes with its own set of challenges caused by georeferencing errors and temporality issues. These issues prompted OSM to come up with guidelines for tracing features [92]. Over the years, various custodians of aerial and satellite imagery—including Bing, Esri, Digital Globe, and Mapbox—have made their data available for tracing in OSM. A comprehensive list of imagery providers is maintained on the OSM Wiki [93]. Making satellite images available post-disaster has been critical in the usability of OSM for disaster response [124]. This has particularly aided the OSM community in quickly creating data for areas that lack good geospatial data during times of need [125]. Projects such as HOT and Missing Maps leverage the image tracing function to mobilize armchair mappers to contribute data for vulnerable places that lack geospatial data.

Second, large data contributions have significantly increased the map data available and overall map usability. A landmark contribution of government data to OSM was the uploading of the Topologically Integrated Geographic Encoding and Referencing (TIGER) dataset produced by the U.S. Census Bureau starting in September 2007. The Automotive Navigation Data (AND) was also uploaded at a similar time, adding the road network for the Netherlands along with parts of India and China [91]. Several organizations, groups, and individuals have since contributed to OSM through large data imports. Such imports of bulk data are valuable for increasing the data volume, though integrating them with existing OSM data is challenging. For example, after the TIGER import, several compatibility errors were noted because the TIGER dataset and OSM do not follow the same road classification [143]. For managing the challenges of data integration, the community has come up with guidelines for importing government data [94]. The OSM wiki maintains a list of ‘large-scale’ data imports and potential data sources for import and use [95, 96].

Several governments are both using and contributing to OSM. The World Bank has supported

development of OSM data for both humanitarian crisis purposes and also as an ongoing effort for places that lack capabilities to develop geospatial data [47]. Government entities including the City of New York and Portland’s Traffic Authority have dedicated teams responsible for improving OSM data in their jurisdictions [77]. Previous research has described government contributions and usage of OSM data in greater detail [47, 77, 55]. Corporate entities such as Mapbox, Stamen, and Geofabrik also use OSM data and make active contribution to the database and community through various services they provide [77]. Corporate contributions to OSM data in small cities that lack good geospatial data has also been noted [110]. Even though focused attention on corporate editing by the OSM community is reaching new, visible heights, the OSM contributor network has been historically comprised of public and private entities that have participated for various reasons in a shared vision of an open map of the world. This report therefore focuses on the apparent growth of corporate involvement in the past few years, and why their growing participation through map editing may be fraught, and what this might mean for the future of OSM.

9.2 Materials and Methods

The companies examined in this report were identified through either their longtime involvement in the OSM community, noted by their continued sponsorship of the Foundation and/or conferences, or their current transparency in publicly revealing their involvement in editing the map. This comprehensive sample was made by those with the most editing activity (Apple, Mapbox, and Kaart) along with seven other corporations that the authors were able to identify through their conference participation and their publicly visible list of paid editors. In total, we identified 954 usernames associated with corporate editing. At the time of writing, we are unaware of other corporations with as much editing activity as those identified here. It is possible that there are other companies employing teams of editors, but have yet to disclose this information.

We used two types of data sources to then further examine the role of corporate editorship: public articles and data about corporate involvement in OSM, and the geospatial data created by corporate editors.

For the first source, we identify information across websites and media outlets to help trace the interest expressed by corporate editors for using and editing OSM. This information links also to publicly-available data that lists usernames of editors associated with each corporate team. It also lends insight into the motivations, the nature of edits, and the mode of edits because these companies both list and discuss specific mapping projects and their progress. The OSM sponsors list was used as the starting point for assembling a list of companies interested in OSM. Media articles were obtained when developments regarding this new phenomenon of corporate editing occurred. The authors' long-time experience in the OSM community, including personal observations at State of the Map conferences, informed the formation of the questions and interpretations.

For the second source, we use historical quarterly-snapshot OSM-QA-Tiles for quantifying where and what the corporate editors are editing on the map. OSM-QA-Tiles are vector tiles containing object level editing behavior for the vast majority of OSM data: roads, buildings, points-of-interest, etc. in an efficient, accessible form. For example, a recently modified building will exist in an OSM-QA-Tile as a polygon object with metadata including the name of the mapper that most recently edited it, the timestamp of this edit, and the current version number of the building: denoting whether this user created the building (version = 1) or edited an existing object (version > 1). We find this to be more accessible than the standard OSM data-model which requires first reconstructing the building by identifying the individual nodes associated with the object. However, an analytical weakness of the standard OSM-QA-Tiles is that map objects are unique (one version of each object), so other than knowing their current version number, objects are unaware of their own editing histories. Thus, these tilesets can only represent a snapshot in time: the most recent version of the map data. For this historical analysis, we used the historical quarterly-snapshot OSM-QA-Tiles. These tiles represent the map at the end of every quarter since 2005. Historical OSM data analysis is possible by iterating through these tiles to get quarterly development of the map. For example, if a road was created in January, and then subsequently edited three more times in April, August, and December of that year, then each of the quarterly snapshots will include this road along with the metadata corresponding to each of these edits (e.g. usernames and date of each

of these four changes). An annual snapshot, in contrast, would only include metadata for the latest edit occurring in December. Objects are edited at all frequencies, but quarterly snapshots give a finer resolution of the evolution of the map while still making global-scale analysis computationally efficient. We use the open-source Javascript framework `tile-reduce` (github.com/Mapbox/tile-reduce) to efficiently process these historical vector tilesets, following the same methodology as previous work by Anderson et al. [5].

Thus, the initial analysis of media articles, blog posts, and wiki pages enables us to position corporate editors in the context of the larger OSM community, while the evaluation using OSM-QA-Tiles quantifies the impacts to the map. To label edits as corporate, we match the usernames associated with edits with the publicly disclosed lists of usernames associated with each company. In the event a mapper edited before and/or after being employed by a company, we filter by time to count only the edits that occurred during the mapper's employment on a corporate data team.

9.3 Results

9.3.1 Observational Analysis of Corporate Involvement

We focused on the ten corporate entities that have shown significant interest in editing OSM. We highlight the announcements and coverage of this phenomenon in different news media in the past two years, then we discuss the visibility and contributions of these companies in the OSM community. We then examine the traces of these companies in the OSM data itself. Table 1 highlights how the quantity and variety of contributions from each varies dramatically.

9.3.1.1 Tracing Corporate Interest Through Media

Bing, a subsidiary of Microsoft, has contributed 125 million building footprints in the U.S. to OSM, which they extracted from aerial imagery through deep learning algorithms [132]. In addition to contributing automatically extracted and generated data, Microsoft has also assembled a team of editors to contribute to OSM. The aim of their Open Maps Team is to work closely with the

Corporation	OSM Foundation Engagement	Team URL	Team Size	Number of Edits	KM of Roads Edited	Bldgs Edited
Amazon	Gold Corporate Sponsor (Amazon Web Services) SOTM 2013	wiki.openstreetmap.org/wiki/Amazon_Logistics	110	388,000	120,000	1000
Apple		github.com/osmlab/appledata/wiki/Data-Team	342	3,944,000	1,643,000	1,156,000
Development Seed		wiki.openstreetmap.org/wiki/DevSeed-Data	8	488,000	62,000	269,000
Facebook	Gold Corporate Member 2018 Gold Sponsor-SOTM 2018 Silver Sponsor-SOTM 2017 Bronze Sponsor-SOTM 2016	wiki.openstreetmap.org/wiki/AI-Assisted_Road_Tracing	87	1,106,000	821,000	1000
Grab	Gold Corporate Member	github.com/GRABOSM/Grab-Data	124	1,593,000	300,000	63,000
Kaart	Bronze Corporate Member Bronze Sponsor-SOTM 2018	wiki.openstreetmap.org/wiki/Kaart #Kaart_Data_Team	93	2,887,000	484,000	702,000
Mapbox	Gold Corporate Member 2018, 2017, 2016, 2014, 2013 Silver Sponsor- State of the Map 2012	wiki.openstreetmap.org/wiki/Mapbox #Mapbox_Data_Team	40	4,483,000	1,694,000	1,088,000
Microsoft (Bing)	Gold Corporate Member 2018, 2017 Platinum Sponsor-SOTM 2011, 2010	github.com/Microsoft/Open-Maps/wiki/Open-Maps-Team-at-Microsoft	29	643,000	458,000	52,000
Telenav	Silver Sponsor-SOTM 2017, 2016 Platinum Sponsor- State of the Map 2012	wiki.openstreetmap.org/wiki/Telenav #Telenav_folks _on_OSM	30	963,000	336,000	5000
Uber		github.com/Uber-OSM/DataTeam	91	464,000	32,000	349,000

Table 9.1: Known Corporate editing teams active in OSM. The column OSM Foundation Engagements shows the current affiliation with OSM Foundation and their sponsorship of State of the Map conferences. Data for edits as of January 2019 (since 2014), rounded to the nearest thousand.

OSM community to improve data quality in places of strategic importance to Microsoft [52]. The team coordinates their activities through GitHub where their team members and projects are listed. Each project is thoroughly described there. In addition, GitHub issues-tracker offers a place for community feedback and questions, which supports transparent, documented issue resolution for each mapping project (github.com/Microsoft/Open-Maps/issues). Mapbox/DevSeed, Apple, Kaart, Telenav, & Grab also use GitHub in the same way to track their projects and answer community questions. Microsoft’s commitment to OSM is an extension of their support of open source projects

[2, 71].

Facebook's OSM contributions to date have mostly been through supervised automated contributions. They use machine learning to detect road networks from satellite imagery which are then validated and reviewed by their OSM editors who work closely with the local OSM communities. All machine-identified roads are reviewed by a human editor before being imported into OSM. Their efforts were initially focused on mapping Thailand; they have completed editing road data for all 79 provinces, adding a total of 515,306 km of road to the map [97]. They have used similar infrastructure in collaboration with HOT to contribute in the aftermath of the 2018 floods in Kerala, India [98].

One of the most valuable aspects of digital maps are navigation capabilities. However, ensuring topological and semantic rules is tedious [41, 23, 142, 83]. Government and corporate data contributions, coupled with the efforts of the community to clean and integrate these data into OSM, have ensured that road network data in many places in Europe and North America are suitable for navigation. However, this is not the case for OSM data in Asia. Many Asian countries have emerged as big markets for ride sharing services such as Uber and Grab [63]. Uber has announced that it wants to migrate its mapping service to OSM; New Delhi, India will be the first city where this OSM-based service will be rolled out [54]. Uber also announced through a community posting in an OSM forum that they will involve a team of editors to improve map data specifically for navigation by modifying and adding turn restrictions, directionality, and road geometry [57]. Grab has dedicated considerable amount of effort into improving OSM data for Southeast Asia. In addition to having a team of editors, Grab has organized several mapathons in many countries for wider community engagement [136]. They have also partnered with HOT for the mapathons to ensure their edits are relevant in crisis situations [122].

Until 2018, the Mapbox data team was the most active team of corporate editors in the OSM community. Mapbox was one of the first companies to employ a team of OSM-specific editors, starting as early as 2014. In late 2017, a large part of the Mapbox data-team merged with the Development Seed data team, creating DevSeed Data [119]. Like Facebook, this team is also heavily

invested in machine-assisted mapping: using machine learning to help their data team identify features to map. Kaart drives vehicles all over the world to capture road networks and ground-level imagery to improve OSM [99].

The OSM community has been divided about its policies to enforce transparency and accountability for what they refer to as “organized editing,” which captures mapping activities by both nonprofit (e.g., HOT humanitarian mapping) and for-profit groups (i.e., the groups described here). In a 2017 survey by the OSM Foundation, 43% of paid editors—compared to 17% of all respondents—opposed having policies that guide editing activities [101]. Ultimately the OSM Foundation produced the Organized Editing Guidelines in November 2018, the goal of which is to ensure meaningful, transparent participation from large editing teams [37].

Though these ten corporations have been transparent about their editing activities of OSM, there have been mishaps regarding editing conflicts with the community. For example, Grab was in the spotlight in late 2018 for the oversight by their outsourced editors for overriding volunteer contributions with incorrect edits in Thailand . This incident brought unresolved attention about why companies (like Grab) which do not seem to be using OSM in their product are interested in contributing and improving OSM. One Bangkok-based OSM enthusiast speculated that Grab (and Uber) were using OSM data for improved routing in their applications without attribution [80, 79].

9.3.1.2 Contributions to the Larger Ecosystem and Community Participation

The involvement of these corporations in OSM extends beyond editing the map. Many of the open source software tools in the OSM ecosystem are developed and maintained by their employees. For example, iD—the user-friendly, in-browser editor incorporated into openstreetmap.org—was initially developed by Mapbox and DevelopmentSeed with funds from the Knight Foundation with the explicit purpose of improving core OSM infrastructure. Today, iD is a successful open-source software project, with more than 10,000 code commits on GitHub—whose core “maintainer” (that is, the lead developer) is a Mapbox employee. Other Mapbox-maintained tools include the OSM validation utility, OSMcha, and a number of OSM data-processing tools available for anyone

working with OSM data. Bing is the primary imagery provider for OSM. Telenav maintains the website improveosm.org, which boasts the tagline: “Tools and Data from Telenav, built for the OpenStreetMap Community.” Some of these data are pre-processed datasets of potentially missing features identified by machine learning on telemetry data. These are just a few examples of corporations participating in OSM in addition to their paid editing-teams. This is far from a comprehensive listing of which companies have contributed useful utilities to the project. Tracing the decade-long involvement of developers, their employers, and the variety of funding sources (corporate, donation, NGOs) is beyond the scope of this article, but it is safe to say that corporate involvement in OSM has shaped and maintained the project as it exists today.

Corporations also have access to rich geographic data from their customers and operations. For example, telemetry data (typically location data from mobile devices) can be used to identify missing roads, turn-restrictions, one-ways, and more. Mapbox compiles these data internally to assist their data-teams (mapbox.org/Telemetry). Amazon Logistics reports that they use their delivery driver’s GPS traces in conjunction with driver feedback to help improve OSM [89]. Grab also states on their wiki page that their data-team process begins by downloading their internal GPS traces [100]. In terms of improving the OSM road network, there are few substitutes to such telemetry data. These datasets leverage a massive number of sensors, obtaining more ground-reference GPS traces than any number of individual OSM contributors could possibly acquire as hobbyists.

