Analysing Supplier Locations Using Social and Semantic Data: A Case Study Based on Indonesian Factories

Lisa Madlberger
Vienna University of Technology, Austria
Vienna Phd School of Informatics
lisa.madlberger@tuwien.ac.at

Heidelinde Hobel
Vienna University of Technology, Austria
Doctoral College Environmental Informatics
hobel@geoinfo.tuwien.ac.at

Andreas Thöni
Vienna University of Technology, Austria
Institute of Software Technology and Interactive Systems
andreas.thoeni@tuwien.ac.at

A Min Tjoa
Vienna University of Technology, Austria
Institute of Software Technology and Interactive Systems
amin@ifs.tuwien.ac.at

ABSTRACT

Many international corporations have globally distributed supply chains exposing their operations to various local risks, e.g., natural disasters. To facilitate assessment of these risks, corporations have to identify geographic locations of their suppliers. However, automated identification of supplier locations is problematic for areas where geocoding of addresses is not effective. In this paper, we present a method to infer location information from user-generated geographic information retrieved from Wikimapia and Foursquare. Using a sample of 139 Indonesian factories supplying large international corporations, we compared results from our approach with locations retrieved from four widely-used geocoding services. We found that best results could be achieved using data from Foursquare, where we retrieved a location within 1km for 73% of the factories. Given that coordinates are only an input for decision making, we linked retrieved locations exemplary to semantic data to determine the risk exposure due to volcanic eruptions for each factory. Both steps combined present an approach for automated supplier risk assessment based on social and semantic data.

Categories and Subject Descriptors

H.4.2 [**Information Systems Applications**]: Types of Systems—*Decision support*

General Terms

Design

i-KNOW '14, September 16 - 19 2014, Graz, Austria

ACM 978-1-4503-2769-5/14/09 ...$15.00.
http://dx.doi.org/10.1145/2637748.2638418

Keywords

Remote Supplier Risk Analysis, Social Media, Semantic Web

1. INTRODUCTION

Today, many corporations have a significant multinational footprint as their own operations, customers and suppliers are internationally distributed. Particularly, a company's supplier base is often very large and globally spread [14].

The geographic dispersion enables companies to leverage local benefits, e.g. lower cost of labor, but also exposes their operations to a higher risk of interruptions triggered by the local environment or society, e.g natural disasters or political unrest [25]. Moreover, corporations face an increased pressure of society and governments to take over responsibility for the local environmental and social impacts caused by their own or their suppliers' operations (e.g. [19]).

In order to assess and monitor these local risks, it is important for a corporation to be able to identify its suppliers and their accurate geographic locations. However, this requires a reliable translation of supplier factory names and addresses into geographic coordinates. Problems can exist for locations in areas which have not yet been well mapped, which is often the case for e.g. industrial areas in emerging countries. Especially for companies with hundreds of suppliers it is a work-intensive task to manually determine factory-locations in potentially far distant countries.

In contrast to this lack of geographic information from authorized sources, there is an increased interest in using the web to create, assemble, and disseminate information provided voluntarily by individuals to websites such as Wikimapia or OpenStreetMap[9]. Besides this intentionally contributed information, geographic data is also generated on social-networks like Twitter or Foursquare through geotagging of messages or directly in status updates, as a side

effect of online communication [20]. Furthermore, the Semantic Web and in particular the Linked Open Data cloud have emerged as valuable sources of open geographic and environmental data over the last years [3, 6, 5], representing a further source of decentralized gathered information.

The goal of this study is to explore whether user-generated geographic information (UGI) and open semantic data can be used to increase the knowledge of supplier locations. For this purpose we infer location information for a sample of 139 supplier factories of multi-national corporations using two platforms providing UGI (Wikimapia, Foursquare) and four widely used geocoding services (Google, Bing, Nokia Here, Nominatim). We evaluated retrieved results, using a manually developed ground truth data set and found that UGI increases the number of correctly inferred locations.

Given that the exact supplier location is only a first step in any risk assessment scenario, this paper further extends the analysis by linking the obtained location data to open environmental data, turning this raw location data into valuable business knowledge. Considering the regional focus of our study, and the fact that natural disasters are seen as the most important trigger for supply chain disruptions [25], we selected volcanic activity as a relevant risk to illustrate our approach in a business use case. We used data from two open sources (OpenStreetMap, DBpedia) to identify the factories located within a specified range of active volcanoes.

The biggest advantage of UGI as well as of Linked Open Data is that everyone can contribute and generate information, which comes however with the drawback of potentially low(er) data quality and credibility [8].

