Future Internet **2012**, *4*, 285-305; doi:10.3390/fi4010285

OPEN ACCESS

future internet

ISSN 1999-5903

www.mdpi.com/journal/futureinternet

Article

Characteristics of Heavily Edited Objects in OpenStreetMap

Peter Mooney [1,*] and Padraig Corcoran [2]

[1] Department of Computer Science, National University of Ireland, Maynooth, Ireland

[2] School of Computer Science and Informatics, University College Dublin, Dublin, Ireland; E-Mail:padraig.corcoran@ucd.ie

[*] Author to whom correspondence should be addressed; E-Mail: peter.mooney@nuim.ie; Tel.: +353-1-268-0100.

Received: 30 January 2012; in revised form: 7 March 2012 / Accepted: 16 March 2012 / Published: 20 March 2012

Abstract: This paper describes the results of an analysis of the OpenStreetMap (OSM) database for the United Kingdom (UK) and Ireland (correct to April 2011). $15,640$ OSM ways (polygons and polylines), resulting in $316,949$ unique versions of these objects, were extracted and analysed from the OSM database for the UK and Ireland. In our analysis we only considered "heavily edited" objects in OSM: objects which have been edited 15 or more times. Our results show that there is no strong relationship between increasing numbers of contributors to a given object and the number of tags (metadata) assigned to it. 87% of contributions/edits to these objects are performed by 11% of the total 4128 contributors. In 79% of edits additional spatial data (nodes) are added to objects. The results in this paper do not attempt to evaluate the OSM data as good/poor quality but rather informs potential consumers of OSM data that the data itself is changing over time. In developing a better understanding of the characteristics of "heavily edited" objects there may be opportunities to use historical analysis in working towards quality indicators for OSM in the future.

Keywords: OpenStreetMap; collaborative editing; volunteered geographic information; spatial data

---

1. Introduction

Volunteered Geographic Information (VGI), the term coined by Goodchild [1], is the recent empowerment of citizens in the collaborative collection of geographic information. He argues that VGI has enormous potential to become a "significant source of geographers' understanding of the surface of the Earth". Crucially, "by motivating individuals to act voluntarily, it is far cheaper than any alternative, and its products are almost invariably freely available". OpenStreetMap (OSM) is a collaborative project to create a free editable map database of the world as is probably the most well known example of VGI [2]. Spatial data is contributed to OSM from: portable GPS devices, tracing shape outlines from aerial photography, import of free spatial data, or simply from local knowledge [3]. Ciepluch *et al*. [4] and Haklay and Weber [5] provide detailed introductions to the OSM project. Real world geographic objects are represented in OSM as points, lines, and polygons. In OSM these are referred to as nodes and ways. Ways is a collective term for both polylines and polygons. Spatial attributes for these objects are stored as *tags*. An object can be tagged with any number of tags. On the OSM wiki [6] there is a community maintained page (see [7]) detailing the most-popular tags. This amounts to something close to an OSM-community generated ontology for the spatial objects in the OSM database. Volunteers, who collect and contribute data to OSM, have the freedom to add their own arbitrary tags if necessary to annotate their data. However it is usually only the tags listed on the *map features* list [7] that are supported by GIS software capable of consuming OSM data and cartographic software for rendering OSM data as map image tiles. The growing spatial coverage and high-quality content in OSM [8,9] has branched beyond "the converted" and has gained enthusiastic endorsement from the likes of Yahoo, ESRI, MapQuest [10], and Microsoft [11]. Yahoo! and Bing have agreed to let OSM use their aerial imagery for the purposes of volunteers tracing the outline of objects. However, the concept of local knowledge is at the very core of VGI and OSM. De Leeuw *et al*. [12] describe results in their paper which shows volunteers with local knowledge classified roads, from high resolution aerial imagery, with over 92% accuracy on average, irrespective of surveying background and always better than professional surveyors without local knowledge. The combination of highly motivated volunteers with detailed local knowledge [13] has been instrumental in seeing OSM grow to become a very large global database of spatial data, frequently edited, and continually growing in size.

We feel that one of the most exciting aspects of OSM is the collaborative editing and development of the OSM database by contributors. This provides motivation for this research work. As will be outlined in Section 2 all current studies into OSM, in the literature, only consider the most currently available version of the OSM database. This literature discusses: quality evaluation, accuracy measurement, or applications. Every few months OSM makes the "planet.osm/full" file available which includes almost all OSM data ever collected [14]. The key motivation of this paper is to use this historical data to investigate if there are any special characteristics that may be observed from analysis of "heavily edited" objects. After the literature review section of the paper (Section 2) we show how the full history for spatial objects can be extracted and processed from the "planet.osm/full" file (Section 3). In this section we also outline carefully how we selected our dataset of "heavily edited" objects. In Section 4 we provide the results of analysis performed on these objects. This includes: rates of contribution (Section 4.2), editing and tagging (Section 4.3 and Section 4.5), changes to object geometry (Section 4.6). Access to

the historical trial of edits for objects allows use to investigate how these features have evolved from their first version to their current version. The full OSM history data for the UK and Ireland is extracted from the "planet.osm/full" file and is used as the case-study data. Section 5 closes the paper with some conclusions from the analysis performed and opportunities for further research.

2. Literature Overview

Currently there is only a small body of literature published on analysis of the spatial data contents of the OSM database. However these publications have delivered important contributions on many issues related to OSM in general. In classical GIS methodologies there have been several accuracy and ground-truth comparisons of OSM data with authoritative sources of spatial data. These include Haklay [15] with the Ordnance Survey UK, Zielstra and Zipf [16] compared OSM data with Teleatlas Data in Germany, Girres and Touya [17] with the French OS dataset IGN), and Mooney *et al*. [18] with land cover features in Ordnance Survey Ireland datasets. In terms of using OSM as a primary source of geographic information there are several examples in the literature. Goetz and Zipf [19] present an extensive 3D building ontology based on OSM making it possible to map indoor spaces in addition to the outdoor environment. The richness of spatial data for urban areas in VGI, particularly for buildings, has seen some authors (such as Goetz and Zipf ([20]) use VGI for the creation of 3D building models for the purposes of building virtual city models. Over *et al*. [8] also discuss the generation of 3D models. Ciepluch *et al*. [4] present a framework for distribution of environmental information using OSM. Some work has been presented on the motivations of those volunteers who contribute to VGI projects [9,21]. Finally some authors have investigated the development of applications based upon the local spatial knowledge inherent in VGI. Pultar *et al*. [22] develop software for wildfire evacuation modelling and travel scenarios of urban environments using VGI as an input data source. De Leeuw *et al*. [12] argues that local knowledge demonstrated in VGI (coupled with the enthusiasm and dedication of contributors as emphasized by Goodchild [23]) could potentially be very rich extending to the point "where there is reason to consider engaging local expertise in the production and updating of NMA topographic maps".

