

Defining Fitness-for-Use for Crowdsourced Points of Interest (POI)

Journal Article**Author(s):**

Jonietz, David; Zipf, Alexander

Publication date:

2016

Permanent link:

<https://doi.org/10.3929/ethz-b-000119533>

Rights / license:

[Creative Commons Attribution 4.0 International](#)

Originally published in:

ISPRS International Journal of Geo-Information 5(9), <https://doi.org/10.3390/ijgi5090149>

Communication

Defining Fitness-for-Use for Crowdsourced Points of Interest (POI)

David Jonietz ^{1,*} and Alexander Zipf ²

¹ Institute of Cartography and Geoinformation, ETH Zurich, Stefano-Franscini-Platz 5, Zurich 8093, Switzerland

² Institute of Geography, University of Heidelberg, Im Neuenheimer Feld 368, Heidelberg 69120, Germany; zipf@uni-heidelberg.de

* Correspondence: jonietzd@ethz.ch; Tel.: +41-44-633-3277

Academic Editor: Wolfgang Kainz

Received: 1 June 2016; Accepted: 19 August 2016; Published: 24 August 2016

Abstract: (1) Background: Due to the advent of Volunteered Geographic Information (VGI), large datasets of user-generated Points of Interest (POI) are now available. As with all VGI, however, there is uncertainty concerning data quality and fitness-for-use. Currently, the task of evaluating fitness-for-use of POI is left to the data user, with no guidance framework being available which is why this research proposes a generic approach to choose appropriate measures for assessing fitness-for-use of crowdsourced POI for different tasks. (2) Methods: POI are related to the higher-level concept of geo-atoms in order to identify and distinguish their two basic functions, geo-referencing and object-referencing. Then, for each of these functions, suitable measures of positional and thematic quality are developed based on existing quality indicators. (3) Results: Typical use cases of POI are evaluated with regards to their use of the two basic functions of POI, and allocated appropriate measures for fitness-for-use. The general procedure is illustrated on a brief practical example. (4) Conclusion: This research addresses the issue of fitness-for-use of POI on a higher conceptual level by relating it to more fundamental notions of geographical information representation. The results are expected to assist users of crowdsourced POI datasets in determining an appropriate method to evaluate fitness-for-use.

Keywords: VGI; data quality; points of interest; fitness-for-use

1. Introduction

Points of Interest (POI) are zero-dimensional features which refer to specific locations or real-world entities in geographical space, such as historical sites, landmarks, public services, shops, restaurants, or bars [1]. Providing a main data resource for numerous web-based services and commercial geo-spatial applications, POI are collected by companies such as Garmin, Facebook or Yelp, often via crowdsourcing, for a variety of purposes, including navigation and routing, providing spatial recommendations, or enabling users to share location-based information such as place reviews, check-ins, or geo-tagged photographs. In a research context, POI are put to additional uses, including the analysis of population dispersion [2], urban social activity in the city [3], or the perceived location of the city center [4].

Compared to traditional spatial data, as provided by commercial vendors or the authorities, Volunteered Geographic Information (VGI) is in particular need of adequate methods for data quality assessment, a fact which is due to its contributors being untrained and heterogeneous, a lack of formal specifications and the potential effects of social factors [5–8]. Since in this context, quality assurance is challenging at best [6], it can be argued that the task of quality assessment has been somewhat shifted from the producer to the users of the data, who are required to evaluate its appropriateness with regards to their specific motive, or its fitness-for-use [9]. According to the International Organization

for Standardization (ISO), the quality of geographic information should generally be related to its suitability with reference to “specific application needs or requirements” [10] (p. 1). The relevance of such task-dependent measures of fitness-for-use is further emphasized by the fact that recently, the Open Geospatial Consortium (OGC) has proposed candidate standards for collecting user experiences with datasets, and providing them as metadata [11]. A quality assessment of a geo-spatial dataset, accordingly, should not occur independently from its intended usage, but rather closely related to its expected suitability for this particular purpose [12]. Whereas the appropriateness of OpenStreetMap (OSM) data has been analyzed with regards to specific application tasks such as navigation [13,14], 3D-building reconstruction from building footprints [15], or bicycle-related mapping and analysis tasks [16], to the best of our knowledge, there is currently no study which explicitly evaluates the fitness-for-use of POI datasets.

Thus, whereas POI are highly used for research- and application-driven purposes, there is still a lack of knowledge with regards to general quality-related characteristics of such datasets as well as potential methods to evaluate their fitness-for-use. Although not focusing primarily on POI, but on VGI in general, [12], for instance, have explicitly recommended for future research activities “to develop a systematic framework that provides methods and measures to evaluate the fitness for purpose of each VGI type” [12] (p. 21). As a first step in that direction, therefore, this work focuses on crowdsourced POI and proposes a generic approach for assessing their fitness-for-use regarding different tasks. With this framework, we aim to assist users of a POI dataset in choosing from the range of available quality measures and assessment methods those which are appropriate for evaluating its fitness-for-use with regards to their specific use case or application task. For this, since POI serve as one of many possible ways to represent geographical entities in Geographical Information Systems (GIS), we approach the topic of POI-based geographic information representation on a higher conceptual level by relating POI to the fundamental notion of geo-atoms proposed by [17]. This allows us to identify and distinguish geo-referencing and object-referencing as the two basic functions of POI, which, as we further argue, each require specific quality measures. With a focus on positional and thematic accuracy, and based on prior research on data quality, we propose a set of according methods. This is relevant for data users since, as we argue, in order to assess the fitness-for-use of POI for a particular use-case, it is necessary to identify the one of the two basic POI functions which is called upon in the respective task, and choose the according quality measures. This is demonstrated on a practical example.

This paper is structured as follows: First, based on a brief review of relevant literature, potential ways for assessing POI data quality are identified. Then, our approach towards evaluating fitness-for-use of POI data is explained and demonstrated on a practical example. Finally, our work is discussed and concluded.