In order to address this challenge in our study, we developed a measure to assess the credibility of user-generated location records. We apply this measure to be able to automatically retrieve the most credible location for a given factory. For the risk assessment step, we accordingly employed a cross-validation mechanism to complement and validate information retrieved from OpenStreetMap with information retrieved from DBpedia in order to retrieve the most recent, credible information.

Our approach to analyze supplier locations using social and semantic data enables automated inference of factory locations, which is beneficial for companies with a high number of suppliers where the initial identification, the continuous tracking of changes, and the identification process of possible future supplier locations are time-intensive. Furthermore, using open data sources for risk assessment provides new ways to link supplier locations to various kinds of contextual information, such as the distance to the nearest volcano or hospital or to local economic indicators like the corruption index, which can be important inputs in risk assessment scenarios (see e.g. [21]).

The contributions of this study are as as follows:

- Comparison of location information retrieved for a sample of Indonesian factories from four widely-used geocoding services and from two platforms providing UGI.
- A method for automated location inference from user-generated data incorporating a credibility measure.
- A supplier risk assessment approach based on user-generated and semantic data.

The remainder of this paper is structured as follows: In Section 2, we discuss related work in both the areas of User-Generated Information (UGI) and Linked (Open) Data (LOD). In Section 3, we describe the data sample used for this study. Subsequently, we present two different ways of inferring location information for this data set (1) from existing geocoding services (2) using a novel method to infer credible location information from UGI, in section 4. In Section 5, we evaluate the locations obtained by different methods using a manually developed ground truth data set. We present our approach for automated risk assessment based on LOD in Section 6 and conclude with a discussion of the results in Section 7.

2. RELATED WORK

Since Internet-based technologies are becoming more and more pervasive in our lives in terms of ubiquitous computing and the current trend of sharing experiences and content enriched with location information, citizens all over the world are turning into social sensors [9]. This information could be used to derive up-to-date and detailed insights into far-distant areas. For instance, Roche *et al.* [17] presented the idea of "GeoWeb" in the context of user-generated and geolocation-based Web content for crisis management. Geospatial data mining for user-generated Web content is currently focusing on identifying geographical references, which are embedded in the metadata or text of the provided resources [10].

The need for user-generated spatial data has gained momentum as can be seen in the emergence of geo-mashups. According to [2], in August 2008 a mashup platform[1] offered 1740 spatial mashups and in February 2010 the amount had increased to 2153. Spatial data required to build sophisticated tools can be retrieved through crowdsourcing and geo-networking [2]. The quality of spatial data is steadily increasing due to the integration of automatically extracted geospatial entities, e.g., spatial objects extracted from Wikipedia [23, 24].

LOD is a relatively new concept that emerged within the Semantic Web. It was first officially defined by Tim Berners-Lee who published several rules that together define LOD in 2006 [4]. LOD is based on URIs that uniquely define data items. It uses HTTP for dereferencing and builds on the Semantic Web standards RDF and SPARQL. Finally, links between data sets are established through cross-linkages expressed by using the URIs of the different sets. Altogether, a vast amount of data is available as Linked Data. In September 2011 the LOD cloud contained 31.6 triples [5]. Eventually, the data available in the LOD cloud could be provided by anyone. This leads to potential concerns about the reliability, integrity, and usability of the data. Nevertheless, in domains where knowledge is publicly double-checked and where more data is better than no data at all, LOD can be

[1] http://www.programmableweb.com/tag/mapping/

utilized to enrich existing applications. A particularly interesting resource is the machine readable LOD conversion of Wikipedia called DBpedia, which is interlinked with a large amount of the LOD cloud [5]. In 2013 it contained 3.8 million concepts – 62 percent of them being classified based on an ontology [7]. Some authors have already discussed the potential use of LOD in the context of supply chain management. While Hofman et al. were more concerned with Linked Data in the supply chain itself [12, 11], Hulstijn et al. saw it as a mean for data publication [13]. Thöni [21] presented LOD as an option of Supply Chain Risk Management (SCRM) with regard to social sustainability. The ideas presented here go in line with Hofman who first suggested LOD in the context of SCRM – here a particular use-case based on geographical supplier data is proposed.

The combination of spatial data and LOD allows to facilitate the development of applications with local or global geographic perspective. From recommendation systems to event-visualization, the opportunities of the application areas are versatile, e.g., Sakaki et al. [18] exploited tweets to detect center and trajectory of events and developed a reporting system for earthquakes.