Some members of corporate editing teams are not new to OSM. At least 14 members of the corporate editing teams were actively editing before 2014 as individual contributors. Collectively, they represent 1% of all corporate editors, but their total edits to the map are equal to about 4% of all corporate edits since 2014, suggesting they are heavy corporate editors. In a 2017 OSM Foundation survey, 55% of respondents that were associated with an organization engaged in paid editing were with OSM for 3 years or more before joining said organization [101].

Corporations often sponsor and participate in OSM conferences. From our observations, their presentations are some of the best attended talks at State of the Map conferences, especially if it is the corporation’s first talk to the OSM community. They can also be some of the most contentious,

prompting aggressive remarks and questions from the community.

9.3.2 Quantitative Evaluation Using Historical Quarterly-Snapshot OSM-QA-Tiles

Next, we use historical quarterly-snapshot OSM-QA-Tiles to visualize and understand the global footprint of the 10 corporate editing teams and the features they are editing. Since mappers may have been active before being employed by a data team, knowing the date when a mapper becomes a corporate editor is an important detail. The resolution of this detail is limited to the responsiveness of the company itself to update their publicly-facing list of data-team employees. For example, if a mapper stops mapping as an employee in January, but the editor list is not updated on the wiki until February, there is no publicly observable method to know that any edits between these times were not corporate edits. At the moment, manually tracking and updating these lists for this type of research is possible, but as these teams continue to grow, this task will become too burdensome for individuals to do manually. Beyond measuring where corporate editors are active and what they are mapping, we show distinct temporal editing patterns that characterize these teams. Specifically, corporate editing teams appear to follow a Western five-day work week, where the activity of these teams is punctuated starkly with periods of little to no editing every five days by apparent weekends as days off. Later we discuss how corporate editing may be identified in the future expressly from these distinct temporal patterns.

9.3.2.1 Global Footprint

Figure 9.3 shows the global footprint of the 10 corporate editing teams. This map is produced by plotting the location of every edit by a member of a corporate team (denoted by color). High concentrations of edits appear to glow white. Together, the power of corporate editing is globally reaching. Mapbox and Apple have the largest footprints with edits on all six populated continents. Telenav and Amazon mostly edit in North America and parts of Europe whereas the Microsoft team is focused mostly on North America and Australia. Grab predictably focuses only on Southeast Asia. While Uber does have some edits all over the globe, they are primarily active in New Zealand. As

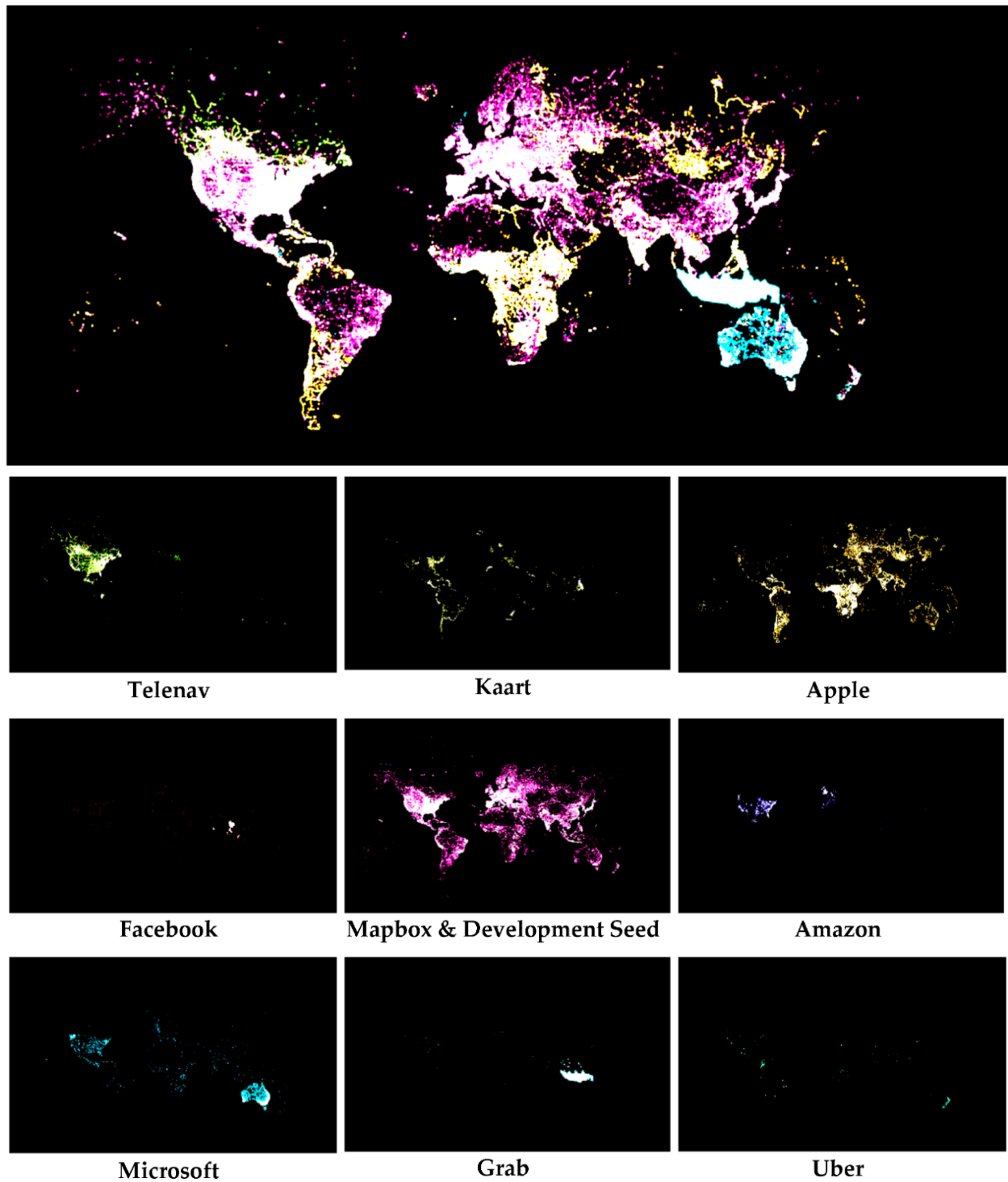


Figure 9.3: Where corporate editors are editing. The main map shows an aggregated view for all 10 companies. The sub figures show where each company is editing. In this map, we have combined the Mapbox and Development Seed teams because they merged in late 2017.

mentioned earlier, Facebook’s work is heavily focused in Thailand. Overall, Figure 9.3 shows that corporate editing is a global phenomenon with specific regions of more interest to some companies than others.

9.3.2.2 What Are Corporate Editors Mapping?

Table 2 highlights the increasing activity of corporate editors over the last 4 years. 2018 stands out as a remarkable year as it seems to indicate a change in collective focus towards editing road networks and building data. Figure 9.4 shows the relative quantity of edits to buildings, kilometers of road, points of interest, and amenities per team per year, compared to the total number of edits in the area. Thus, the radar charts highlight the main features of focus for the teams in the areas they are editing. For example, in 2018, Apple editors, on average, were responsible for nearly 80% of all the edits to existing roads and 70% of all the new roads created in the areas where they were active, defined by zoom level 12 map tiles (about 95 km² at the equator, the size of a small city). Generally, companies have a preference for a particular edit type. Telenav and Grab, which focus on navigation, are primarily editing roadways. In the case of both corporations, they are editing existing roads more often than they are creating them.

Apple, Microsoft, and Facebook also have a massive imprint on the road networks in the areas their data teams are active. In 2017 and 2018, these teams were responsible for creating more than half of the new roads and editing more than half of the existing roads. Compared to all editors, Uber

Year	Features	New KMs of Road	Edited KM of Existing Road	New Buildings	Edits to Existing Buildings	Amenities	POIs
2015	1,703,107	96,604	660,591	321,535	47,730	13,892	40,096
2016	2,251,615	87,321	677,795	308,785	198,366	69,949	214,087
2017	3,121,727	179,256	591,627	632,859	305,665	58,616	178,887
2018	9,925,463	682,938	2,982,248	1,709,935	176,113	33,845	61,238

Table 9.2: Total number of features, kilometers of roads, number of buildings, amenities, and points of interest edited per year by all corporate editors. The increase in number of features edited since 2015 shows the overall rise in corporate editing.

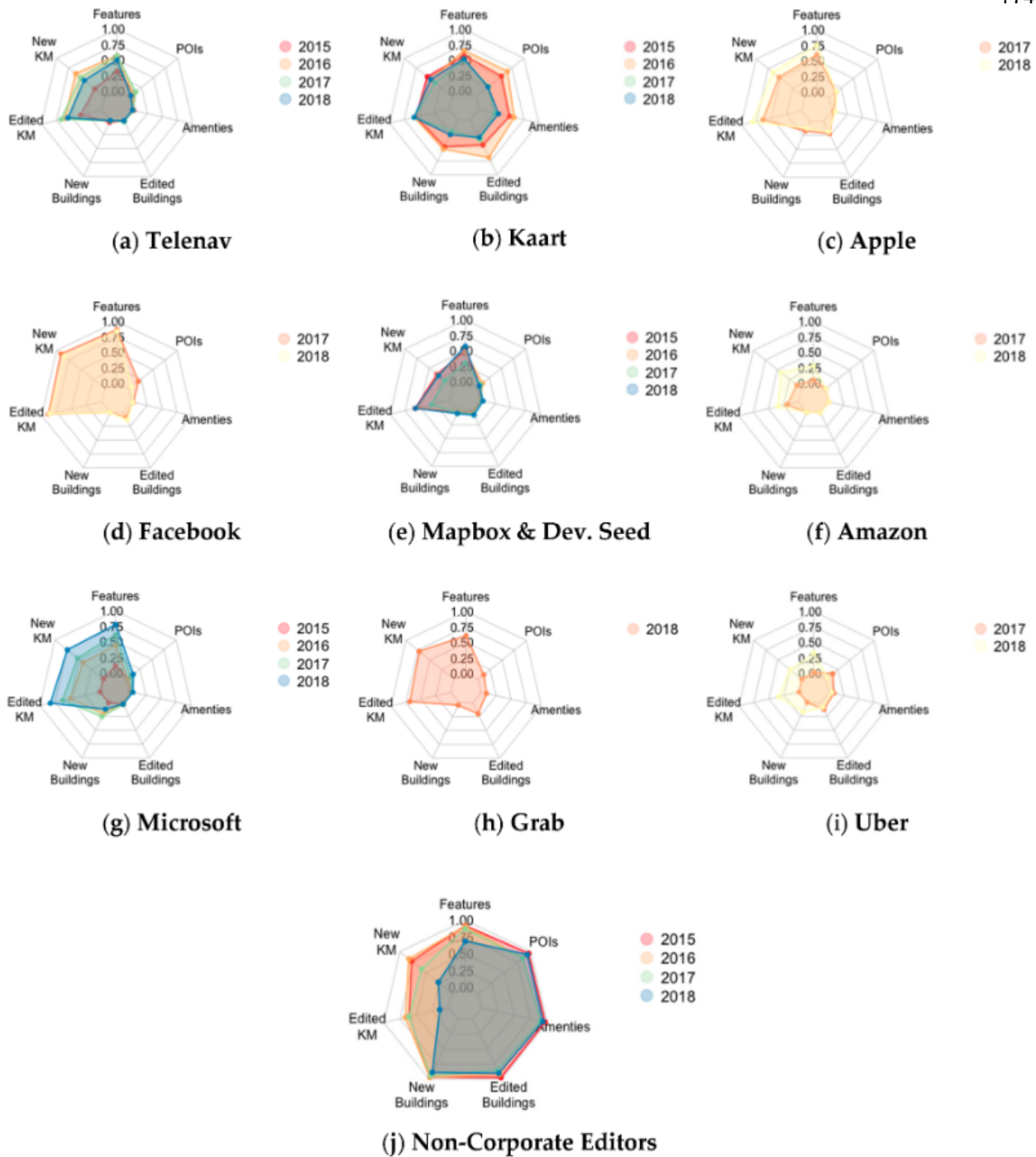


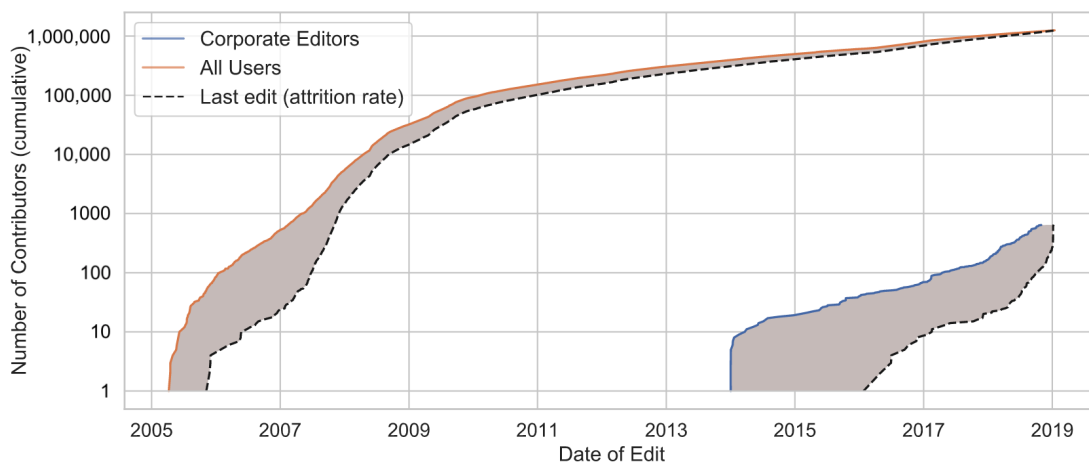
Figure 9.4: Each figure shows the types of edit these companies performing, relative to the total editing activity where they are active. These are annual averages over all of the zoom level 12 map tiles where a company is active. “Features” refers to editing any feature (all types of edits). The final figure (j) represents the activity of non-corporate editors in areas where (any) corporate-editors are active.

never dominates editing in regions where they have been active in the two years they have been involved in corporate editing. We see that in 2017, they were more focused on editing buildings, amenities, and points-of-interest; they did not focus on road editing until 2018. In recent years, Kaart continues to be responsible for over half of the total road edits in regions where they are active, but the percentage of buildings, amenities, and points-of-interest they are mapping has been decreasing, on average. Grab, which has only been active in the last year, has been predominantly mapping the roads in the areas in which they operate, making them responsible for nearly 75% of both new roads and edited roads in the map.