2. Methods for POI Quality Assessment

There are several studies in which the quality of VGI, in most cases with a strong focus on OSM, was examined (e.g., [18–20]). For many of these, an orientation towards the relevant data characteristics was provided by a set of quality measures which have been defined by ISO: completeness, logical consistency, positional accuracy, temporal quality, thematic accuracy, and usability [10]. Due to the fact that this list was developed in the context of geographical data in general, however, its validity for VGI has also been challenged [5], with alternative approaches placing extra emphasis on the trustworthiness and credibility of the generated information [21,22].

With regards to previous work on geo-spatial data quality assessment, two basic approaches can be distinguished depending on whether the data is compared to a reference dataset, which is assumed to be of highest quality standards and therefore used as ground truth, or not. In the first case, an extrinsic evaluation is conducted, while the second type of quality assessment is termed intrinsic, and relies on a selection of data characteristics which serve as quality indicators, such as the sampling ratio or tag density [23], the temporal development of the dataset [24], but also demographic or socio-economic

attributes of the specific geographical area of interest (e.g., [25]). In a recent study, [18] provide a conceptual framework for intrinsic quality assessment of VGI.

In this work, the focus is put on extrinsic approaches to quality assessment. When relating the crowdsourced data to a reference dataset, in most cases obtained from authoritative or commercial sources, for each ISO quality measure, numerous assessment methods have been proposed. Thus, for instance, data completeness can be assessed by comparing the total number of features in both datasets [26] or deriving a completeness index based on the number of features represented in both datasets in relation to their total number [27]. There are also several potential methods to assess the logical consistency of a crowd-sourced dataset, such as topology checks [28], however, this aspect is more closely related to the intrinsic quality of a dataset. Typical indicators for positional accuracy are Euclidean distances between co-referent points (e.g., [26,27]), distance deviations on the X- and Y-axis [29], or the evaluation whether, and in case of line or polygon features how much of, a feature from a crowdsourced dataset is located within a certain buffer zone computed around a reference feature (e.g., [20,30]). The temporal accuracy can be related to the actuality of the data, but is, similar to the logical consistency, also more relevant with regards to intrinsic quality evaluations. Finally, the thematic accuracy is examined by deriving the percentage of correct feature classifications (e.g., [15,29]), the Levenshtein distance which expresses the similarity of strings (e.g., [26]), or the number of features which have certain attributes [15,30].

In the past, several studies have specifically focused on or at least included an assessment of the quality of POI datasets. In [26], for instance, POI from OSM are compared to IGN BD topo, and the positional accuracy (Euclidean distance method), semantic accuracy (Levenshtein distance and attribute frequency) and currentness (internal assessment of the update frequency) are assessed. [30] reports on work which used Teleatlas data as ground truth for a comparison to OSM POI and focuses on positional accuracy (buffer method) and the completeness (comparison of the number of features per raster cell). [31] examines the positional accuracy of geo-tagged photographs from Flickr and Panoramio by assessing the Euclidean distance to the position from which the picture was most likely taken, as reckoned by an expert. [27] combines POI obtained from Navteq and Yelp to a reference dataset against which OSM data is compared, and focus on positional (Euclidean distance method) and semantic accuracy (Levenshtein distance), as well as completeness (completeness index).

The mentioned studies operationalize data quality as the degree of similarity to the ground truth, as represented by a reference dataset. Thus, this work also relates to a different stream of research, which focuses not on data quality per se, but on an assessment of the similarity of POI obtained from different datasets for the purposes of conflation or data fusion. This is mostly done with the goal of detecting co-referent POI, which represent the same real-world entity and should, therefore, be avoided or matched in an integrated POI database. [1], for instance, uses fuzzy set theory to detect and merge co-referent POI based on their syntactical similarity with regards to the name, locational correspondence under consideration of the scale at which they were digitized, and the semantic closeness of their allocated category. Another example is provided by [32], who base their matching algorithm on the Euclidean distance, the name similarity, and the website similarity of two POI. [33] matches POI obtained from different social network sites by comparing their geographic distance as well as the string similarity of selected semantic attributes. Aiming to develop an assistive system for data editing, [34] compute the similarity of POI in OSM based on the change history of their respective tags. Apart from POI, there has been work focusing on matching co-referent geo-objects of linear (e.g., [35–37]) or polygonal (e.g., [15,38]) geometry types. In a combined evaluation of quality control measures and data conflation from different VGI sources, [7] state that in practice, the two steps are often entangled, which, according to the authors, limits the possibilities to evaluate the fitness-for-use of such data.

3. Defining Fitness-for-Use for POI

As previously mentioned, the quality of a geospatial dataset should be assessed in terms of its fitness-for-use [9]. To the best of our knowledge, however, there is currently no work on the specific case of assessing the fitness-for-use of POI datasets. At present, therefore, a prospective user of a POI dataset is faced with the full variety of existing quality measures and assessment methods, as presented in the previous chapter, and is required to develop a task-specific, appropriate strategy for assessing the fitness-for-use of the dataset without any assistance or guidelines to refer to. Accordingly, as it has already been stated, the development of a corresponding orientation framework has been identified as a pressing research need [12]. Thus, in this chapter, a generic approach to define the fitness-for-use of POI data is presented. For this, based on a formal notion of POI, we first identify and distinguish geo-referencing and object-referencing as the two major functions of POI. In a second step, we further argue that each task which involves the use of POI calls upon one or the other of these two functions, and accordingly requires different methods to be applied for evaluating the dataset's fitness-for-use. Consequently, we develop suitable quality indicators for each of these functions, which are then related to typical use-cases of POI data.

3.1. The Bi-Functionality of POI

The purpose of POI has been stated as describing “geographic locations or entities at geographic locations” [1] (p. 2). While this statement already hints at a certain functional duality, it makes sense to approach the notion of POI in a more formal way.