In our study, we aim to demonstrate the potential of user-generated information and LOD being used together in a business scenario, to identify locations of factories and assess associated risk factors. While many studies focused on the development of technical methods of location extraction, our goal is to demonstrate the applicability of such methods in a business scenario using a case study relying on a real-world data sample.

3. DATA SAMPLE

As a first step to increase the transparency in their supply chains, several multi-national companies recently published supplier lists on their websites, including the names of the factories and suppliers from which they are sourcing products or components. As a basis for our studies we used information about suppliers retrieved from the websites of four international companies from different industries.

Given that we rely on user-generated data to gain knowledge about suppliers and factories, we chose to focus our experiments on a region with high social media usage and a high occurrence of factories supplying the international market. Indonesia accounts for the third largest number of Twitter users (over 50 million users) [16] with Jakarta being the city where most tweets originate from [15] globally. With 54 million Facebook users, Indonesia ranks number four in the absolute number of Facebook users. [1] Hence, we identified Indonesia as an appropriate candidate for our studies due to its position as a country with a high social media penetration as well as an important status as a sourcing country for various products. To this end, from all retrieved records, we selected those located in Indonesia, resulting in a list of 163 factories including factory names and location information. While in 102 of 163 cases location information included the complete addresses, in the remaining 61 records only the name of the city was given. We detected 16 duplicate entries resulting from factories supplying several of the four initial companies. Furthermore, we deleted 7 entries where several buildings of the same factory with the same address were stated as separate factories. For the factories with missing address information we added address information manually as found in online address directories or on the factories' websites. The resulting data sample consists of 139 unique factories defined by their names and addresses.

4. LOCATION RETRIEVAL

In this section, we present two different ways to infer location information. First, we present the traditional way of using geocoding services to decode addresses into geographic coordinates. Second, we present a method for automated location inference based on user-generated geographic information retrieved from Foursquare and Wikimapia.

4.1 Location Retrieval based on Geocoding

A natural way to determine physical locations is to use address information and transform them into geographical coordinates, a process referred to as geocoding. We use the address information contained in our data sample to query four different geocoding interfaces:

1. `nominatim.com` Nominatim is the geocoding tool of OpenStreetMap
2. `developer.nokia.com` Nokia HERE
3. `msdn.microsoft.com` Bing Maps
4. `developers.google.com` Google Maps

While Bing and Nominatim returned hardly any results for our set of Indonesian addresses, the Google Maps API could decode 10 and the interface of Nokia Here 13 addresses to the point of a housenumber or a rooftop. The number of results increased with addresses geocoded on a lower level of detail, i.e., street, postal code or city. However, as streets in Indonesia are often several kilometers long, especially in industrial areas, we only consider information below this level as sufficiently located. Only four factories could be geocoded by two different services. Therefore, in total, we could retrieve coordinates for 20 of 139 factories based on address information. Results are outlined in Table 1.

Table 1: Locations Resulting from Geocoding

	Rooftop	House Number	Street	Postal Code	District	City	No Result	Total
Nokia Here		13	6	7	46	19	48	139
Bing							139	139
Google	4	6	19			78	32	139
Nominatim			1				138	139

4.2 Location Retrieval based on UGI

In a next step, we used user-generated geographic information from two platforms to infer factory locations:

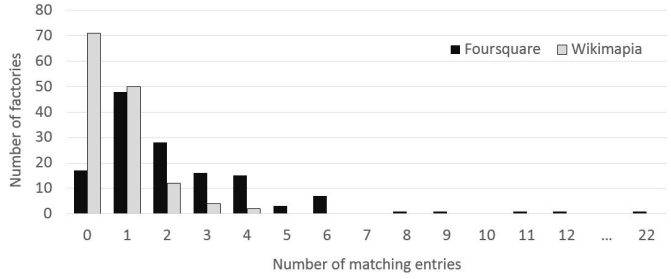
1. `foursquare.com` (FS) is a social network, where users can "check-in" to locations called "venues" and thereby indicate their current location and share it with friends on FS and other social networks. If a venue does not exist yet, users can add this venue to the directory.

2. `wikimapia.org` (WM) is a collaborative mapping project aimed at marking all geographical objects in the world and providing a useful description of them. It combines an interactive web map with a wiki system.

Both platforms have in common that they encourage users to tag places identified by geographic coordinates with labels and optionally additional information, e.g. categories, extended descriptions or pictures. Furthermore, both platforms provide the opportunity to query their location database through Web services. Given a location name and information about the city (in FS the city name, in WM the coordinates of the city) the services return a list of matching location records.