The issue of the spatial data quality in OSM is dominant amongst the available literature. Goodchild [23] argues that in the case of OSM the rather unique task of compiling independently contributed pieces of a (geographic) patchwork necessarily imposes some degree of quality control. He adds that "one might term this process structured, to distinguish it from the essentially unstructured process by which entries in Wikimapia and other VGI gazetteers are compiled". The quality of VGI is now a hot-topic in GIS [18]. Qian *et al*. [24] argue that in VGI "since general users can add and change data, the stored data should be updated frequently, resulting in an abundant and updated geographic dataset". This has "reversed the traditional top-down flow of information" [25]. Qian *et al*. [24] conclude by arguing that one of the most serious disadvantages of OSM is that the underlying data is acquired by non-professionals with non-professional equipment, meaning that there is no guarantee of quality about the data unless it can be compared to some other source. Flanagan and Metzger [26] argue that as the amount of VGI continues to grow "the issues of credibility and quality should assume a prominent place on the research agenda". This will require a multi-disciplinary approach combining knowledge from geography, computer science, social sciences, to understand the credibility of VGI. Metadata is also an issue with Bulterman [27] suggesting that the "complete disregard for documentation of data

resources" has made it almost impossible for one to perform a fitness for use or fitness for purpose evaluation on available data resources. Without some quantitative measures of accessing the quality of the OSM data the GIS community has been slow to consider OSM as a serious source of data [18]. Imports of government or National Mapping Agency (NMA) data into OSM, as mentioned in Section 1 is the exception rather than the rule [28]. The "contributors are spontaneous and the density of data is unpredictable" and consequently the spatial distribution of the data itself will continue to be uneven and inconsistent. Flanagan and Metzger [29] remark that for VGI in general the "professional and scientific gate-keeping that usually filters and reviews digital information may not be present (in sufficient forms or structures)" and subsequently can lead to information which is prone to being "poorly organized, out-of-date, incomplete, or inaccurate". Ballatore and Bertolotto [30] calls OSM "spatially-rich and semantically-poor". Brando and Bucher [31] suggest that the quality of VGI is enhanced if proper metadata is created and maintained which details: types of changes and edits, methods of survey and collection, and finally a fitness for purpose statement. The recent study by Zielstra and Zipf [16] of OSM and TeleAtlas for Germany shows that "while professional data is not without its faults the coverage of OSM in rural areas is too small to be seriously considered a sophisticated alternative for *any* applications". However the study does conclude that for larger cities (Berlin, Frankfurt, Munich) the data diversity is so rich that "OSM is replacing proprietary data for many projects". Heterogeneity of the spatial data coverage in OSM is a real barrier. Neis *et al*. [32] show that the difference between the OSM street network for car navigation in Germany and a comparable proprietary dataset was only 9% in June 2011.

In this study we analyse "heavily edited" objects in OpenStreetMap. We are unaware of any similar example in other sources of crowdsourced spatial data or VGI. In the next section of this literature review we make connections to research work carried out in Wikipedia. Similar to OSM it is possible to extract the entire history of edits to articles in Wikipedia. A growing body of literature is available on this topic. The crowdsourced collection and collaborative editing of spatial data in OSM is unique and is certainly non-traditional for geographical data. The literature available dealing with these topics in relation to OSM is still rather limited. There is a very substantial collection of literature on Wikipedia where the equivalent to the OSM collaboratively edited object is the *article* [33]. We feel that it is useful to briefly introduce some key outcomes from research studies of the collaborative editing nature of Wikipedia. Heavily edited articles in Wikipedia are usually those that gain the status of "featured article". Featured articles are recognized as articles of high quality, with a long history of collaborative editing, and have become relatively stable (no major recent edits) [34]. In Korfiatis *et al*. [35] the authors introduce a measure of article quality for featured articles based on the succeeding edits by other users. Subsequent edits (deletions) and roll-backs to earlier versions are considered as disapproval whereas maintaining the edits of previous editors is deemed as a sign of approval. Wesler *et al*. [33] analyzed the history of edits of Wikipedia and identified a number of specific users, most notably *Substantive* editors. Substantive editors contributed, at minimum, between 30 and 80 percent of all content edits to pages. Overall research appears to indicate that there does not appear to be a "standard contributor" to Wikipedia or standard pattern of contribution. Yang and Lai [36] conclude that frequent Wikipedia users (like *Substantive* users by Wesler *et al*. [33]'s definition) may contribute knowledge by making minor changes to Wikipedia entries. Conversely, some users who contribute infrequently may provide

extremely rich content. Antin [37] reports in a survey of Wikipedia contributors that many contributors are worried about how "individual agendas could shape editing behaviors" particularly those edits from very frequent contributors. This is supported by Hecht and Gergle [38] who provide evidence that many Wikipedia users continually re-edit their "pet pages" very frequently.

3. Experimental Setup

In this section we describe the experimental setup for the analysis performed. It is necessary to introduce how OSM data is obtained and how we process this data and prepare it for analysis. OSM data can be downloaded in OSM-XML format and Shapefile (SHP) format from services such as GeoFabrik [39]. An important characteristic of these downloads is that the OSM spatial data contained within the OSM-XML files or SHP files are close to real-time representations of the spatial data stored in the global OSM database. GeoFabrik state that "essentially any change made to the global OSM database is usually reflected in the data packages available for download within 24 h" [39]. Crucially, the OSM XML data available for download from GeoFabrik, similar services, or indeed the OSM web API contains only the most recent version of each object in the selected region. These XML files contain none of the historical data.

3.1. Processing Historical OSM Data

OSM XML data can be processed in a number of ways. Two of the most popular methods in the OSM community involve the usage of command line tools such as `osm2pgsql` [40] and `osmosis` [41]. However, these tools are developed to process only the most recent version of the OSM data for a given region. Mooney and Corcoran [42] describe a process using Linux command line tools and the OSM web API to extract the history of selected objects in OSM. However this process takes prohibitively long due and can only be used for a small number of features to prevent overloading the OSM web API servers. In this paper we have updated the approach in Mooney and Corcoran [42] and use an alternative method, which has greater efficiency. Every few months OSM makes the "planet.osm/full" file available which includes almost all OSM data ever collected [14]. This file is currently 26 Gb of compressed OSM-XML expanding to 600 Gb uncompressed. We use the open source *osm-history-splitter* tool [43] to extract the history of edits from the full history dump. The *osm-history-splitter* allows splitting of the full history dump based on bounding rectangles or polygons. Using a predefined polygon for the UK and Ireland, from the country clipbounds polygons from Geofabrik [44]), the history for all nodes and ways in this region is extracted saved in OSM-XML format. The final step of the process involves loading the history of the objects into a PostgreSQL PostGIS database. The structure of the tables in this database are very similar to the standard way that OSM data is usually stored in databases by tools such as `osm2pgsql`. We developed two Python scripts to process the OSM history file by firstly inserting all nodes and then inserting all ways (polygons and polylines). The process took 305 h. The Python script must load every node (each node has multiple versions potentially) into the Postgis database. Then for each way (polygon, polyline) the Python script must reference the correct node (with the correct version) for each way version. An edit or contribution is an action which causes a new version of an object to be created. An edit or contribution can include: adding or deleting nodes, moving nodes, editing nodes,

adding tags, editing existing tags, *etc*. The contributor must then submit or commit these to the OSM database. Formally, the following information is stored for every object. Suppose that we are storing the object P at the n^{th} version. P is stored as the tuple as follows

$$P = \Big(u, v, \tau, G, T\Big)$$

where the elements of the tuple P are as defined as follows: u is the user id of the OSM contributor who created this version of P, v is the version of the OSM object, τ is the timestamp for the edit, G is the geometry of P, and T is the set of tags (keys,values) assigned to this version of P_i which are stored as a comma-separated list

