Freedom versus Standardization: Structured Data Generation in a Peer Production Community

Andrew Hall*, Sarah McRoberts*, Jacob Thebault-Spieker*, Allen Yilun Lin†, Shilad Sen‡, Brent Hecht†, Loren Terveen*

*GroupLens Research, University of Minnesota; †Northwestern University; ‡Macalester College
{hall,mcroberts,thebault,terveen}@cs.umn.edu, {allen.lin, bhecht}@eecs.northwestern.edu, ssen@macalester.edu

ABSTRACT

In addition to encyclopedia articles and software, peer production communities produce *structured data,* e.g., Wikidata and OpenStreetMap's metadata. Structured data from peer production communities has become increasingly important due to its use by computational applications, such as CartoCSS, MapBox, and Wikipedia infoboxes. However, this structured data is usable by applications only if it follows *standards*. We did an interview study focused on OpenStreetMap's knowledge production processes to investigate how – and how successfully – this community creates and applies its data standards. Our study revealed a fundamental tension between the need to produce structured data in a standardized way and OpenStreetMap's tradition of contributor freedom. We extracted six themes that manifested this tension and three overarching concepts, **correctness**, **community**, and **code**, which help make sense of and synthesize the themes. We also offered suggestions for improving OpenStreetMap's knowledge production processes, including new data models, sociotechnical tools, and community practices (e.g. stronger leadership).

Author Keywords

OpenStreetMap; peer-production communities; structured data; standardization

ACM Classification Keywords

H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous;

INTRODUCTION

Peer production communities such as Wikipedia and OpenStreetMap (OSM) are among the most successful applications of social computing. One major factor in the success of these communities is their widespread adoption of a "be bold" contribution principle [28,29]. This principle encourages contributors to quickly "make changes as they see fit" [10] and elevates contributor freedom to the status of a core community value.

While contributor freedom within the community helped Wikipedia and OSM create millions of encyclopedia articles and hundreds of gigabytes of geospatial data, it is not clear whether it will be as helpful as both communities begin major pushes into producing *highly standardized structured data*. Wikipedia's structured data efforts are manifest in Wikidata, a major Wikimedia Foundation project described as "a free linked database that can be read and edited by both humans and machines." [30] OSM's structured data creation occurs in its tagging infrastructure, which lets editors specify the semantics of a geospatial entity using key-value pairs (e.g. its name, whether it is a restaurant or a hospital, etc.). Both these initiatives are motivated by a desire to help computers understand real world semantics (e.g., moving towards computing over "things not strings", as Google puts it [31]).

For Wikidata and OSM to support computing applications most effectively, their structured data must have a high degree of standardization. For example, popular tools developed to make maps from OSM data (e.g. MapBox, CartoCSS) will not use the right color or icon for entities if the entities are not tagged properly. Similarly, standardizing entities like roads is especially valuable for projects such as the Humanitarian OpenStreetMap Team (HOT) [32], which seeks to provide aid in humanitarian disasters. The story is similar for Wikidata. Tools developed to generate natural language Wikipedia articles from Wikidata (e.g. Reasonator [33]) work properly only if the data conforms to standards. The same is true for Wikipedia itself, which pulls in data from Wikidata for language alignment and infoboxes [34].

There is an inherent tension between the standardization needs of structured data and peer production communities' ethos of contributor freedom, which encourage contributors to just "go for it" and employ "trial and error" [29]. OSM emphasizes contributor freedom even more than Wikipedia. Where Wikipedia tempers contributor freedom with a set of policies (such as "Neutral Point of View") that are strictly enforced by the community, OSM's "Good Practice" [28] says "Nobody is forced to obey [the OSM guidelines], nor will OSM ever force any of its mappers to do anything."

CHI 2017, May 06-11, 2017, Denver, CO, USA
 ISBN 978-1-4503-4655-9/17/05...$15.00
DOI: http://dx.doi.org/10.1145/3025453.3025940

This tension led us to articulate a central research question: **How does OSM's strong commitment to contributor freedom affect its efforts to produce standardized data?**

To answer this question, we performed an interview study of OSM contributors. The study focused on the production of OSM metadata ("tags"), investigating community practices that lead to standardization successes and failures.

Our contributions are as follows:

1. We show *why* the large degree of contributor freedom affects the ability of peer production communities to be standardized. For example, some contributors – through greater technical skill or dedication to a cause – were able to influence standards. Cultural differences also caused standardization problems – for example, a "highway" can have different definitions in different regions.
2. Based on our results, we offer several new sociotechnical strategies and tools to improve standardized data creation in peer production communities. For example, our work problematizes OSM's 1:1 tagging structure, motivates the need to be able to link similar entities, and informs the design of tools that can improve standardization without increasing the effort required to contribute.

We next introduce OSM's mechanisms for standardizing its metadata, followed by a discussion of related research and then our research methods. The heart of the paper consists of our results, interleaved with discussion of their meaning and implications. We conclude with a synthesis and brief summary.

SPECIFYING DATA STANDARDS IN OSM

OpenStreetMap contributors apply metadata to characterize the semantics of geographic entities. Specifically, they apply *tags* – tags are key-value pairs, where the *key* represents a concept and the *value* indicates a specific *instance* of the concept. For example, a fast food restaurant might include the tag "cuisine = burger".

The OSM community also created and maintains a wiki[1] that identifies "certain key and value combinations for the most commonly used tags, which act as informal standards." In other words, the wiki tells contributors how to tag entities. The wiki specifies information such as relationships between tags – for example, that an entity tagged as "amenity = restaurant" also should include a value for the key "cuisine" to indicate the type of food served. It also characterizes appropriate values for a key – http://wiki.openstreetmap.org/wiki/Key:cuisine recommends the value "burger" for "e.g. McDonald's, Wendy's, Jollibee (Philippines)" and the value "coffee_shop" for places that serve "mainly coffee, [and] may have some light cold snacks such as cakes".

Thus, to the extent that OSM contributors can produce standardized metadata, it is by consulting the OSM wiki and understanding how it applies to the geographic entities they are editing. In addition, several editing tools (e.g. JOSM, Potlatch, iD) help editors by suggesting metadata to apply.

RELATED WORK

Data Standardization in OSM

OSM has been the subject of considerable research since its inception in 2004. Much of this research has focused on comparing OSM spatial entities to ground truth data (e.g. [5,6,21]). This work mostly has looked at spatial dimensions such as positional accuracy and completeness of entities mapped.