Querying the FS API we have retrieved 766 location records and after manual inspection we have identified 331 venues matching to entries in our sample of 139 factories. For 122 factories we found at least one matching entry, for other factories we found more than ten and up to 22 matching entries and for 17 factories we did not retrieve any matching entry (See also Figure 1).

Figure 1: Number of retrieved entries



The Wikimapia API returned over 200.000 entries for single locations. We limited the results to the top 5 entries where we could manually identify 94 matching entries. We could find at least one entry for 68 out of our set of 139 factories. From both sources together, we could retrieve location information for 136 of 139 factories.

These results suggest that UGI may contain valuable location information for most of the factories in our sample. However, in an environment where any user can contribute and change information with little verification mechanisms, obviously information credibility is a critical issue. Provided the fact that for most of the factories we retrieve multiple different location results, we cannot assume to retrieve unique, objective results from these platforms. Therefore, in order to use this information for automated location inference, we remain with the task to select the most credible location record and to judge whether this entry is credible enough to infer location information from it.

4.2.1 Credibility Assessment of UGI

The overall goal of the automated location inference process is to determine from a list of location records the one with the minimal geographic distance to the real factory location given a factory name. Since the real factory location is unknown, we have to rely on the credibility of a record to accurately correspond to the actual factory location. In order to assess this credibility, we make use of three features of a location record l_i: (1) a label $name_l$ (2) a category $cat_l \in \{0, 1\}$ (3) the number of checkins (only available for FS-data) $checkin_l$. We assume that the credibility for a record is high when the location label is similar to the factory name (i.e., the Levensthein-distance between the strings is small), when the record is tagged with the category "Factory" or "Production", and when it has a high number of checkins, which means that this location has been frequently confirmed by users.

We developed a measure of credibility (see Eq. (1)) for a location record $l = \langle name, cat\{, checkin\}\rangle$, to correspond to the queried location $q = \langle name\rangle$. Please note in Eq. (1), in case that checkin is not set, which is the case for data from WM, the last term (in curly brackets) will not be considered in the equation.

$$\begin{aligned} &Credibility(q, l) = \\ &\quad = w_{cat} \times cat_l + w_{dist} \times [1 - dist(name_q, name_l)] \\ &\quad \{+ w_{checkin} \times normalize(checkin_l, limit_{checkin}) \} \end{aligned} \quad (1)$$

Weights (w_{dist},w_{cat},$w_{checkin}$): We use weights to account for the influences of the different features to the total credibility. Note that the weights have to be chosen such that:

$$w_{dist} + w_{cat} + w_{checkin} = 1$$

We calibrated the weights based on manual inspection of the calculation results. The weights might have to be adjusted depending on the application scenario, in our study we used for FS: $w_{dist} = 0.8$, $w_{cat} = 0.1$, $w_{checkin} = 0.1$ and for WM: $w_{dist} = 0.8$, $w_{cat} = 0.2$.

Distance ($dist(name_q, name_l)$) (see Eq. (2)): We preprocessed both input parameters ($clean()$) by removing parentheses and the letters "PT" at the beginning and at the end of a name (a task specific to Indonesian company names e.g. PT Industry Indonesia), as well as by transforming each string to lower case characters. Subsequently, we calculated the Levensthein-distance ($lsd()$), which returns the number of insert, delete and replace operations to transform the first term into the second one. We divided this measure by the length of the longer term to normalize the distance.

$$\begin{aligned} &dist(name_q, name_l) = \\ &\qquad = \frac{lsd(clean(name_q), clean(name_l))}{max(length(name_q), length(name_l))} \end{aligned} \quad (2)$$

Normalization ($normalize(checkin_l, limit_{checkin})$) (see Eq. (3)): The parameter $limit_{checkin}$ introduces an upper limit

to the feature checkins in order to prevent that popular public places, like airports with millions of checkins, are assigned a too high credibility. We set $limit_{checkin} = 300$ since the number of checkins is below this threshold for most factories.

$$normalize(checkin_l, limit_{ch}) = \\ = \frac{min(checkin_l, limit_{ch})}{limit_{ch}} \quad (3)$$

Using this formula, we can obtain the location record with the highest credibility using the following algorithm.

For each factory in our data sample:

1. Query WM / FS using the factory name and the city.
2. Calculate the *Credibility* for each location record in the result.
3. Select the record with the highest credibility measure as the best match. If the credibility value is under a defined limit $limit_c$, no result is returned.
4. Sort the results according to their credibility and select those records that exceed a credibility threshold. Here, we set a threshold of 0.7 as most records under this threshold seemed to be too different to be considered as credible.