3.2. Selection of Study Area and Historical Objects

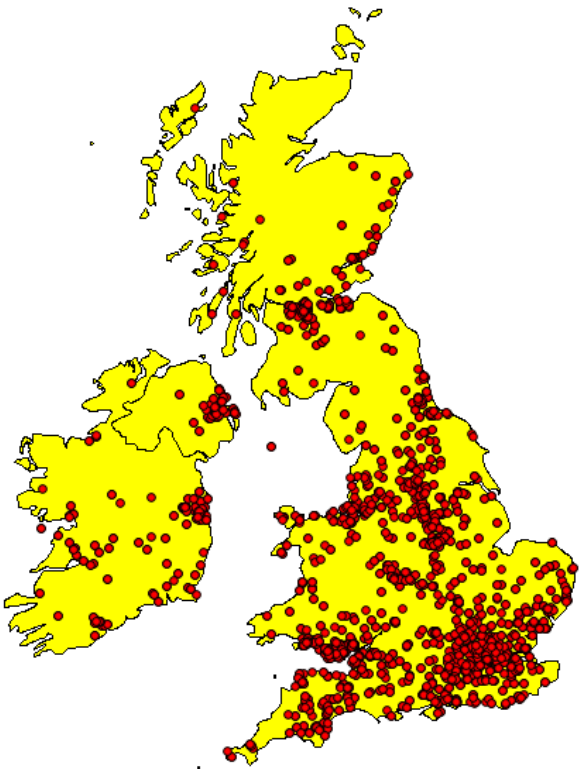
The UK and Ireland is the study area for this research. The authors are working in a university in Ireland and also have extensive knowledge of the UK and the OSM communities in both regions. We feel that the approach in this paper is sufficiently generic and flexible that it could be applied, in the future, to any OSM region(s). The OSM history data used here was extracted from a full planet history dump for April 2011. There is no guidance from the literature as to the most suitable value for the lower threshold of version for a "heavily edited" object. As described in the literature review the concept of "heavily edited" objects in OSM is loosely based upon the concept of "featured articles" in Wikipedia. We sought to identify objects in OSM which exhibited characteristics of collaborative editing: multiple editors, changes and reverts to tags and geometry, and a long edit history. An analogous concept is available in Wikipedia by comparing "heavily edited" objects in OSM to featured articles in Wikipedia. At the time of writing, that there are 3377 featured articles out of $3,744,295$ articles on the English Wikipedia [45] corresponding to 1 from every 1000 articles. Wesler *et al*. [33] explain that usually heavily edited articles in Wikipedia gain the status of "featured article" and are subsequently recognized as articles of high quality. Korfiatis *et al*. [35] based their analysis of quality of Wikipedia articles on successive edits and therefore focused on articles with a long edit history. Hecht and Gergle [38] focus on articles which have been edited frequently, particularly those by the same contributor. Nemoto *et al*. [46] indicates that quality increases, and stabilizes, the more contributors work on a given article. Table 1 shows the distribution of version numbers amongst all objects in the UK and Ireland databases. In total there are $3,793,813$ objects in the OSM database for the UK and Ireland. It is evident from the table that objects with low version numbers are the most frequently occurring. Just under 60% of all objects in the UK and Ireland OSM database exist having only a single version. There are $3,617,350$ objects (equivalent to just over 95% of all objects) in the UK and Ireland OSM database with a current version number of between 1 and 5 inclusive. In Wikipedia the collaboratively edited object is an *article* or knowledge artefact [33] while in OSM the object can be a polygon or polyline representing some real world geographic feature. We decided to analyse those "heavily edited" objects in our OSM dataset with 15 or more versions of editing. This represents approximately 0.4% of all available objects. The total number of objects with 15 or more versions of editing is $15,640$. This criteria allows us to discard analysis, for this study, of objects in OSM with a very low number of edits which would in all likelihood exhibit little collaborative editing behaviour. Our software and subsequent analysis is flexible enough to

allow this threshold value of 15 be increased or decreased. The centroids of these objects are plotted on the map of UK and Ireland in Figure 1.

Table 1. Version number distribution for all objects in the OSM dataset for the UK and Ireland.

Version	Number of Objects	Total %	Cumulative %
1	2,246,369	59.211	59.211%
2	780,320	20.568	79.779%
3	329,342	8.681	88.460%
4	169,831	4.477	92.937%
5	91,488	2.412	95.349%
6	72,347	1.907	97.256%
$7 \rightarrow 9$	68,485	1.805	99.061%
$10 \rightarrow 14$	19,991	0.527	99.588%
$15 \rightarrow 20$	11,210	0.295	99.883%
$21 \rightarrow 30$	3,381	0.089	99.972%
$31 \rightarrow 40$	706	0.019	99.991%
>40	343	0.009	100%
Total	3,793,813	100%	100%

Figure 1. A map showing the location of the centroids of all heavily edited OSM objects in our UK and Ireland case-study dataset.



4. Experimental Analysis and Results

In this section we outline the experimental analysis of the selected dataset of "heavily edited" objects.

4.1. Characteristics of the Study Area

As outlined in Section 3 $15,640$ "high edit" polygons and polylines from the United Kingdom in Ireland were chosen. These polygons and polylines represent the following features. "Amenity" are represented by polygons and usually represent places such as schools, hospitals, *etc*. "Highway" represents all types of roads, streets, laneways, highways, *etc*. Polygons marked exclusively with "Landuse" tags can represent many different types of polygons such as forest, woodland, tillage, grass, *etc*. "Natural" usually indicate lakes, ponds, and other bodies of water but can also represent features such as grassland, scrub, woods, and beaches. Finally "waterway" represent canals, rivers, streams, *etc*. These five are chosen as they represent five of the most "popular" features in OSM in terms of user contributions and edits. One can easily obtain information about the number of features with these tags in the global OSM database from services such as TagInfo ([47]). In February 2012 there were: 48 million "highway" ways, 6.37 million "waterway" ways, 5.8 million "natural" ways, and 1.1 million "amenity" ways in the OSM database.

These $15,640$ geographic features yield a total of $316,949$ unique versions. "Highway" is the dominant feature amongst the five chosen and it represents 87% of the data. 994 of the objects are from the island of Ireland while the remaining $14,646$ are from the United Kingdom. The earliest contribution is May 2007 with the most recent contribution occurring in April 2011 when the data was downloaded. There are 4128 unique contributors. There are 58 contributors who have edited all five chosen "popular" feature types.

