Exploring the Relationship Between "Informal Standards" and Contributor Practice in OpenStreetMap

Andrew Hall[*], Jacob Thebault-Spieker[†], Shilad Sen[‡], Brent Hecht[§], Loren Terveen[*]
[*]GroupLens Research, University of Minnesota; [†]Virginia Tech; [‡]Macalester College; [§]Northwestern University;
{hall, terveen}@cs.umn.edu, jthebaultspieker@vt.edu, ssen@macalester.edu, bhecht@eecs.northwestern.edu

ABSTRACT

Peer production communities create valuable content such as software, encyclopedia articles, and map data. As part of the creation process, these communities define production standards for their content, e.g., semantic and syntactic requirements. We carried out a study in OpenStreetMap to investigate the role of that community's standards for geographic metadata. We found that most applied metadata was consistent with the community's standards; however, we also found that the standards identified many opportunities for applying metadata that were not achieved. In addition, when we situated the standards in the context of OpenStreetMap's data model, we found a significant amount of ambiguity; the syntax allowed only one value, but everyday meaning -- and the standards themselves -- called for multiple values. Our results suggest significant opportunities for OpenStreetMap to produce additional valuable open source content to power applications.

CCS CONCEPTS

• Human-centered computing → Human computer interaction (HCI);

KEYWORDS

OpenStreetMap, peer production, volunteered-geographic information

1 INTRODUCTION

Over the last 15 years, peer production has seen great success, producing content that has gained extensive use. *Wikipedia*, the most well-known peer production system, currently is the world's fifth-most-visited website [2]. Wikipedia content is also used in the Google Knowledge Graph and other third-party services [27]. Knowledge Graph and Apple's voice assistant Siri use peer-produced structured data from Wikipedia's sister project *Wikidata*. Further, *OpenStreetMap (OSM)* is a peer production community focused on providing open, structured map content. Applications that have taken advantage of OSM's freely available content include Craigslist, Foursquare, and many others.

Several core peer production community principles (e.g. Wikipedia's "Five Pillars" [44]) have been driving forces in ensuring peer production's success. In particular, the principle of **contributor freedom** empowers contributors to edit Wikipedia articles "as they see fit" [15] without being overburdened with complicated rules. OpenStreetMap, "The Wikipedia of Maps" [29], shares a similar attitude towards contributor freedom: *"Nobody is forced to obey [community mapping guidelines]…nor will OSM ever force any of its mappers to do anything."* [13]

In addition to core principles, peer production communities establish guidelines and rules to promote quality and consistency. For example, the pages in the OSM wiki[1] serve as "informal standards" for OSM contributors. Given the prevailing ethos of contributor freedom, these guidelines and rules often require interpretation and generally do not *have* to be followed. There is good reason for such an attitude: when rules have proliferated and their enforcement has grown strict, productivity of peer production communities tends to decline [16]. Given the non-binding nature of the "informal standards", we posed a question:

(RQ) How do OpenStreetMap's 'informal standards' relate to actual contributor practice?

We use the term *standardization* to refer to the process by which OSM contributors orient their practice with the "informal standards" of OpenStreetMap.

We first investigated standardization by analyzing the extent to which OSM practice is consistent with the guidelines in the OSM wiki. We found that most applied metadata (or "tag" data) is in fact consistent with the wiki. However, this analysis revealed a second important observation: due to properties of the OSM data model, many of the guidelines cannot be complied with fully. Specifically, in a number of cases, the wiki accurately identifies multiple appropriate values for a given attribute: for example, a "Dairy Queen" serves both "ice cream" and "burgers". However, the OSM data model restricts each attribute to have a single value. This *ambiguity* is a problem for applications that use OSM data because entities are only partially described. Our analysis also led to a third observation about the OSM standardization process: the wiki guidelines reveal many unmet opportunities for applying metadata. For example, operating hours and phone number data

OpenSym '18, August 22–24, 2018, Paris, France

ACM ISBN 978-1-4503-5936-8/18/08…$15.00
https://doi.org/10.1145/3233391.3233962

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

are identified as relevant to many business entities, but are rarely applied. These data are commonly used when available by popular mapping services such as Yelp and Google Maps, which indicates their user demand.

Our work contributes by shedding light on the nature of standardization in OpenStreetMap as follows:

- Most applied metadata is consistent with the standard.
- The constraints of the OSM data model lead to a large amount of ambiguous metadata.
- The informal standard of the OSM wiki defines large unmet opportunities to apply useful metadata.

In the remainder of the paper, we first discuss related work and provide helpful background about tagging in OSM. We then discuss our specific research context and methods. We next present our results. We conclude by discussing how our findings motivate changes to OSM and peer production more generally. Specifically, we discuss changes related to sociotechnical tools, data model structure, and community informal standards.

2 RELATED WORK

2.1 Contributor Freedom and Community Rules in Peer Production

As we noted, contributor freedom is crucial for the success of peer production. However, the amount of contributor freedom varies by community. In Wikipedia, contributor freedoms have become more limited over time. For example, Wikipedia's response to the expansion that occurred in its first several years was the "creation and clearer articulation of policies" [22]. These came in the form of various types of rules seeking to enable a broad set of views. [22] As Halfaker et al. put it: "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'." [16] While these changes addressed real problems, they also had the harmful side effect of reducing contributor retention [16].

OSM follows the principle of contributor freedom even more than Wikipedia. While Wikipedia has sought to maintain quality content through rules, OSM has instead sought diversity by holding "bureaucracy at bay" with "social and technical approaches" [31]. Fewer rules facilitate OSM newcomer participation [24]. This level of freedom has led some OSM contributors to express a preference for OSM over Wikipedia: "'Wikipedia…feels like Germany, too many rules and regulations.'" [17]. Hence, our work exploring informal standards and their use in OpenStreetMap takes place in a context with an especially strong adherence to the contributor freedom principle.

2.2 Tagging Research in OpenStreetMap

Our study of informal standards focuses on OSM metadata or *tags* [37]. Prior research characterized OSM as "spatially rich, but semantically poor" [4]. For example, similar entities are tagged inconsistently, resulting in "semantic heterogeneity" [39]. One way to improve tag application consistency has been through the use of tag recommenders, e.g. [21, 39].