The algorithm returns an inferred location for each factory in our data set obtained by credible location records retrieved from Wikimapia and Foursquare.

5. EVALUATION

In order to evaluate both the results from geocoding as well as our method to infer location information from UGI, we manually developed a ground truth data set.

5.1 Ground Truth

Since we do not possess information about the real locations of the factories in our data sample, we manually reviewed the results from all geocoding services, FS, and WM in order to manually determine the location which seemed most plausible as a factory's actual location. We viewed the locations on different internet map services and determined whether the street name is aligned with the retrieved locations. Furthermore, we analysed the locations on satellite pictures, as factories can be easily differentiated from residential buildings by their relatively big rooftops and as they are typically not located within residential areas, in forests, water etc. In case of doubt, we also compared information and pictures from factory websites to the satellite pictures. Our confidence increased when information from multiple independent sources pointed to the same location. If no concrete location could be found, then this data record was omitted from our ground truth set and therefore also from further evaluations in the following sections.

As a result of this manual analysis process we compiled a list where we could match 119 of the 139 factories to geographic coordinates with very high confidence.

5.2 Location Retrieval from Wikimapia and Foursquare

We queried the WM and the FS API using names and cities of 119 factories and applied the algorithm described above to the location entries retrieved by FS (766 records) and to the top 5 results/factories returned by WM (671 records). Next, we computed the geographic distances between the automatically detected locations and the locations in our ground-truth data, using the Greater Circle Distance[2], a common measure to determine the distance between two points on a sphere.

If the retrieved location was within 1km distance of the location stated in our ground truth data, we considered the retrieved location as correct.

Applying the location inference algorithm to both WM and FS showed that FS covers a bigger range of factories, but the chance of retrieving a correct result was comparable in both sources (with precision: 0.73 (FS) and 0.69 (WM)), for the basic case, in which we used the manually determined weights for the attributes checkin and category. In subsequent experiments we excluded these attributes from the calculations and found that for FS information about the number of checkins improves the result, whereas the property category decreases the performance. For Wikimapia the inclusion of the category feature could slightly improve the result (See Table 2). Note that the number of retrieved results differs, as the inclusion of additional properties changes the credibility values and we used a static value for the credibility measure to decide whether a record is included in the result (see Section 4.2.1).

Table 2: Comparison of FS and WM results

	Foursquare				Wikimapia	
Category	✓	✓	-	-	✓	-
Checkin	✓	-	✓	-	-	-
Total queried	119	119	119	119	119	119
Total retrieved	108	107	107	114	51	58
<1km (Correct)	79	76	86	83	35	38
>1km (Incorrect)	29	31	21	31	16	20
Precision	0.73	0.71	0.80	0.73	0.69	0.66

5.3 Comparing User-Generated Information to Geocoding

In the next step, we compared the location inference performance from the geocoding services (Nokia Here, Google) to the one that was found based on UGI (Foursquare, Wikimapia). For each source and factory, we calculate the geographic distance between the returned location and our ground truth data. We compute three different performance measures, considering that results within (1) 1km (2) 5km (3) 10km of the actual location are considered as correct. The results are stated in Table 3. We found that for accurate location inference with low divergence the methods based

[2]We used Sedgewick's java function for the Greater Circle Distance `http://introcs.cs.princeton.edu/java/12types/GreatCircle.java.html`

on UGI performed significantly better (Precision FS: 0.73, WM: 0.69; basic case) than the geocoding services (Nokia: 0.22, Google: 0.25). The higher divergence can be explained by the fact that the geocoding services could only locate few factories to a high level of detail, e.g. house numbers (See Table 1). If we tolerate a higher divergence (<5km or <10km), the precision of geocoding improves, such that 50% of the factories could be located within a range of 10km compared to our ground-truth data. Which means that if one is interested into an approximate location, the use of geocoding services is a viable option.

The number of retrieved results differs between different sources, while Nokia (76%) and Foursquare (91%) returned results for most of the factories, Google (64%) and Wikimapia (43%) returned only fewer results. Considering the absolute number of correctly retrieved locations Foursquare clearly reached a better result than the other three information sources, about 73% of the factory locations could be determined within 1km.

Note that we used the basic case for FS and WM for this comparison, since we want to avoid overfitting, caused by optimization and evaluation performed on the same data sample. The results from the previous section however suggest, that the retrieval performance of FS could be even further improved, when the attribute category is excluded from calculations.