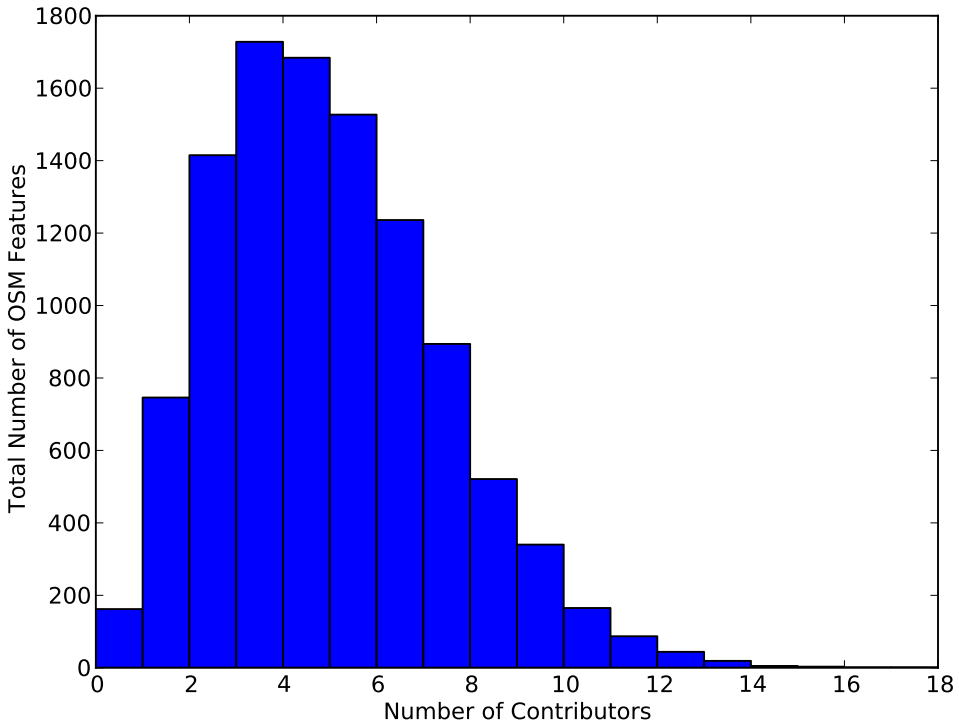
4.2. Contribution Rates

In Table 2 a grouped distribution of the number of edits or contributions performed by all of the contributors in our case-study area is shown. Figure 2 shows a graphical representation of this distribution with $\mu = 4.54$ and $\sigma = 2.39$. The value of $\mu = 4.54$ is the mean number of unique contributors to the objects in our case-study with a standard deviation of 2.39. This indicates that there is collaborative editing occurring for the majority of objects in our case study. There are also some interesting observations from the distribution of contribution effort amongst the contributors to these objects. We see from Table 2 that approximately 11% (10.779%) of contributors have contributed 100 or more edits (substantial editing efforts). This corresponds to $275,744$ edits from a total of $316,949$ or 87%. Just under 52% of contributors contributed 5 edits or less (2128 from 4128 users) with 74% providing less than 20 edits. Of course these users may have edited other features which were not included in this case study. In total 1096 users contributed only one edit corresponding to 26% of contributors. At the opposite end of the contribution scales just over 2.5% of users contributed large bodies of edits (over 500 edits) whilst 3 users stand out as very high volume contributors with over 5000 edits.

Table 2. The number of edits or contributions performed by all of the OSM contributors in our case-study area.

Contributions	#Unique Contributors	% Contributors	Total %
≤5	2128	51.55	51.55
5–10	521	12.62	64.17
10–20	404	9.78	73.95
20–50	378	9.15	83.11
50–100	252	6.10	89.22
100–200	175	4.23	93.46
200–500	163	3.94	97.41
500–1000	67	1.62	99.03
1000–2000	20	0.48	99.51
2000–5000	18	0.43	99.95
≥5000	3	0.048	99.99

Figure 2. The distribution of the number of unique contributors to each geographic object in our OSM study area.



4.3. Editing the Name Tag Attribute

We have analysed the objects in our case-study which have "name" attributes. In Table 3 an example is given where five different name values are assigned to the "name" attribute or key. The current version of this object is version 26 edited on 04/12/2010 with the name tag assigned the value of "Old Crosby". In

our case study 9837 objects have a "name" attribute. There are 412 (4.1%) objects where the assignment to the "name" key changes 3 or more times (as shown in the example in Table 3). We calculated the correlation between the number of changes of the "name" attribute and the number of editors who had edited the corresponding object. However the result is inconclusive. A very weak negative correlation of -0.17 was calculated which does not allow us to make any assumptions about the reasons for this behaviour. A total of 43 objects had their "name" attribute changed 3 or more times by the same contributor. On the other end of the scale 112 of the objects had their "name" attribute changed 3 or more times in the presence of 6 or more unique contributors. As shown by Korfiatis *et al.* [35] in Wikipedia edits (particularly deletions) of content by other contributors are considered as disapproval whereas maintaining the edits of previous editors is deemed as a sign of approval. In the case of "name" attribute changes in the presence of multiple authors a deeper analysis is required as to why these changes occur. This may possibly involve asking contributors to fill in a questionnaire regarding their decision making processes for changing attributes such as "name".

Table 3. An example of the "name" tag changing on a street polyline in the UK.

Version	Name Tag	User ID	Date of Edit
v1	NULL	11895	06/07/2008
v2	Station Road	11895	06/08/2008
v8	Oswald Road	11985	07/08/2008
v11	Frodingham Road	26825	10/10/2008
v23	Ferry Road	11985	02/06/2009
v25	Old Crosby	11985	17/06/2009

Whilst our study area, for this paper, is confined to the UK and Ireland we feel that it is useful to the discussion to show an example from the German OSM. We chose this example because of the unusually high number of edits to the "highway" tag for this feature. A "tag war" has broken out amongst OSM contributors over what they individually perceive as the correct hierarchy value for the highway tag. Up to mid-February 2011 88 versions of this polyline have been created. OSM contributor ID 7070 seems to intend on ensuring that the highway tag *always* has the value of `trunk` whereas all other contributors who have supplied edits to this polyline believe the correct tag value is `construction`. Viewing of aerial imagery of this feature in Google Maps and Bing Maps is inconclusive and potential erroneous given that the aerial imagery may be old and out-of-date. A segment of the history trail of edits for this object is shown in Table 4. We investigated the values assigned to tags and analysed if edits to those tags resulted in reverts or the assignment of new values. In total 64% of tag edits resulted in the values of tags being reverted to a previous value. In 32% of tag edits the value(s) assigned to a tag where updated or changed to a new value. It is difficult to comment if these reverts to old values or assignment of new values are strictly correct. To do so would require us to compare the tagging against some ground-truth database.

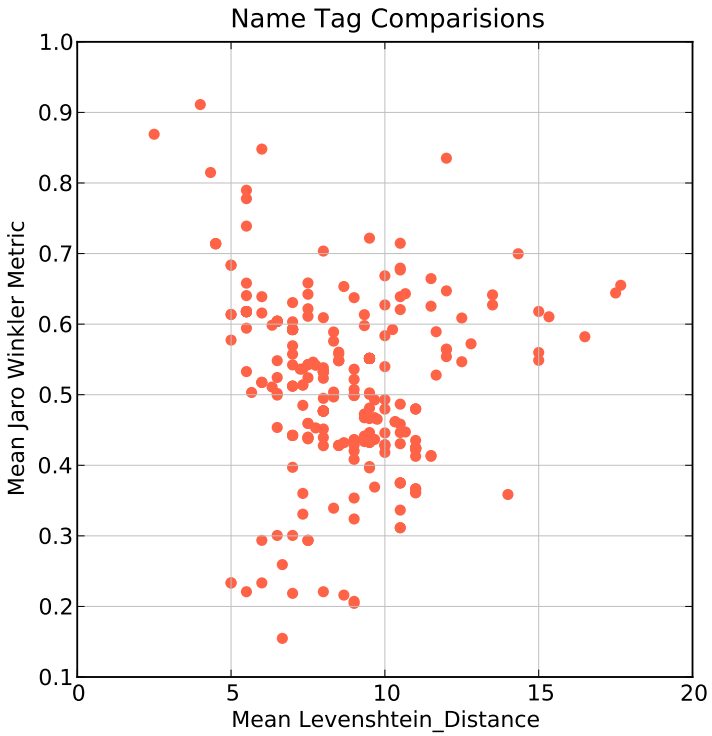
Table 4. Example of a "tag war" among OSM contributors in Germany. The value of the HIGHWAY tag representing a road feature is frequently reverted back by its original contributor. The OSM-ID of this feature is $23,704,273$.