Some work has sought to measure the quality of the outcome of the OSM tagging process. A frequent technique has been to compare tags to government and commercial data sources, e.g. [12, 25, 41]. For example, several OSM tags were compared with French BD TOPO® highway ground truth data across records for a region of the country [12]. Unfortunately, comparing OSM tag data with sources like this is not always possible and does not scale well. Hence, other work has sought to develop *intrinsic* measures of the quality of tagging, e.g. [6, 19]. One such method [6] considers the mean tag count per OSM record. With this metric, higher averages indicate higher quality tagging and vice-versa.

In our research, we studied tags from a different perspective: we examined how well tagging practice and the OSM wiki align. Several previous tag standardization studies have considered the wiki, e.g. [9, 28, 34]. Our work differs from such studies by systematically analyzing a substantial portion of the wiki to extract tag application "guidelines" and determining the adherence of each tag to the guidelines over a large number of OSM records. For example, we consider whether the tags applied across thousands of McDonald's OSM records are each applied in accordance with wiki guidelines.

Further, an important part of the wiki standards are the prose-heavy descriptions that describe what entity characteristics can be represented through tags. Our robust, large-scale *qualitative and quantitative* approach involved analyzing and interpreting the wiki instructions, including this prose. This analysis aimed to follow the same process that OSM contributors can (if they choose) follow when mapping. This analysis approach is novel in the context of OSM research and led us to identify data standard and data model issues that were not discovered or analyzed in prior work.

We also build upon work identifying challenges in creating or following the OSM wiki [5, 17]. Specifically, the current study complements our own prior work [17], which found that OSM struggles to craft the wiki to represent the views of all its contributors. This is due to problems such as cultural differences and toxic behavior by some contributors. We quantify the effect that such problems have on data standardization. Our prior work also identified the data model/data standard issue that results in what we call *ambiguous* data. We quantify this issue here.

3 OPENSTREETMAP TAGS AND TAGGING STANDARDS

As stated, OpenStreetMap refers to its metadata as *tags* [37]. Similar to other peer production communities such as Wikipedia and Wikidata, OSM implements tags as key-value pairs [37]. These pairs are used for mapping real-world entities such as railroads, businesses, and rivers. For example, consider the tag "amenity=fast_food". The key ("amenity") refers to a specific attribute of a mapped entity and the value ("fast_food") refers to the attribute's value. OSM's data model effectively limits contributors to applying one tag with a given key (e.g. one "amenity" key) for any given mapped entity. Although technically semi-colons can be used to separate multiple values for a given key, semi-colon use is discouraged because applications don't always handle this syntax appropriately [35]. A proposed solution is to provide only the '"primary"' value for an attribute [35]. However, as we will see (e.g., for Dairy Queen cuisine), often there is no single '"primary" value.

We refer to the OSM wiki as the OSM tagging **standard**. As stated, it provides informal tagging guidelines, offering guidance on how to apply tags via pages often consisting of significant amounts of unstructured text. Many such pages are specific to a given key or tag. For example, the wiki describes the tag "amenity=fast_food" as appropriate "for a place concentrating on very fast counter-only service and take-away food." [36]

As we noted, contributors are not required to follow the wiki guidelines. However, the wiki represents common community tagging practices and consensus on how tags are intended to be used. As the community changes and grows, the wiki evolves. Such wiki modifications are often performed in accordance with a tag proposal and voting process [33]. This process determines what new tagging-related content should be added to the wiki.

As mentioned, in our work, we define tag **standardization** as the process of orienting contributor tagging practice with the informal standard of the OSM wiki. "Standardization" refers to the extent that tagging practice unambiguously adheres to the wiki guidelines.

4 MOTIVATION FOR ANALYZING CHAIN BUSINESS STANDARDIZATION

Since the OSM wiki consists of significant amounts of prose tag descriptions and application instructions, comparing this informal standard to tagging practice is hard. It is not practical to manually compare every wiki page with each of the more than 500 million OSM records in our dataset. We therefore needed to find a tractable approach to measuring standardization.

We did this by identifying a large and interesting subset of entities with substantially similar structure, specifically *chain businesses* such as McDonald's, Starbucks, Safeway (a major U.S. supermarket chain), and Wal-Mart. Each individual McDonald's restaurant, Starbucks coffee shop, or Safeway supermarket is similar to every other one. This contrasts to boutique or "one-off" businesses, where each instance is potentially unique, and thus manual information lookup and analysis would be needed to determine whether an OSM representation follows a standard. Because of the standardized nature of chain businesses, all OSM records for say, U.S. McDonald's restaurants (likewise for other chain businesses) should be tagged substantially the same. These observations lead to a tractable process that yields a conservative estimate of standardization: group all OSM records for a chain business; identify the "substantially similar" metadata instances of a chain business should have; and determine whether individual OSM records have that metadata.

Focusing on chain businesses yields an approach that scales, but analyzing these businesses also has additional value. First, they are popular: McDonald's accounts for over 17% of the fast food market share in the United States [11], and over 40% of Americans visit a Wal-Mart each week [1]. Second, chain businesses have been under-studied by geographic HCI researchers and the broad social computing community, with most projects focusing on the discovery of boutique venues. Third, fast food restaurants and convenience stores (both of which we analyzed) are more prevalent in low SES areas [23]. Chain businesses are therefore important to the populations of those areas, who tend to be underserved in peer production [14, 20]. Finally, and critically, if OSM contributors cannot apply tags in a standardized way to real-world entities that are in fact highly standardized, it is unlikely that less similar real-world entities will be standardized well in OSM either.

5 METHODS

5.1 Clustering Algorithm

OSM does not provide a widely adopted formal means to link together different instances of the same business (or any other conceptual category). Therefore, to analyze standardization of chain businesses, we first needed to *extract* chain businesses from the OSM dataset, which we did by developing a clustering procedure. We handled inconsistencies in OSM representations of businesses through a hybrid clustering approach that combined automated algorithms with manual verification and coding. We detail our procedure next.

5.1.1 Selecting OSM Instances for Analysis. We used United States OSM data records from February 2014 that were available from [18]. Although our initial data contained records from outside the U.S., we used a U.S. census Tiger shapefile [8] to filter out these records. Our dataset contained the current state of all OSM data records (node and way objects) in the 50 U.S. states. This included roads, bodies of water, and other entities. We limited records to the U.S. because our manual coding process required familiarity with the business data, and all our coders are from the United States. We removed non-business records by filtering based on tags. For example, we identified non-business tags (e.g., "amenity=university") through manual inspection of the dataset and removed all records with these tags. This initial filtering step did not remove all irrelevant data; subsequent normalization and clustering steps were necessary.