In their work on the automated cleansing of POI databases, [1] provide a fundamental definition of POI: “A point of interest (or POI) is axiomatically understood as a piece of data that describes a geographic entity in the real world that is modelled by E ” [1] (p. 6). With E , in this case, the authors refer to the appropriate universe of properties of an entity which, in order to achieve a successful reference between POI and real-world entity, must be linked with the universe O_{POI} of the corresponding POI via a surjective function $\rho: O_{POI} \rightarrow E$. The universe O_{POI} is typically compound and therefore consists of non-compound universes $O_{POI} = U_1 \times U_2 \times \dots \times U_n$. These can for instance include the following classes with the according labels: *name* (U_1), *latitude* (U_2), *longitude* (U_3), and *category* (U_4). In this case, U_2 and U_3 can be logically grouped together since in combination they provide the *location* of the POI [1]. The referencing relation between POI and real-world entity, therefore, is established via its corresponding universes and the according values, e.g., *name* = ‘*St. Paul’s Cathedral*’, *latitude* = 51.51382, *longitude* = -0.09850 and *category* = ‘*place of worship*’, which must correspond to the characteristics of the matching real-world entity, or in other words, its compound universe E .

As it has been stated in the introduction, POI serve to represent geographic information, and can, therefore, be related to the more abstract notion of geo-atoms [17]. Aiming to provide a general theory of GIS-based geographical representation, geo-atoms have been introduced as abstract primitives which underlie both continuous fields and discrete objects. A geo-atom, accordingly, has been defined as “an association between a point location in space-time and a property. We write a geo-atom as a tuple $\langle x, Z, z(x) \rangle$ where x defines a point in space-time, Z identifies a property, and $z(x)$ defines the particular value of the property at that point” [17] (p. 243). Based on this reduced, atomic form of geographic information, higher-level objects or fields can be constructed. With regards to discrete objects, thus, whereas a point object consists only of one geo-atom, linear and polygonal objects can be conceptualized as aggregations of geo-atoms which have common, specified values for certain properties, such as *name* = ‘*St. Paul’s Cathedral*’ [17].

A geo-atom, therefore, has a dual function. On the one hand, its fundamental purpose is to associate a certain location with a particular property. At this stage, there is no particular need for this property to relate to any kind of super-ordinate geo-object. On the other hand, however, there is also a derived mereological function of being part of a higher-level geo-object.

The same functions, we argue, apply for POI. Usually, a POI refers to a real-world entity, such as a building or a place, which could also be represented in a GIS as a higher-level geo-object, for instance, a polygon feature. In this context, the POI can be understood as a reduced model of the original geo-object, a selected geo-atom which has been picked as a representative from the larger set of geo-atoms which constitute this particular geo-object. Thus, for example, if a real-world building is represented as a footprint polygon, this representation can be further reduced to only one of its geo-atoms, for instance the one which is located precisely at the geometrical centroid of the polygon. The exact location of the selected geo-atom, however, is not essential, since the semantic relation to the corresponding geo-object and ultimately the real-world entity is rather established via the POI *name* or another unambiguous identifier, as for instance done in a gazetteer, which typically links a distinct coordinate pair with a place name and, if available, additional information. In other words, the surjective function $\rho: O_{\text{POI}} \rightarrow E$ which maps between the compound universes of the POI and the real-world entity depends on a correspondence between universes such as *name* = 'St. Paul's Cathedral', and not necessarily the *location*. In the following, this is referred to as the object-referencing function of a POI.

Similar to a geo-atom, however, a POI is first and foremost an association of a particular location with a property. We refer to this as the geo-referencing function of a POI. Thus, on the one hand, it is possible that a POI is independent of a superordinate geo-object as, for instance, in the case of a POI which represents a suitable viewpoint on a hiking trail, and thus merely provides the information that at this particular *location*, a property *visibility* = *high*. On the other hand, in several of their use cases, POI which do in fact refer to a certain geo-object and the according real-world entity are reduced to their geo-referencing function, as for instance when counting the number of POI in a study area, computing POI density surfaces or calculating the shortest path to a POI. In such cases, the focus is clearly put on the locational information of the POI, whereas the semantic connection to a particular real-world entity is of less importance. With regards to its geo-referencing function, therefore, the functionality of a POI depends on the universes *latitude* and *longitude* which define its *location*, rather than its *name* or *category*.

3.2. Function-Dependent Quality Measures for POI

Having identified these two fundamental functions of POI, we further argue that, when assessing their fitness-for-use, it is necessary to identify which basic function is addressed in the particular task. There are for instance several studies which analyze the spatial distribution of the POI in combination with their allocated information, in particular their assigned category, to infer other environmental properties, such as population estimates [2], urban land use [32], employment size [39], social hubs in the city [3], or the perceived boundaries of the city center [4]. In such contexts, as stated before, the reference to a particular real-world object is of minor importance, it is rather the *location* and, to a lesser degree, the *category* of a POI which determines the quality of the received results. Thus, for instance, in order to receive accurate results when classifying an area within a city as a shopping district based on the *category* of crowdsourced POI located within its boundaries, a high positional accuracy of the involved point features is required. Otherwise, POI located on the periphery of the area could be falsely allocated to neighboring regions. Furthermore, the POI need to be correctly classified as *category* = *shop*. Important, however, is the fact that uncertainty with regards to which exact shop a POI is referring to would in this case not have any effect on the analysis results. In order to assess the fitness-for-use of a POI dataset for this purpose or similar ones, therefore, a prospective data user should place the focus on quality measures and assessment methods related to the geo-referencing function rather than the object-referencing function.