While the predominant focus in the OSM literature is on the spatial entities themselves, some researchers have examined OSM metadata. Some of this work (e.g. [5,21,26]) touched on challenges in ensuring metadata standardization; these challenges helped motivate our research. For example, Girres and Touya compared OSM highway tag data to French BD TOPO ground truth data in a small portion of France and found roadway standardization problems [5].

External OSM tag editing applications and the algorithms they leverage have played an increasingly important role in shaping the OSM tagging folksonomy. These algorithms do not necessarily enforce the 'informal standards' of the wiki since they factor in observed tagging practice (which may or may not follow the wiki standards). Vandecasteele and Devillers [26] point out the large degree of "semantic heterogeneity" in OSM and propose a solution to mitigate this problem through a tag recommender, OSMantic. This tool uses existing community tagging practices such as tag co-occurrence to recommend new tags for OSM records being edited. Karagiannakis et al. created a similar recommender, OSMRec [13]. While these systems may well help contributors follow community practice, there is no guarantee that practice actually follows standards; if not, these systems end up reinforcing suboptimal practices. Determining how standardization fails is thus of great importance, and we seek to understand it. Our study builds upon standardization work in OSM and seeks to understand the *causes* of standardization problems in data.

A small amount of work has considered the OSM community's process of standard (as specified in the wiki) creation. Ballatore and Mooney considered the negotiation process of the creation of standards found on the wiki [1]. They did so by textual analysis of the OSM wiki and OSM mailing lists. They find some similar impediments to standardization attempts that we find (but through a different method). For instance, they found that cultural differences in road representation have resulted in the community's inability to arrive at a global roadway tagging standard.

[1] http://wiki.openstreetmap.org

Freedom versus Constraint

Particularly relevant to this study is work that has shown the contrast in how OSM and Wikipedia treat community norms. The Wikipedia community has developed hundreds of policies (including essays, guidelines, and "official policies") to "[manage…] diverse views" [15]. Palen et al. [23] discussed how OSM differs from Wikipedia: OSM seeks to keep "bureaucracy at bay" as an effort to support "diverse participation". They also note that as OSM grows, "social and technical approaches" for governance are preferred over bureaucratic ones, while Wikipedia managed "community growth through the creation and clearer articulation of policies". While policy enforcement varies by language edition in Wikipedia (see [25]), Wikipedia, in general, is more policy-oriented than OSM[2].

Policies can have a negative effect on a community. Halfaker et al. wrote that "Wikipedia has changed from 'the encyclopedia that anyone can edit' to 'the encyclopedia that anyone who understands the norms, socializes him or herself, dodges the impersonal wall of semi-automated rejection and still wants to voluntarily contribute his or her time and energy can edit'." [9] Strictly enforced norms certainly have positive effects, e.g. managing vandalism. However, they also may have negative side effects; for example, Halfaker et al. found that Wikipedia's strong and strictly enforced norms reduced retention of new editors. Moreover, Lin [20] has noted that OSM's "lack of established rules" has advantages: it makes for a low "entry barrier" and provides a chance to become more of a community member. Here, we explore how OSM's substantial scope for contributor freedom affects its ability to produced standardized data, where following standards is essential for automatic processing.

METHOD

To answer our research question, we performed 15 semi-structured interviews. Our questions focused on the process of creating OSM's metadata standard and applying it while entering specific geographic data, with an eye to identifying reason for standardization failures. The interviews occurred during March and April 2016 via Skype or Google Hangouts. Recruitment was conducted in three ways: two OSM mailing lists ("Tagging" and "Talk"), an OSM forum, and snowball sampling. Two of the authors performed the interviews (mostly separately, but occasionally together) in English following predefined protocols. All participants consented electronically and verbally in line with our IRB approved-protocol. We compensated participants with a $10 USD Amazon gift card. Participants came from a number of countries, mainly from Western and developed countries (~33% from each of North America, Europe, and Asia). This distribution is generally consistent with OSM demographics [3]. There was a wide-range of OSM experience, from several months to 10 years (see Table 1). 80% of participants were male, broadly consistent with general OSM demographics [7]. Most participants edited OSM in a non-professional capacity (i.e. did not edit OSM as part of their jobs); a minority edited due to their role in a humanitarian or other organization. See Table 1 for additional participant information gathered from a web tool called "How did you contribute to OpenStreetMap?" [35].

Participant	Sex	Time in OSM	Map Changes
P1	M	4 years	>1,500,000
P2	M	7 years	>250,000
P3	F	3 years	>1000
P4	M	10 years	>2,500,000
P5	M	3 years	>1000
P6	M	2 years	>1,000,000
P7	M	8 years	>250,000
P8	M	7 years	>250,000
P9	M	< 1 year	>250,000
P10	M	7 years	>500,000
P11	F	< 1 year	>100,000
P12	M	6 years	>2,500,000
P13	M	5 years	>1,500,00
P14	M	6 years	>100,000
P15	F	1 year	>100,000

Table 1: Participant Information

To analyze the interviews, we first transcribed the audio recordings and then employed a Grounded Theory approach [22]. Three of the authors applied open coding to the transcripts. Then four of the authors collaboratively reached consensus on general themes of the codes using an affinity-diagramming approach, involving constant comparison amongst codes. This resulted in six themes we describe next.

RESULTS AND INTERPRETATIONS

Our themes all relate to how the OSM community's attempts to produce standardized data mesh with its commitment to contributor freedom. To help understand the themes more holistically, we introduce a set of concepts that cut across many of the themes.

- **Correctness**. For metadata to be standardized, there must be definitions of "correct" and "incorrect" metadata. We observed two types of standardization issues associated with the notion of correctness: (a) how *complete* must metadata for an entity be for it to be considered correct? (b) how *consistent* must metadata be across different contexts for it to be considered correct?
- **Community**. How does the OSM community manage itself to produce standardized metadata? To what extent

[2] For the discussions in this paper, we focus on English Wikipedia.

are standards enforced? How does the community reach *consensus* on standards? As we consider these questions, it is instructive to compare OSM to the most prominent peer production community, Wikipedia.

- **Code**. To what extent do OSM's data ontology and data entry tools facilitate – or impede – the production of standardized data?