5.1.2 Normalizing Instance Names. A naive approach to clustering would group all records with the same value for the "name=" tag. However, a "name" can appear inconsistently; for example, McDonald's locations have names ranging from the standard "McDonald's" to "McDonald's – East Liberty Station" to misspellings and variations in capitalization. We reduced these inconsistencies by 1) normalizing tag case, and 2) using

Wikipedia redirects to help recognize naming variations of the same business. Wikipedia redirects link common name variations in Wikipedia searches to a central article. For example, Wikipedia redirects a search for "Chik-fil-A" (a fast food restaurant) to the article entitled "Chick-fil-A". Our method sought to capture naming variations similar to these by making the "normalized" name field available to the automated clustering algorithm discussed next.

5.1.3 Automated Clustering + Post-Processing. We used a semi-supervised clustering algorithm [7] to further improve clustering results. The algorithm clustered instances based on their tags (including the "name=" tag and the redirect name created in the previous step). We manually selected the 50 largest business clusters, which represented some of the most common businesses in the United States.

We next performed a series of manual steps to ensure the precision of clusters. First, we combined several clusters that represented the same business (e.g. three McDonald's clusters, two CVS pharmacy clusters, etc…). Second, we only retained instances that had the 'standard' name for a business ("mcdonald's"), or small variations ("mcdonalds" or "mcdonald's - east liberty station"). This process resulted in the 42 distinct clusters shown in Table 1.

We explicitly highlight that our clustering process reflected a need for high cluster precision that was essential to the accuracy of our tagging standardization analysis. This is because the goal of the analysis was to compare tags applied to the instances of a business cluster – say McDonald's – to the wiki instructions that describe when those tags are appropriate. The comparison only made sense if all instances in the cluster did in fact represent McDonald's instances. If say 25 instances in the "McDonald's" cluster should belong to a "Safeway" cluster instead, we might falsely conclude that the tag "shop=supermarket" applied to those instances was applied in a way that was misaligned with wiki instructions. As noted previously, to avoid this problem we manually inspected "name" tags in our clusters to ensure instances were placed in appropriate clusters. Although we prioritized precision over recall, we note that our clustering approach identified some of the most common tagging practices for each business we analyzed.

After clustering was complete, our largest cluster was McDonald's, with 3424 instances. Our smallest was Sonic (a fast food restaurant), with 169. The mean number of instances across all clusters was 672 (s.d. = 674) and the median was 343. Across all clusters, there were 28,420 business instances total.

As mentioned, chain businesses are inherently standardized in the real world because instances of a given business share many characteristics (e.g. all Dairy Queen locations serve ice cream, all Starbucks have operating hours, many McDonald's have a drive through). In our analysis, we focused on the tags corresponding to these inherent similarities. By focusing on the metadata that represents inherently standardized attributes of entities, our analyses should provide an upper bound of their standardization. Given these considerations, we removed, for example, tags related to the specific address of an instance (street address, city name, etc.) and miscellaneous notes pertaining to the instance (e.g. "created_by", "note", "attribution", etc.).

Table 1: Chain Businesses Used in Standardization Analyses

Chain Business						
7-Eleven	Best Buy	CVS	Home Depot	Panda Express	Sam's Club	Taco Bell
Applebee's	Burger King	Dairy Queen	IHOP	Panera Bread	Sonic	Wal-Mart
Arby's	Chevron	Denny's	Jack in the Box	Pizza Hut	Staples	Walgreens
AutoZone	Chick-fil-A	Dollar Tree	KFC	RadioShack	Starbucks	Wells Fargo
Bank of America	Circle K	Dunkin' Donuts	McDonald's	Rite Aid	Stewart's	Wendy's
Barnes & Noble	Culver's	H-E-B	Olive Garden	Safeway	Subway	Whataburger

Further, to ensure manual coding was tractable, we selected the 10 most applied keys for each business and their associated values; this resulted in 41 distinct keys and 416 distinct business and key combinations, or "business-key pairs", collectively comprising over 94% of the remaining metadata in our clusters. Since we chose the most applied keys, this data also represented the most common tagging norms in terms of key applications in each respective business cluster.

5.2 Determining a Metadata Taxonomy

Different tags in the OSM wiki serve different descriptive roles. Certain tags are *appropriate for all instances of a given business*. Examples include tags like "amenity= fast_food" for McDonald's. Other tags contain a key that is *appropriate for all instances of a given business, but whose value is instance-specific*. This includes tags such as "opening_hours=<some operating hours value>" in the case of many businesses. Finally, other tags are *appropriate for some – but not all – instances of a given business*. This includes tags such as "drive_through=yes" to indicate the presence of a drive through at McDonald's. We developed a taxonomy to account for these different types of metadata. This taxonomy provides a foundation for evaluating the community's standardization process. We defined three classes of metadata:

- *Universal* metadata describe key-value pairs appropriate for all instances of a business. All U.S. Starbucks have the same brand, so all Starbucks instances can be tagged "brand=starbucks". The "brand" attribute has one value for all Starbucks instances.
- *Universal-Varying* metadata describe keys appropriate for all instances of a business, but whose values are instance-specific. All McDonald's locations have an operating hours attribute which can be denoted in OSM with the "opening_hours" key. The specific value appropriate for the key representing that attribute varies across instances of McDonald's.
- *Contingent* metadata describe real-world variation, i.e., keys that may or may not apply to any given instance of a chain business since the attribute they represent may

or may not be present (we discuss metadata describing nonexistent attributes later). For example, some McDonald's locations have drive-through windows, and some do not, some are wheelchair accessible, and some are not, etc.

We categorized each key associated with a business as Universal, Universal-Varying, or Contingent. We did this categorization for keys (not tags), since a key can have only one value for a given OSM record, so, for example, the "amenity" key could not be both Universal and Universal-Varying.

To categorize keys, we systematically analyzed the OSM wiki page for each key, keeping in mind the context of each business it was applied to. Each key was placed into a single category (Universal, Universal-Varying, or Contingent) based on its role for the business. To ensure reliability of this qualitative process, the first and second authors classified the keys independently and then resolved disagreements. See Table 2 for the results[2].

6 RESULTS

6.1 Measuring Standardization

We next systematically compared tag data in each of our clusters to corresponding pages in the OSM wiki, thus assessing standardization. The first and second authors carried out the coding procedures for this process. The procedure varied by metadata type.