There are, however, also use cases of POI which call upon the latter function, and therefore require an unambiguous semantic reference to a particular real-world object, for instance in the context of check-ins at distinct places (e.g., Facebook), place reviews (e.g., Yelp), or place recommendation systems [40]. Here, the exact location of a POI is not critical, as long as it is located within an acceptable threshold distance from the place's true location, however, an unambiguous reference to a particular

real-world entity must be provided, normally by means of a place name or some other attribute which serves as an identifier. To give an example, if a customer uses a web-based service to attribute a rating to a POI which represents a restaurant he or she has previously visited, it is necessary for the POI to unambiguously refer to the correct real-world entity, since otherwise, the rating could be falsely attributed to a neighboring establishment. Also of relevance in this case is the question how a suitable distance threshold measure could be determined. Apart from the simplest case of subjectively defining an absolute value, e.g., 50 m, another strategy could include the use of relative distances which depend on the actual size of the represented real-world object. Thus, for instance, a deviation of 50 m should certainly be evaluated as more drastic in the case of a POI referring to a small restaurant rather than a football stadium or an airport. In addition, acceptable threshold values could be related to typical spatial patterns of POI of similar categories. Accordingly, for instance, whereas cafes can usually be found closer to each other, police stations are intentionally located in a more dispersed manner, which in turn allows for higher positional inaccuracies of the corresponding POI to still be acceptable. Finally, apart from Euclidean distance measures, topological relationships should not be neglected, such as co-referent POI which are, due to positional error in the crowdsourced dataset, located in different administrative areas or on different sides of major roads or rivers.

Hence, there is indeed a need for suitable quality indicators which explicitly note the difference between the geo-referencing function as opposed to the object-referencing function of POI. In our view, of the quality measures defined by ISO, completeness, logical consistency and temporal quality affect the fitness-for-use of a POI dataset on a higher conceptual level, meaning that corresponding quality-related problems would equally affect its fitness-for-use regardless of the POI function in question. Thus, to refer back to the exemplary cases used previously, the use of an incomplete POI dataset would negatively influence the results of a land use classification procedure due to, e.g., missing objects of *category = shop* in an area, but equally decrease the practical usefulness of a place review service. As it has been shown, however, quality issues with regards to the positional and the thematic accuracy can indeed differ in their effect depending on the respective task and, therefore, need to be addressed in a task- and, ultimately, function-specific manner. For this reason, they are of particular interest for our approach.

With regards to the geo-referencing function of a POI, the emphasis is clearly put on its positional accuracy and absolute location. Therefore, a suitable method for quality assessment would be to compute the deviation, e.g., in terms of the Euclidean distance, from the correct reference location (e.g., [27,35]). Despite the focus on the position, however, the thematic accuracy cannot be neglected completely. Instead, the correctness of all semantic attributes should be provided, yet there is no difference to be made between them. Thus, since the reference to a particular real-world entity is of secondary importance here, as it has been shown on the example of a POI-based land use classification task, there is no obligatory distinction to be made between identifiers, such as the universe *name*, and other properties like the POI *category*. It can be the case, though, that certain universes are of a higher relative importance, such as *category* in the used example, but this is highly specific to the respective task. Accordingly, in general, the accuracy of all attributes of the POI can be evaluated equally, for instance by calculating their semantic distance to the corresponding reference dataset by means of lexical ontologies such as WordNet [41] or by computing the Levenshtein distance (e.g., [26]).

In contrast, the object-referencing function does not depend primarily on the accuracy of the exact location, but rather on the possibility to infer an unambiguous reference to a real-world entity from the compound universe of the POI, as it has been discussed on the example of a web-based place rating service. The positional accuracy, therefore, is sufficient if the POI is included within the set of geo-atoms which constitute the super-ordinate geo-object, e.g., the footprint of the respective restaurant, but does not depend on the Euclidean distance to a reference location, for instance, the object's centroid. Therefore, an appropriate quality assessment method includes testing this topological relation via an intersect operation with the respective geo-object, for instance, the building footprint. With regards to the semantic accuracy, however, an unambiguous identifier is needed in order to establish a clear

mapping between the compound universes of the POI and the real-world entity, such as a *name* attribute with an accurate value, the quality of which can be tested by computing the Levenshtein distance. If there is no *name* provided, however, a reference to a specific real-world entity can also partly be established in case of a high positional accuracy and no conflicting semantic information. For instance, if a POI is missing a *name* attribute, but is located at the exact *location* of a restaurant, and is accurately labeled as *category = restaurant*, then the *location* can partly take over the role as object identifier, however, with remaining uncertainty since it could for instance also be thinkable that in fact, the POI refers to another restaurant which used to be in the same *location* before closing down. Accordingly, if there is no *name* provided, the semantic accuracy of the remaining attributes as well as the positional accuracy can be appropriately assessed by computing the semantic distance via WordNet or the Levenshtein distance, and using one of the methods proposed for computing an acceptable distance threshold proposed before.

4. An Exemplary Application to POI Obtained from Facebook

For illustrative purposes, the developed concept is applied to the hypothetical example of a researcher who aims to use a crowdsourced POI dataset for two different purposes, and is now facing the challenge of assessing its fitness-for-use. In our simple example, the POI dataset consists only of one point feature, namely Facebook's POI representation of St. Paul's Cathedral in London, which was obtained via Facebook's Graph API. Since a reference dataset is always required for an extrinsic quality assessment, the co-referent POI is taken from Factual, a commercial vendor of quality-ensured POI datasets. In general, however, when choosing appropriate ground truth data sources, one should acknowledge the fact that nowadays, the traditional assumption of authoritative or commercial datasets being of a higher quality compared to VGI is no longer fully reliable. Thus, reference datasets must be chosen with care and in the specific context of the particular case study, a challenge which, however, is inherent to extrinsic quality assessment methods in general and, therefore, exceeds the scope of this work.