We conclude each of our themes with a discussion to situate it with respect to these concepts and draw implications for design, practice, or research.

Theme 1: Freedom vs. Metadata Completeness

Freedom is fundamental. A number of contributors (P1, P4, P7, P8, P10, P11) explicitly noted that lack of rules is fundamental to OSM's ethos and differentiates it from Wikipedia. While freedom may harm standardization by letting contributors vary the amount of content they provide, contributor freedom often is precisely why participants preferred OSM (P8, P10). Four participants noted that – compared to Wikipedia, where articles must adhere to strict quality standards such as verifiability and neutrality – OSM has less rigid standards and less enforcement (P4, P7, P8, P10). Participant P10's remark was illustrative:

> *Wikipedia to me feels like Germany, too many rules and regulations. (P10)*

However, too much freedom may cause problems. One participant characterized another OSM contributor as taking freedom to an absurd degree by proposing excessive detail:

> *...he suggested that certain stores like supermarkets, have a list of things they sell, brands, and did we wanna put the brands they sell in OpenStreetMap. Well, for Christ's sake. Imagine that! How many brands? You'd have a list of thousands, and they would change all the time! (P1)*

Several other participants explicitly described the lack of clearly defined and enforced standards as problematic. For example, P11 expressed confusion over what tags should be applied to a McDonald's fast food restaurant.

> *Do you need to list, like everything that a McDonald's will ever give you? (P11)*

While OSM seeks to minimize use and enforcement of policies [23], and the community values this, we see that too little direction also can be problematic. Thus, a balance must be struck. We next suggest several possibilities for doing so.

Region-specific information maturity can suggest appropriate contributor freedom. One participant (P4) noted that actual contributor freedom in OSM was a function of the completeness of the map in a given region.

> *...both in Wikipedia and OSM…freedom allows growth to happen quickly, allows iteration to happen fast. With the iteration, you improve and you figure out a system for that place…And this freedom is kind of contextual. Like in the UK for instance [in OSM], it doesn't seem like there's much freedom for things new, and it would be very wrong if they would impose such restrictions in growing communities like India and other places where they say "oh this is how it should be done and this is the right thing". That'll completely stifle growth of the community…(P4)*

Of course, a notion of *localized* freedom also can be problematic; as we discuss later, others desired *global* standards. However, P4 raises an issue that has become urgent for Wikipedia's health – an increase in policies and restrictions aimed at standardization can effect community growth negatively [9]. Palen et al. characterized OSM as of 2015 as similar to the Wikipedia of 2007: it is growing fast and looking to "find ways to cope with and maximize this opportunity." [23] They argue that it is too early to tell if OSM's less policy-oriented approach to handle growth "will address problems of policy rigidity and unfriendliness to newcomers that Wikipedia has faced." (e.g. those pointed by Halfaker et al. in [9]).

Contributors look to "satisfice" – apply "basic" tags and do a "good enough" job (P1, P2, P3, P4, P7, P10, P13, P15). Participants had different definitions of what they took to be "complete enough" when mapping. Adding basic tags to characterize entities was a suggested heuristic. For example:

> *...I always think the main thing is to just get it [a business] on there [OSM], and sort of basically what it is, if it's a restaurant, or you have a cuisine, maybe a website or something, the phone number, or the address. (P7)*

And while contributors believed they put in good faith effort to tag entities, they also acknowledged that they sometimes skipped details – it was too much work or too time consuming to apply all possibly relevant tags to an entity.

> *Yeah, it's too much work to add everything. I just add...minimum I feel is required to uniquely describe, or not uniquely but, to a good extent describe the object to somebody who is new. (P4)*

Discussion

Freedom is highly valued by the OSM community and plays a major role in defining and differentiating the community, particularly in contrast to Wikipedia. This is reflected in relaxed attitudes about the degree of metadata completeness required. Participants resolved this in a rough and ready manner, by suggesting a "good enough" heuristic and noting that regions of differing information maturity rightly should expect different levels of completeness.

Theme 2: Project-Specific Freedom and Metadata Correctness: Humanitarian OpenStreetMap (HOT)

HOT is a special interest group in OSM that works to create maps "when relief organizations are responding to disasters or political crises" [36]. HOT and other humanitarian activities play a prominent role in the OSM community: HOT serves as "a driver of OSM's evolution" [23].

At least half our participants (P2, P6, P7, P9, P11, P12, P15) did HOT work, often in Africa. HOT mapping is directed by the HOT Tasking Manager, a tool that lets volunteers do

"armchair mapping" (mapping using remote imagery) for predefined regions throughout the world [37] [38].

HOT prioritizes only metadata needed for humanitarian efforts, not metadata completeness. P11 noted that HOT's focus is achieving humanitarian goals, not producing detailed maps. Specifically, providing detail not needed for humanitarian efforts slows contributors down, resulting in fewer objects mapped per unit time. Highways, tracks, and paths were noted as particularly important to HOT, leading to metadata like "highway=path", "highway=track", etc.

> *I feel like I've been told to calm down and like hold back on some of the detail and I think it's because it's specifically for that project [HOT]; like I wanted to [map] every single house, and I wanted to tag whether I thought it was a house or whether it was a commercial building. I was going into all this detail…And someone was like 'just calm down, we're trying to save lives here we don't need like the most, like intricate map'. But I also really want to put in pretty much every bit of information. My theory is that eventually someone's gonna need that information… (P11)*

> *…[HOT wants her to] 'tell us how big the village is and tell us where the roads are'. They're really particular about 'we only want highways and paths and tracks and that's it' and I'm like 'there's so much more detail' (P11)*

HOT may redefine metadata meaning, leading to inconsistency with global standards. Participants observed that HOT contributors both invent new tags and redefine global definitions of tag meaning to achieve their goals.

> *The NGOs in HOT…invent their own tags...so they sort of define what they want (P12)*

> *…we do bend the rules slightly with Humanitarian OpenStreetMap for the humanitarian work. For instance, in the earthquake in Nepal, we were trying to help the aid agencies to reach remote places. So what we were doing online was trying to identify helicopter landing sites. And what we did was, we found the feature for a helicopter-landing site. And what we would do is look for an open area about 30 meters...A clearance of 30 meters that was level land near a village, and we will label that as a helicopter landing site. (P6)*

In other words, HOT appropriated and redefined the tag for a "helicopter landing site" to meet their immediate needs (for the April 2015 Nepal Earthquake). Interestingly, in less than a month, the OSM wiki was updated with a new tag for this purpose, "emergency=landing_site". This is a vivid example of HOT bending OSM's global rules, leading to (at least temporarily) inconsistent semantics, and eventually driving the evolution of the global ontology.