6.1.1 Universal Metadata. For each tag in each cluster, the coders analyzed corresponding wiki key and tag page descriptions. The coders performed this process to consider the tags' appropriateness for the business instances they were applied to. For example, the wiki indicated that the tag "amenity=fast_food" would be appropriate for McDonald's cluster instances but not for Safeway cluster instances.

Although 1592 distinct key-values for Universal keys were applied in our dataset, we narrowed our focus by selecting the 10 most common values for each key for our coding process[3]. Remaining values were considered applications that did not align with wiki instructions. We believe selecting the 10 most common values was reasonable, since this included all key-values that appeared more than once within a business (with two exceptions: one 11th-most-popular value was applied twice, the other was applied thrice)[4]. This coding process identified 133 business-Universal key pairs with at least one appropriate (according to the wiki) value. We used metadata associated with these pairs for Universal metadata standardization analyses.

This process showed that some applied metadata did not align with the wiki. For example, "shop=supermarket" was applied to 8 instances of the pharmacy CVS. The wiki states that "shop=supermarket" is for "a full service grocery store" [38]. Given this, and given the coders' knowledge of CVS locations in the United States, it was clear that this tag was not appropriate for CVS instances. We classified such applications as **misaligned** since they did not align with wiki instructions. We consider *misaligned* metadata to be *unstandardized* metadata.

Table 2: Chain Business Metadata (Key) Role Classes

Universal	Universal-Varying	Contingent	
shop	ref:store_number	drive_through	delivery
amenity	building:levels	fax	smoking
contact:website	opening_hours	wifi	outdoor_seating
alt_name	contact:phone	area	contact:fax
cuisine	phone	operator	wheelchair
drive_in	building	fuel:diesel	atm
website		motorcar	dispensing
brand		fuel:octane_91	landuse
url		highway	internet_access
takeaway		entrance	

Many other tag applications were in alignment with the wiki instructions. For example, we observed both "amenity=fast_food" and "amenity=cafe" applied to different Panera Bread restaurants. Careful reading of the wiki suggested that both tags were appropriate. The wiki page for "amenity=fast_food" says that this tag should be used "for a place concentrating on very fast counter-only service and take-away food." and "They usually, but not always, have sit-down facilities ranging from two or three to many easy-to-clean chairs and tables." The wiki page for "amenity=cafe" describes a café as "a generally informal place with sit-down facilities selling beverages and light meals and/or snacks." Both tags provide accurate and useful descriptive information about Panera Bread instances and were applied consistently with the wiki instructions. However, due to OSM's one-key-one-value data model, only one of the values could be applied to a given Panera Bread instance. Hence, applications of either of these tags were considered **ambiguous**. More generally, whenever at least two distinct instances of the same business had different values for the same key and each value aligned with the wiki instructions, we considered those tag applications to be **ambiguous**. We consider *ambiguous* metadata to be *unstandardized* metadata.

6.1.2 Location-Specific Metadata: Universal-Varying and Contingent. We found that very few **Universal-Varying keys** actually were applied to appropriate business instances. For example, "opening_hours" was Universal-Varying for Walgreens and other businesses, and thus was appropriate for all of them. However, only 3% of Walgreens had this key applied, and this trend was consistent for other businesses, too. A similar scenario played out for phone number metadata. Across all Universal-Varying metadata, 88% of *potential* metadata was unapplied. Note

[2] 5 keys were removed because both coders agreed they were not relevant (3 were not in the wiki, 1 was not business related, the final key "ref:arbys", was removed since it was for Arby's restaurants only).
[3] We coded *all* tags for website-related keys. Further details of website analyses are discussed in Detailed Results.
[4] Regardless, most data aligned with the wiki anyway.

that mapping applications such as Google Maps provide this data when it is available, indicating there is user demand for this location-specific content. OSM severely lacks this type of metadata, limiting its utility as a data providing source.

A likely reason for the lack of Universal-Varying metadata is that applying it requires more work from contributors than business-wide (Universal) metadata does; contributors have to look up information for each individual business location. This extra work may be more than contributors are willing to do; our prior research has shown that contributors limit their effort when tagging individual records [17].

Contingent metadata was even more rarely applied than Universal-Varying metadata. 94% of *potential* metadata was unapplied. Determining this was more complex than for Universal-Varying metadata since Contingent metadata only applies to some instances of a given business. (Although metadata is sometimes applied to indicate the lack of an attribute's existence, this was not common in our dataset.) Hence, the effort of determining if an attribute represented with Contingent metadata *is present* is possibly a reason why even less was applied. Given our need to look up location-specific information about Contingent metadata, we sampled an important and representative subset. For more details of this sampling process and of our rationale, see the Appendix.

Given that Universal-Varying and Contingent metadata was so rarely applied when it was appropriate, we focused our remaining analysis on Universal metadata – 38,220 Universal business-key-values. We return to Universal-Varying and Contingent metadata when discussing important opportunities for the community to improve the number of tag applications.

6.2 Detailed Results

6.2.1 Universal Metadata Standardization. Recall that Universal metadata were key-values (tags) that were universal to instances of a given business. Figure 1 illustrates the results for Universal metadata standardization. There were 38,220 applied Universal business-key-value triples. Only 3706 business-key-value triples did *not* align with wiki instructions. Thus, 90% of applied metadata aligned with wiki instructions.

However, out of the remaining 34,514 aligned triples, 76 of 133 Universal key-business pairs were ambiguous, leading to 18,841 ambiguous triples (49% of all triples). Thus, while most tag applications complied with the wiki, a significant amount of applied metadata was ambiguous. The result was that 15,673 triples were aligned and not ambiguous: that is, *only 41% of metadata did not have standardization issues*.

6.2.2 Universal Standardization Failures through Different Lenses. We found that standardization of keys varied quite a bit, with a common pattern: keys whose OSM specifications are less clear are more likely to be misaligned. We discuss details next. We also observed that standardization of businesses depends largely on the keys applied to them; if keys are problematic, the businesses will be, too. Thus, analyzing standardization by businesses provided little new insight, so we do not discuss that dimension further.