Figure 1 shows the geo-located POI from both data sources together with their thematic attributes. It can be seen that differences exist in terms of both the location and the thematic information attributed to the POI, or, using the terminology provided by [1], the surjective function $\rho: O_{\text{POI}} \rightarrow E$ which establishes the reference between POI and real-world entity differs regarding the universes $O_{\text{POI}} = U_1 \times U_2 \times \dots \times U_n$. Thus, a clear deviation can be seen in terms of the universe *location*, and slight variances of the allocated *category* labels. Apart from differing values of shared attributes, however, there are also variations with reference to the constituting universes of O_{POI} . Thus, whereas Facebook provides additional attributes *general info* and *about*, Factual gives more place-related information related to the *region* or *neighbourhood* as well as other facts such as the *hours of operation*.

In the first scenario, our researcher aims to evaluate the fitness-for-use of Facebook's POI representation of St. Paul's Cathedral for the purpose of developing a web-based system which allows checking-in at places. In accordance to our proposed concept, as a first step, he or she must identify which one of the two fundamental POI functions the intended use case calls upon. Due to the fact that checking-in at a place requires an unambiguous semantic reference to exist between the POI and the specific real-world entity, in this case it would clearly be the object-referencing function. This insight provides the information which is necessary for the next step, namely to choose the appropriate quality assessment methods. Thus, in the specific case of object-referencing, it should be evaluated how well the referencing relation between POI and real-world entity is established with regards to the Facebook POI in comparison to the data obtained from Factual, which, in its role as a reference dataset, is assumed to be of the highest quality. Therefore, with reference to the positional accuracy, as it has been discussed, it must be tested whether the x, y-coordinate pair of Facebook's POI qualifies it to be a member of the set of geo-atoms which constitute the super-ordinate geo-object, thus, the polygonal footprint of St. Paul's Cathedral. Accordingly, a second reference dataset is needed here, which, in this case, can be obtained from Ordnance Survey. With a simple intersect operation,

our researcher can check whether the Facebook POI spatially intersects with the footprint of St. Paul’s Cathedral which, in this case, returns true. With regards to the semantic accuracy, an unambiguous identifier is needed to establish the reference between the POI and the real-world object. Thus, if a *name* attribute is provided, as it is the case here, it is compared to the corresponding *name* of the reference dataset via computing the Levenshtein distance, which, due to a lack of spelling errors, returns 0 here. With regards to the object-referencing function, therefore, the other attributes are not needed, and, due to perfect locational and thematic accuracy, our researcher receives the highest possible value, for instance 1, to express the fitness-for-use of the Facebook dataset for the purpose of developing a place check-in system.

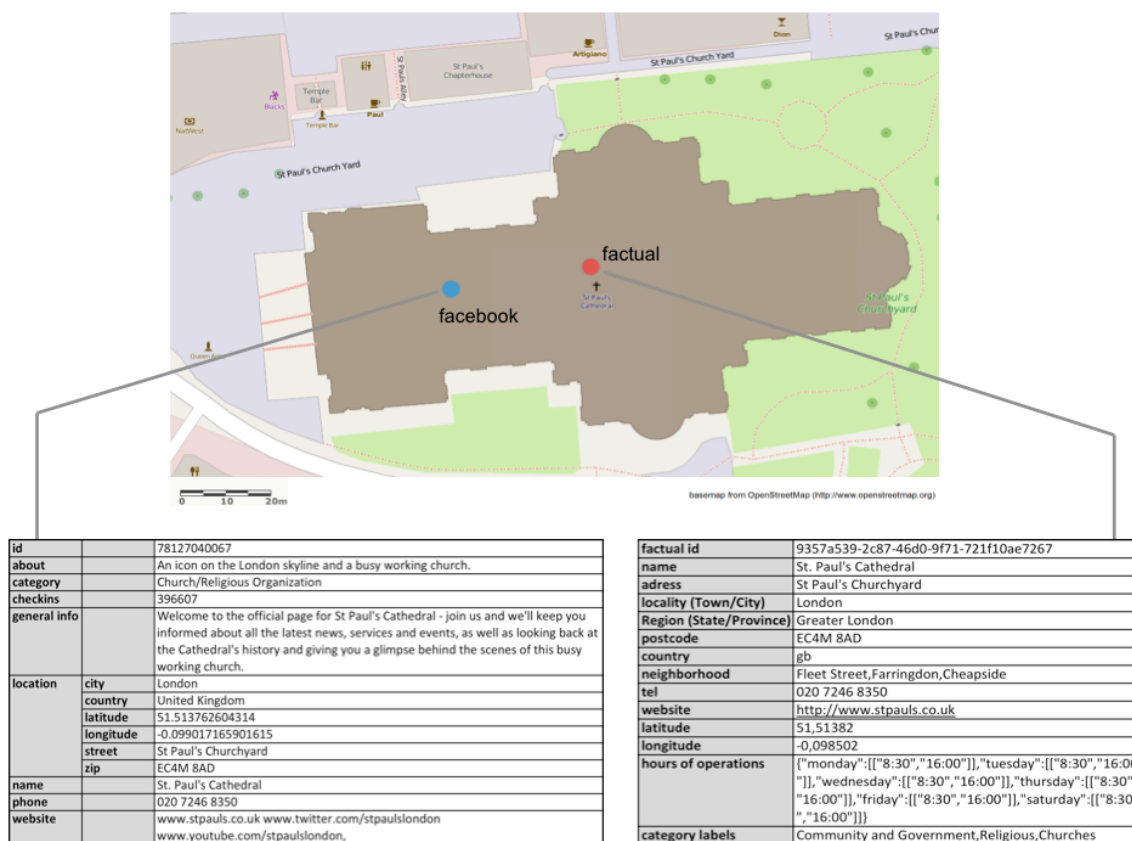


Figure 1. POI Co-referents from Facebook and Factual.