Version	User	HIGHWAY	Edited on
1	7,070	Trunk	15/11/09
4	20,510	Construction	16/11/09
5	7,070	Trunk	17/11/09
7	19,889	Construction	17/11/09
10	7,070	Trunk	17/11/09
11	19,889	Construction	17/11/09
12	7,070	Trunk	19/11/09
⋮	⋮	⋮	⋮
78	206,986	Construction	19/12/09
79	7,070	Trunk	20/12/09
80	206,986	Construction	20/12/09
81	210,596	Trunk	20/12/09
88	145,231	Construction	16/02/11

4.4. Using String Matching to Understand Tag Changes

Using the notation from earlier in the paper each object P has n versions $(0,(n-1))$. Then T_i is the set of tags (keys, values) assigned to P at version i. Then $T_i = \{t_0^i, t_1^i, \ldots, t_{n-1}^i\}$ where t_0^i is the set of tags (possibly empty) at the first version of P and t_{n-1}^i is the set of tags (possibly empty) at the final version of P. For each object with 3 or more changes to the "name" tag attribute we clustered the assigned name tags into chronological groups and then compared the transformation of tags into one another using two well known string matching metrics to quantify how similar the name tags were. The Levenshtein distance is defined as the minimal number of characters you have to replace, insert or delete to transform from one string to another [48]. The JaroWinkler distance [49] is a similar metric used mostly for duplicate detection in databases. The metric is normalized such that 0 equates to no similarity and 1 is an exact match between the two strings. In Figure 3 we show a plot of the mean Levenshtein distance against the mean JaroWinkler distance for 412 objects. Most objects are clustered around a mean Levenshtein distance of 10 and mean JaroWinkler distance of 0.5 which indicates that the changes from one name tag to the next name tag are substantially different. This is potentially caused by contributors: spelling placenames incorrectly, providing local variations on official placenames, incorrect naming of streets, performing correction of placename spellings.

Figure 3. Using the Levenshtein distance and JaroWinkler distance metrics to visualize changes to "name" attribute tags of 412 objects in our OSM case-study area.



4.5. General Tagging of Objects

We will now discuss the tagging of the objects in the case-study. The advent of Web 2.0 with its desktop-like experience and focus on social applications helped tagging be established as a concept worthy of being considered as an open decentralized way of structuring and sharing information and meta-data in the knowledge society [50]. Morrison [51] even suggests that the "folksonomies" of tags created from social and collaborative projects on the web could be effective tools for Information Retrieval on the Web and "could be studied in that same way that search engines have been studied in the past". In collaborative projects, such as OSM, contributors can add, edit, delete tags easily [52]. For objects in OSM these tags can be updated to reflect changes over time and space. For example: changes in landcover type, building usage, highway designation, *etc*. In Figure 4 the distribution of the number of tags at the final current version of each object in our case study is shown. The mean number of tags assigned to features is 3.45. In the case of many edits to features in OSM some editor software automatically embeds their own "tag" mark to indicate which editor software performed the edit. In some cases we found this automatic tag as the only tag existing. Without tags objects cannot be properly rendered by map-tile generation software or queried from the OSM spatial database. In Figure 5 a scatter-plot shows the relationship between the number of unique contributors to each object against the number of tags associated with that object at its final or current version. The correlation (0.17) is very weak and statements about statistical significance could not be made based on this. In a similar

fashion Figure 6 plots the number of versions of each object against the number of tags associated with that object at its final or current version.

Figure 4. Distribution of the number of tags from all objects in the case study.

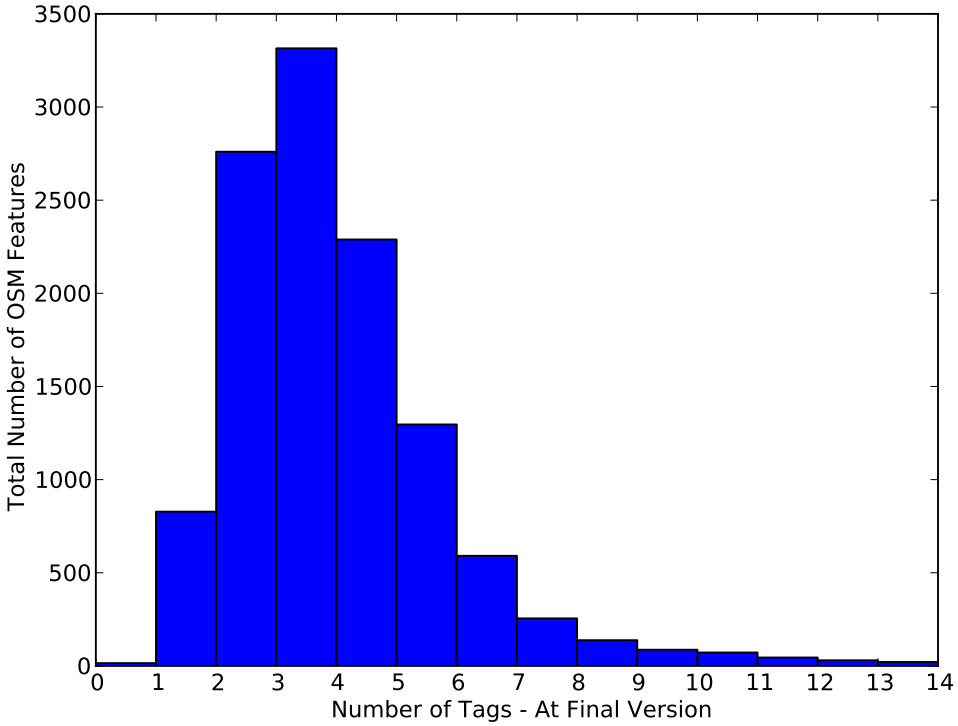


Figure 5. Scatter plot of the number of tags against the number of contributors for all objects in the case study area.

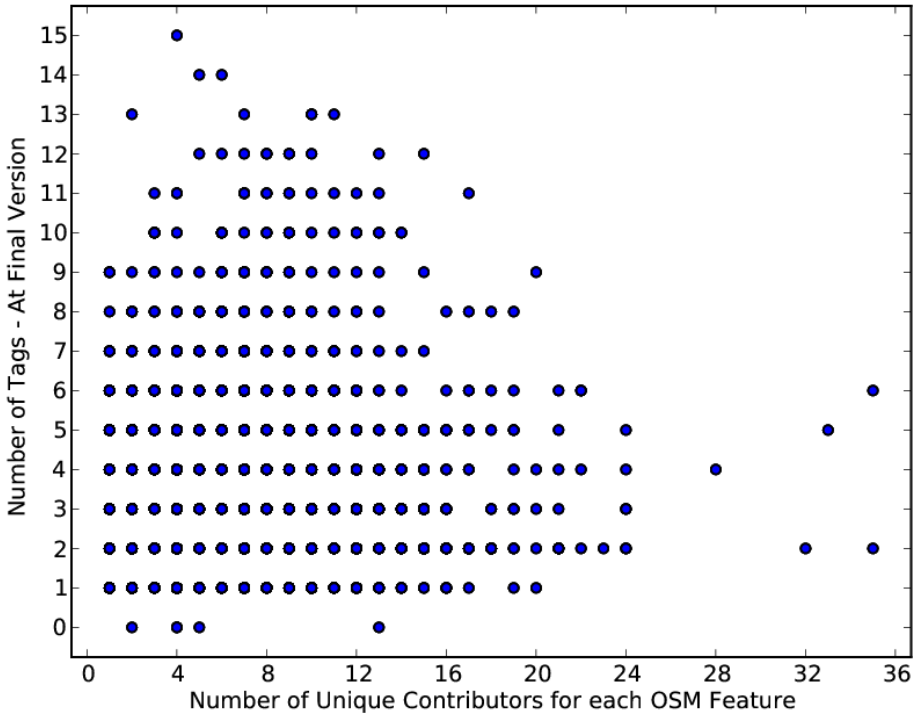
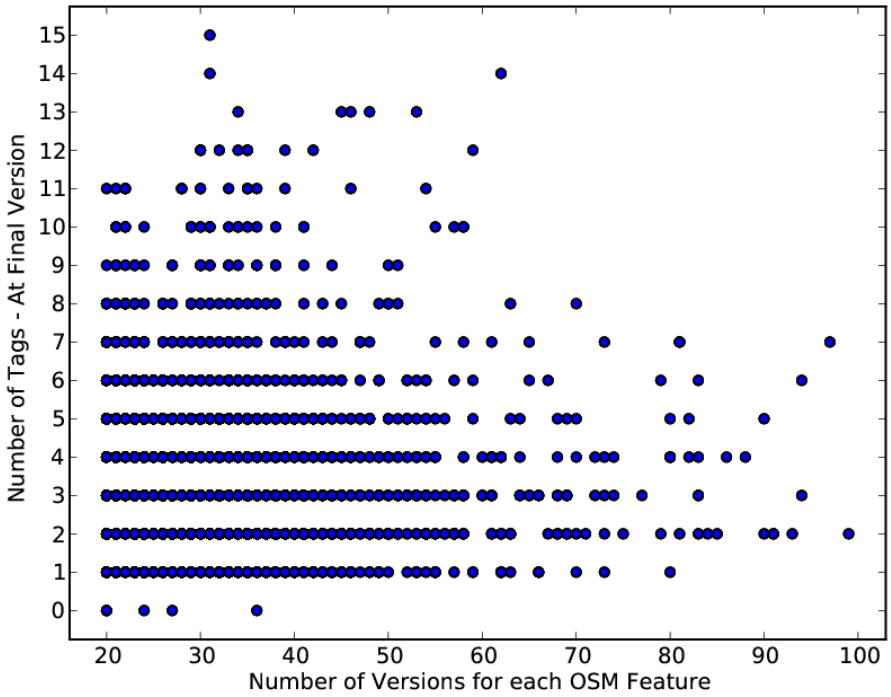