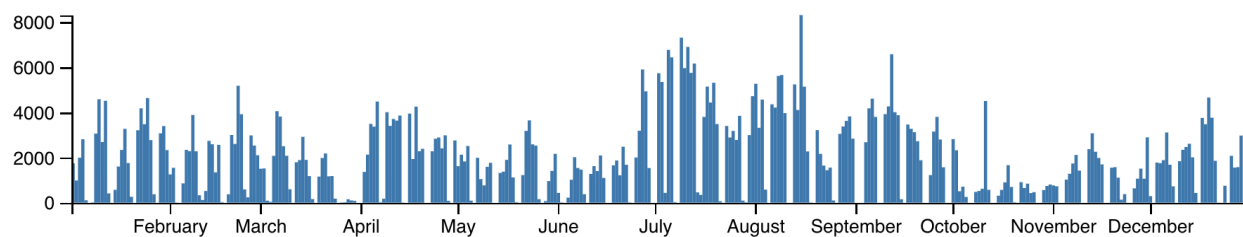
9.3.2.3 Characterizing Corporate Editing Patterns

Corporate editors also leave a distinct Monday through Friday time-signature in the database. In addition, Figure 9.5a shows the difference between corporate and non-corporate editors in terms of their lifespans of active editing. The solid line represents the number of editors starting from their first edit, while the dotted line represents the number of editors on the day for which they made their last edit. Depicted this way, the area between the two lines represents the size of the active community at any given time as people join and leave. This figure shows that even though there are less than 1000 corporate editors, the relative size of the active community is larger than the number of total mappers when OSM first began, and generally more stable in terms of ongoing contribution. The primary difference is that there are very few “one-time contributors” on corporate data teams; this one-time contribution behavior, which is more common in the general OSM community, drives the two lines closer together as the first edit and last edit of a mapper are on the same day. Instead, the slope of the dotted line is steep at the end of 2018, showing that many editors are still active right up to when we pulled the data.

A signal of possible corporate editing is the apparent weekly pattern of editing activity expressed in Figure 9.5b. Each of the data-teams explored here maintain this same pattern: editing consistently during the week throughout the year with few-to-no edits on weekends. This also suggests that these unique temporal signatures could be used to identify corporate editing activity by



(a) The rate of growth of all OSM editors compared to corporate editors. The solid lines represent number of contributors denoted by the day of their first edit. The dotted lines represent the number of users denoted by the day of their last edit. The shaded area between the solid lines and the dotted lines could be thought of as the relative size of the “active” community. These two lines converge at the end because those are the most recent edits in our data. The steep slope in the corporate-editors dotted line shows that these editors have been active recently (not one-time contributors)



(b) Edits per day by the Facebook team in 2018. Consistent activity throughout the year showing 52 weeks of relatively consistent work five days of the week, with no editing on weekends. This pattern of consistent weekday editing is present across all of the data teams we have examined.

Figure 9.5: Characteristics of Corporate editors

teams that have not yet disclosed a list of users or come forward publicly as using and contributing to OSM. This could be quite volatile if these editors are found to be in violation of the organized editing guidelines. Preliminary analysis identifies another 3000 active mappers that exhibit similar editing patterns in temporality and volume: many of whom are involved in import efforts and humanitarian mapping tasks. However, there is currently no evidence suggesting these are undisclosed corporate editors and, moreover, it is difficult to validate such accusations unless the mapper self-reports an employer in their user profile—as is common for currently known corporate editors.

9.4 Discussion

The growing phenomenon of corporate editing is the latest evolution of corporate involvement in OSM. Of specific interest is the massive growth in the number of corporate editors and the apparent investment that corporations are making in OSM. Though prolific, corporate editing varies in geographic reach, objects edited, and volume across corporations.

While Figure 9.3 may initially appear to present corporate editing as dominating the map, Table 2 and Figure 9.4 explores the impact of these edits. Though there is disproportionate impact across the globe, it appears that corporate editors have the largest impact on the road networks in areas where they are active (compared to buildings, amenities, points-of-interest). This is not a surprise given the value of a routable road network, but also not out of character for the evolution of the map without these editors: the map often evolves first from the road network [23]. This does raise further questions about longevity of corporate interest once the road networks are complete in these areas: will there be motivation for these corporations to map buildings or points-of-interest?

Figure 9.4j shows while there has been a consistent rise in corporate editing as a percentage of the total edits, non-corporate editors are still the dominant force who are responsible for nearly 70% of all features edited in 2018 (averaged globally in areas where corporate editors were active). Meanwhile, the percentage of the road network edited by non-corporate-editors is under 30% for these areas, on average. This means that corporate editing is having a significant impact in the regions where it is happening, but it is not currently dominating the global map. However, the motivations of corporate involvement and their long-term impact on the OSM data and community require further research.

In terms of motivations for mappers, Budhathoki and Haythornthwaite found that, among other factors, learning OSM to demonstrate proficiency to future employers was a potential financial benefit of contributing to OSM [18]. Published in 2013, this study predates the rise in corporate editing and found that while these financial benefits and career outcome were relatively low motivators for contributors, the notion of such financially motivated mapping was present. The increasing

interest of various corporations (Table 1) and the growth of the number of corporate editors (Figure 9.5a) highlights the evolution of OSM and may change the motivations for contributors. With regards to OSM as a VGI platform, these corporate editors exacerbate the double-edged sword conundrum highlighted by Seiber and Haklay [121]. On one hand, compensation typically means some level of expert or professional involvement, indicating high quality data and validation. On the other, if contributions are paid for, the data can be seen as coerced, and perhaps even disqualify as VGI, taking away the benefits of crowd-wisdom and local knowledge for which VGI is recognized. While OSM contributions are still majorly volunteered (Figure 9.5a), the prolific activity of corporate editors pushes the threshold of OSM's status as a VGI project. Regardless of who contributes data, as long as the quantity and quality of the data improves, OSM will continue to be a valuable open data platform.

9.5 Conclusions and Future Research

In what we believe to be a first report of the phenomenon of corporate editors in OSM, we have highlighted corporate editors' place in the community and their visible footprint on the map. Our analysis addresses some of the current tension in the OSM community regarding this new phenomenon. The historical context and observational analysis also highlight the multi-faceted involvement of the companies in OSM which go beyond just editing the map. Corporations appear to have their own map editing agendas that are probably aligned with corporate interests. We also note that other organized groups as well as individuals have been cited as having particular interests that drive their contributions to the map. The contributions are further shaped by the values embedded in the technology which drive who can participate and how [78, 34]. Thus, the combined effort of all groups driven by their own set of values, motivations, and goals, mediated by the OSM platform produce what is perceived as the unified map of the world. With the ongoing growth and map spread of corporate editing in OSM, it is too early to draw conclusions about the lasting impact of this new iteration of corporate involvement through paid mappers. Instead, we raise questions for consideration about how OSM might evolve.

First, how does corporate editing activity affect the map data? We might wonder if we can separate ideologies of the sources of the data from its presence. One might argue that the uneven coverage of data in OSM and the large-scale edits that corporations are capable of making can close this gap in the database. Additionally, the data added by corporate editors will probably be of good quality because these editors are trained, have economic incentive and managerial oversight, and because editing in these areas will bring attention to the map via a variety of edit-monitoring services. Data are more likely to improve in areas where some data already exists [22, 26]. Thus, these activities, especially in developing nations, may be looked on as map seeding which prompts growth of the OSM community and densification of the data. In developed nations, the editing activities are probably going towards progressive data updates and quality improvements, and thus are more in line with map gardening [73]. However, another important argument comes from the point of view of bias toward self-serving interests: are corporations introducing geographic bias into the map? As the map continues to be filled in, will corporate interests have too much voice in what and where gets mapped?

Second, how does corporate editing affect local communities? Historically, the attitude towards large corporations have been contentious with avid mappers being more congenial [18]. One concern is that corporate editing is squeezing out existing “local” mappers. The organized editing guidelines advocate strongly for working with local communities to avoid this. While the data shows that corporate editing is certainly prolific and is found to be the largest editing force in many places, it is unclear what the relationship is or may become with local mapping communities. Empirically, we observed through the wiki and Github repositories that the reported corporations are currently cooperating with the organized editing guidelines and reaching out to local mapping communities. The community has also arranged itself in such a way as to monitor if corporations overstep in ways that the community can currently foresee, but as the OSM landscape evolves, additional mechanisms might need to be put in place.

Third, are corporations acting reciprocally with the OSM community, and offering as much or more as they are getting from their OSM involvement? Some corporations have access to large,

rich datasets (telemetry) that no one else has, which could in turn improve the map if shared—but how much do corporations share? It says something about the value of geospatial data when we observe that achieving a more complete map is driving corporations to collaborate. Furthermore, global-scale validation and monitoring is difficult for individuals because of the sheer volume of edits. We know from conference presentations and the production of tools that corporate editors actively monitor map changes for vandalism and accuracy at scales beyond the abilities of individual contributors.

Fourth, what is the best way for the community to monitor and support corporate editing, assuming that it does have collective value? Mechanisms put in place such as the organized editing guidelines are primarily based on self-reporting, which is what assessment is then based. However, as the number of corporate entities continues to grow, maintaining lists of usernames and corresponding edits could become onerous. Further mechanisms may be needed so that the community can hold corporate editors accountable, ensuring that (1) their community engagement is proportional to their impact and subsequent benefits from the data, and (2) that their impact is constructive, and in keeping with shared goals of the OSM community.

Ultimately, consequences that stem from the publicized activity of corporations' data production might have yet different effects on market and corporate behavior. Having quantified and contextualized the current footprint and involvement of corporate editing, our hope is that new research about OSM can arise as this vibrant community continues to evolve.

Funding

This work is made possible by the U.S. National Science Foundation, Grant IIS-1524806.

Acknowledgments

The authors thank colleagues both at the University of Colorado Boulder and in the OSM community for their valuable feedback and deeper understanding of the long, nuanced history of corporate involvement in OSM. Additionally, we thank the various custodians of data-team user

lists that made conducting this research possible and their willingness to help gather and curate a master list of corporate editors. Additional thanks to colleague Kenneth M. Anderson and the U.S. National Science Foundation for funding support.

Part V

Full-History Tile-Based Analysis

Chapter 10

Full-History Tile-based OSM Data Analysis

10.1 Moving to Full-Historical OSM-QA-Tiles

To fully understand *how* the map evolves, we need to capture the full editing history of each map object. While *Part IV* demonstrated that annual and quarterly snapshots allow us to ask questions about how the map has changed over time, these snapshots lack the true resolution to answer the question: “Who changed what?” Exposing these exact edit-level behaviors requires transitioning away from *snapshots* of the data to create new representations of the entire history of an object *in a form that allows for effective data analysis*. Section 5.6 introduced two data schemas for historical OSM data that allow us to encode an object’s complete history into a vector tile:

- (1) **Embedded Object History** Distinct objects, with historical versions embedded in their own `@history` object.
- (2) **Individual Versions** Distinct object versions, each version with a `@validSince` and `@validUntil` attribute to distinguish when each version was current on the map.

This chapter discusses the current state of full-historical tile-based OSM analysis, the tools myself and others have developed to enable this type of work, and the projects, workshops, and presentations that have utilized these. First, I introduce the `osm-wayback` utility, the tool I have been developing for the past 2 years to create full-historical OSM-QA-Tiles.

10.1.1 Recreating Histories with OSM-Wayback

The `osm-wayback` utility is a program written in C++ and node.js that uses `libosmium`¹ and `RocksDB`² to transform an OSM history file into a stream of history-enriched GeoJSON objects. In the context of this chapter, history-enriched refers to a representation of an OSM object that has been enriched with the full metadata associated with each of the previous versions, telling the analyst who edited the object when, and what was changed. Where applicable, minor versions and the metadata associated with these geometry changes should also be included in history-enriched objects. `osm-wayback` can be configured to either ignore geometries and handle only the versions known to OSM (the `version` attribute), or save all node locations and perform an additional step to compute the minor versions and historic geometries for all previous versions of an object.

The program runs in three parts, as shown in Figures 10.1 and 10.2. First, it converts an OSM history file into a RocksDB index (with an optional node-location index to reconstruct historical

¹ osmcode.org/libosmium: C++ library for working with OSM data at any scale

² github.com/facebook/rocksdb: Persistent, on-disk key-value storage optimized for very fast lookups. Based on LevelDB

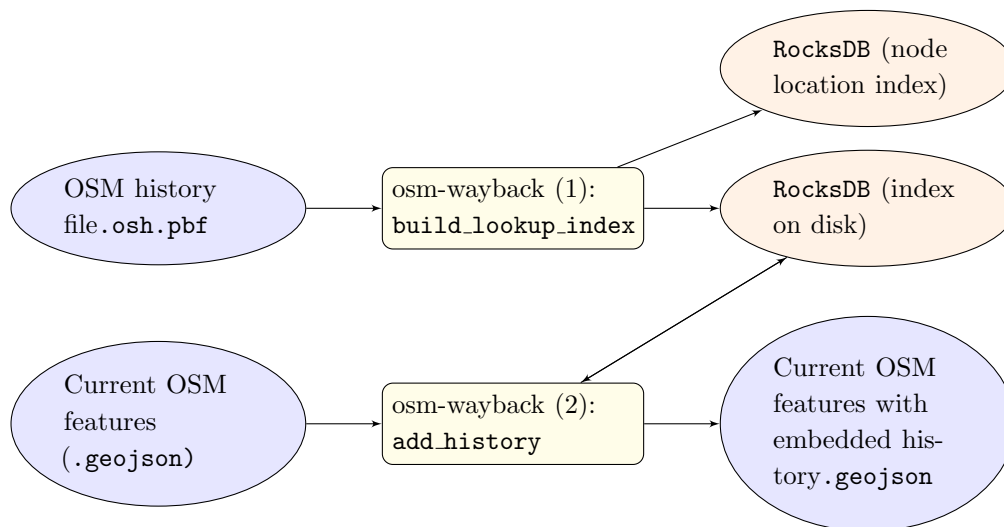


Figure 10.1: `osm-wayback` first ingests an OSM history file and converts it into a rocksDB index on disk with its `build_lookup_index` function. Second, a stream of current OSM objects (from `osmium-export`) is fed into the `osm-wayback add_history` function which reads each object and looks up all possible historical versions in `RocksDB`. After computing the differences, it adds the `@history` attribute to the OSM object, producing a stream of *history enriched* GeoJSON objects.

geometries in an additional step). This step essentially transforms a compact and complex-to-parse history file into a larger, on-disk persistent key-value store where keys reference OSM object IDs and the values are OSM elements (re-encoded as PBFs to be compact). This enables the utility to quickly look up all previous versions of an OSM element by its ID.