Table 3: Comparison Geocoding and UGI

		Nokia	Google	FS	WM
Total queried		119	119	119	119
Total retrieved		90	76	108	51
<1km	# Correct	20	19	79	35
	Precision	0.22	0.25	0.73	0.69
<5km	# Correct	51	52	91	37
	Precision	0.57	0.68	0.84	0.73
<10km	# Correct	60	61	92	37
	Precision	0.67	0.80	0.85	0.73

6. USE OF LOCATION INFORMATION IN SUPPLIER RISK MANAGEMENT

After presenting how locations of supplier factories can be automatically inferred from UGI, we show in this section how this information can be further enriched with semantic information to be used in a business scenario.

Knowledge about the physical locations of suppliers is an important input for multiple decisions in Supply Chain Management (SCM). Particularly risk is strongly associated with the geographic environment a supplier is embedded in. For example this is true for natural disasters such as earthquakes, hurricanes or volcanic eruptions which are generally considered as a major source of supply chain disruptions [25]. Given that Indonesia has around 400 volcanoes within its border of which at least 90 are still considered active [26], volcanic activity could pose a significant risk to supply lines in Indonesia. Consequently, we use the exposure to volcanos as a practical example in order to showcase how a basic natural disaster risk assessment of supplier locations could be performed relying on UGI and semantic information.

In a first step, we used OpenStreetMap (OSM)[3] to retrieve all volcanoes within 20km of a supplier location. For the fast retrieval of data we used the "Overpass API"[4] which is optimized for OSM data consumers and querying of geospatial data. Using the query presented in Listing 1 combined with the factory locations obtained in the first part of this paper, we found that 54 out of 139 factories are located within 20km distance of a volcano.

```
<osm-script>
  <query into="_" type="node">
    <around from="_" into="_" lat="<lat>"
        lon="<lon>" radius="20000"/>
    <has-kv k="natural" v="volcano"/>
  </query>
  <print from="_" limit="" mode="body" order="id"/>
</osm-script>
```

Listing 1: Volcano Overpass Query in XML Form

OSM returns the physical locations along with additional meta attributes including the name, the elevation, and the last eruption date. Among other information, scientists use historic eruptions in order to predict future eruptions (e.g., [22]). Moreover, the date of a volcanic eruption can give insights in whether a volcano may have caused problems recently. While this is an important information we found that information on the last eruption date provided by OSM is partly incomplete.

Therefore, this paper suggests improving the data quality available from OSM through augmenting and comparing it with data available from LOD. Dealing again with UGI we have to make assumptions about the credibility of the data. Since in our perception in a crowdsourced environment a missing update to metadata is more likely than an incorrect information, we replace older eruption dates with the most recent last eruption date available.

The update process takes place in multiple steps:

- First, the XML data containing the volcanos from OpenStreetMap are loaded and parsed including the eruption dates from OSM d_{OSM} if available.
- Second, the `eruptionYear` and the `lastEruption` properties are loaded for all objects of the type `dbpedia-owl:Volcano` that contained the volcano's Indonesian name in its label. The later year of the two dates was set as the new candidate year d_{NCY}. Listing 2 displays the SPARQL query used.
- Third, if the date from LOD d_{NCY} was later than the date in OSM d_{OSM} (if available), the new date was chosen. In case a year could be retrieved from DBpedia without an old date from OSM available, the new information was added.
- Finally, the XML structure was accordingly updated and saved to a file.

[3] http://www.openstreetmap.org
[4] `http://wiki.openstreetmap.org/wiki/Overpass_API`

```
PREFIX dbpedia-owl: <http://dbpedia.org/ontology/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX dbprop: <http://dbpedia.org/property/>

SELECT DISTINCT ?volcano ?eYear ?leYear
WHERE {
 ?volcano rdf:type dbpedia-owl:Volcano.
 ?volcano rdfs:label ?l.
 OPTIONAL {?volcano dbpedia-owl:eruptionYear ?eYear.}
 OPTIONAL {?volcano dbprop:lastEruption ?leYear.}

 FILTER (REGEX(STR(?l), "[VOLCANO_NAME_HERE]", "i"))}}
```

Listing 2: SPARQL to query DBpedia and return eruption dates

Altogether, 15 volcanoes have been retrieved from OSM. For 6 of these, the year of the latest eruption was provided. These could be complemented by two further eruption dates from LOD and an additional record for the volcano Galunggung. Details can be seen in Table 4. Particularly interesting is the update for the volcano "Tangkuban Perahu" which erupted in 2013. This could be an interesting detail for companies sourcing from the 5 factories located nearby this active volcano.