Discussion

HOT exploits OSM's freedom to meet its own needs for metadata correctness: only as complete as necessary for humanitarian work and inconsistent with global standards if needed. It is instructive to compare HOT to projects within Wikipedia; Wikipedia includes many *WikiProjects*, dedicated to improving Wikipedia coverage of specific topics. Like HOT (and other OSM projects), WikiProjects can define new information structures as needed. However, WikiProjects must abide by global Wikipedia standards such as the "five pillars" – not doing so, even temporarily, is considered a policy violation [39] and is grounds for immediate reverting. In contrast, as we have seen, HOT may choose not to follow global OSM recommendations for metadata completeness and may adopt inconsistent semantics for global metadata concepts.

HOT's approach helps it achieve important humanitarian aims. However, as P11 noted, OSM data is persistent and potentially of interest over time and in multiple contexts. Thus, there is a tension between HOT's need for focus and speed and desire for generally useful and globally consistent metadata. We will revisit this tension in the next themes.

Theme 3: Cultural Differences Make Global Metadata Correctness Standards Difficult to Achieve and Maintain

Since OSM is a global community, contributors come from different cultures and speak different languages. Our participants often mentioned how language and culture differences make it difficult to achieve globally consistent definitions of metadata correctness.

Must Western definitions be followed globally in OSM? (P1, P6, P10, P11, P15). OSM started in London, and participants noted the strong Western influence on the OSM metadata standards. This was reflected in tag naming conventions, available tags, and tags considered important.

> *Why should we be using British terminology just 'cause it started in Britain? (P1)*

> *…a lot of the tags are very Western world-centric, you know, so there's lots of amenities and services for people in the African countries where volunteers are serving and tags don't exist for those kind of things yet...one of our volunteers in Zambia has told me that in Zambia they have what's called a maternity waiting shelter and it's right next to their rural health centers so it's a space for women to stay at the shelter until they give birth... (P15)*

Prior research also found that global road standardization in OSM is a complex problem that has not been solved satisfactorily [1]; our participants likewise complained that the global road classification system did not accurately describe roads in all areas, particularly outside the global West. For example:

> *[In the context of humanitarian mapping in Africa] Someone says "this is a highway", and I'm like I disagree. And I'm really afraid to map that as a highway if I think, for example, like a vehicle can't go down it and that sort of thing. And I think there's a lot of conflict between what sort of roads people use to differentiate between those things and what roads people think are important or not…people see things differently. (P11)*

Different language editions differ in metadata correctness standards (P3, P5, P6, P7, P10). The OSM wiki (like Wikipedia) has language-specific versions. Some tags exist, or are documented, only in certain language versions.

> [P5 noted that in Japan, navigation can occur by using neighborhood police boxes] *This [system] is, as far as I can tell, a totally ubiquitous part of the urban landscape in Tokyo and most Japanese cities, and a really important way-finding mechanism...I think that's because it's just not a phenomenon that exists in too many other countries, and as a result the tagging documentation only exists in Japanese.*

Further, a tag might have different descriptions in different language versions of the wiki, thus, there is no globally agreed upon definition of correctness.

> *…there are Dutch translations, or French translations, or Spanish translations…they don't say the same things (P10)*

As a specific example, P7 (an American) noted that different language versions specified wholly different tagging conventions (e.g., McDonald's could be "amenity= fast_food" in one, and "amenity=restaurant" in another).

> *…last year I found that there's this whole parallel tagging scheme, so I tried to sort of communicate with the guy who has been working on that page, I think it was in German or something, and sort of, we sort of, if you think one of these is better than the other, and we didn't really come to anything… (P7)*

It is interesting to note that all these phenomena also have been observed across Wikipedia language editions [4,11,12,19]. OSM tags that occur in only some languages correspond to "concept-level diversity" [12] in Wikipedia. and differences in tag descriptions correspond to "sub-concept-level diversity" [12]. The parallels with Wikipedia engender promising research possibilities. For instance, the methods used by Hecht and Gergle [12] and Bao et al.[2] could be leveraged to assess and visualize the differences in spatial attributes in different parts of the world. Similarly, work on Wikipedia has shown that the diversity between language editions surfaces in algorithms that leverage Wikipedia data [12]. It would be interesting to see if that also occurs with algorithms that utilize OSM data.

Local tagging practices are preferred over (conflicting) global correctness standards (P1, P3, P5, P7, P8, P9, P15). P3 (an American) noted that the wiki contained conflicting descriptions of what constitutes "correct" metadata. When she sees multiple recommendations for tagging an entity, she looks at how others have mapped that same entity in the area she is mapping to select "locally appropriate" tags.

> *Yeah, I know I have [seen conflicting advice on the wiki pages]…when that happens I guess I just pick a way, whatever I guess, if there's one way that seems to be prevalent in this area. (P3)*

Another American participant noted that she believes that the tags she applies are correct per wiki guidelines, but that she also values *"the local knowledge of the people involved…" (P15)*; thus, her tags correspond to *local participants'* ideas of metadata correctness (this mapping philosophy also has been reported in prior work [14]). In general, our participants – who typically were *not local to the areas they mapped* – emphasized the value of mapping for the sake of local accuracy. P8, who had spent time mapping locally in Thailand, said just this – while also reporting that everyone does not take this approach (similar to the issues Wikipedia has had representing "Indigenous knowledge" [27]).

> *I believe the map has to be usable by locals, so for Thailand, it has to be Thai script [the name tag], but it's not a map for foreigners, made by foreigners, but it should be a map for locals. So that is some deviation, likely other mappers don't do. (P8)*

Working towards a global consensus for tag definition (P1, P4, P5, P6, P7, P8, P12, P14, P15). In tension with the previous point, many participants expressed a desire to develop a globally consistent consensus on tag meanings. This is a challenge, particularly given different language representations of the same entity.