Second, a GeoJSON stream of the current version of OSM objects is ingested by the `add_history` function. These streams of OSM objects as GeoJSON are easiest to produce using the `osmium export` command of the `osmium-tool`, as first discussed in Section 4.3.2. As `osm-wayback` reads each OSM object, it checks the current (latest) version number. If `@version > 1`, then there should be a previous version of this object saved to RocksDB in the previous step because it existed in the history file. The utility then sequentially queries RocksDB for all entries with keys: `<@id>!<possible previous versions>`. For example, if the current version of the Way element with ID=123 is version 3, then `osm-wayback` queries the *way column family* for the keys: `123!2` and `123!1`. If these versions are present, `osm-wayback` computes the differences between them, storing the diffs as shown in Section 5.6: *new_tags*, *deleted_tags*, and *modified_tags*. The metadata and diffs for each of these versions are then added to the `@history` attribute, which is embedded into the original GeoJSON object. The final result is an output stream of *history enriched* GeoJSON objects, though these do not have historical geometries yet.

Figure 10.2 shows how `osm-wayback` reconstructs historical geometries for each previous version and any minor versions. First, this requires that all of the node locations were written to a separate RocksDB column family during the first step (a configurable option). If this index exists, the `add_geometry` function builds a list of every node that has ever been associated with any version of the object by looking through the `@history` attribute added in the previous step. Then, it looks up all of these nodes in the RocksDB node location index, adding them to the GeoJSON object as a separate attribute called `nodeLocations`.

Finally, this *history enriched with nodeLocations* collection of GeoJSON OSM objects is processed by the `geometry-reconstruction` node script. This script utilizes the same architecture as `tile-reduce` by invoking another program I created from the `tile-reduce` code-base called

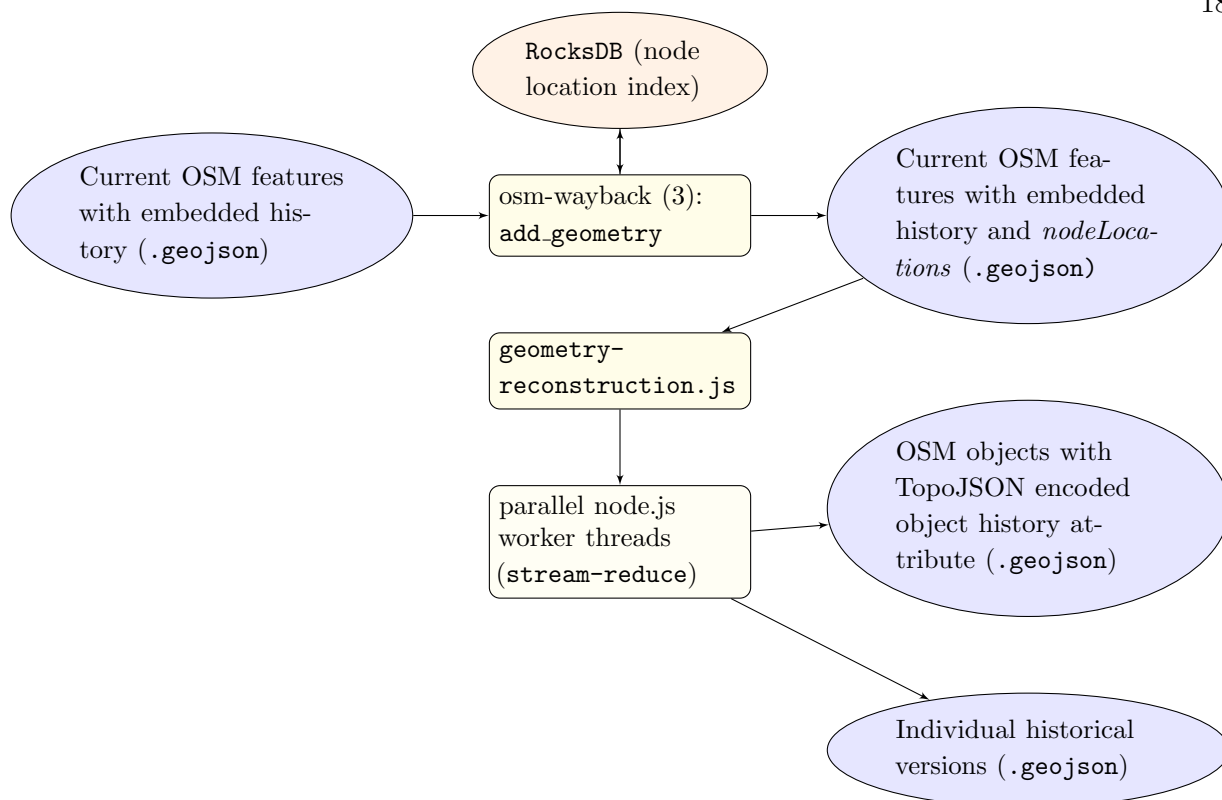


Figure 10.2: Adding historical geometries requires passing the *history enriched* stream of GeoJSON objects produced in the previous step through the `add_geometry` function which identifies every node present across all versions of an object. It looks up all of the possible versions for each of these nodes in the node location index and then embeds them into a `nodeLocation` attribute. The `geometry-reconstruction.js` node script then uses another utility I developed by forking the `tile-reduce` utility called `stream-reduce` which ingests lines of GeoJSON and distributes them to worker functions so that historical geometries can be computed in parallel.

`stream-reduce`.³ This Javascript program reads lines of JSON and distributes them to parallel worker threads. Each worker ingests an OSM object, its history, and the list of all previous node locations, and then computes all of the possible historic geometries for all versions and minor versions of the object. Once an object’s historical geometries are calculated, the GeoJSON is written back out to the stream in one of the two formats: (1) embedded object history, or (2) individual historical versions.

³ github.com/jenningsanderson/stream-reduce is a simplified map-reduce Javascript utility for large files of line-delimited JSON, capable of performing operations on each line of JSON in parallel and returning the results to the main thread. I created this utility for these geometry reconstruction tasks, but designed it to function with any arbitrary line-delimited JSON file, such as extracting and processing geolocated tweets from a massive collection of tweets, for example.

Calculating historical geometries is computationally expensive for objects with many nodes. There are a few tricks to help restrict the total number of possibilities, but ultimately parallelizing the computation with `stream-reduce` was the largest performance improvement I achieved. One trick is to first group nodes by their changeset ID, assuming that any objects modified in the same changeset should belong to the same version or minor-version.⁴ Following this, the total number of possible minor versions should be equal to the number of distinct changeset IDs across the history of all of the referenced nodes. Knowing this helps bound the space of possible historic geometries to compute, but these computations remain a complex task due to the many different editing practices present in the data.

10.1.1.1 Advantages of `osm-wayback`

The `osm-wayback` utility was designed with a number of self-imposed constraints to integrate the system easily with other OSM data analysis tools, specifically into OSM-QA-tile-based analysis workflows. First, the utility has relatively low memory requirements and scales vertically as processing power is added. Since everything (including the node locations) is stored in RocksDB, nothing is kept in memory for very long. The entire planet can be processed on a modest machine as long as there is enough disk space available for the RocksDB index.⁵ The process runs in multiple steps, each producing valid, human-readable GeoJSON representations of OSM data. Not only does this make debugging and iteration easier, but these files can be shared or used in different types of analysis. This record of historical versions could be read into Python or R, if, for example, an analyst was more familiar with questioning the data with these analysis tools. Additionally, since the input is any OSM history file, `osm-wayback` can be run at any scale an analyst chooses, defined by the geographic extent of the history file. This allows the workflow to scale horizontally by first segmenting a history file into smaller chunks and distributing each piece to a different machine (or

⁴ Since changesets can be open for hours it is possible that minutes or hours pass between the timestamps in adjacent nodes; this is entirely dependent on how the mapper works. This can be simplified by assuming that when the mapper pressed “save” is the true time of “a version” (or minor version), denoting what the mapper intended during this editing session.

⁵ An index of the entire planet history is about 1TB. The main problem with large indexes is raising the Operating System’s limits on open files because rocksdb keeps all of the SST files open when connected.

running these in batch if resources are limited).

Evaluating performance is difficult because it heavily depends on the amount of computational resources one throws at this process.⁶ Additionally, other utilities do not match this process end-to-end (computing differences, TopoJSON encoding, etc., so it is difficult to compare with other approaches). For an idea of performance: In Fall 2018, I geographically segmented the planet into 64 chunks along Zoom level 3 tile boundaries, and then ran OSM-wayback on each segment. In this way, I was able to keep the size of each index and resulting files relatively small, satisfying the limitation of a 250GB SSD on the machine I was using.⁷ Running these jobs in batch and uploading each section to Amazon S3 buckets took about two days. Considering that creating the latest snapshot OSM-QA-Tile on the same machine takes about 16 hours, 2-days to generate full-history tiles for the entire planet are acceptable. Since these are for historical analysis, they really only need to be generated once.

As a more practical example, running this workflow locally⁸ with the entire history of Nepal (350MB history PBF file) runs in about 30 minutes, broken down as follows: The history file has 52M node elements, 6.7M way elements, and 42K relation elements. `osm-wayback` then processes all of these elements, creating indexes from these 59.3M objects. Locally, this took 10M, processing an average of 100k nodes/second. Converting the most recent version of each of these objects to GeoJSON with `osmium export` yields 5.3M OSM objects. Reading from RocksDB to compute the attribute changes for all of these 5.3M objects took 3 minutes and enriched these 5.3M objects with an additional 6.2M historical versions. Performing the geometry look up and computing the historical versions took another 17 minutes.

10.1.1.2 Limitations of `osm-wayback`

`osm-wayback` works by enriching a stream of GeoJSON objects, therefore, only those OSM objects that can be converted into GeoJSON may be turned into a stream of *history enriched* objects

⁶ Running primarily on leased machines through ChameleonCloud, there is no guarantee that the same resources are available twice, so these are always changing.

⁷ 64G ram, 48 vcpus, provided by Chameleon Cloud.

⁸ 2018 Macbook Pro i9

with this utility. Additionally, historical geometry processing is only available for way elements, so though the `osmium-tool` is now⁹ capable of processing many relation elements into MultiPolygons, such as complex buildings or administrative boundaries, `osm-wayback` cannot compute the historic geometries for these more geometrically complex objects.

Additionally, `osm-wayback` is designed to run within a larger processing pipeline specifically to create full-history OSM-QA-Tiles. Accessing the full history of individual objects is then only possible through processing the entire history enriched GeoJSON output, easiest done through a tile-reduce job against the final tileset, or visualizing the tileset with a utility like the `wayback-viewer`: A visualization utility to easily render the contents of full-history OSM-QA-tiles for easy debugging and exploration of a particular dataset.¹⁰ This makes the `osm-wayback` processing pipeline most useful to those looking to work with a medium to large amount of OSM history data, not specific objects.

Regarding size, however, another limitation is the complexity of handling a large tileset. Since the size of each feature increases when history is added, the byte-size of the tiles can be reduced by increasing the zoom from the standard OSM-QA-Tile zoom level 12 to 15 to cover a smaller area per tile, thereby decreasing the total number of objects. This also improves the resolution of analysis, as done in Section 8.2. Generating zoom level 15 tiles for full-history enables close to 1 sq. km areas of analysis, but produces 64 times as many tiles. In implementation, this has caused failures at the planet-scale, but works well for Country-sized tilesets. The continental US, for example, is 6M tiles and the full-history can be processed in can be processed 10-20 minutes, depending on the complexity of the analysis.

10.1.2 Other Concurrent Development

`osm-wayback` was developed to implement and test full-history OSM schemas for tile-based analysis, and has been successful in this implementation. It cannot, however, compute geometries

⁹ The support for MultiPolygons in the `osmium` export (previously a stand-alone tool called `minjur`) has dramatically improved since development of `osm-wayback` began.

¹⁰ Available at github.com/jenningsanderson/wayback-viewer; currently only used internally, hope to release to broader community in Fall 2019.

for relations, and therefore can only handle geometries for a subset of what the `osmium-export` utility can. While I will continue to use this tool in this particular data processing pipeline, I do not plan to develop it any further. Today, there are two other utilities: `OSMesa` and `OHSOME` that have been implemented at larger scales and support creating similar history-enriched OSM objects that can be adapted to fit into my pipeline.

A primary reason to move away from this batch-processing model is that these other tools can offer more on-demand processing. The `osm-wayback` pipeline presented here was built to turn a given history file into a full-historical tileset. To update the tileset then requires re-processing the entire dataset to compute all of the differences. A different database backend that could be kept up to date more efficiently and offer on-demand object histories for a given region would significantly reduce the processing overhead of the current pipeline.

10.1.2.1 OSHDB + OHSOME

OHSOME, the new Java-based OSM historical data analysis system built atop the OSHDB database is capable of performing many of the same tasks as `osm-wayback`. OSH takes a different approach from `osm-wayback` by delta encoding all versions of an OSM object into new objects called *OSM Entities*, this decreases the total amount of data stored [112]. OSHDB also handles the concept of a minor version. Supported by an active developer team of OSM researchers, the OSHDB offers a more robust and complete approach from the `osm-wayback` utility and can be integrated into a processing pipeline that will match my full-history schemas for tile-based analysis.¹¹ Future work includes collaborating directly with the team at HeiGIT to investigate such an integration.

10.1.2.2 OSMesa + Auroria

OSMesa offers a powerful cloud-optimized approach to recreating the full-editing history of OSM. Already in production, this infrastructure is robust and built to scale to any quantity of OSM data (especially the full planet history). OSMesa can therefore act as the full-history processing

¹¹ The OSHDB + OHSOME were developed in parallel with `osm-wayback`, which is why I did not build my pipeline on top of this project in the first place. Additionally, they are an active team of developers, I am one person.

engine: computing the version diffs and then exporting the results as GeoJSON to cloud storage that can be retrieved on demand. Currently these results are stored on S3 and can be indexed and queried with Amazon Athena. This allows the arbitrary lookup of the geometry-complete history of any OSM object through familiar SQL queries.¹² At the moment, this only lacks spatial indexing, but it offers a powerful, cloud-based environment for exploring historical OSM data. Initial prototyping shows that Amazon Aurora may be able to provide a spatially-indexed scalable relational database to store the entire editing history, accessible on-demand.¹³ Such a solution could turn the generation of analysis tilesets (such as OSM-QA-Tiles) into on-demand, object-specific tilesets. A tile-server on top of this database could be configured to return vector tiles with all editing metadata and complete editing histories for any region, any time period, for either all data, or only specific edit types. This would solve the concerns about one-size-fits-all OSM-QA-Tiles put forth in Section 5.3. Moving forward, I personally see this Amazon Web Services based workflow as the most feasible option forward for development of all OSM-QA-Tile derivatives, from object-specific (roads or buildings only) to full-history tilesets.