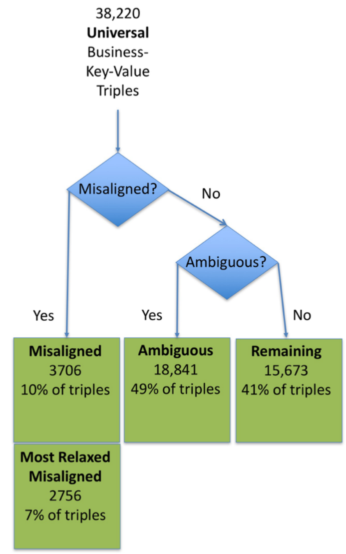

Figure 1: Universal Business-Key-Value Triples

Misalignment by key. Business website information (represented in our dataset by the keys "contact:website" and "website") is prevalent in various mapping applications including those from Google, Bing, and Yelp – indicating its demand. Our initial analysis found that 63% of the website data was misaligned. As we coded the wiki, however, we observed that the wiki specifications for how to enter URLs were hard to interpret and contained very specific formatting instructions. Hence, many URLs were close to being aligned, with only small syntax problems, and most URL variations were infrequently applied. Although many URLs did not align precisely with the wiki, web browsers and other applications can handle a range of variations and still retrieve the appropriate web page. We applied some simple normalizations to the URLs in our dataset, which reduced the misalignment. We then took a further step: checking whether URLs navigated to a working website. This was the case for 93% of URLs in our dataset. Of the remaining 7%, most generated an HTTP 403 or 404 error, likely indicating that these URLs were not kept up-to-date as of the time we checked (April 2017). Figure 1 includes this most relaxed version of misalignment for website metadata. To sum up, from a strict syntactic perspective, most website data was misaligned, but from a practical perspective, nearly all website metadata was in fact *aligned*.

We also further examined non-website-related misaligned metadata to check if these tags were simply typos (i.e. slight and obvious misspellings of *aligned* metadata). 98% of misaligned metadata were not a result of typos; *instead, the errors were due to substantive misalignments with wiki instructions. These tags were intentionally applied when the wiki does not indicate that they should be*.

Ambiguity by key. Four Universal keys were sometimes ambiguous: "amenity", "cuisine", "shop", and "website". Table 3 provides details about the businesses each key was ambiguous for.

“amenity” was ambiguous for 9 of 28 businesses it was applied to. “cuisine” was ambiguous for almost all – 21/22 – the businesses it was applied to. “website” was ambiguous for 35 out of 41 businesses it was applied to.

There are a couple of interesting implications from these results. First, the one-key-one-value data model restriction appears to be particularly incompatible with attributes such as restaurant cuisine. Essentially all restaurants had multiple types of cuisine, *but a given instance could only have one of the types specified*. Thus, applications using the data may be unaware that a given restaurant has multiple types of cuisine. Second, it is possible that some ambiguity (especially in the case of website metadata) is due to the hard-to-interpret instructions in the wiki that were discussed previously. Thus, clarifying wiki instructions is important if the community would like to improve metadata consistency. Performing work focused on understanding the degree to which the OSM wiki is universally understood (and informed by sociolinguistic theory on achieving common ground) is an important first step in this direction. Improving consistency would also allow the various applications using OSM data to more easily process data.

6.2.3 Missed Opportunities to Apply Metadata. In addition to studying standardization, we noticed significant missed opportunities to apply useful metadata. We discuss these next.

Universal Metadata. If all appropriate Universal metadata was applied to the businesses in our dataset (e.g., if all Dairy Queens had cuisine information or all Wal-Marts had website information), the amount of applied Universal metadata would increase from 38,220 to 95,926 Universal business-key-values, or by over 250%.

The amount of missed opportunities for Universal metadata applications varied widely across keys: mean = 76%, s.d. = 33%. Only the “amenity” key was applied with great consistency; just 4% of business instances where the wiki deemed this metadata appropriate (e.g. “amenity=fast_food” for McDonald’s) did not have it. There are several possible reasons why “amenity” is applied consistently: 1) “amenity” is the “primary” point of interest (e.g. chain business) key according to Over et al. [30], 2) applications such as OsmAnd[5] appear to use “amenity=” tags to render icons, and 3) our method of filtering records (discussed in the Clustering Algorithm section of Methods) may have favored records with this key. Specifically, records chosen for further analysis either had “amenity=”, “shop=”, or “cuisine=” applied to them, or had the same “name=” tag as a record that did. We did this to remove the large amount of irrelevant records from our sample.

As our previous research has shown [17], OSM contributors have said that they just “basically” characterized objects with the “minimum” information; “it’s too much work to add everything”. Our results align with this observation: “amenity” is precisely the type of “minimum” “basic” information likely to be provided for an entity. The other Universal keys all had substantial missed opportunities to apply metadata; for example, 94% of *potential* “website” key applications did not exist. In the Discussion section, we consider ways to improve metadata application while still respecting OSM contributor values and attitudes.

Table 3: Business-Universal Pairs with 2 or More Aligned Values

amenity	Bank of America	Dairy Queen	Denny's	Dunkin' Donuts	IHOP	McDonald's
	Panera Bread	Starbucks	Wells Fargo			
cuisine	Applebee's	Arby's	Burger King	Chick-fil-A	Culver's	Dairy Queen
	Denny's	Dunkin' Donuts	IHOP	Jack in the Box	KFC	McDonald's
	Olive Garden	Panda Express	Panera Bread	Pizza Hut	Sonic	Starbucks
	Subway	Taco Bell	Wendy's			
shop	Chevron	CVS	Dairy Queen	Dollar Tree	Home Depot	Radio Shack
	Rite Aid	Sam's Club	Staples	Walgreens	Wal-Mart	
website	7-Eleven	Applebee's	Arby's	AutoZone	Bank of America	Barnes & Noble
	Burger King	Chick-fil-A	Circle-K	CVS	Dairy Queen	Denny's
	Dollar Tree	Dunkin' Donuts	Home Depot	IHOP	Jack in the Box	McDonald's
	Olive Garden	Panda Express	Panera Bread	Pizza Hut	Radio Shack	Rite Aid
	Safeway	Sam's Club	Staples	Starbucks	Subway	Taco Bell
	Walgreens	Wal-Mart	Wells Fargo	Wendy's	Whataburger	

Universal-Varying Metadata. Recall that Universal-Varying metadata were keys that were universal to instances of a given business, along with values that were location-specific. If all Universal-Varying metadata was applied to every instance of the respective businesses they belonged to (e.g. if all Olive Garden restaurants or Walgreens had operating hours information), the amount of applied Universal-Varying metadata would increase from 9,319 to 75,591 business-key-value, an increase of over 810%.

There was less variation in metadata application between different Universal-Varying keys compared to different Universal keys: nearly all appropriate Universal-Varying metadata was left unapplied. For example, two Universal-Varying keys – for operating hours (“opening_hours”) and for phone number (“phone”) – were applied to fewer than 5% of businesses they could be applied to. The information this metadata provides is very useful for potential customers, as evidenced by its use when available in applications such as Google and Bing Maps and Yelp. The absence of this data reduces the usefulness of mapping applications that use this information.