> *...there was a discussion on tagging the streets in a way that, for shop=marine, and people in the US it made sense that that's a place that would sell boating supplies, or gear, things like that, but people in Europe, and the UK were like no, that's chandlery…So there was some discussion, and I think settled on boat supplies or something, it was something that would be easily understandable around the world, so until you get that sense of how worldwide it is, and how diverse all the people are around the world, it's easy to just assume I know the right way, this is it, and completely not understand that there's other concepts, there are other things to call them. (P7)*

P7 noted that the community insists on globally applicable rather than region-specific tags, regardless of the entity being mapped, resulting in these different interpretations of entities across cultures/languages. Similar to the "culture clashes" noted by Lin [20], P6 stated that defining a general tag that makes sense on a global scale is difficult.

> [related to tagging apartment buildings] *The result is that conflicts can occur because a [tag for an apartment] can be different things in different countries…What generally happens [hypothetical scenario] is that the French, if they're trying to set up a specific local feature, would actually come on to the OSM and discuss it. They'd put a proposal on the wiki page, and open it up for discussion on the tagging forum. And then everyone would come in and discuss the possibilities, how that would fit in…and how it can be used. And then after the proposal, they go for a vote to actually vote whose version is going to come up as being the accepted version. And after the vote takes place, then a wiki page is set up which describes that feature. (P6)*

Discussion

We saw that the OSM community insists on a global set of tags (i.e. the French do not have their own French-specific tags), reflecting a desire for consistent global standards. However, there also is a preference for local tagging schemes to be given priority over other, more global schemes (i.e. if both "highway=path" and "highway=primary" make sense for a road in Africa, the one preferred by locals should be used). This tension can result in a local-versus-global tug of war over metadata correctness that can be confusing to contributors. This is analogous to the distinction between personal and public tags in the folksonomy literature. In studying the movie recommendation site MovieLens[3], Sen et al. [24] found that personal tags (analogous to locally appropriate OSM tags) were not as valuable to the community (analogous to OSM as a whole) as they are to a given person (or, for OSM, a local community).

This observation is consistent with the commentary of Ballatore and Mooney [1]. They stated that OSM's "global, universalistic scope…clashes with the heterogeneity of its contributors and objects of interest" and also noted that "Recurrently, contributors set off to find a universalistic conceptualisation and, after encountering insurmountable problems, resorted to more contingent and localised approaches."

There are two possible paths for resolving the tension participants reported between global standards and local knowledge. First, OSM could privilege the local, allowing language-specific tags and tagging schemes analogous to language-specific versions of Wikipedia. This would help purely local applications, say routing applications within a specific country or region. On the other hand, intra-regional comparisons and applications would suffer. Second, OSM could insist on a global set of tags and tagging standards. Several possible steps could make this effective across regions and contexts, including: (a) clearer definitions of tag meanings and conditions for applying them, (b) (as suggested by P4) use of pictures on the wiki to help contributors understand the physical entity corresponding to a given tag and minimize misinterpretations due to cultural differences (the use of pictures has been successful for similar purposes in other multi-national online systems [8]).

Theme 4: Community-Management Obstacles to Achieving Consensus

The OSM community uses various online media to discuss proposals for new tags and related topics. These include many forums and mailing lists. According to our participants, these forums have problems that make it hard for the community to achieve a clear and unambiguous metadata standard, and thus make it hard for contributors to produce correct metadata.

Standards proposals lack authority. Some participants did not like the voting process for proposed tags (P3, P10). Specifically, since so few people participated, our participants did not consider the process to be a valid way to reach a community consensus.

> *Think about the thousands of people who contribute to OpenStreetMap and think like it shouldn't just be 12 people deciding on this [proposed tag]. (P3)*

Online discussions are often unproductive (P2, P4, P5, P8, P13). Five participants described the online communication among community members as unproductive. A veteran contributor from Thailand was particularly frustrated with discussions related to how to map in the US.

> *So I tell ya, I see a lot of discussion, particularly in the US map the discussions are just endless and they talk and talk and talk and they don't actually do anything. The US map, I shouldn't say this but I think it's in terrible condition. (P13)*

P5 felt that stronger leadership was important for OSM:

> *[Wikipedia has] arrived at a stronger set of norms, governance. They definitely have their own problems, but they have a strong central leadership organization and have professionalized in a way that OpenStreetMap has not…I think [OSM should be more similar to Wikipedia]. (P5)*

Hostility and toxic behavior online (P3, P4, P5, P9, P15). Five participants said that hostility/toxicity was a problem in the mailing lists and other communication media.

> *Oh boy so the lists can be really great and an awesome way to keep the community connected and supportive of each other…but they [emails] can also be like horrible…I know there's been rifts in the past and sometimes the email list can be very toxic…people feel like they can say things that they wouldn't say to someone else's face. (P15)*

P3 also mentioned that she had "heard of women not being listened to or respected".

Another participant noted that small sets of contributors with rather extreme viewpoints were given more credence than they should. In a particular instance, the effectiveness of HOT was brought into question.

> *That is a fringe minority opinion that HOT could be doing bad work, and everyone's entitled to spout off their beliefs, but without the kind of strong leadership that can establish a vision for the project, these ideas get way more credence than they deserve. (P5)*

Discussion

Lam et al. noted that psychological research has shown that while large groups are more prone to conflict, "large and diverse groups can make better decisions than individuals or

[3] https://movielens.org

experts" [16]. Their study of Wikipedia found that more contributors increased quality while warning to "be wary of decisions that are made by groups that are very small" [16]. This research reinforces the idea that OSM decision processes would benefit by involving more people.

Further, as in many online communities, toxic behavior and inefficient communication reduce community effectiveness and member satisfaction. Similar issues exist in Bitcoin, another open source community that lacks strong leadership and enforced norms [40]. Crucial updates needed to handle the increased popularity of Bitcoin were hindered due to ineffective leadership [40]. As P5 stated above, strong, proactive leadership as in Wikipedia could improve communication, likely resulting in increased productivity. Improved productivity could reduce the number of people who are frustrated with the decision-making process (like P13), and make them more inclined to participate in that process, which could increase its authority.

Theme 5: Data Representation Prevents Conceptual Correctness

In the OSM data model, only one value is allowed per key for any record; for example, a record representing a Dairy Queen shop cannot be tagged both "cuisine=burger" and "cuisine=ice_cream". A possible workaround is to concatenate multiple values with semicolons (e.g. "cuisine=burger;ice_cream"). However, this workaround is rarely used: it breaks most map rendering applications, and the OSM wiki recommends that contributors apply only the "primary" attribute of an entity [41].