¹² aws.amazon.com/athena: This was used to reconstruct the history of the Taj Mahal in Section 4.3.1.

¹³ aws.amazon.com/rds/aurora: This prototype was a recent collaborative effort led by Seth Fitzsimmons.

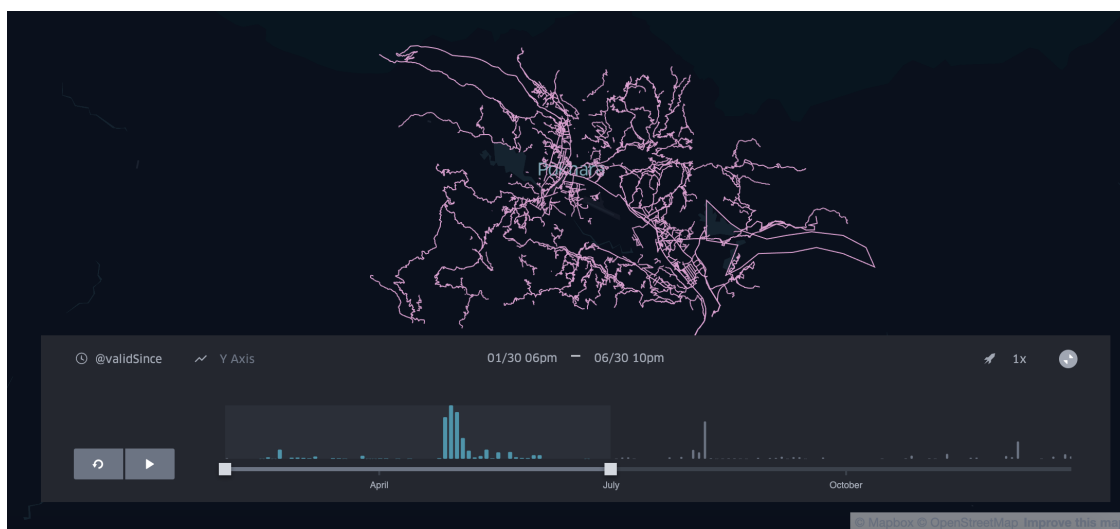


Figure 10.3: Roads edited in Pokhara, Nepal in the first half of 2015. These data represent the successful extraction of OSM data from of a spatially-indexed OSM history-enriched database with historical geometries running on Amazon Aurora. Visualized here with kepler.gl

10.2 Full Historical Analysis To Date

To date, I have implemented full-history OSM-QA-Tiles in an OSM data analysis workshop, ongoing research about the development of the map in the US, as well as various examples and presentations. This section will share some of the results of these implementations of full-history OSM-QA-Tiles.

10.2.1 State of the Map US 2018: OpenStreetMap Data Analysis Workshop

To show the analytical abilities of full-history data analysis, I collaborated with OSM data expert Seth Fitzsimmons to organize an OSM data analysis workshop at the State of the Map US conference in Detroit, Michigan in October, 2018. The workshop attracted about 20 attendees. This next section contains both excerpts from our OSM diary post¹⁴ about the workshop and additional commentary to describe the main takeaways from the workshop. The workshop was advertised in the conference program with the following description:

With an overflowing Birds-of-a-Feather session on “OSM Data Analysis” the past few years at State of the Map US, we’d like to leave the nest as a flock. Many SotM-US attendees build and maintain various OSM data analysis systems, many of which have been and will be presented in independent sessions. Further, better analysis systems have yet to be built, and OSM analysis discussions often end with what is left to be built and how it can be done collaboratively. Our goal is to bring the data-analysis back into the discussion through an interactive workshop. Utilizing web-based interactive computation notebooks such as Zeppelin and Jupyter, we will step through the computation and visualization of various OpenStreetMap metrics.

The purpose of this workshop was two-fold: first, we wanted to take the OSM data analysis discussion past the “how do we best handle the data?” to actual data analysis. We had previously observed that the OSM data analysis conversations often get stuck at the data-handling and data-wrangling step when working with historical data. Rather than go through this again, we implemented our full-history analysis workflows to overcome this obstacle and provided participants

¹⁴ Jennings Anderson and Seth Fitzsimmons. OSM Data Analysis Workshop. State of the Map US 2018. Detroit, Michigan. October 5, 2018. [https://www.openstreetmap.org/user/Jennings Anderson/diary/47133](https://www.openstreetmap.org/user/Jennings%20Anderson/diary/47133). Excerpts republished here with permission from my coauthor.

with pre-processed results so they could move right into the visualization and interpretation steps of analysis. Second, we hoped that providing such an environment to explore the data would in turn generate more questions around the data: What is it that people want to measure? What are the insightful analytics?

A third, internal goal of the workshop, was to compare OSM histories as computed by OSMesa, maintained by Seth and my utility: *osm-wayback*. Since these two infrastructures approach reconstructing the OSM history completely differently, comparing the results across the two offered the first form of external validation. As expected, there were differences in the edit counts between the two approaches that were explained mostly by the tracking of deleted objects (OSMesa does, *osm-wayback* cannot). However, it was validating to see that many of the other numbers, including *minor versioning* were similar. This was particularly inspiring because we had been talking about this problem and the complexities for the past few years and each set about implementing it within our own infrastructures.

10.2.1.1 Preparing for the Workshop

Intentionally, we concealed the entire data-preparation part of the workshop to achieve the goal of actually getting to the *analysis*. To do this, we precomputed the editing histories for 40 different major North American cities and produced files for data analysis at three granularities:

- (1) **Per Edit:** TSV file with individual edits per line describing the change that occurred, by who, when, and where, and to what object. These files were generated with `osm-wayback` and a tile-reduce analysis workflow. First, I extracted the OSM history for a city and then ran it through OSM-Wayback to generate full-history OSM-QA-Tiles. Then, using tile-reduce, I calculated and extracted per-edit statistics, such as the length of road edited or which tags were added, modified or deleted: exporting these edits as individual rows to a TSV file. For a large city, these files were a few hundred MB.
- (2) **Per Changeset:** CSV file with a set of 70 descriptive statistics calculated per distinct

changeset for a region. These include quantities of roads, points of interest, buildings, addresses, sidewalks, parking lots, and more. Computed by `OSMesa`

- (3) **Per User:** CSV file with the same 70 descriptive statistics above, but aggregated per contributor to easily compare activity between mappers.

After computing these editing summaries, I created sample **Jupyter Notebooks*** to act as OSM data analysis tutorials. These notebooks included all of the Python code necessary to read in the editing history for any of the cities and compute and generate a series of graphs, such as identifying the top 15 editors by editing volume or visualize the number of buildings added over time. Workshop participants could then open these notebooks and only need to change a few variables to load data for another city or alter the attributes being visualized to generate new figures. Additionally, participants could export a subset of the editing history to visualize on an interactive map, filterable by time (such as creating an interactive map of when and where the *name attribute* was added to map objects).

10.2.1.2 Running the Workshop

Using Jupyter notebooks running in the cloud allowed us to host a single analysis environment for the workshop such that each participant did not have to install or run any analysis software on their own machines: This was critical because the workshop was only 90 minutes. The notebooks ran on a single cloud machine that was provided by ChameleonCloud.org, an NSF funded cloud-computing infrastructure for computer science research.

One anecdotal analysis enabled by the workshop included a mapper quantifying and visualizing the evolution of parking lots in Chicago. He had been mapping parking lots in Chicago for years and using the per-edit dataset, was able to identify all of the parking lot edits to both visualize his edits, and see all of the other parking lot editors, only some of which he was previously aware. The workshop content, example analysis notebooks, city-level data, and instructions to get up and running locally are all available on Github: github.com/jenningsanderson/sotmus-analysis

10.2.2 Analysis of the US

Using the full-history of editing in the Continental US, I have been exploring the mapping patterns of the US community, specifically investigating the concept of “local knowledge” in the map. In collaboration with OSMUS, we¹⁵ are using demographic information from the 2016 OSMUS Community Census. This was an online survey distributed through social media, mailing lists, and advertised at conferences. One of the questions asked respondents to identify their “local tile” on the map. With this information collected at zoom-level 12, we can connect the 250 respondents with edits that occurred in their local section of the map; as well as categorize their non-local edits.

One hypothesis upon starting this analysis was that local editing—edits one makes on the tile they marked as their local tile—will be distinguishably different from non-local edits: Even possibly allowing us to build machine learning models to classify edits as local or not. One question we looked at it was, “are local users more likely to improve attribute information, such as changing an existing building object from `building=yes` to `building=school`. Initial analysis, however, does not find significant differences; so far there *does not appear* to be major differences in how many mappers edit on local or non-local tiles in terms of attribute-changes. Instead, this initial analysis in the US appears to highlight general mapping experience as the most salient difference: Mappers with more experience do more detailed work, regardless of locality. While more analysis is needed here to find significant differences, such a finding has the potential to redefine the concept of “local knowledge” in the map.

With new support from the executive team, OpenStreetMap US is hoping to conduct another survey in Fall 2019 to identify new needs of the growing community. This survey will include more questions about the community’s editing activities. I plan to advise the design of the survey and then collaborate in the data analysis process, using the information to further this particular analysis of editing patterns in the US.

¹⁵ Current, ongoing work with Robert Soden and OpenStreetMap US

10.3 Full Historical Analysis: Future Work

10.3.1 Paid Editing Interactions and Map Seeding

The root question here is: *What is the interaction between paid and non-paid editors?* Answering this requires that we use the complete editing history of the map, especially the computed differences versions of an object. While the quarterly historic snapshots allowed us to quantify the activity of the paid mappers in Chapter 9, the interaction between these editors was not explored because the quarterly snapshots did not include the version diffs. As put forth in the discussion, understanding these interactions is the next step in that research.

However, this potentially incendiary question needs to be approached with care. As noted in Chapter 9, tensions on the mailing lists and at conferences lead me to suspect that there are some that hope this research will find instances of paid editors deleting and overwriting the work of local mappers. These results could then be used as an argument for banning or reverting paid edits. However, we know these editing patterns *have already occurred*, such as with Grab [79], or even a local case in Denver.¹⁶ My suspicion is that where this has happened, the community has already spoken out and reached resolution. The next question is then, what is the interaction like when paid editors perform data validation or vandalism detection? Now that Chapter 9 quantified when and where paid editing is happening, we can investigate the complete editing histories of these places to look into these patterns.

Additionally, of particular interest to me is checking if *map seeding* exists? Introduced in Chapter 9, map seeding could occur once an area is mapped for the first time by one group of mappers and then another group forms to maintain and grow the data. One hypothesis is that the maintenance—or map gardening [72]—and extension of existing data may have a lower barrier to entry than adding new data to a “blank” map. This may even have the potential to change how people perceive paid editing: If a corporate editing team is the first to map an area, does this

¹⁶ A paid editor was splitting existing roads into smaller segments to add turn restrictions; instead of making a new version, the editing software deleted the original road and created two new objects in its place. This made it look like a paid editor was coming in and intentionally deleting, then replacing existing work.

activity inhibit a localized community of mappers from developing? Or does the addition of such data make the map more valuable and therefore cared for and maintained in that region?

10.3.2 Validation Behaviors

When a geometry change that results in a new minor version occurs, what precisely gets changed? Empirical observations suggest that many minor versions of buildings involve *squaring the corners*. This is most typically done by a more experienced mapper adjusting the corners of a building originally created by a less experienced mapper. This is most common in response to a humanitarian mapping task where an abundance of new mappers add many buildings to the map in accordance with the task's objective, yet are unfamiliar with the community norms of square corners on buildings.

These activities, however, have yet to be fully quantified. The frustrations often expressed on the mailing list make it sound like these non-square buildings are an epidemic across the map. Therefore, actually quantifying these validation/corrective edits will lend some clarity to the discussion.

Furthermore, these types of validation/corrective edits indicate work being done by experienced mappers that is likely going hidden in the minor-version of the object (and therefore not present on the way element itself). The presence of experienced mappers in these areas has intrinsic quality implications. For example, if there are five buildings and a mapper modifies the geometry of only two of these buildings to "square them up," then we may surmise two things: First, this mapper is likely a more experienced mapper than the previous, because they have additional knowledge of community norms regarding buildings, and they know how to create square corners on buildings with the map editor (knowledge of the specific menu option or keyboard-shortcut). And Second, data validation has now occurred for all five of these buildings. Because this mapper only corrected two of them, they are likely implying that the other three buildings are correct; otherwise, this mapper would have adjusted their geometries as well. While the metadata of these other buildings will not reflect this validation because nothing was changed, the nearby (corrective) edit implies

this validation has occurred.

This extends especially to *professional editing*. As Chapter 9 mentions briefly, corporate editing teams often have more regimented, supervised editing practices with specific data validation steps. If the metadata shows that a corporate editing team was active in a given area, can a general level of data validation be assumed to have occurred? While these editors may not have touched every object, does their editing presence in the area offer some level of quality guarantee? Broadly I consider all of these editing patterns to fall into the category of *validation behaviors*, and am curious if these can be generalized to create new intrinsic quality indexes for the map.

10.3.3 Scaling to Real-Time

The ability to provide closer to real-time analysis of mapping activity is another goal for future work to better provide the community with the most relevant and actionable analysis. As Section 3.10 discussed, there is a real need and community desire for real-time analysis, especially in disaster-mapping activations. This work, however, focused on scaling in the opposite direction, by first incorporating the complete history of the map at the global scale. Instead, the currently implemented real-time systems such as the missing maps leaderboard supported by *OSMesa* or *osm-analytics.org* (updated daily, tile based analysis) were developed by the larger OSM community and represent two different approaches to real-time analysis. With the successful implementation of full-historical tile-based analysis, there is now opportunity to revisit the implications of these approaches on real-time analysis, allowing easier quantification of what is changing on the map at any given moment with the additional context of the contributor-centric history of the region. For example, knowing not only if an editor is active in a region, but knowing who was active before them, their expertise, and the types of editing they were doing. This has major implications for better contributor-feedback, validation, and vandalism detection. A number of validation and vandalism-detection bots currently offer some of this as automated services within a day of the activity, but they are still very rudimentary, and do not embrace the context of the contributions or history.