It makes sense that less Universal-Varying data would be applied than Universal data: determining the proper value for a Universal-Varying key for a given instance is a non-trivial task since location-specific information is needed. To illustrate the

[5] http://osmand.net

effort required, determining the opening hours for a specific Starbucks location requires looking up the information on the web. Obtaining other location-specific information may even require physically visiting the actual location (which is not part of OSM's common remote "armchair mapping"[6]).

Contingent Metadata. Recall that Contingent metadata are key-values describing attributes that are *not* universal to instances of a given business. As mentioned previously, we sampled important and representative Contingent metadata. Specifically, we considered the "internet_access" key for McDonald's and Starbucks and the "drive_through" key for McDonald's, Starbucks, and Walgreens. Google Maps and Yelp also use these types of information when available – again, indicating this data is important and in demand. Based on the results of our sampling process, if all Contingent metadata was applied to every instance of the respective businesses they belonged to (e.g., if all McDonald's containing a drive-through or all Starbucks with internet access had corresponding metadata applied), the amount of applied Contingent metadata in our dataset would increase from 14 to 245 Contingent business-key-values, an increase of 1750%.

Each business-key sampled had missed opportunities to apply metadata at least 90% of the time that it was appropriate. As mentioned previously and as discussed in the Appendix, our samples were likely among the most applied Contingent metadata. The key "internet_access" for McDonald's locations had the largest amount of missed opportunities, missing them 98% of the time. All but one of the McDonald's sampled had internet access in reality – but there was no metadata to show for it.

7 DISCUSSION

We summarize our core findings here. We found:

- The OSM community does a good job of applying data that is aligned with the wiki instructions.
- The one-key-one-value OSM data model restriction results in a very significant amount of ambiguous applied metadata.
- A significant number of opportunities to apply metadata are missed.

Based on these findings, we next provide implications for OSM and peer production more generally.

Increasing precision in data standards. Related to ambiguity and misalignment**,** the discrepancy in the level of "structure" between the OSM standard (wiki) and the data itself leaves room for interpretation. Similar to the observations of prior work [3, 32], sometimes wiki descriptions are quite general and hard to interpret (as was seen for website metadata). This issue may be due to an effort to make the definitions globally applicable and relatable across languages; after all, contributors try to create one global tagging standard. However, this leaves room for contributors to tag the same thing in different ways. It may make sense for the community to consider alternative definitions and descriptions for metadata. Increased use of tagging examples (e.g. of business-specific examples) could leave less room for interpretation. Further, more pictures in wiki descriptions, as suggested in our previous work [17], may mitigate this problem. Here, it's worth stating that since Wikidata is also a data repository with a single language-independent version, it may have similar problems and could potentially benefit from similar solutions.

OSM data model. While the above data standard changes might help reduce metadata misalignment/ambiguity, ambiguity stemming from the data model is still problematic. As we noted previously [17], the data model could change to account for entities in the real-world that have multiple values for different attributes. A Dairy Queen specializes in both burgers and ice cream for its cuisine, and a contributor should not need to choose just one. This data model change would improve end-user experience since applications would have access to all information on OSM entities. This proposed data model has been shown to work in similar peer production communities: both Wikipedia and Wikidata have avoided OSM's data model issue by opting for multiple values per key in their structured data.

Metadata that is harder to apply or requires frequent maintenance is less likely to be applied. Based on our data and analyses of missed opportunities to apply metadata, 60% of *potential* Universal metadata, 88% of *potential* Universal-Varying metadata, and 94% of sampled *potential* Contingent metadata was not applied. This suggests that as metadata becomes more variable -- and thus, requires more work to apply – it will be applied less. Additionally, location-specific metadata requires more frequent updates than Universal metadata. For example, it is more likely that one specific Subway shop's operating hours will change than it is for the cuisine of all Subways to change. Likewise, it is more likely that a Walgreen's location will add or remove a drive-through than it is for all Walgreen's to become something other than a pharmacy. Indeed, discussion with OSM contributors has indicated that the need to maintain metadata is a deterrent from applying it in the first place. Enabled by the core community value of contributor freedom, OSM contributors limit their tagging effort [17], and this shows in the form of lower percentages of potential location-specific (Universal-Varying and Contingent) metadata being applied.

Given these considerations, it might be worth pursuing new ways to use automation for tagging. A possible option that was also discussed by the authors in previous work [17] would be to integrate data entry tools with businesses' databases. While prior research (e.g. [10, 20, 42]) has shown the negative effects data imports and remote or non-local work can have on data quality, businesses are naturally incentivized to input and maintain accurate and detailed metadata for their locations. Of course, creating the code to facilitate business data imports would put an initial added burden on OSM contributors; however, we believe that in the end, this approach would ease the burden of getting business metadata into OSM.

[6] https://wiki.openstreetmap.org/wiki/Armchair_mapping

Interestingly, the idea of businesses updating their own data has been considered in other peer production contexts, including Wikipedia and Wikidata. In fact, Wikipedia has a "conflict of interest"[7] policy preventing businesses from doing this. However, this is not true for Wikidata, a community that, like OSM, focuses on producing structured data instead of prose. Discussion in Wikidata around creating a similar conflict of interest policy to Wikipedia indicated a feeling that "since Wikidata does not allow for natural language, a lot of nuance and opportunity for bias goes away." [40] Given the similarities between Wikidata and OSM and the views that the Wikidata community has, it might be reasonable for OSM to follow suit and allow businesses to update their own metadata.

8 LIMITATIONS AND FUTURE WORK

Our analysis of OSM wiki pages focused on the key and tag pages corresponding to metadata applied to instances of our chain business clusters. We note that eight of the businesses in our study currently have their own OSM wiki pages with business-specific tagging instructions. However, these pages are not widely adapted across businesses. Second, none of these pages actually existed when we obtained our dataset – hence, contributors could not follow them. Third and most importantly, these business pages provide specific tagging instructions, typically providing only one appropriate value for a given key. Given the one-key-one-value data model, this makes sense to do. However, the advantage to our approach is that it helps provide an understanding of the extent to which this data model constraint results in incomplete representations of the attributes of businesses. Our approach also more conservatively estimates misalignment, only considering a tag application to be misaligned if it is not helping to describe the entity that it is applied to (based on information regarding that tag in the wiki). Hence, we gave the benefit of the doubt to contributors when calculating misalignment.