However, the obvious problem is that many real-world entities have multiple valid values for certain attributes – a Dairy Queen does serve **both** ice cream and burgers. But OSM's data model forces contributors to choose one, meaning that the entity cannot be represented correctly. Participants (P10, P14) described examples of this problem:

> *...you have to map the hotel, and the restaurant, but sometimes it's just a restaurant with a hotel upstairs, so it isn't really two things, it's just one thing. So you have to say it like this is a restaurant with rooms available or something like that. (P10)*

Discussion

To make sense of the problems noted here, it is useful to refer to Lessig's notion of *code* [18]. He observed that the code of a system constitutes an architecture that constrains human behavior by allowing, forbidding, encouraging, or discouraging certain actions. Specifically, the OSM data model – adopted for certain technical reasons – constrains contributors' ability to represent the real world correctly. And thus, changes to the code – the data model – are needed to enable correct representation.

An existing OSM technical solution is to use the concept of "relations" to link two geometries (e.g. the restaurant and hotel geometries) into one entity. However, relations are rarely used in OSM, and do not accurately capture the real world semantics anyway. A more fundamental change to OSM's code, namely to allow multiple values per key, would solve the "multiple values per key" problem directly. Of course, it is likely that other parts of the OSM code base and applications that use OSM data rely on the one-to-one assumption, so such a change would have to be designed and implemented carefully.

Theme 6: Data Entry Tools May Harm Metadata Correctness and Privilege Certain Users

As we mentioned, OSM contributors use a number of tools to facilitate the data entry process. Despite their benefits, they also raise problems of metadata correctness.

Data import tools. The OSM community is wary of bulk import of data from external sources "because poor imports can have significant impacts on both existing data and [the] local mapping community." [42] Many participants (P1, P4, P5, P6, P8) expressed similar concerns that data imports can cause problems of metadata correctness.

> *They did a big data import from the New York State Department of Environmental Conservation and areas that are wilderness [they incorrectly tagged those areas]. "Landuse forest" [a tag], they're totally wrong… (P1)*

> *…any import is viewed extremely skeptically, and the only ones that happen as a result are the ones that nobody talks about and that are done stealthily and improperly. (P5)*

> *...if someone is importing a bogus tag...it comes up with a high number of occurrences, so just counting the number is not the correct method to say this is widely in use. So you would have to count a little bit more, you would have to count how many individual mappers have used it. (P8)*

Discussion

Searching for places of interest is a common use case for applications built on OSM data. Business places specifically have an incentive for accurate data representations in OSM, since they want to make it easy for potential customers to find them. Two participants (P4, P12) suggested that businesses update their own metadata in OSM – P12 specifically called for OSM to make it easier for businesses to do this. Since businesses maintain their own databases, this would require OSM to extend its sociotechnical code by developing data import tools and data correctness monitoring tools that enable it to trust bulk imported data.

However, there is another technical problem that would have to be solved to enable this. With current OSM tagging practices, it is not easy to identify accurately and reliably the OSM entities that correspond to instances of a given business, e.g., Starbucks. Instead, user-assigned entity names (which often are inconsistent with official business names) and other entity attributes must be examined to infer that an OSM instance actually is (say) a Starbucks coffee shop. Modifying the OSM code to provide unique identifiers for real-world entities (say, a single ID for all Starbucks), for example, is a reasonable approach to solve this problem

(which is somewhat similar to the suggestion by Girres et al. [5]). P5 explicitly noted that this approach could facilitate metadata import by businesses and other organizations.

Data Entry Tools. Widely used OSM editing tools such as JOSM and Potlatch facilitate the metadata entry process by suggesting 'preset' tags for entities. For example, a preset for tagging fast food restaurants suggests some relevant keys and likely values for the keys. From the perspective of metadata correctness, however, we note that (a) each tool implements different presets – P9 and P2 mentioned this; and (b) the source for the suggested presets is unclear; for example, P6 thought they were based on actual tagging practice. More fundamentally, our participants were unsure whether these presets were consistent with the metadata standard laid out in the OSM wiki (P1, P2, P4, P5, P7, P8, P9, P10, P12, P14). Some thought they were; others did not:

> *…there is not a single source for those presets…it's all manually done by the developers of the tool. (P8)*

> *[JOSM is] pretty much a direct implementation [of the wiki]* (P9)

Some of our participants noted another aspect of these tools: the *power* of their code [18]. While OSM has little central structure and the community's ethos champions contributor freedom, the data entry tools contributors use constrain their actions while elevating some contributors' positions:

> *If you really want to push for specific tagging…you have to get the tool [that uses the desired tagging] in [use in OSM]. So if something is a preset in JOSM, if something is a preset in iD, and the stand up map is rendering it, you have a very high chance of that being actively used (P8)*

Discussion

These observations strongly remind us of the power of code [18]. Of course, it is natural and useful for data entry tools to suggest plausible metadata for entities as they are entered. However, to the extent that the community has a standard for metadata correctness, the tools should base their suggestions on that, rather than on observed practice (which may or may not accurately follow the standard).

Further, P8's quote makes explicit the 'politics' of code, "the conditions of code and software in relation to power, capital, and control." [17] Those who create (or influence the creation of) OSM tools are exerting control over the community [17]. This perspective is consistent with that of Lin, who argued that "technical skills…are interlinked with the roles one holds" in OSM [20]. Another example of this in OSM can be seen in several participants (P1, P7, P15) mentioning that editing the wiki requires technical expertise, and that not all OSM contributors felt comfortable doing so. If editing the wiki were easier, more contributors might be able and willing to participate creating and maintaining the standard for metadata correctness. Once again, this could be done by modifying the OSM sociotechnical code, for example, by training contributors in use of visual editing tools.

REFLECTING ON CORRECTNESS, COMMUNITY, AND CODE

We now take the opportunity to summarize and reflect on how the concepts of **correctness**, **community**, and **code** let us make sense of our results.

Theme 1 emphasized how OSM contributor freedom ("Nobody is forced to obey") may result in incomplete (and thus incorrect) metadata. Contributors may settle for their own notion of "good enough", with no assurance that they supplied sufficient metadata for applications that use the data.