10.4 Conclusion

This dissertation has traced the development and implementation of contributor-centric analysis approaches across four iterations of OSM data analysis systems. This began with an infrastructure to perform historical analysis of specific disaster mapping events in support of crisis informatics research, and culminates in parallelized planet-scale analysis of the complete evolution of the map.

The first system, *EpicOSM*, was built to answer questions of user collaboration in OSM, requiring us to design new infrastructure that prioritized the metadata about the changes to the map over the map data itself. This includes not just looking at the current state of the data, but also including the complete editing history of the map up to that point. The development of new data processing systems also prompted the development of a new visualization and presentation tool known as *osmdown*. This tool created easy-to-share results in the form of interactive maps and graphs. This system was successfully implemented in support of two research papers and the tracking of one real-time disaster mapping event.

It quickly became evident, however, that with the rapid growth of the OSM community, the sheer quantity of data produced in disaster mapping activations required new systems that could scale to handle future events with more mappers producing significantly more data. Moreover, there are no geographic bounds on where contributors edit: An individual's contributions can be worldwide. Systems that analyze only the current state of the map can look at specific regions in isolation, but when asking questions about editing patterns and interactions among contributors, analysis systems need to account for an individual's edits across the entire globe to capture the full context. The implementation of *EpicOSM* as a real-time analysis system was useful in tracking the mapping response, but added new constraints and responsibilities to the system that were ultimately deemed out-of-scope as the types of questions we hoped to answer look at the evolution of the map over time.

The transition to vector tile based analysis embraced the latest in planet-scale OSM data

processing systems. This also involved transitioning away from the standard OSM data model to the simpler GeoJSON object representation, a more effective and communicable unit of analysis when classifying and measuring edits to individual objects. By extending an existing parallel-processing analysis workflow to better incorporate the editing metadata, I was able to perform planet-scale analysis of the evolution of the map at an annual resolution. Computing individual contributor histories in this way provided the full editing-history context of each contributor, allowing for more informed approaches to intrinsic quality analysis.

With sustained increase in the amount of editing to the map, annual resolution was deemed insufficient to adequately represent the editing history as too many edits that occurred between the years were not being counted. This led to the third iteration, vector tile based analysis based on quarterly snapshots of the map. This optimized an increase in resolution with the increase in processing time by a factor of four, and identified another 8M annual contributions. These datasets supported global analysis of the impact that corporate editors are having on the map, identifying the location and type of over 17M edits made by paid editors. To investigate the localized impact that these types of edits have on the map, however, requires more than quarterly resolution, prompting the final iteration: full-history tile based analysis.

The development of full-history vector tiles of OSM data for large-scale analysis of the evolution of the map at the edit-level interaction between contributors marks the fourth and final iteration of this work. This required the creation and implementation of new data processing tools to capture the complete editing history as well as new data schemas that can efficiently and accurately capture the evolution of the map at the individual object level, allowing analysts to break down and classify the specific change to each object, when, where, and by whom.

10.4.1 Contributions

Contributor-centric approaches to OSM data analysis are the result of researching disaster mapping in OSM from a background in HCI rather than Geography or GIS. This involved first considering OSM as a site of online collaboration and second as a map. These HCI sensibilities

prioritize the people responsible for the data before the data itself, yet never one without the other. This involves a re-orientation of previous analysis methods to first consider the metadata that reveals who made what changes, when, and where. From there, we can extend existing intrinsic data analysis methods such as those presented in Chapter 7 to better incorporate the data provenance to enrich these methods with more context. For example, learning not only how many contributors were active in an area or how recent the data is, but instead who those contributors are (professional or hobbyists), their mapping expertise, and what the actual change to the map was. From this, we can learn if the edit was the creation of new data, the addition of localized knowledge, a correction or validation, or even all of these. I consider this ability to extend and improve existing data analysis methods as the methodological contribution of this work.

To implement these methodologies, however, I first had to develop new and extend existing analysis infrastructure, beginning at the bottom of the stack with the data representation to embrace the data provenance. Additionally, the creation of OSM-QA-Tiles-Plus allows edits to abstract or invisible objects to be accounted for. These tilesets are a path forward to enabling researchers to perform more types of intrinsic quality analysis at the planet-scale by making visible the hidden and more complex editing activity. The development of `EpicOSM`, `osmdown`, `quarterly-osm-qa-tiles`, `osm-qa-tiles-plus`, `osm-wayback`, and `stream-reduce` represent open-source technical contributions of this work, as does the the implementation of two new historical data schemas and innovations to the existing `tile-reduce` processing workflows.

10.4.2 Final Remarks

OSM data analysis will always be a moving target. Unraveling the history of the map and extracting meaningful insights will become more complex as both the communities within OSM and the way they edit the map continue to evolve. As this happens, our approaches and analysis infrastructures must also evolve. The next major phase of OSM editing, machine-assisted mapping, is about to be upon us. Machine-assisted mapping will use machine learning to aid mappers in object identification. This will create yet another distinct signature in the data that future systems

will need to learn to interpret to better contextualize the editing activity, as Chapter 7 identified such a need for in disaster mapping. Additionally, paid editing as shown in Chapter 9 will continue to be a dominant force in both mapping and data validation. Each of these editing behaviors has or will have a profound impact on the data that that future analysis systems need to learn to interpret. Contributor-centric approaches can help with this contextualization by enabling data analysts to see the activity at the individual edit level and reveal interactions between mappers.

Chapter 9, and then reiterated in the Section 10.3, outline a research agenda moving forward for contributor-centric OSM data analysis. Only through this work has the importance of extracting the *who* from the contributions become so obvious to me as it enables us to distinguish between the many OSM communities at the level of the individual edit. Only once we see this distinction in the editing history can we know what types of editing behaviors we can identify and then learn what the pertinent questions are. Digging deeper into the interactions between paid and volunteer editors and investigating these interaction patterns as forms of data validation or map seeding will shed some of the first light onto the impact that paid editing is really having on the map and the community. Additionally, I will continue to develop front-end data science environments powered by these rich editing histories to lower the barrier to entry for all OSM researchers, so that the first step of OSM data analysis no longer needs to be about convoluted data-wrangling.

Continuing to make the activity underneath the map visible should be a priority for all OSM researchers, as this historical record of the project traces not only the map, but the community as well. Building systems that elucidate *who* is changing the map and *how* ensures that analysts can continue to ask any range of questions about the evolution of both the map and the community. Contributor-centric approaches to OSM data analysis help grow the intersection between the questions that we *want* to ask of the data behind the map and the questions that we actually *can*.

Bibliography

- [1] Maik Anderka and Benno Stein. “A breakdown of quality flaws in Wikipedia”. In: **Proceedings of the 2nd Joint WICOW/AIRWeb Workshop on Web Quality**. New York, New York, USA: ACM, 2012, pp. 11–18. DOI: 10.1145/2184305.2184309.
- [2] Erich Andersen. **Microsoft joins Open Invention Network to help protect Linux and open source**. 2018. URL: <https://azure.microsoft.com/en-us/blog/microsoft-joins-open-invention-network-to-help-protect-linux-and-open-source/> (visited on 03/11/2019).
- [3] Jennings Anderson, Dipto Sarkar, and Leysia Palen. “Corporate Editors in the Evolving Landscape of OpenStreetMap”. In: **ISPRS International Journal of Geo-Information** 8.5 (May 2019), p. 232. DOI: 10.3390/ijgi8050232.
- [4] Jennings Anderson, Robert Soden, Kenneth M. Anderson, Marina Kogan, and Leysia Palen. “EPIC-OSM: A software framework for OpenStreetMap data analytics”. In: **Proceedings of the Annual Hawaii International Conference on System Sciences**. Vol. 2016-March. 2016, pp. 5468–5477. DOI: 10.1109/HICSS.2016.675.
- [5] Jennings Anderson, Robert Soden, Brian Keegan, Leysia Palen, and Kenneth M. Anderson. “The Crowd is the Territory: Assessing Quality in Peer-Produced Spatial Data During Disasters”. In: **International Journal of Human-Computer Interaction** 34.4 (Apr. 2018), pp. 295–310. DOI: 10.1080/10447318.2018.1427828.

- [6] Kenneth M Anderson. “Embrace the Challenges: Software Engineering in a Big Data World”. In: **Proceedings - 1st International Workshop on Big Data Software Engineering, BIGDSE 2015**. 2015, pp. 19–25. DOI: 10.1109/BIGDSE.2015.12.
- [7] Jamal Jokar Arsanjani, Christopher Barron, Mohammed Bakillah, and Marco Helbich. “Assessing the Quality of OpenStreetMap Contributors together with their Contributions”. In: **16th AGILE International Conference on Geographic Information Science**. (2013). DOI: 10.1080/14498596.2014.927337.
- [8] Andrea Ballatore. “Defacing the map: Cartographic vandalism in the digital commons”. In: **Cartographic Journal** 7041.May (2014). DOI: 10.1179/1743277414Y.0000000085. arXiv: 1404.3341.
- [9] Mario Barrenechea, Kenneth M. Anderson, Leysia Palen, and Joanne White. “Engineering crowdwork for disaster events: The human-centered development of a lost-and-found tasking environment”. In: **Proceedings of the Annual Hawaii International Conference on System Sciences**. 2015. DOI: 10.1109/HICSS.2015.31.
- [10] Christopher Barron, Pascal Neis, and Alexander Zipf. “A Comprehensive Framework for Intrinsic OpenStreetMap Quality Analysis”. In: **Transactions in GIS** 18.6 (Dec. 2014), pp. 877–895. DOI: 10.1111/tgis.12073. arXiv: 9605103 [cs].
- [11] Daniel Bégin, Rodolphe Devillers, and Stéphane Roche. “The life cycle of contributors in collaborative online communities -the case of OpenStreetMap”. In: **International Journal of Geographical Information Science** 32.8 (2018), pp. 1611–1630. DOI: 10.1080/13658816.2018.1458312.
- [12] Yochai. Benkler. **The wealth of networks: How social production transforms markets and freedom**. Yale University Press, 2006.
- [13] Melissa Bica, Julie L. Demuth, James E. Dykes, and Leysia Palen. “Communicating Hurricane Risks”. In: **Proceedings of the 2019 CHI Conference on Human Factors in**

- Computing Systems - CHI '19**. New York, New York, USA: ACM Press, 2019, pp. 1–13. DOI: 10.1145/3290605.3300545.
- [14] Christian Bittner. “OpenStreetMap in Israel and Palestine – ‘Game changer’ or reproducer of contested cartographies?” In: **Political Geography** (2017). DOI: 10.1016/j.polgeo.2016.11.010.
- [15] Joshua E Blumenstock. “Size matters: Word count as a measure of quality on Wikipedia”. In: **WWW**. 2008.
- [16] M. Bostock, V. Ogievetsky, and J. Heer. “D3 Data-Driven Documents”. In: **IEEE Transactions on Visualization and Computer Graphics** 17.12 (2011), pp. 2301–2309. DOI: 10.1109/TVCG.2011.185.
- [17] Adam R. Brown. “Wikipedia as a Data Source for Political Scientists: Accuracy and Completeness of Coverage”. In: **PS - Political Science and Politics** (2011). DOI: 10.1017/S1049096511000199.
- [18] Nama R. Budhathoki and Caroline Haythornthwaite. “Motivation for Open Collaboration: Crowd and Community Models and the Case of OpenStreetMap”. In: **American Behavioral Scientist** 57.5 (Dec. 2013), pp. 548–575. DOI: 10.1177/0002764212469364.
- [19] Sébastien Caquard. “Cartography II: Collective cartographies in the social media era”. In: **Progress in Human Geography** 38.1 (2014), pp. 141–150. DOI: 10.1177/0309132513514005.
- [20] Valentina Carraro and Bart Wissink. “Participation and marginality on the geoweb: The politics of non-mapping on OpenStreetMap Jerusalem”. In: **Geoforum** (2018). DOI: 10.1016/j.geoforum.2018.02.001.
- [21] Steve Chilton. “Crowdsourcing Is Radically Changing the Geodata Landscape : Case Study of Openstreetmap”. In: **24th International Cartographic Conference, Chile**. 2009.

- [22] Błażej Ciepluch, Peter Mooney, and Adam C Winstanley. “Building Generic Quality Indicators for OpenStreetMap”. In: **19th annual GIS Research UK (GISRUK)** (2011), p. 5.
- [23] Blazej Cipeluch, Ricky Jacob, Adam Winstanley, Peter Mooney, Blazej Ciepluch, Ricky Jacob, Adam Winstanley, and Peter Mooney. “Comparison of the accuracy of OpenStreetMap for Ireland with Google Maps and Bing Maps”. In: **Ninth International Symposium on Spatial Accuracy Assessment in Natural Resuorces and Enviromental Sciences** (2010), p. 337.
- [24] Steve Coast. **The Book of OSM**. CreateSpace Independent Publishing Platform, 2015.
- [25] David J Coleman, Yola Georgiadou, Jeff Labonte, Earth Observation, and Natural Resources Canada. “Volunteered Geographic Information : the nature and motivation of producers”. In: **GeoInformation Science** 4.Special Issue GSDI-11 (2009), p. 20. DOI: 10.2902/1725-0463.2009.04.art16.
- [26] Pdraig Corcoran, Peter Mooney, and Michela Bertolotto. “Analysing the growth of OpenStreetMap networks”. In: **Spatial Statistics** 3 (Feb. 2013), pp. 21–32. DOI: 10.1016/j.spasta.2013.01.002.
- [27] Michael Crutcher and Matthew Zook. “Placemarks and waterlines: Racialized cyberscapes in post-Katrina Google Earth”. In: **Geoforum** 40.4 (2009), pp. 523–534. DOI: 10.1016/j.geoforum.2009.01.003.
- [28] Jeffrey Dean and Sanjay Ghemawat. “MapReduce: Simplied Data Processing on Large Clusters”. In: **Proceedings of the Symposium on Operating Systems Design and Implementation** (2004).
- [29] Martin Dittus, Giovanni Quattrone, and Licia Capra. “Analysing Volunteer Engagement in Humanitarian Mapping: Building Contributor Communities at Large Scale”. In: **Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work &**

- Social Computing - CSCW '16**. ACM, Feb. 2016, pp. 108–118. DOI: 10.1145/2818048.2819939.
- [30] Martin Dittus, Giovanni Quattrone, and Licia Capra. “Mass Participation During Emergency Response: Event-centric Crowdsourcing in Humanitarian Mapping”. In: **Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing**. 2017. DOI: 10.1145/2998181.2998216.
- [31] Line Dubé, Anne Bourhis, and Réal Jacob. “Towards a typology of virtual communities of practice”. In: **Interdisciplinary Journal of Information, Knowledge, and Management**. Vol. 1. New York, New York, USA: ACM Press, Feb. 2006, pp. 69–93. DOI: 10.1145/2441776.2441845.
- [32] Melanie Eckle and João Porto de Albuquerque. “Quality Assessment of Remote Mapping in OpenStreetMap for Disaster Management Purposes”. In: **Proceedings of the ISCRAM 2015 Conference - Kristiansand, May 24-27**. Kristiansand, 2015.
- [33] Umberto Eco. **How to Travel with a Salmon & Other Essays**. Houghton Mifflin Harcourt, 1995, pp. 95–106.
- [34] Sarah Elwood. “Geographic information science: Emerging research on the societal implications of the geospatial web”. In: **Progress in Human Geography** (2010). DOI: 10.1177/0309132509340711.
- [35] Sarah Elwood. “Volunteered geographic information: Future research directions motivated by critical, participatory, and feminist GIS”. In: **GeoJournal** 72.3-4 (July 2008), pp. 173–183. DOI: 10.1007/s10708-008-9186-0. arXiv: arXiv:1002.2562v1.
- [36] Sarah Elwood, Michael F Goodchild, and Daniel Z Sui. “Researching Volunteered Geographic Information: Spatial Data, Geographic Research, and New Social Practice”. In: **Annals of the Association of American Geographers** 102.3 (2012), pp. 571–590. DOI: 10.1080/00045608.2011.595657.