Figure 6. Scatter plot of the number of tags against the number of versions (final) for all objects in the case study area.



4.6. Changes to Object Geometry

In Table 5 we present the summary of the number of consecutive versions with the combinations of invalid or valid geometries. We analyze consecutive versions v_i and v_j of the object P where $i < j$ and $i \geq 1$. The number of consecutive version pairs where the same contributor u created both v_i and v_j is also included. The validity of geometries is tested in the PostGIS database using the `ST_IsValid()` function. This PostGIS function is compliant with the Open Geospatial Consortium's (OGC) OpenGIS Specifications where according to these specifications, a valid geometry is one such that there are "no anomalous geometric points, such as self intersection or self tangency and primarily refers to 0 or 1-dimensional geometries". All spatial databases implementing the OGC OpenGIS Specifications, such as Oracle Spatial or Microsoft SQL Server, will provide a function with this functionality. Therefore geometries invalid in PostGIS will also be returned as invalid in OGC OpenGIS compliant databases. We used this PostGIS function to investigate if invalid geometries were being created by contributions and edits to objects. In Table 5 there are 91% of versions (v_i followed by v_j) which are both valid. In 83% of these cases the same contributor edited the object. However the number of versions with consecutive invalid geometries is a non-trivial 8%. Of this 8% $21,843$ (or 87%) of these consecutive versions were created by the same contributor. This indicates that either these contributors were not aware of the invalidity of the polygons/polylines they had created or for some other reason had not fixed these problems. In the current versions v_n of all $15,640$ objects there are $14,891$ valid polygons/polylines while 749 are invalid. In Table 6 we present the summary of the number of consecutive versions with node edits. We look at consecutive versions v_i and v_j of the object P where $i < j$ and $i \geq 1$. The number of consecutive version pairs where the same contributor u created both v_i and v_j is also included.

Remarkably, the contribution of additional spatial detail occurs in 79% of edits. Of these edits 90% were carried out by the same contributor which could potentially indicate that contributors incrementally contribute spatial data to objects over time. This could potentially be the result of contributors "saving" work as they work through a series of edits or returning at a later time to incrementally edit their data.

Table 5. The validity status of consecutive versions of the same objects. The percentages in brackets indicate the number of cases where consecutive versions were edited by the same contributor.

Consecutive Versions	Total	(Same User)
Invalid,Invalid	25,107(8%)	21,843(87%)
Valid,Valid	285,598(91%)	237,046(83%)
Invalid,Valid	1685(<1%)	1145(68%)
Valid,Invalid	1455(<1%)	1091(75%)

Table 6. Summary of the number of consecutive versions where nodes were added, deleted, or left unchanged. The percentages in brackets indicate the number of cases where consecutive versions were edited by the same contributor.

Node Edit Action	Total	(Same User)
Nodes Unchanged	50,216(16%)	41,678(83%)
Nodes Deleted	15,692(5%)	9,729(62%)
Nodes Added	247,937(79%)	223,143(90%)

5. Conclusions and Future Work

In this paper 15,640 ways (polygons and polylines) resulting in 316,949 unique versions of these objects were analyzed from the OSM database for the UK and Ireland. In our analysis we only considered "heavily edited" objects in OSM: objects which have been edited over 15 times. We motivated the selection of this threshold in Section 3.2 and we feel that this provided us with a good representative sample of OSM activity in the UK and Ireland. There is good spatial distribution of the selected objects as illustrated in Figure 1. As stated by Zielstra and Zipf [16] OSM data is found in the largest quantities and coverage in urban areas. The map of the locations of our "heavily edited" objects shows greater concentration of these objects around the cities of Dublin, London, Belfast, Cardiff, and Glasgow. To our knowledge, and following an extensive literature search, this is the first study of its type of historical OSM data. Kessler *et al.* [53], Roick *et al.* [54], and van Exel *et al.* [55] consider version history but only for visualization purposes.

5.1. Conclusions

Our analysis of OSM history data has given us some interesting research results. In Section 4.2 we showed that 11% of contributors created or edited 87% of the spatial data in the 15,640 "heavily

edited" objects. Assignment of values to attributes or tag keys is another area where a historical analysis of edits demonstrates issues in the collaborative nature of OSM. 4.1% of objects have the assigned value to their "name" attribute changed 3 or more times. This rises to 25% of objects which have the assigned value to their "name" attribute changed 2 or more times. Disputes and disagreements occur. Table 4 shows a long and protracted dispute in Germany over the assignment of a classification to the highway tag. Table 3 shows an example from the UK where a street is assigned 5 different names. The use of two well known string matching metrics in Figure 3 shows that changes to "name" attributes are not subtle single character changes but major edits to the value assigned to "name". This will form a useful basis for future work to investigate if this behaviour extends to other OSM regions and communities. The uncertainty introduced by frequent changes to "name" or "highway" attributes has implications for the development of gazetteers from OSM and location-based services (LBS) which will need to be evaluated [56]. In Over *et al*. [8] the authors comment that the quality control of OSM differs fundamentally from professionally edited maps. The community-based approach allows anyone to upload and alter map data. However, due to the huge number of editors, errors and conflicts are usually quickly resolved. A long-term historical analysis, following on from this work, will provide evidence to support this hypothesis that eventually OSM data "stabilizes" for an area. Some tools are beginning to appear for visualization of the history of Wikipedia pages [57]. Some tools are beginning to emerge for OSM history but they are still in their infancy and lack powerful information visualisation functionality. Section 4.5 discusses the number of tags, or metadata, assigned to each object at its final (current) version. The mean number of tags assigned to features is 3.45. There is no discernible statistical relationship between increasing numbers of contributors and number of tags nor is there a statistical relationship between the number of versions created for an object and the number of tags. This could indicate that new contributors to an object passively accept the current set of tags without adding any additional tags. The increase in versions, without an apparent correlated increase in the number of tags, is probably a result of the "tag flip-flopping" we discussed in Section 4.3 or edits to the geometry of the object (Section 4.6). As Table 6 shows in 79% of edits nodes are added to objects. In Section 4.6 we showed in Table 5 that consecutive edits to the same object create and maintain valid spatial geometries in 91% of cases. However it is worth noting that in 8% of cases an object with an invalid geometry is edited and this invalidity problem is not fixed. In 87% of these cases the same contributor is responsible. This raises potential issues surrounding the understanding these contributors have of the need for valid geometries (avoiding self intersections, *etc*.) in a spatial dataset.