In a second scenario, however, our data user aims to deploy the same POI dataset for a different purpose, namely to analyze the spatial distribution of different types of POI in London. Here, as it has been discussed, the reference to a particular real-world entity, e.g., St. Paul’s Cathedral, is not relevant, it is rather the accuracy of the locational information, and therefore the geo-referencing function, which determines the dataset’s fitness-for-use. Thus, the positional accuracy needs to be assessed in terms of the exact deviation to the position of its co-referent in the reference dataset. An appropriate method for quality assessment would therefore be calculating the deviation in terms of the Euclidean distance. Accordingly, in our example, a deviation of roughly 36 m between Facebook’s and Factual’s POI for St. Paul’s Cathedral is computed, a value which can then be translated to a normalized quality index which expresses the positional accuracy. With regards to the thematic accuracy in the context of the geo-referencing function of POI, as it has been said, there is no difference to be made between the semantic attributes, which is due to the fact that since the reference to a particular real-world entity is irrelevant, the *name* universe is of equal importance compared to the other attributes. Thus, all semantic attributes can be compared in a piece-wise manner for their semantic similarity. For attributes such

as the *name*, *address*, *telephone number*, *website*, or *email address*, the Levensthein distance can be used, whereas the values allocated for *category*, for instance, are more appropriately compared via computing the WordNet semantic distance [41]. In our case, due to a lack of spelling mistakes and the existence of largely corresponding category labels, the highest quality index value is attributed for semantic accuracy by our researcher. The fact that there is a different number of constituting universes of $O_{POI} = U_1 \times U_2 \times \dots \times U_n$ provided in the Facebook versus the Factual dataset is neglected here, but could for instance be noted by a simple comparison of their total count in both datasets.

By following our generic approach, and as a result of the preceding analyses, our researcher would therefore find that the fitness-for-use of the Facebook dataset is relatively higher for use cases which relate to its object-referencing function than to its geo-referencing function, a fact which is due to the different procedure of calculating positional accuracy in this particular example. Thus, the dataset would be better suited for developing a place check-in system than for analysing the spatial distribution of POI in the study area.

5. Discussion

Despite its exploratory character, the previous exemplary application demonstrates the practical usefulness of our proposed approach. Thus, it was shown how, by following the strategy of first identifying the relevant one of the two basic POI functions, and then selecting appropriate quality assessment methods, a user of a POI dataset is assisted in assessing its fitness-for-use with regards to different tasks. In fact, the dataset used in the example was evaluated differently with regards to its fitness-for-use for the two usage scenarios, in our case resulting in relatively higher values for the one related to the object-referencing function than to the geo-referencing function. Apart from providing information about the quality of a POI dataset with regards to specific use cases, as it has been demonstrated, our framework could also be used to compare POI obtained from different sources for their appropriateness depending on the specific use case in question, and guide the choice of the most suitable alternative. This is especially useful for POI datasets which, as we have argued in this paper, are on the one hand often the result of crowdsourcing and, therefore, prone to issues related to data quality, and on the other hand used for a large variety of tasks.

With the assistive framework proposed in this paper, we provide a clear strategy for identifying an appropriate way for evaluating the fitness-for-use of POI datasets, which is based on relating different positional and thematic quality measures, as well as their appropriate assessment methods, e.g., computing the Euclidean distance deviation or the semantic distance of attributes, to the two basic functions of POI, geo-referencing and object-referencing. These, in turn, are then related to corresponding use cases for POI datasets such as web-based place check-in services or POI density calculations. Without our systematic approach, users of POI datasets are required to develop a work flow for evaluating fitness-for-use without any assistance, and directly relate their particular use case to the appropriate quality measures and assessment methods. In our opinion, however, this greatly increases the risk of making arbitrary, sub-optimal choices as well as decreases the comparability of studies which assess fitness-for-use and their results. Thus, if there is no clear motivational link between the given use case on the one hand, and the chosen quality measures and assessment methods on the other hand, it is hard to explain clearly why e.g., the focus was placed on positional and not thematic accuracy for assessing fitness-for-use, and which exact factors motivated the selection of the particular quality assessment methods. At the same time, different approaches for assessing fitness-for-use of POI for the same task would likely differ in terms of the applied method and, therefore, be only partly comparable. By following the sequence of steps as it was proposed here, which lead from the particular use case via the corresponding POI function to the appropriate quality measures and assessment methods, such problems can be avoided since the choice of the latter is clearly motivated, directly linked to the respective task, and fully transparent.

Apart from providing guidance for selecting an appropriate strategy for assessing fitness-for-use, as stated above, the proposed concept could also assist the crowdsourcing process for data generation

itself. Thus, if the primary purpose of a POI dataset to be created was already known prior to the data collection process, one could develop task-specific quality control measures in a similar manner, namely by first identifying the corresponding POI function, and then accordingly placing the focus on one or the other indicator for thematic or positional accuracy.

Owing to the early stage of this research, however, there are still several limitations of our approach. Thus, the validity of our proposed method for assessing fitness-for-use has not yet been evaluated with larger datasets and actual use cases. It might be the case that the use of more complex quality indicators is needed in practice. Further, although we sketched potential ways of deriving thresholds for tolerable positional deviation, choosing appropriate values is certainly challenging, and might to a certain degree reduce the intended comparability of different methods which was mentioned before. Another critical issue is our simplifying assumption that use cases clearly address one or the other of the two basic POI functions, whereas we have mentioned numerous examples where this is actually the case, in practice there are also use cases which require a combination of both geo-referencing and object-referencing, e.g., navigation to a specific POI, which would require a combinatory approach for assessing fitness-for-use. In general, a more comprehensive overview of potential use cases would be worthwhile.