- [37] OpenStreetMap Foundation. **Organised Editing Guidelines**. 2018. URL: https://wiki.osmfoundation.org/wiki/Organised_Editing_Guidelines (visited on 03/21/2018).
- [38] R. Stuart Geiger and Aaron Halfaker. “When the levee breaks: Without bots, what happens to Wikipedia’s quality control processes?” In: **Proceedings of the 9th International Symposium on Open Collaboration - WikiSym ’13** (2013). DOI: 10.1145/2491055.2491061.
- [39] Melissa Gilbert. “THEORIZING DIGITAL AND URBAN INEQUALITIES”. In: **Information, Communication & Society** 13.7 (2010), pp. 1000–1018. DOI: 10.1080/1369118x.2010.499954.
- [40] Jim Giles. “Internet encyclopaedias go head to head”. In: **Nature** (2005). DOI: 10.1038/438900a.
- [41] Jean François Girres and Guillaume Touya. “Quality Assessment of the French OpenStreetMap Dataset”. In: **Transactions in GIS** 14.4 (2010), pp. 435–459. DOI: 10.1111/j.1467-9671.2010.01203.x.
- [42] Michael F. Goodchild. “Citizens as sensors: The world of volunteered geography”. In: **GeoJournal** 69.4 (2007), pp. 211–221. DOI: 10.1007/s10708-007-9111-y. arXiv: arXiv:1404.3341.
- [43] Mark Graham, Bernie Hogan, Ralph K Straumann, and Ahmed Medhat. “Uneven Geographies of User-Generated Information: Patterns of Increasing Informational Poverty”. In: **Annals of the Association of American Geographers** 104.4 (2014), pp. 746–764. DOI: 10.1080/00045608.2014.910087.
- [44] Mark Graham, Matthew Zook, and Andrew Boulton. “Augmented reality in urban places: Contested content and the duplicity of code”. In: **Transactions of the Institute of British Geographers** 38.3 (2013), pp. 464–479. DOI: 10.1111/j.1475-5661.2012.00539.x.

- [45] Anita Graser, Markus Straub, and Melitta Dragaschnig. “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”. In: **Transactions in GIS** (2014). DOI: 10.1111/tgis.12061.
- [46] Mordechai Haklay. “How good is volunteered geographical information? A comparative study of OpenStreetMap and ordnance survey datasets”. In: **Environment and Planning B: Planning and Design** 37.4 (July 2010), pp. 682–703. DOI: 10.1068/b35097.
- [47] Mordechai Haklay, Vyron Antoniou, and Sofia Basiouka. **Crowdsourced Geographic Information Use in Government**. Tech. rep. London: GFDRR, 2014, pp. 1–79.
- [48] Mordechai Haklay, Sofia Basiouka, Vyron Antoniou, and Aamer Ather. “How many volunteers does it take to map an area well? The validity of Linus’ law to volunteered geographic information”. In: **The Cartographic Journal** 47.4 (2010), pp. 315–322. DOI: 10.1179/000870410X12911304958827.
- [49] Muki Haklay. **Geography and HCI**. Tech. rep. University College London, 2013, pp. 1–3.
- [50] Alexander Halavais and Derek Lackaff. “An analysis of topical coverage of Wikipedia”. In: **Journal of Computer-Mediated Communication** (2008). DOI: 10.1111/j.1083-6101.2008.00403.x.
- [51] Aaron Halfaker, R. Stuart Geiger, Jonathan T. Morgan, and John Riedl. “The Rise and Decline of an Open Collaboration System: How Wikipedia’s Reaction to Popularity Is Causing Its Decline”. In: **American Behavioral Scientist** (2013). DOI: 10.1177/0002764212469365.
- [52] Oisín Herriot. **Building up the Microsoft Open Maps Team**. State of the Map 2018. 2018.
- [53] Desislava Hristova, Giovanni Quattrone, Afra Mashhadi, and Licia Capra. “The life of the party: impact of social mapping in OpenStreetMap”. In: **Proc. ICWSM ’13**. 2013, pp. 234–243. DOI: 10.1145/1296951.1296960.

- [54] Mayank Jain. **Uber May Bid Adieu to Google Maps; Move to Open Source Ones Instead**. URL: https://www.business-standard.com/article/companies/uber-to-soon-let-you-crowdsource-your-commute-via-openstreetmap-project-118072501638_1.html (visited on 11/27/2018).
- [55] Peter A. Johnson. “Models of direct editing of government spatial data: challenges and constraints to the acceptance of contributed data”. In: **Cartography and Geographic Information Science** (2017). DOI: 10.1080/15230406.2016.1176536.
- [56] Gerald C. Kane. “A multimethod study of information quality in wiki collaboration”. In: **ACM Transactions on Management Information Systems** (2011). DOI: 10.1145/1929916.1929920.
- [57] Suneel Kaw. **Uber planning to explore and contribute to OpenStreetMap in Delhi**. July 9, 2018. URL: <https://forum.openstreetmap.org/viewtopic.php?id=62986> (visited on 04/19/2019).
- [58] Brian C. Keegan. “Breaking news on wikipedia”. In: **Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work Companion - CSCW '12**. New York, New York, USA: ACM Press, Feb. 2012, p. 315. DOI: 10.1145/2141512.2141609.
- [59] Rob Kitchin and Martin Dodge. “Rethinking maps”. In: **Progress in Human Geography** 31.3 (2007), pp. 331–344. DOI: 10.1177/0309132507077082.
- [60] Aniket Kittur, Bongwon Suh, and Ed H. Chi. “Can you ever trust a wiki?” In: **Proceedings of the ACM 2008 conference on Computer supported cooperative work - CSCW '08** (2008), p. 477. DOI: 10.1145/1460563.1460639.
- [61] Marina Kogan, Jennings Anderson, Leysia Palen, Kenneth M. Anderson, and Robert Soden. “Finding the Way to OSM Mapping Practices”. In: **Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16**. 2016, pp. 2783–2795. DOI: 10.1145/2858036.2858371.

- [62] Alfred Korzybski. **Science and Sanity: An introduction to non-Aristotelian systems and general semantics**. Institute of General Semantics, 1958.
- [63] KrASIA. **Asia Is The World’s Largest Ride-Hailing Market With Over 70% Share - Grab Dominates SEA**. Sept. 7, 2018. (Visited on 02/03/2019).
- [64] David Laniado and Riccardo Tasso. “Co-authorship 2.0: patterns of collaboration in Wikipedia”. In: **Proceedings of the 22nd ACM conference on Hypertext and hypermedia - HT ’11**. 2011. DOI: 10.1145/1995966.1995994.
- [65] Yang W. Lee, Diane M. Strong, Beverly K. Kahn, and Richard Y. Wang. “AIMQ: A methodology for information quality assessment”. In: **Information and Management** 40.2 (2002), pp. 133–146. DOI: 10.1016/S0378-7206(02)00043-5.
- [66] Wen Lin. “Volunteered Geographic Information constructions in a contested terrain: A case of OpenStreetMap in China”. In: **Geoforum** (2018). DOI: 10.1016/j.geoforum.2018.01.005.
- [67] Thomas Ludwig, Christoph Kotthaus, Christian Reuter, Sören van Dongen, and Volkmar Pipek. “Situated crowdsourcing during disasters: Managing the tasks of spontaneous volunteers through public displays”. In: **International Journal of Human Computer Studies** (2017). DOI: 10.1016/j.ijhcs.2016.09.008.
- [68] Thomas Ludwig, Christian Reuter, and Volkmar Pipek. “Social Haystack”. In: **ACM Transactions on Computer-Human Interaction** 22.4 (2015), pp. 1–27. DOI: 10.1145/2749461.
- [69] Brendan Luyt and Daniel Tan. “Improving wikipedia’s credibility: References and citations in a sample of history articles”. In: **Journal of the American Society for Information Science and Technology** (2010). DOI: 10.1002/asi.21304.
- [70] Mikel Maron. **Haiti OpenStreetMap Response**. Jan. 14, 2010. URL: brainoff.com/weblog/2010/01/14/1518 (visited on 05/20/2015).

- [71] Neil McAllister. **Redmond top man Satya Nadella: 'Microsoft LOVES Linux'**. Oct. 20, 2014. URL: https://www.theregister.co.uk/2014/10/20/microsoft_cloud_event/ (visited on 11/21/2018).
- [72] Alan McConchie. **Introducing “map gardening”**. URL: <http://mappingmashups.net/2013/05/25%20/introducing-map-gardening>.
- [73] Alan McConchie. **Map Gardening in Practice: Tracing Patterns of Growth and Maintenance in OpenStreetMap**. Apr. 22, 2015.
- [74] Peter Mooney and Padraig Corcoran. “Analysis of interaction and co-editing patterns amongst openstreetmap contributors”. In: **Transactions in GIS** 18.5 (Oct. 2014), pp. 633–659. DOI: 10.1111/tgis.12051.
- [75] Peter Mooney and Padraig Corcoran. “How social is OpenStreetMap?” In: **Conference on Geographic Information Science**. 2012, pp. 282–287.
- [76] Peter Mooney, Padraig Corcoran, and Adam C. Winstanley. “Towards quality metrics for OpenStreetMap”. In: **Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems - GIS '10** (2010), p. 514. DOI: 10.1145/1869790.1869875.
- [77] Peter Mooney and Marco Minghini. “A Review of OpenStreetMap Data”. In: **Mapping and the Citizen Sensor**. Ed. by G Foody, L See, S Fritz, Peter Mooney, A-M Olteanu-Raimond, C C Fonte, and Vyron Antoniou. London: Ubiquity Press, 2017. Chap. 3, pp. 37–59. DOI: 10.5334/bbf.c.
- [78] Mordechai (Muki) Haklay. “Neogeography and the Delusion of Democratisation”. In: **Environment and Planning A** (2013). DOI: 10.1068/a45184.
- [79] Mishari Muqbil. **Grab is using OSM data**. May 9, 2018. URL: <https://archive.mishari.net/en/2018/12/grab-osm-data/> (visited on 02/03/2019).

- [80] Mishari Muqbil. **Uber most likely using OSM data**. July 26, 2016. URL: <https://archive.mishari.net/th/2016/07/uber-using-osm-data/> (visited on 02/03/2019).
- [81] Pascal Neis, Marcus Goetz, and Alexander Zipf. “Towards Automatic Vandalism Detection in OpenStreetMap”. In: **ISPRS International Journal of Geo-Information** 1.3 (2012), pp. 315–332. DOI: 10.3390/ijgi1030315.
- [82] Pascal Neis and Dennis Zielstra. “Recent Developments and Future Trends in Volunteered Geographic Information Research: The Case of OpenStreetMap”. en. In: **Future Internet** 6.1 (Jan. 2014), pp. 76–106. DOI: 10.3390/fi6010076.
- [83] Pascal Neis, Dennis Zielstra, and Alexander Zipf. “The Street Network Evolution of Crowdsourced Maps: OpenStreetMap in Germany 2007–2011”. In: **Future Internet** (2011). DOI: 10.3390/fi4010001.
- [84] Pascal Neis and Alexander Zipf. “Analyzing the Contributor Activity of a Volunteered Geographic Information Project — The Case of OpenStreetMap”. In: **ISPRS International Journal of Geo-Information** 1.3 (2012), pp. 146–165. DOI: 10.3390/ijgi1020146.
- [85] Jakob Nielsen. **The 90-9-1 Rule for Participation Inequality in Social Media and Online Communities**. 2006. URL: <http://www.nngroup.com/articles/participation-inequality/> (visited on 04/16/2019).
- [86] **OpenStreetMap Statistics**. URL: http://www.openstreetmap.org/stats/data_stats.html (visited on 11/21/2018).
- [87] **OpenStreetMap Tag Info**. URL: <https://taginfo.openstreetmap.org> (visited on 06/10/2019).
- [88] **OpenStreetMap Statistics**. URL: http://www.openstreetmap.org/stats/data_stats.html (visited on 06/14/2015).
- [89] **Amazon**. URL: <https://wiki.openstreetmap.org/wiki/Amazon> (visited on 02/04/2019).