5.2. Future Work

While the paper does not specifically provide "measurements" of quality of the OSM data we believe that this work could provide a platform for future studies on OSM data quality which would consider the lineage or evolutionary history of the OSM data as part of quality assessments. A survey of contributors to OSM, particularly large scale contributors, is required to gain a better understanding of the rationale behind some of the tagging behaviour we have observed in this paper. We have used the threshold of 15 versions as the qualifying criteria for "heavily edited" objects in OSM. As part of ongoing work we are investigating the effects of revising this threshold downwards whilst attempting to understand the factors that are related to new versions of objects being created in OSM. Relations in OSM are one of

the code data elements. They consist of one or more tags and an ordered list of one or more nodes and/or ways as members. They are used to define logical or geographical relationships between elements. We decided against investigating relations in this research as we wanted to maintain focus on the editing of ways as a first step towards understanding the characteristics of heavily edited objects in OSM. Just under 30% of the $15,640$ ways we analysed in this research were marked explicitly as members of relations. As part of future work we intend to carry out an analysis of the characteristics of relations in OSM. In this paper we considered an edit as a composite record of edits to an object's geometry and tags. We shall be investigating the types of edits recorded: edits to geometry and then edits to tagging. We believe this will help us understand the editing behaviour of contributors—do some contributors contribute geometry but never perform tagging or do some contributors only correct or update tagging? Finally, an analysis of a larger number of "heavily edited" objects is required to validate our findings here to show the existence of other characteristics (spatial autocorrelation and spatial interaction). This will be the subject of our immediate future work.

References

1. Goodchild, M. Citizens as sensors: The world of volunteered geography. *GeoJournal* **2007**, *69*, 211–221.
2. OpenStreetMap. An Introduction to OpenStreetMap. 2011. From the OSM Wiki. Available online: http://wiki.openstreetmap.org/wiki/Main_Page (accessed on 15 March 2012).
3. Coast, S. How OpenStreetMap is Changing the World. In *Proceedings of the 10th International Symposium on Web & Wireless GIS*; Tanaka, K., Fröhlich, P., Kim, K.S., Eds.; Springer: Berlin, Germany, 2010; Volume 6574, p. 4.
4. Ciepluch, B.; Mooney, P.; Jacob, R.; Winstanley, A.C. Using OpenStreetMap to deliver location-based environmental information in Ireland. *SIGSPATIAL Spec.* **2009**, *1*, 17–22.
5. Haklay, M.; Weber, P. OpenStreetMap: User-generated street maps. *IEEE Pervasive Comput.* **2008**, *7*, 12–18.
6. OpenStreetMap. Getting Started with OpenStreetMap. 2010. Available online: http://wiki.openstreetmap.org/wiki/Beginners_Guide_1.1 (accessed on March 2010).
7. OpenStreetMap. The *Map Features* page. 2010. Available online: http://wiki.openstreetmap.org/wiki/Map_Features (accessed on March 2010).
8. Over, M.; Schilling, A.; Neubauer, S.; Zipf, A. Generating web-based 3D City Models from OpenStreetMap: The current situation in Germany. *Comput. Environ. Urban Syst.* **2010**, *34*, 496–507.
9. Haklay, M.; Basiouka, S.; Antoniou, V.; Ather, A. How many volunteers does it take to map an area well? The validity of linus' law to volunteered geographic information. *Cartogr. J.* **2010**, *47*, 315–322.
10. MapQuest. AOL Corporate Communications—MapQuest "Opens Up" in Europe with Open-Source Mapping with UK Launch. AOL Online Blog report. 2010. Available online: http://corp.aol.com/2010/07/09/mapquest-opens-up-in-europe-with-open-source-mapping-with-uk-l/ (accessed on 15 March 2012).

11. Bing Maps. Bing Engages Open Maps Community. Microsoft Bing Maps Online Blog report. 2010. Available online: http://www.bing.com/community/site_blogs/b/maps/archive/2010/11/23/bing-engages-open-maps-community.aspx (accessed on 15 March 2012).
12. de Leeuw, J.; Said, M.; Ortegah, L.; Nagda, S.; Georgiadou, Y.; DeBlois, M. An assessment of the accuracy of volunteered road map production in Western Kenya. *Remote Sens.* **2011**, *3*, 247–256.
13. Coleman, D.; Georgiadou, P.; Labonte, J. Volunteered geographic information: The nature and motivation of produsers. *Int. J. Spat. Data Infrastruct. Res.* **2009**, *4*, 332–358.
14. OSM History. The Full OpenStreetMap History Dump. 2012. Available online: http://wiki.openstreetmap.org/wiki/Planet.osm/full (accessed on March 2012).
15. Haklay, M. How good is volunteered geographical information? A comparative study of OpenStreetMap and ordnance survey datasets. *Environ. Plan. B Plan. Des.* **2012**, *37*, 682–703.
16. Zielstra, D.; Zipf, A. A Comparative Study of Proprietary Geodata and Volunteered Geographic Information for Germany. In *Proceedings of the 13th AGILE International Conference on Geographic Information Science*; Painho, M., Santos, M.Y., Pundt, H., Eds.; Springer Verlag: Guimarães, Portugal, 2010.
17. Girres, J.F.; Touya, G. Quality assessment of the french OpenStreetMap dataset. *Trans. GIS* **2010**, *14*, 435–459.
18. Mooney, P.; Corcoran, P.; Winstanley, A.C. Towards Quality Metrics for OpenStreetMap. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*; ACM: New York, NY, USA, 2010; pp. 514–517.
19. Goetz, M.; Zipf, A. Extending OpenStreetMap to Indoor Environments: Bringing Volunteered Geographic Information to the Next Level. In *Proceedings of the UDMS: Urban and Regional Data Management: 2011*; Rumor, M., Zlatanova, S., LeDoux, H., Eds.; Delft: Delft, The Netherlands, 2011; pp. 47–58.
20. Goetz, M.; Zipf, A. Towards defining a framework for the automatic derivation of 3D CityGML models from volunteered geographic information. *Int. J. 3-D Inf. Model.* **2012**, *1*, 496–507.
21. Budhathoki, N.R.; Nedovic-Budic, Z.; Bruce, B. An interdisciplinary frame for understanding volunteered geographic information. *Geomat. J. Geospat. Inf. Sci. Technol. Pract.* **2010**, *64*, 14–29.
22. Pultar, E.; Raubal, M.; Cova, T.J.; Goodchild, M.F. Dynamic GIS case studies: Wildfire evacuation and volunteered geographic information. *Trans. GIS* **2009**, *13*, 85–104.
23. Goodchild, M. Commentary: Whither VGI? *GeoJournal* **2008**, *72*, 239–244.
24. Qian, X.; Di, L.; Li, D.; Li, P.; Shi, L.; Cai, L. Data cleaning approaches in Web 2.0 VGI application. In *Proceedings of 2009 17th International Conference on Geoinformatics*; Fairfax, VA, USA, 12–14 August, 2009; pp. 1–4.
25. Fritz, S.; McCallum, I.; Schill, C.; Perger, C.; Grillmayer, R.; Achard, F.; Kraxner, F.; Obersteiner, M. Geo-Wiki.Org: The use of crowdsourcing to improve global land cover. *Remote Sens.* **2009**, *1*, 345–354.
26. Flanagin, A.J.; Metzger, M.J. The credibility of volunteered geographic information. *GeoJournal* **2008**, *72*, 137–148.
27. Bulterman, D.C.A. Is it time for a moratorium on metadata? *IEEE Multimed.* **2004**, *11*, 10–17.