Theme 2 showed how an OSM project, Humanitarian OpenStreetMap (HOT), uses OSM's freedom to ignore global standards of metadata correctness, both completeness and consistency. They do this because they prioritize achieving their humanitarian goals; however, we also saw that their efforts help drive the global evolution of OSM. We also contrasted OSM's community management with Wikipedia by comparing the latitude given to HOT and WikiProjects.

Theme 3 examined a set of culture- and language-based issues that impact metadata correctness and related community practices. The Western origin of OSM has led to the OSM ontology embodying Western concepts, for example, of what constitutes a "highway" or "helicopter landing site". This reminds us of the power of code to constrain behavior. As contributors map in non-Western regions, they may have to appropriate and redefine the meaning of these concepts to make sense in their context, resulting in inconsistencies between "global" (more accurately, Western) and local standards. Moreover, different language versions of the OSM wiki – its metadata standard – may differ in their treatment of concepts, again putting a truly global correctness standard out of reach.

Theme 4 showed that the large amount of freedom OSM contributors enjoy and its weak community management can exacerbate common online community problems of inefficient communication and toxic behavior. Particularly important for our concerns is the effect of the problems on the development and propagation of metadata correctness standards. Some of our participants expressed concern that only a small, but very active group of people participates in the creation of these standards. In addition, toxic behavior, including sexism, can systematically exclude many contributors from participating in this process. This may result in standards that reflect the preferences of only a small group and not those of the entire OSM community.

Themes 5 and 6 emphasized the power of code in shaping OSM and OSM contributors' behavior. OSM's data model makes it impossible to represent certain real world entities – for example, a fast food place that serves both ice cream and burgers – in an intuitively correct way. OSM data entry tools

facilitate metadata entry, but can propagate metadata practices inconsistent with the OSM standard. In effect, because they are *used* by OSM contributors, they may define a second, unofficial standard of more power than that laid out in the wiki. We also saw that OSM community members with greater technical ability and access – for example, knowledge of how to edit the OSM wiki or ability to modify a data entry tool – had greater power to influence the nature of OSM data. *Code* is especially powerful in OSM because *community management* is (intentionally) weak.

SUMMARY

We carried out an interview study to understand how the OSM community created its metadata standard and applied it to represent geographic data, paying particular attention to understanding standardization failures. Our data analysis identified six major themes, all reflecting tensions between the need to produce structured data in a standardized way and OSM's tradition of contributor freedom. We also identified three overarching concepts, **correctness**, **community**, and **code**, which helped us make sense of and synthesize these themes.

ACKNOWLEDGEMENTS

We would like to thank our interviewees for volunteering to participate in this study. We would also like to thank Morten Warncke-Wang for his feedback on our work. This work was funded in part by the U.S. National Science Foundation (NSF IIS-1552955; NSF IIS-1526988; NSF IIS-0964695; NSF IIS-1218826; NSF IIS-1111201) and the U.S. Department of Education (USDE GAANN-p200A100195).

REFERENCES

1. Andrea Ballatore and Peter Mooney. 2015. Conceptualising the geographic world: the dimensions of negotiation in crowdsourced cartography. *International Journal of Geographical Information Science* 0, 0: 1–18. https://doi.org/10.1080/13658816.2015.1076825
2. Patti Bao, Brent Hecht, Samuel Carton, Mahmood Quaderi, Michael Horn, and Darren Gergle. 2012. Omnipedia: bridging the wikipedia language gap. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1075–1084. Retrieved September 19, 2016 from http://dl.acm.org/citation.cfm?id=2208553
3. Nama R. Budhathoki and Caroline Haythornthwaite. 2013. Motivation for Open Collaboration Crowd and Community Models and the Case of OpenStreetMap. *American Behavioral Scientist* 57, 5: 548–575. https://doi.org/10.1177/0002764212469364
4. Ewa S. Callahan and Susan C. Herring. 2011. Cultural bias in Wikipedia content on famous persons. *Journal of the American society for information science and technology* 62, 10: 1899–1915.
5. Jean-François Girres and Guillaume Touya. 2010. Quality Assessment of the French OpenStreetMap Dataset. *Transactions in GIS* 14, 4: 435–459. https://doi.org/10.1111/j.1467-9671.2010.01203.x
6. Mordechai Haklay. 2010. How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environment and planning. B, Planning & design* 37, 4: 682.
7. Muki Haklay and Nama Budhathoki. OpenStreetMap–Overview and Motivational Factors. *ResearchGate*. Retrieved September 18, 2016 from https://www.researchgate.net/publication/44295974_OpenStreetMap-Overview_and_Motivational_Factors
8. Scott A. Hale. 2012. Net Increase? Cross-Lingual linking in the blogosphere. *Journal of Computer-Mediated Communication* 17, 2: 135–151.
9. Aaron Halfaker, R. Stuart Geiger, Jonathan T. Morgan, and John Riedl. 2012. The rise and decline of an open collaboration system: How Wikipedia's reaction to popularity is causing its decline. *American Behavioral Scientist*: 0002764212469365.
10. Aaron Halfaker, Aniket Kittur, and John Riedl. 2011. Don't bite the newbies: how reverts affect the quantity and quality of Wikipedia work. In *Proceedings of the 7th international symposium on wikis and open collaboration*, 163–172. Retrieved January 5, 2016 from http://dl.acm.org/citation.cfm?id=2038585
11. Brent Hecht and Darren Gergle. 2009. Measuring self-focus bias in community-maintained knowledge repositories. In *Proceedings of the fourth international conference on Communities and technologies*, 11–20. Retrieved August 4, 2016 from http://dl.acm.org/citation.cfm?id=1556463
12. Brent Hecht and Darren Gergle. 2010. The tower of Babel meets web 2.0: user-generated content and its applications in a multilingual context. In *Proceedings of the SIGCHI conference on human factors in computing systems*, 291–300. Retrieved September 21, 2016 from http://dl.acm.org/citation.cfm?id=1753370
13. Nikos Karagiannakis, Giorgos Giannopoulos, Dimitrios Skoutas, and Spiros Athanasiou. 2015. OSMRec Tool for Automatic Recommendation of Categories on Spatial Entities in OpenStreetMap. In *Proceedings of the 9th ACM Conference on Recommender Systems*, 337–338. Retrieved October 27, 2015 from http://dl.acm.org/citation.cfm?id=2796555
14. Marina Kogan, Jennings Anderson, Leysia Palen, Kenneth M. Anderson, and Robert Soden. 2016. Finding the Way to OSM Mapping Practices: Bounding Large Crisis Datasets for Qualitative Investigation. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2783–2795. Retrieved May 26, 2016 from http://dl.acm.org/citation.cfm?id=2858371
15. Travis Kriplean, Ivan Beschastnikh, David W. McDonald, and Scott A. Golder. 2007. Community, consensus, coercion, control: cs* w or how policy mediates mass participation. In *Proceedings of the 2007 international ACM conference on Supporting group*

work, 167–176. Retrieved September 11, 2016 from http://dl.acm.org/citation.cfm?id=1316648