- [90] **CloudMade**. URL: <https://wiki.openstreetmap.org/wiki/CloudMade> (visited on 04/07/2019).
- [91] **History of OpenStreetMap**. URL: https://wiki.openstreetmap.org/wiki/History_of_OpenStreetMap (visited on 11/21/2018).
- [92] **Using Imagery**. URL: https://wiki.openstreetmap.org/wiki/Using_Imagery (visited on 11/21/2018).
- [93] **Aerial Imagery**. URL: https://wiki.openstreetmap.org/wiki/Aerial_imagery (visited on 11/21/2018).
- [94] **Importing Government Data**. URL: https://wiki.openstreetmap.org/wiki/Importing_Government_Data (visited on 11/21/2018).
- [95] **Import/Catalogue**. URL: <https://wiki.openstreetmap.org/wiki/Import/Catalogue> (visited on 11/21/2018).
- [96] **Potential Datasources**. URL: https://wiki.openstreetmap.org/wiki/Potential_Datasources (visited on 11/21/2018).
- [97] **AI Assisted Road Tracing**. URL: https://wiki.openstreetmap.org/wiki/AI-Assisted_Road_Tracing (visited on 04/18/2018).
- [98] **Kerala Road Import**. URL: https://wiki.openstreetmap.org/wiki/Kerala_Road_Import (visited on 04/18/2018).
- [99] **Kaart**. URL: <https://wiki.openstreetmap.org/wiki/Kaart> (visited on 11/30/2018).
- [100] **Grab**. URL: <https://wiki.openstreetmap.org/wiki/Grab> (visited on 11/30/2018).
- [101] **Results of Organised Editing Survey 2017**. URL: https://wiki.osmfoundation.org/wiki/Data_Working_Group/Results_of_Organised_Editing_Survey_2017 (visited on 04/18/2019).
- [102] **About OpenStreetMap Foundation**. URL: <https://wiki.osmfoundation.org/wiki/About> (visited on 05/07/2019).

- [103] YouthMappers Organization. **YouthMappers**. URL: <https://www.youthmappers.org/> (visited on 06/01/2019).
- [104] Leysia Palen and Sophia B Liu. “Citizen communications in crisis: Anticipating a future of ICT-supported public participation”. In: **Natural Hazards**. 2007, pp. 727–736. DOI: 10.1145/1240624.1240736.
- [105] Leysia Palen, Robert Soden, Jennings Anderson, and Mario Barrenechea. “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 - CHI '15**. New York, New York, USA: ACM Press, Apr. 2015, pp. 4113–4122. DOI: 10.1145/2702123.2702294.
- [106] Jeffrey Parsons and Roman Lukyanenko. “Contributor-centric Information Quality for Crowdsourcing”. In: **Workshop on Openness and Transparency Research (ICIS 2013)**. Milan, Italy, 2013. DOI: 10.13140/2.1.2348.6248.
- [107] Paulo Perrotta. **Metaprogramming Ruby 2**. The Pragmatic Programmers, LLC, 2014, p. 262.
- [108] Thiago Henrique Poiani, Roberto Dos Santos Rocha, Livia Castro Degrossi, and Joao Porto De Albuquerque. “Potential of Collaborative Mapping for Disaster Relief: A Case Study of OpenStreetMap in the Nepal Earthquake 2015”. In: **2016 49th Hawaii International Conference on System Sciences (HICSS)**. Vol. 2016-March. IEEE, Jan. 2016, pp. 188–197. DOI: 10.1109/HICSS.2016.31.
- [109] Amir Pourabdollah, Jeremy Morley, Steven Feldman, and Mike Jackson. “Towards an Authoritative OpenStreetMap: Conflating OSM and OS OpenData National Maps’ Road Network”. In: **ISPRS International Journal of Geo-Information** 2.3 (2013), pp. 704–728. DOI: 10.3390/ijgi2030704.
- [110] Sterling Quinn. “Using small cities to understand the crowd behind OpenStreetMap”. In: **GeoJournal** 82.3 (2017), pp. 455–473. DOI: 10.1007/s10708-015-9695-6.

- [111] Sterling D. Quinn and Doran A. Tucker. “How geopolitical conflict shapes the mass-produced online map”. In: **First Monday** 22.11 (2017). DOI: 10.5210/fm.v22i111.7922.
- [112] Martin Raifer, Rafael Troilo, Fabian Kowatsch, Michael Auer, Lukas Loos, Sabrina Marx, Katharina Przybill, Sascha Fendrich, Franz-Benjamin Mocnik, and Alexander Zipf. “OS-HDB: a framework for spatio-temporal analysis of OpenStreetMap history data”. In: **Open Geospatial Data, Software and Standards** 4.1 (Dec. 2019), p. 3. DOI: 10.1186/s40965-019-0061-3.
- [113] Lucy Holman Rector. “Comparison of Wikipedia and other encyclopedias for accuracy, breadth, and depth in historical articles”. In: **Reference Services Review** (2008). DOI: 10.1108/00907320810851998.
- [114] Christian Reuter, Marc-André Kaufhold, and Thomas Ludwig. “End-User Development and Social Big Data – Towards Tailorable Situation Assessment with Social Media”. In: **New Perspectives in End-User Development**. Cham: Springer International Publishing, 2017, pp. 307–332. DOI: 10.1007/978-3-319-60291-2_12.
- [115] Cindy Royal and Deepina Kapila. “What’s on wikipedia, and what’s not... ?: Assessing completeness of information”. In: **Social Science Computer Review** (2009). DOI: 10.1177/0894439308321890.
- [116] Anna Samoilenko and Taha Yasseri. “The distorted mirror of wikipedia: A quantitative analysis of wikipedia coverage of academics”. In: **EPJ Data Science** (2014). DOI: 10.1140/epjds20.
- [117] Manuela Schmidt and Silvia Klettner. “Gender and Experience-Related Motivators for Contributing to OpenStreetMap”. In: **International workshop on action and interaction in volunteered geographic information** (2013).
- [118] Aaron Schram and Kenneth M. Anderson. “MySQL to NoSQL: Data Modeling Challenges in Supporting Scalability”. In: **Proceedings of the 3rd annual conference on Systems,**

- programming, and applications: software for humanity - SPLASH '12.** Tucson, Arizona: ACM, 2012. DOI: 10.1145/2384716.2384773.
- [119] Ian Schuler. **Announcing DevSeed Data.** Dec. 21, 2017. URL: <https://medium.com/devseed/announcing-devseed-data-1a3d8102cb23> (visited on 03/21/2019).
- [120] Sukhjit Singh Sehra, Jaiteg Singh, and Hardeep Singh Rai. "Assessing openstreetmap data using intrinsic quality indicators: An extension to the QGIS processing toolbox". In: **Future Internet** 9.2 (2017), p. 15. DOI: 10.3390/fi9020015.
- [121] Renée E. Sieber and Mordechai Haklay. "The epistemology(s) of volunteered geographic information: a critique". In: **Geo: Geography and Environment** (2015). DOI: 10.1002/geo2.10.
- [122] Biondi Sanda Sima. **GrabBike Drivers Support Disaster Managers Identify IDP Camps in Bali.** Oct. 24, 2017. URL: https://www.hotosm.org/updates/2017-10-24-grabbike_drivers_support_disaster_managers_identify_idp_camps_in_bali (visited on 11/27/2018).
- [123] R Soden, N Budhathoki, and L Palen. "Resilience-building and the crisis informatics agenda: Lessons learned from open cities Kathmandu". In: **ISCRAM 2014 Conference Proceedings - 11th International Conference on Information Systems for Crisis Response and Management.** 2014.
- [124] Robert Soden. **4 Years On, Looking Back at OpenStreetMap Response to the Haiti Earthquake.** URL: <http://blogs.worldbank.org/latinamerica/4-years-looking-back-openstreetmap-response-haiti-earthquake> (visited on 11/21/2018).
- [125] Robert Soden and Leysia Palen. "From Crowdsourced Mapping to Community Mapping: The Post-earthquake Work of OpenStreetMap Haiti". In: **Conference on the Design of Cooperative Systems.** 2014, pp. 311–326. DOI: 10.1007/978-3-319-06498-7_19.

- [126] Robert Soden and Leysia Palen. “Infrastructure in the Wild: What Mapping in Post-Earthquake Nepal Reveals About Infrastructural Emergence”. In: **Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16** (2016). DOI: 10.1145/2858036.2858545.
- [127] Patricia Solis. **YouthMappers Presentation**.
- [128] Monica Stephens. “Gender and the GeoWeb: Divisions in the production of user-generated cartographic information”. In: **GeoJournal** (2013). DOI: 10.1007/s10708-013-9492-z.
- [129] Besiki Stvilia, Les Gasser, Michael B. Twidale, and Linda C. Smith. “A Framework for Information Quality Assessment”. In: **Journal of the American Society for Information Science and Technology** 58 (2007), pp. 1720–1733. DOI: 10.1002/asi.20652.
- [130] Besiki Stvilia, Michael B. Twidale, Linda C. Smith, and Les Gasser. “Information Quality Work Organization in Wikipedia”. In: **Journal of the Association for Information Science and Technology** 59.6 (2008), pp. 983–1001. DOI: 10.1002/asi.v59:6.
- [131] Daniel Sui, Sarah Elwood, and Michael Goodchild. “Crowdsourcing geographic Knowledge: Volunteered geographic information (VGI) in theory and practice”. In: **Crowdsourcing Geographic Knowledge: Volunteered Geographic Information (VGI) in Theory and Practice** 9789400745 (2013). Ed. by Daniel Sui and Michael F. Goodchild, pp. 1–396. DOI: 10.1007/978-94-007-4587-2. arXiv: arXiv:1011.1669v3.
- [132] Bing Maps Team. **Microsoft Releases 125 million Building Footprints in the US as Open Data**. June 28, 2018. URL: <https://blogs.bing.com/maps/2018-06/microsoft-releases-125-million-building-footprints-in-the-us-as-open-data> (visited on 11/26/2018).
- [133] Jochen Topf. **Modding the OSM Data Model**. State of the Map 2018.

- [134] W Ben Towne, Aniket Kittur, Peter Kinnaird, and James Herbsleb. “Your process is showing: Controversy management and perceived quality in Wikipedia”. In: **Proceedings of CSCW ’13**. 2013.
- [135] Arnaud Vandecasteele and Rodolphe Devillers. “Improving volunteered geographic information quality using a tag recommender system: The case of OpenStreetMap”. In: **Lecture Notes in Geoinformation and Cartography** (2015). DOI: 10.1007/978-3-319-14280-7_4.
- [136] Pimlada Veerapongwatta. **Grab work with the OSM Thailand community (Mapathon Kick-off in BKK)**. May 9, 2018. URL: <https://forum.openstreetmap.org/viewtopic.php?id=62282> (visited on 11/28/2018).
- [137] Yair Wand and Richard Y. Wang. “Anchoring data quality dimensions in ontological foundations”. In: **Communications of the ACM** (1996). DOI: 10.1145/240455.240479.
- [138] Barney Warf and Daniel Sui. **From GIS to neogeography: Ontological implications and theories of truth**. 2010. DOI: 10.1080/19475683.2010.539985.
- [139] Morten Warncke-Wang, Vivek Ranjan, Loren Terveen, and Brent Hecht. “Misalignment Between Supply and Demand of Quality Content in Peer Production Communities”. In: **ICWSM**. 2015.
- [140] Harry Wood. **1 million map contributors!** URL: <https://blog.openstreetmap.org/2018/03/18/1-million-map-contributors/> (visited on 04/15/2019).
- [141] Eti Yaari, Shifra Baruchson-Arbib, and Judit Bar-Ilan. “Information quality assessment of community generated content: A user study of Wikipedia”. In: **Journal of Information Science** (2011). DOI: 10.1177/0165551511416065.
- [142] Hongyu Zhang and Jacek Malczewski. “Accuracy Evaluation of the Canadian OpenStreetMap Road Networks”. In: **International Journal of Geospatial and Environmental Research** 5.2 (2018).

- [143] Dennis Zielstra, Hartwig H. Hochmair, and Pascal Neis. “Assessing the effect of data imports on the completeness of openstreetmap - A United States case study”. In: **Transactions in GIS** 17.3 (2013), pp. 315–334. DOI: 10.1111/tgis.12037. arXiv: 9780201398298.
- [144] Dennis Zielstra and Alexander Zipf. “Quantitative studies on the data quality of OpenStreetMap in Germany”. In: **Proceedings of GIScience** (2010).
- [145] Matthew Zook, Mark Graham, Taylor Shelton, and Sean Gorman. “Volunteered Geographic Information and Crowdsourcing Disaster Relief: A Case Study of the Haitian Earthquake”. In: **World Medical & Health Policy** 2.2 (Jan. 2010), pp. 6–32. DOI: 10.2202/1948-4682.1069.

Appendix A

Glossary

The following is a brief description of a number of terms used throughout the document. Terms introduced with an asterisk in the main document can be found defined here.

GeoJSON: The standard specification for representing Geo-Spatial data in JSON form. Objects are represented as individual features. Here is an example of a point at coordinates 0,0 with two attributes:

```
{"type": "Feature",  
  "geometry": { "type": "Point", "coordinates": [0,0] },  
  "properties": { "hello": "world", "value": 100 } }
```

Jupyter Notebook: Browser-based Interactive Computational Notebook. jupyter.org

mapbox-gl: Javascript library built on WebGL that can efficiently render vector tiles according to a specific, user-defined style, which can include data-driven styling and filtering.

mbtiles: SQLITE database of vector tiles in the mapbox vector tile format.

minor version: An intermediate version of a way or relation that is not recorded on the object itself. It is the product of a change to a child object (typically a node) that inherently changes the parent object but the change is only recorded on the child object. Minor versions are identified through calculating the position of all child elements of an object throughout the history of the object. See Section 2.2

OHSOME / OSHDB: OpenStreetMap History Data Analytics Platform / OpenStreetMap History Database. An OSM full-history database with analytics platform and API built and maintained by Heidelberg Institute for GeoInformation Technology [112]. heigit.org

OSMesa: Large-Scale OSM data processing optimized for cloud-distribution. Capable of ingesting OSM-history files and computing full object histories. github.com/azavea/osmesa

osmium: Developer tools for working with OSM data. The `osmium-tool` offers `osmium export` and `osmium tags-filter`, while *libosmium* is a C++ library for efficient OSM data processing. osmcode.org/

Overpass API: Publicly available at overpass-turbo.eu, the Overpass API offers a fully featured query language for extracting data from OSM. It is a primary data source for many analysts pulling specific data types from OSM.

tile-reduce: An open-source javascript library owned by Mapbox that reads a `mbtiles` file and performs a map-reduce job by distributing individual tiles to worker threads that execute specific user-defined javascript code against the objects in each vector tile.

tippecanoe: A command-line utility to turn GeoJSON into vector-tiles. Written in C++, this open-source utility is owned by Mapbox and maintained by employee Eric Fischer. It is capable of handling millions of GeoJSON objects and writing very large `.mbtiles` files. It is used by people all around the world and is consistently being updated.

TopoJSON: Topologically encoded geometries in JSON format. Stores points and ‘arcs’ and references to them instead of encoding complete geometries. github.com/topojson/topojson.

vector tile: A collection of map objects that exist within a specific bounding box at a specified zoom level. The size and location of a tile is defined by a specific grid location and zoom level.