6. Conclusions

This work was motivated by the discrepancy between the frequent use of crowdsourced POI for a variety of commercial and research-driven purposes on the one hand, and the lack of work on assessing the fitness-for-use of such datasets on the other hand. Before the background of quality-related problems of VGI in general but with a particular focus on crowdsourced POI, this research addressed the research need for a systematic framework to assist data users in choosing quality measures and methods to evaluate fitness-for-purpose. Based on the conceptual closeness of the notions of POI and geo-atoms, we theorized about a dual functionality of POI, namely geo-referencing as opposed to object-referencing, and argued that in order to assess the fitness-for-use of POI datasets, one should first identify the basic function which is of relevance with regards to the respective task or use case, and then apply an appropriate, since clearly motivated, method for quality assessment, as it has been demonstrated on a practical example.

From the perspective of the individual data user, potential advantages arise from an increased comparability of methods for assessing fitness-for-use and their results as well as a useful guideline which is provided for the process of selecting quality indicators which are clearly motivated and chosen in a transparent manner. Further, in our opinion, our work contributes to enhancing the potential of crowdsourced POI datasets as a valuable data source for various applications in the geo-spatial domain. Thus, the proposed framework is expected to help gaining a deeper understanding of quality-related issues of crowdsourced POI, for instance with regards to typical quality-related problems, quality differences between various sources of POI data, or potential use cases for which crowdsourced POI are more or less suited. Further, it is possible to develop quality improvement measures which are explicitly targeted to intended use cases on the basis of such task-specific quality analyses.

Our initial work opens up possibilities for future work, such as the comparative analysis of POI obtained from different data sources, e.g., Facebook, Foursquare or OSM, with regards to their relative fitness-for-use concerning different use cases. Such analyses would provide useful information for future data users. Further, one could develop an exhaustive framework for assessing fitness-for-use for a variety of different typical use cases which involve POI. This could possibly be implemented as an automated assistive system providing support for deciding between different POI datasets. A prerequisite for this, however, involves the testing and validation of our concept with a large dataset of crowdsourced POI, and a critical comparison of the results to other existing approaches for geo-spatial data quality assessment. Additionally, as it has already been mentioned, one could develop task-dependent quality control measures for the data collection process of crowdsourced datasets.

Author Contributions: David Jonietz and Alexander Zipf conceived and designed the experiments; David Jonietz wrote the paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. De Tré, G.; Van Britsom, D.; Matthe, T.; Bronselaer, A. Automated cleansing of POI databases. In *Quality Issues in the Management of Web Information*; Springer: Berlin, Germany; Heidelberg, Germany, 2013; pp. 55–91.
2. Bakillah, M.; Liang, S.; Mobasheri, A.; Arsanjani, J.J.; Zipf, A. Fine-resolution population mapping using OpenStreetMap points-of-interest. *Int. J. Geogr. Inf. Sci.* **2014**, *28*, 1940–1963. [[CrossRef](#)]
3. Bawa-Cavia, A. Sensing the urban: Using location-based social network data in urban analysis. In Proceedings of the Workshop on Pervasive Urban Applications (PURBA) 2011, San Francisco, CA, USA, 12 June 2011.
4. Huang, H.; Gartner, G.; Turdean, T. Social media as a source for studying people's perception and knowledge of environments. *Mitteilungen der Österreichischen Geographischen Gesellschaft* **2013**, *155*, 291–302. [[CrossRef](#)]
5. Antoniou, V.; Skopeliti, A. Measures and indicators of VGI quality: An overview. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2015**, *II-3/W5*, 345–351. [[CrossRef](#)]
6. Fonte, C.C.; Bastin, L.; Foody, G.; Kellenberger, T.; Kerle, N.; Mooney, P.; Olteanu-Raimond, A.-M.; See, L. VGI quality control. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2015**, *II-3/W5*, 317–324. [[CrossRef](#)]
7. Leibovici, D.G.; Evans, B.; Hodges, C.; Wiemann, S.; Meek, S.; Rosser, J.; Jackson, M. On data quality assurance and conflation entanglement in crowdsourcing for environmental studies. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2015**, *II-3/W5*, 195–202. [[CrossRef](#)]
8. Roick, O.; Heuser, S. Location based social networks—Definition, current state of the art and research agenda. *Trans. GIS* **2013**, *17*, 763–784. [[CrossRef](#)]
9. Veregin, H. Data quality parameters. In *Geographical Information Systems: Principles and Technical Issues*, 2nd ed.; John Wiley & Sons: New Jersey, USA, 2005; pp. 177–189.
10. International Organization for Standardization (ISO). *ISO/TC 211 19157: Geographic Information—Data Quality*; No. ISO 19157:2013; ISO: Geneva, Switzerland, 2013.
11. Open Geospatial Consortium (OGC). OGC Seeks Public Comment on Candidate Geospatial User Feedback Conceptual Model and xml Encoding Standard. Press Release 2016. Available online: <http://www.opengeospatial.org/pressroom/pressreleases/2356> (accessed on 20 July 2016).
12. Senaratne, H.; Mobasheri, A.; Ali, A.L.; Capineri, C.; Haklay, M. A review of volunteered geographic information quality assessment methods. *Int. J. Geogr. Inf. Sci.* **2016**. [[CrossRef](#)]
13. Mondzech, J.; Sester, M. Quality analysis of OSM data based on application needs. *Cartographica* **2011**, *46*, 115–125. [[CrossRef](#)]
14. Zhang, X.; Tinghua, A. How to model roads in OpenStreetMap? A method for evaluating the fitness-for-use of the network for navigation. In *Advances in Spatial Data Handling and Analysis*; Springer: Berlin, Germany; Heidelberg, Germany, 2015; pp. 143–162.
15. Fan, H.; Zipf, A.; Fu, Q.; Neis, P. Quality assessment for building footprints data on OpenStreetMap. *Int. J. Geogr. Inf. Sci.* **2014**, *28*, 700–719. [[CrossRef](#)]
16. Hochmair, H.; Zielstra, D.; Neis, P. Assessing the completeness of bicycle trail and lane features in openstreetmap for the United States. *Trans. GIS* **2015**, *19*, 63–81. [[CrossRef](#)]
17. Goodchild, M.; Yuan, M.; Cova, T.J. Towards a general theory of geographic representation in GIS. *Int. J. Geogr. Inf. Sci.* **2007**, *21*, 239–260. [[CrossRef](#)]
18. Ballatore, A.; Zipf, A. A conceptual quality framework for volunteered geographic information. In Proceedings of the 12th International Conference on Spatial Information Theory, COSIT 2015, Santa Fe, NM, USA, 12–16 October 2015.
19. Dorn, H.; Törnros, T.; Zipf, A. Quality evaluation of VGI using authoritative data—A comparison with land use data in southern Germany. *Int. J. Geo-Inf.* **2015**, *4*, 1657–1671. [[CrossRef](#)]
20. Haklay, M. How good is volunteered geographic information? A comparative study of OpenStreetMap and ordnance survey datasets. *Environ. Plan. B* **2010**, *37*, 682–703. [[CrossRef](#)]
21. Bishr, M.; Mantelas, L. A trust and reputation model for filtering and classifying knowledge about urban growth. *GeoJournal* **2008**, *72*, 229–237. [[CrossRef](#)]