Table 4: Year of last eruption years for 15 Indonesian volcanoes based on OSM and DBpedia (LOD)

	Last Eruption Year			Factories <20km
Volcano	OSM	LOD	FINAL	
Gunung Ungaran				8
Tangkuban Perahu		**2013**	2013	5
Gunung Kiaraberes-Gagak	1939	1939	1939	7
Gunung Salak	1938	1938	1938	7
Gunung Perbakti		**1699**	1699	7
Gunung Telomoyo				3
Gunung Penanggungan				2
Gunung Linting				2
Gunung Ringgit				2
Gunung Pangrango				5
Gunung Merapi	2010	2011	2011	1
Gunung Merbabu	1797	1797	1797	1
Gunung Galunggung	1982	**1984**	1984	1
Gunung Telagabodas				1
Gunung Gede	1957	1957	1957	2

6.1 Overall Process

Figure 2 presents the steps taken in this paper in order to retrieve and integrating data from multiple public data sources to increase knowledge about factory locations. First, supplier data from public websites of four companies was used in order to compile a list of 139 supplier factories. Second, location information for these factories was inferred using UGI retrieved by Wikimapia and Foursquare. Third, inferred locations were enriched with geographic risk information based on data retrieved from OpenStreetMap (in our example volcanoes have been used). Finally, we retrieved Linked Open Data to perform a cross validation of geographical information and to improve the data quality.

Altogether this provides a showcase how open social and semantic data can be utilized to generate business knowledge. In our case, the use case of supplier risk evaluation was used as it is relevant for many multinational companies.

6.2 Limitations and Future Work

We focused on Indonesian factories in this study, future work could evaluate if similar results could be obtained using data from other countries, since both the platforms as well as the way how social media platforms are used differs regionally, results might be specific to a region.

Furthermore, in order to improve the expressiveness of the credibility measure, further features could be included. For example, the occurrence of multiple entries with a similar name located nearby could increase the credibility. The distance to other factories could indicate whether a geographic point is located in an industrial area. In a more extended version, one could also incorporate the results of geocoding services or try to automatically detect the visual patterns of rooftops of industrial buildings on satellite pictures.

Improved matching algorithms to retrieve volcanic metadata and further attributes and datasets could be explored. Overall, the enrichment for Supply Chain Risk Management could include additional (natural) risk factors apart from volcanic activities.

7. CONCLUSION

In this study, we inferred location information for 139 Indonesian factories from four different geocoding services (i.e, Google Maps, Bing, Nokia Here, Nominatim) and two platforms providing user-generated volunteered geographic information (Wikimapia, Foursquare). We evaluated the precision of retrieved locations based on a manually developed ground truth data set and found that UGI provided more precise and accurate information than geocoding services. Using Foursquare we could achieve the best results retrieving a location within 1km for 70% of the factories.

Our findings suggest, that the fact that local employees and residents start collecting and contributing geographic data can offer corporations an opportunity to possess better information about their factory locations. Especially when the number of suppliers and sub-suppliers is high and geocoding of addresses is not effective, automated location inference from UGI can reduce a lot of manual work.

In a second step, we demonstrated how the retrieved location information can be enriched with Linked (Open) Data to showcase how the geographic coordinates can be translated into insights that help in everyday Supply Chain Risk Management. For this purpose we retrieved and validated geographic data from OpenStreetMap and DBpedia and linked it to the inferred factory locations in order to assess their exposure to volcanoes presenting the last eruption year.

With increasing access to the internet and mobile technology, employees local residents in traditional outsourcing countries become users of (geo)-social networks and start gener-

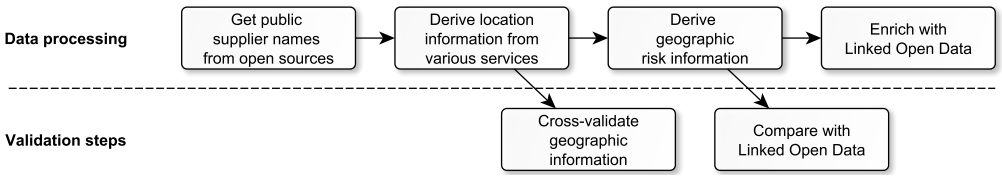


Figure 2: Processing and evaluation steps

ating information. Once this information can be processed and aggregated in a meaningful way, it can be used by corporations, governments or even consumers to gain better information about the physical locations of factory buildings and their environments but potentially also about other local aspects. In the future, we want to explore how social networks, like Twitter, can help to monitor local events like strikes, industry-related protests, and other sustainability risks in real-time, providing corporations, NGOs, and governments with an indicator for social unrest and problems caused by environmental and social impacts of companies.