16. Shyong K. Lam, Jawed Karim, and John Riedl. 2010. The effects of group composition on decision quality in a social production community. In *Proceedings of the 16th ACM international conference on Supporting group work*, 55–64. Retrieved September 13, 2016 from http://dl.acm.org/citation.cfm?id=1880083
17. Ganaele Langlois, Fenwick McKelvey, Greg Elmer, and Kenneth Werbin. 2009. Mapping commercial Web 2.0 worlds: Towards a new critical ontogenesis. *Fibreculture* 14: 1–14.
18. Lawrence Lessig. 1999. *Code and Other Laws of Cyberspace*.
19. Yilun Lin, Bowen Yu, Andrew Hall, and Brent Hecht. 2017. Problematizing and Addressing the Article-as-Concept Assumption in Wikipedia. In *Proceedings of the 20th ACM Conference on Computer Supported Cooperative Work & Social Computing*. https://doi.org/10.1145/2998181.2998274
20. Yu-Wei Lin. 2011. A qualitative enquiry into OpenStreetMap making. *New Review of Hypermedia and Multimedia* 17, 1: 53–71.
21. Peter Mooney, Padraig Corcoran, and Adam C. Winstanley. 2010. Towards Quality Metrics for OpenStreetMap. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems* (GIS '10), 514–517. https://doi.org/10.1145/1869790.1869875
22. Michael Muller. 2014. Curiosity, Creativity, and Surprise as Analytic Tools: Grounded Theory Method. In *Ways of Knowing in HCI*, Judith S. Olson and Wendy A. Kellogg (eds.). Springer New York, 25–48. Retrieved September 18, 2016 from http://link.springer.com.ezp2.lib.umn.edu/chapter/10.1007/978-1-4939-0378-8_2
23. Leysia Palen, Robert Soden, T. Jennings Anderson, and Mario Barrenechea. 2015. Success & scale in a data-producing organization: the socio-technical evolution of OpenStreetMap in response to humanitarian events. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, 4113–4122. Retrieved September 15, 2016 from http://dl.acm.org/citation.cfm?id=2702294
24. Shilad Sen, Shyong K. Lam, Al Mamunur Rashid, Dan Cosley, Dan Frankowski, Jeremy Osterhouse, F. Maxwell Harper, and John Riedl. 2006. Tagging, communities, vocabulary, evolution. In *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work*, 181–190. Retrieved September 19, 2016 from http://dl.acm.org/citation.cfm?id=1180904
25. Besiki Stvilia, Abdullah Al-Faraj, and Yong Jeong Yi. 2009. Issues of cross-contextual information quality evaluation—The case of Arabic, English, and Korean Wikipedias. *Library & Information Science Research* 31, 4: 232–239. https://doi.org/10.1016/j.lisr.2009.07.005
26. Arnaud Vandecasteele and Rodolphe Devillers. 2015. Improving Volunteered Geographic Information Quality Using a Tag Recommender System: The Case of OpenStreetMap. In *OpenStreetMap in GIScience*, Jamal Jokar Arsanjani, Alexander Zipf, Peter Mooney and Marco Helbich (eds.). Springer International Publishing, 59–80. Retrieved March 24, 2015 from http://link.springer.com/chapter/10.1007/978-3-319-14280-7_4
27. Maja van der Velden. 2013. Decentering Design: Wikipedia and Indigenous Knowledge. *International Journal of Human-Computer Interaction* 29, 4: 308–316. https://doi.org/10.1080/10447318.2013.765768
28. Good practice - OpenStreetMap Wiki. Retrieved August 8, 2016 from http://wiki.openstreetmap.org/wiki/Good_practice
29. Wikipedia:Be bold - Wikipedia, the free encyclopedia. Retrieved August 8, 2016 from https://en.wikipedia.org/wiki/Wikipedia:Be_bold
30. Wikidata. Retrieved August 8, 2016 from https://www.wikidata.org/wiki/Wikidata:Main_Page
31. Official Google Blog: Introducing the Knowledge Graph: things, not strings. Retrieved August 9, 2016 from https://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html
32. Humanitarian OpenStreetMap Team. Retrieved August 8, 2016 from https://hotosm.org/
33. Reasonator. Retrieved August 8, 2016 from https://tools.wmflabs.org/reasonator/
34. Category:Templates using data from Wikidata - Wikipedia, the free encyclopedia. Retrieved August 9, 2016 from https://en.wikipedia.org/wiki/Category:Templates_using_data_from_Wikidata
35. How did you contribute to OpenStreetMap ? Retrieved September 18, 2016 from http://hdyc.neis-one.org/
36. About | Humanitarian OpenStreetMap Team. Retrieved August 8, 2016 from https://hotosm.org/about
37. HOT Tasking Manager -. Retrieved August 8, 2016 from http://tasks.hotosm.org/
38. LearnOSM. Retrieved September 15, 2016 from http://learnosm.org/en/coordination/remote/
39. Wikipedia:WikiProject Council/Guide/WikiProject - Wikipedia. Retrieved December 31, 2016 from https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Council/Guide/WikiProject#Violating_policies
40. Episode 708: Bitcoin Divided. *NPR.org*. Retrieved September 13, 2016 from http://www.npr.org/sections/money/2016/06/29/484029238/episode-708-bitcoin-divided
41. Semi-colon value separator - OpenStreetMap Wiki. Retrieved December 5, 2016 from http://wiki.openstreetmap.org/wiki/Semi-colon_value_separator

42. Import/Guidelines - OpenStreetMap Wiki. Retrieved January 6, 2017 from http://wiki.openstreetmap.org/wiki/Import/Guidelines