22. Flanagin, A.; Metzger, M. The credibility of volunteered geographic information. *GeoJournal* **2008**, *72*, 137–148. [[CrossRef](#)]
23. Mooney, P.; Corcoran, P.; Winstanley, A.C. Towards quality metrics for OpenStreetMap. In Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems 2010, San Jose, CA, USA, 2–5 November 2010.
24. Barron, C.; Neis, P.; Zipf, A. A comprehensive framework for intrinsic OpenStreetMap quality analysis. *Trans. GIS* **2014**, *18*, 877–895. [[CrossRef](#)]
25. Zielstra, D.; Zipf, A. A comparative study of proprietary geodata and volunteered geographic information for Germany. In Proceedings of the 13th AGILE International Conference on Geographic Information Science, Guimarães, Portugal, 11–14 May 2010.
26. Girres, J.F.; Touya, G. Elements of quality assessment of French OpenStreetMap data. *Trans. GIS* **2010**, *14*, 435–459. [[CrossRef](#)]
27. Mashadi, A.; Quattrone, G.; Capra, L. The impact of society on volunteered geographic information: The case of OpenStreetMap. In *OpenStreetMap in GIScience*; Springer International Publishing: Cham, Switzerland, 2015; pp. 125–141.
28. Corcoran, P.; Mooney, P.; Winstanley, A. Topological consistent generalization of OpenStreetMap. In Proceedings of the GISRUUK 2010: GIS Research UK 18 Annual Conference, London, UK, 14–16 April 2010.
29. Stark, H.-J. Quality assessment of volunteered geographic information using open web map services within OpenAddresses. In Proceedings of the GI_Forum 2011, Salzburg, Austria, 5–8 July 2011.
30. Arsanjani, J.J.; Barron, C.; Bakillah, M.; Helbich, M. Assessing the quality of OpenStreetMap contributors together with their contributions. In Proceedings of the AGILE' 2013 International Conference on Geographic Information Science, Leuven, Belgium, 14–17 May 2013.
31. Hochmair, H.; Zielstra, D. Positional accuracy of flickr and panoramio images in Europe. In Proceedings of the GI_Forum 2012, Salzburg, Austria, 3–6 July 2012.
32. Jiang, S.; Alves, A.; Rodrigues, F.; Ferreira, J., Jr.; Pereira, F.C. Mining point-of-interest data from social networks for urban land use classification and disaggregation. *Comput. Environ. Urban Syst.* **2015**, *53*, 36–46. [[CrossRef](#)]
33. Scheffler, T.; Schirru, R.; Lehmann, P. Matching points of interest from different social networking sites. In *KI 2012: Advances in Artificial Intelligence 2012*; Springer: Berlin, Germany, 2012; pp. 245–248.
34. Mülligann, C.; Janowicz, K.; Ye, M.; Lee, W.-C. Analyzing the spatial-semantic interaction of points of interest in volunteered geographic information. In *Spatial Information Theory 2011*; Springer: Berlin, Germany, 2011; pp. 350–370.
35. Anand, S.; Morley, J.; Jiang, W.; Du, H.; Jackson, M.; Hart, G. When worlds collide: Combining ordnance survey and open street map data. In Proceedings of the Association for Geographic Information Geocommunity Conference 2010, London, UK, 30 June 2010.
36. Ludwig, I.; Voss, A.; Krause-Traudes, M. A comparison of the street networks of Navteq and OSM in Germany. In *Advancing Geoinformation Science for a Changing World 2011*; Springer: Berlin, Germany, 2011; pp. 65–84.
37. Yang, B.; Zhang, Y.; Luan, X. A probabilistic relaxation approach for matching road networks. *Int. J. Geogr. Inf. Sci.* **2013**, *27*, 319–338. [[CrossRef](#)]
38. Du, H.; Alechina, N.; Jackson, M.; Hart, G. A method for matching crowd-sourced and authoritative geospatial Data. *Trans. GIS* **2016**. [[CrossRef](#)]
39. Rodrigues, F.; Alves, A.; Polisciuc, E.; Jiang, S.; Ferreira, J.; Pereira, F.C. Estimating disaggregated employment size from points-of-interest and census data: From mining the web to model implementation and visualization. *Int. J. Adv. Intell. Syst.* **2013**, *6*, 41–52.
40. Berjani, B.; Strufe, T. A recommendation system for spots in location-based online social networks. In Proceedings of the 4th Workshop on Social Network Systems SNS'11, Salzburg, Austria, 10 April 2011.
41. Princeton University. About WordNet. Available online: <http://wordnet.princeton.edu> (accessed on 20 July 2016).