As mentioned, the OSM wiki represents an ever-changing and expanding community tagging standard. As new keys and tags are added, opportunities to apply metadata increase accordingly. Because of this evolution, it's important to note that not all the identified missed tagging opportunities were necessarily considered missed opportunities at earlier points in OSM's history. For instance, a contributor might have applied much of the relevant metadata to a McDonald's record when mapping it in 2010. However, by 2015, many opportunities might be missed for that record if additional relevant tags were introduced in the wiki but were not applied to the record. Future work should explore how the level of missed tagging opportunities has changed as OSM has matured.

"Coverage", or the degree to which OSM provides data describing the real-world, is a commonly used lens for considering OSM data quality. While both our work and prior work has considered missed coverage opportunities, prior work (e.g. [14, 26, 43]) measured such missed opportunities by considering how often objects (restaurants, highways, etc.) from the real world are represented by objects in OpenStreetMap. We, however, quantified coverage by instead considering missed opportunities to apply metadata for objects that do exist. Some work in OSM has provided evidence that coverage biases exist along dimensions such as population density [26]. Future OpenStreetMap work might consider whether similar biases occur with our definition of coverage, particularly given the substantial impact of these chain businesses in low-SES and rural areas as noted above.

[7] https://en.wikipedia.org/wiki/Wikipedia:Conflict_of_interest

9 CONCLUSION

We studied the relationship between tagging practice and "informal standards" in OSM through a novel approach involving qualitative and quantitative methods. Although the ethos of contributor freedom is strong in OSM, contributors generally do follow standards. However, we uncovered a significant standardization issue largely related to the OSM data model's inability to represent certain entities accurately. Further, we also found many missed opportunities to apply metadata in OSM. Some of these opportunities would help OSM become a better source of open content for applications. We concluded with implications for the OSM community and peer production more generally.

ACKNOWLEDGEMENTS

We thank Allen Lin and Morten Warncke-Wang for assistance and feedback in this research. The research was partially funded by the United States National Science Foundation (NSF IIS-1527173; NSF IIS-1707319; NSF IIS-1218826; NSF IIS-1111201; NSF IIS-964695).

REFERENCES

[1] 19 Facts That Show Just How Massive Walmart Really Is: *http://www.buzzfeed.com/sapna/19-facts-that-show-just-how-massive-walmart-really-is*. Accessed: 2016-08-09.

[2] Alexa Top 500 Global Sites: *https://www.alexa.com/topsites*. Accessed: 2018-01-24.

[3] Ali, A.L. et al. 2014. Ambiguity and plausibility: managing classification quality in volunteered geographic information. *Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (2014), 143–152.

[4] Ballatore, A. and Bertolotto, M. 2011. Semantically Enriching VGI in Support of Implicit Feedback Analysis. *Web and Wireless Geographical Information Systems*. K. Tanaka et al., eds. Springer Berlin Heidelberg. 78–93.

[5] Ballatore, A. and Mooney, P. 2015. Conceptualising the geographic world: the dimensions of negotiation in crowdsourced cartography. *International Journal of Geographical Information Science*. 29, 12 (2015), 2310–2327.

[6] Barron, C. et al. 2014. A Comprehensive Framework for Intrinsic OpenStreetMap Quality Analysis. *Transactions in GIS*. 18, 6 (Dec. 2014), 877–895. DOI:https://doi.org/10.1111/tgis.12073.

[7] Bilenko, M.Y. 2006. *Learnable similarity functions and their application to record linkage and clustering*.

[8] Census.gov: *https://www.census.gov/en.html*. Accessed: 2018-06-05.

[9] Davidovic, N. et al. 2016. Tagging in Volunteered Geographic Information: An Analysis of Tagging Practices for Cities and Urban Regions in OpenStreetMap. *ISPRS International Journal of Geo-Information*. 5, 12 (2016), 232.

[10] Eckle, M. and de Albuquerque, J.P. 2015. Quality Assessment of Remote Mapping in OpenStreetMap for Disaster Management Purposes. *ISCRAM* (2015).

[11] Fast food industry market share in the U.S. 2015 | Statistic: *http://www.statista.com/statistics/196611/market-share-of-fast-food-restaurant-corporations-in-the-us/*. Accessed: 2016-08-08.

[12] Girres, J.-F. and Touya, G. 2010. Quality Assessment of the French OpenStreetMap Dataset. *Transactions in GIS*. 14, 4 (Aug. 2010), 435–459. DOI:https://doi.org/10.1111/j.1467-9671.2010.01203.x.

[13] Good practice - OpenStreetMap Wiki: *http://wiki.openstreetmap.org/wiki/Good_practice*. Accessed: 2016-08-08.

[14] Haklay, M. 2010. How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environment and planning. B, Planning & design*. 37, 4 (2010), 682.

[15] Halfaker, A. et al. 2011. Don't bite the newbies: how reverts affect the quantity and quality of Wikipedia work. *WikiSym* (2011), 163–172.

[16] Halfaker, A. et al. 2012. The rise and decline of an open collaboration system: How Wikipedia's reaction to popularity is causing its decline. *American Behavioral Scientist*. (2012), 0002764212469365.

[17] Hall, A. et al. 2017. Freedom versus standardization: structured data generation in a peer production community. *CHI* (2017), 6352–6362.

[18] Index of /full-history-extracts/latest/continents: *http://osm.personalwerk.de/full-history-extracts/latest/continents/*. Accessed: 2018-06-03.

[19] Jilani, M. et al. 2014. Automated highway tag assessment of OpenStreetMap road networks. *Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (2014), 449–452.

[20] Johnson, I.L. et al. 2016. Not at home on the range: Peer production and the urban/rural divide. *CHI* (2016), 13–25.

[21] Karagiannakis, N. et al. 2015. OSMRec Tool for Automatic Recommendation of Categories on Spatial Entities in OpenStreetMap. *Proceedings of the 9th ACM Conference on Recommender Systems* (2015), 337–338.

[22] Kriplean, T. et al. 2007. Community, consensus, coercion, control: cs* w or how policy mediates mass participation. *Proceedings of the 2007 international ACM conference on Supporting group work* (2007), 167–176.

[23] Lee, H. 2012. The role of local food availability in explaining obesity risk among young school-aged children. *Social Science & Medicine*. 74, 8 (Apr. 2012), 1193–1203. DOI:https://doi.org/10.1016/j.socscimed.2011.12.036.