28. Mooney, P.; Corcoran, P.; Winstanley, A.C. A Study of Data Representation of Natural Features in OpenStreetMap. In *Proceedings of the 6th GIScience International Conference on Geographic Information Science*, Zurich, Switzerland, 14–17 September 2010; p. 150.
29. Flanagin, A.J.; Metzger, M.J. The role of site features, user attributes, and information verification behaviours on the percieved credibility of Web-based information. *New Media Soc.* **2007**, *9*, 319–342.
30. Ballatore, A.; Bertolotto, M. Semantically Enriching VGI in Support of Implicit Feedback Analysis. In *Web and Wireless Geographical Information Systems*; Tanaka, K., Fröhlich, P., Kim, K.S., Eds.; Springer Berlin/Heidelberg: Berlin, Heidelberg, Germany, 2011; Volume 6574, pp. 78–93.
31. Brando, C.; Bucher, B. Quality in User Generated Spatial Content: A Matter of Specifications. In *Proceedings of the 13th AGILE International Conference on Geographic Information Science*; Painho, M., Santos, M.Y., Pundt, H., Eds.; Springer Verlag: Guimarães, Portugal, 2010.
32. Neis, P.; Zielstra, D.; Zipf, A. The Street Network Evolution of Crowdsourced Maps: OpenStreetMap in Germany 2007–2011. *Future Internet* **2012**, *4*, 1–21.
33. Welser, H.T.; Cosley, D.; Kossinets, G.; Lin, A.; Dokshin, F.; Gay, G.; Smith, M. Finding Social Roles in Wikipedia. In *Proceedings of the 2011 iConference*; ACM: New York, NY, USA, 2011; pp. 122–129.
34. Anderka, M.; Stein, B.; Lipka, N. Towards Automatic Quality Assurance in Wikipedia. In *Proceedings of the 20th International Conference Companion on World Wide Web*; ACM: New York, NY, USA, 2011; pp. 5–6.
35. Korfiatis, N.; Poulos, M.; Bokos, G. Evaluating authoritative sources using social networks: An insight from wikipedia. *Online Inf. Rev.* **2006**, *30*, 252–262.
36. Yang, H.L.; Lai, C.Y. Motivations of Wikipedia content contributors. *Comput. Hum. Behav.* **2010**, *26*, 1377–1383.
37. Antin, J. My Kind of People? Perceptions About Wikipedia Contributors and Their Motivations. In *Proceedings of the 2011 Annual Conference on Human Factors In Computing Systems*; ACM: New York, NY, USA, 2011; pp. 3411–3420.
38. Hecht, B.J.; Gergle, D. On the "Localness" of User-Generated Content. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work*; ACM: New York, NY, USA, 2010; pp. 229–232.
39. GeoFabrik. GeoFabrik: Download Server for OpenStreetMap data. 2010. Web Based Download Application: Available online: http://download.geofabrik.de/ (accessed on 19 March 2012).
40. osm2pgsql. Osm2pgsql—An OSM data importer for Postgis databases. 2012. Available online: http://wiki.openstreetmap.org/wiki/Osm2pgsql (accessed on 19 March 2012).
41. OSMOSIS. OSMOSIS—A command line Java application for processing OSM data. 2012. Available online: http://wiki.openstreetmap.org/wiki/Osmosis (accessed on 19 March 2012).
42. Mooney, P.; Corcoran, P. Accessing the History of Objects in OpenStreetMap. In *Proceedings of the 14h AGILE International Conference on Geographic Information Science*; Geertman, S., Reinhardt, W., Toppen, F., Eds.; Springer Verlag: Utrecht, The Netherlands, 2011; p. 155.

43. MaZderMind. osm-History-Splitter: A C++ Tool to Split OSM Full-History-Planet-Dumps into Smaller Extracts Based on Bounding-Boxes or Polygons. 2012. Available online: https://github.com/MaZderMind/osm-history-splitter (accessed on 19 March 2012).
44. Geofabrik. Clipbounds—Boundary *poly* Files for the Extract Of Country Regions From OSM-XML Data. 2012. Available online: http://download.geofabrik.de/clipbounds/ (accessed on 19 March 2012).
45. Wikipedia. Featured Articles in Wikipedia. 2011. Available online: http://en.wikipedia.org/wiki/Wikipedia:Featured_articles (accessed on 15 March 2012).
46. Nemoto, K.; Gloor, P.; Laubacher, R. Social Capital Increases Efficiency of Collaboration Among Wikipedia Editors. In *Proceedings of the 22nd ACM Conference on Hypertext and Hypermedia*; ACM: New York, NY, USA, 2011; pp. 231–240.
47. Topf, J. The TagInfo Webservice—Statistics About Tags in The OpenStreetMap Database. 2012. Available online: http://taginfo.openstreetmap.org/keys (accessed on February 2012).
48. Yujian, L.; Bo, L. A normalized levenshtein distance metric. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *29*, 1091–1095.
49. Bilenko, M.; Mooney, R.; Cohen, W.; Ravikumar, P.; Fienberg, S. Adaptive name matching in information integration. *IEEE Intell. Syst.* **2003**, *18*, 16–23.
50. Derntl, M.; Hampel, T.; Motschnig-Pitrik, R.; Pitner, T. Inclusive social tagging and its support in Web 2.0 services. *Comput. Hum. Behav.* **2011**, *27*, 1460–1466.
51. Morrison, P.J. Tagging and searching: Search retrieval effectiveness of folksonomies on the World Wide Web. *Inf. Process. Manag.* **2008**, *44*, 1562–1579.
52. Overell, S.; Sigurbjörnsson, B.; van Zwol, R. Classifying Tags Using Open Content Resources. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*; ACM: New York, NY, USA, 2009; pp. 64–73.
53. Kessler, C.; Trame, J.; Kauppinen, T. Tracking Editing Processes in Volunteered Geographic Information: The Case of OpenStreetMap. In *Proceedings of the COSIT'11 Workshop: Identifying Objects, Processes and Events in Spatio-Temporally Distributed Data (IOPE)*, Belfast, Maine, USA, 12 September 2011; pp. 17–29.
54. Roick, O.; Loos, L.; Zipf, A. A Technical Framework for Visualizing Spatio-Temporal Quality Metrics of Volunteered Geographic Information. In *Proceedings of the GEOINFORMATIK 2012—Mobility and Environment*, Braunschweig, Germany, 28–30 March 2012.
55. van Exel, M.; Dias, E.; Fruijtier, S. The Impact of Crowdsourcing on Spatial Data Quality Indicators. In *Proceedings of GiScience 2011*, Zurich, Switzerland, 14–17 September 2010.
56. Mooney, P.; Corcoran, P. Using OSM for LBS—An Analysis of Changes to Attributes of Spatial Objects. In *Proceedings of the 8th International Symposium on Location-Based Services*; Gartner, G., Ortag, F., Eds.; Springer: Vienna, Austria, 2011; pp. 165–176.

57. Pirolli, P.; Wollny, E.; Suh, B. So You Know You're Getting the Best Possible Information: A Tool That Increases Wikipedia Credibility. In *Proceedings of the 27th International Conference on Human Factors In Computing Systems*; ACM: New York, NY, USA, 2009; pp. 1505–1508.