[24] Lin, Y.-W. 2011. A qualitative enquiry into OpenStreetMap making. *New Review of Hypermedia and Multimedia*. 17, 1 (2011), 53–71.

[25] Ludwig, I. et al. 2011. A Comparison of the Street Networks of Navteq and OSM in Germany. *Advancing geoinformation science for a changing world*. Springer. 65–84.

[26] Mashhadi, A. et al. 2013. Putting Ubiquitous Crowd-sourcing into Context. *Proceedings of the 2013 Conference on Computer Supported Cooperative Work* (New York, NY, USA, 2013), 611–622.

[27] Medelyan, O. et al. 2009. Mining meaning from Wikipedia. *International Journal of Human-Computer Studies*. 67, 9 (Sep. 2009), 716–754. DOI:https://doi.org/10.1016/j.ijhcs.2009.05.004.

[28] Mooney, P. and Corcoran, P. 2012. The Annotation Process in OpenStreetMap. *Transactions in GIS*. 16, 4 (Aug. 2012), 561–579. DOI:https://doi.org/10.1111/j.1467-9671.2012.01306.x.

[29] OpenStreetMap: "It's the Wikipedia of maps": 2012. *http://www.theguardian.com/theobserver/2012/feb/18/openstreetmap-world-map-radicals*. Accessed: 2018-03-08.

[30] Over, M. et al. 2010. Generating web-based 3D City Models from OpenStreetMap: The current situation in Germany. *Computers, Environment and Urban Systems*. 34, 6 (Nov. 2010), 496–507. DOI:https://doi.org/10.1016/j.compenvurbsys.2010.05.001.

[31] Palen, L. et al. 2015. Success & scale in a data-producing organization: the socio-technical evolution of OpenStreetMap in response to humanitarian events. *Proceedings of the 33rd annual ACM conference on human factors in computing systems* (2015), 4113–4122.

[32] Pazoky, S.H. et al. 2015. YOU DESCRIBE IT, I WILL NAME IT: AN APPROACH TO ALLEVIATE THE EFFECT OF USERS'SEMANTICS IN ASSIGNING TAGS TO FEATURES IN VGI. *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences*. 40, (2015).

[33] Proposal process - OpenStreetMap Wiki: *https://wiki.openstreetmap.org/wiki/Proposal_process*. Accessed: 2018-06-10.

[34] Quattrone, G. et al. 2017. Work Always in Progress: Analysing Maintenance Practices in Spatial Crowd-sourced Datasets. (2017).

[35] Semi-colon value separator - OpenStreetMap Wiki: *http://wiki.openstreetmap.org/wiki/Semi-colon_value_separator*. Accessed: 2016-12-05.

[36] Tag:amenity=fast_food - OpenStreetMap Wiki: *https://wiki.openstreetmap.org/wiki/Tag:amenity=fast_food*. Accessed: 2018-05-28.

[37] Tags - OpenStreetMap Wiki: *https://wiki.openstreetmap.org/wiki/Tags*. Accessed: 2018-06-03.

[38] Tag:shop=supermarket - OpenStreetMap Wiki: *https://wiki.openstreetmap.org/wiki/Tag:shop%3Dsupermarket*. Accessed: 2018-06-03.

[39] Vandecasteele, A. and Devillers, R. 2015. Improving Volunteered Geographic Information Quality Using a Tag Recommender System: The Case of OpenStreetMap. *OpenStreetMap in GIScience*. J.J. Arsanjani et al., eds. Springer International Publishing. 59–80.

[40] Wikidata:Requests for comment/Conflict of Interest - Wikidata: *https://www.wikidata.org/wiki/Wikidata:Requests_for_comment/Conflict_of_Interest*. Accessed: 2018-03-06.

[41] Zhang, H. and Malczewski, J. 2017. Quality Evaluation of Volunteered Geographic Information: The Case of OpenStreetMap. *Volunteered Geographic Information and the Future of Geospatial Data*. IGI Global. 19–46.

[42] Zielstra, D. et al. 2013. Assessing the effect of data imports on the completeness of OpenStreetMap–a United States case study. *Transactions in GIS*. 17, 3 (2013), 315–334.

[43] Zielstra, D. and Zipf, A. 2010. A comparative study of proprietary geodata and volunteered geographic information for Germany. *13th AGILE international conference on geographic information science* (2010).

[44] 2017. Wikipedia:Five pillars. *Wikipedia*.

A APPENDIX: SAMPLING CONTINGENT METADATA

We sampled Contingent keys (specific samples discussed below), and consulted each business's website store locator to determine whether the attribute represented by the key was *present* for a specific business location. If the attribute was present, we checked if metadata indicating that attribute existed. For example, we checked if a particular McDonalds had a drive-through and -- if it did -- if it then had metadata to represent that attribute.

We checked for attribute presence since tags may sometimes indicate a non-existent attribute. For example, the OSM wiki states that it is valid to explicitly indicate that a fast food restaurant does not have a drive through: "drive_through=no". It's important to note that in our analysis of "missed tagging opportunities", we did not consider it to be a "missed opportunity" if a tag for a non-existent attribute was unapplied. This is because this situation appears fairly rare in practice and tag application in this scenario is not necessarily described in the wiki. Thus, we chose a conservative interpretation of OSM's ontology when defining "missed opportunities".

To sample Contingent metadata, we manually selected the key "internet_access" for McDonald's and Starbucks and the key "drive_through" for McDonald's, Starbucks, and Walgreens. We chose these businesses and keys because the businesses were among the United States' leaders in their market categories and because the two keys are important attributes for potential customers of these businesses. Chain businesses are important fixtures in low-income areas, and internet penetration also suffers as well. Whether a given McDonald's has internet access can thus, be very important. When available, these types of metadata are also used by applications such as Google Maps and Yelp – indicating their importance and demand. Given that we looked at prominent businesses and broadly important attributes, we felt the sampled metadata *should* be among the Contingent metadata that is applied the most. We analyzed 60 instances per business for 180 total instances and 300 *potential* business-key-value triples. Out of the 300 triples, 245 of the attributes represented by those triples actually existed in the real-world.

Given our choice of Contingent metadata sampled, the fact that Contingent metadata is laborious to apply, and the fact that even less Contingent metadata was applied in scenarios where it was appropriate relative to Universal-Varying metadata, we had confidence that our sampling approach provided a reasonable “best case” proxy for the degree to which all Contingent metadata is applied when it is applicable. And further, we were confident that that degree of application was very low.