8. REFERENCES

[1] Allfacebook.de. Facebook nutzerzahlen 2013. http://allfacebook.de.

[2] Michael Batty, Andrew Hudson-Smith, Richard Milton, and Andrew Crooks. Map mashups, web 2.0 and the gis revolution. *Annals of GIS*, 16(1):1–13, 2010.

[3] T. Berners-Lee, J. Hendler, O. Lassila, et al. The semantic web. *Scientific american*, 284(5):28–37, 2001.

[4] Tim Berners-Lee. Linked data - design issues, 2006.

[5] Chris Bizer, Anja Jentzsch, and Richard Cyganiak. State of the LOD cloud.

[6] Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked data-the story so far. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 5(3):1–22, 2009.

[7] DBpedia. wiki.dbpedia.org : About, 2013.

[8] Andrew J Flanagin and Miriam J Metzger. The credibility of volunteered geographic information. *GeoJournal*, 72(3-4):137–148, 2008.

[9] M. F. Goodchild. Citizens as voluntary sensors: Spatial data infrastructure in the World of Web 2.0. *International Journal of Spatial Data Infrastructures Research*, 2:24–32, 2007.

[10] Bo Han, Paul Cook, and Timothy Baldwin. Geolocation prediction in social media data by finding location indicative words. In *COLING*, pages 1045–1062, 2012.

[11] W. Hofman. Supply chain visibility with linked open data for supply chain risk analysis. In *Workshop on IT Innovations Enabling Seamless and Secure Supply Chains*, pages 20–31, 2011.

[12] W. Hofman, H. Bastiaansen, J. van den Berg, and P. Pruksasri. A platform for secure, safe, and sustainable logistics, 2012.

[13] J. Hulstijn, S. Overbeek, H. Aldewereld, and R. Christiaanse. Integrity of supply chain visibility: Linking information to the physical world. In *Advanced Information Systems Engineering Workshops, Lecture Notes in Business Information Processing*, volume 112, pages 351–365, 2012.

[14] Beatrice Kogg and Oksana Mont. Environmental and social responsibility in supply chains: The practise of choice and inter-organisational management. *Ecological Economics*, 83:154–163, 2012.

[15] Kalev Leetaru, Shaowen Wang, Guofeng Cao, Anand Padmanabhan, and Eric Shook. Mapping the global twitter heartbeat: The geography of twitter. *First Monday*, 18(5), 2013.

[16] PeerReach Twitter Analytics. http://blog.peerreach.com/. Accessed: 2014-01-28.

[17] S. Roche, E. Propeck-Zimmermann, and B. Mericskay. GeoWeb and crisis management: issues and perspectives of VGI, 2011.

[18] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 851–860, New York, NY, USA, 2010. ACM.

[19] Stefan Seuring and Martin MÃijller. From a literature review to a conceptual framework for sustainable supply chain management. *Journal of Cleaner Production*, 16(15):1699–1710, 2008.

[20] Anthony Stefanidis, Andrew Crooks, and Jacek Radzikowski. Harvesting ambient geospatial information from social media feeds. *GeoJournal*, 78(2):319–338, 2013.

[21] Andreas Thöni. Integrating linked open data for improved social sustainability risk management in supply chains. In Matthias Horbach, editor, *Lecture Notes in Informatics*, pages 916–927. Gesellschaft für Informatik, 2013.

[22] Michael B. Turner, Shane J. Cronin, Mark S. Bebbington, and Thomas Platz. Developing probabilistic eruption forecasts for dormant volcanoes: a case study from mt taranaki, new zealand. *Bulletin of Volcanology*, 70(4):507–515, February 2008.

[23] Jeremy Witmer and Jugal Kalita. Extracting geospatial entities from wikipedia. In *ICSC*, pages 450–457. IEEE Computer Society, 2009.

[24] Daryl Woodward, Jeremy Witmer, and Jugal Kalita. A comparison of approaches for geospatial entity extraction from wikipedia. In *ICSC*, pages 402–407, 2010.

[25] World Economic Forum. New models for addressing supply chain and transport risk.

[26] Worldatlas.com. Indonesia map / geography of indonesia / map of indonesia, 2014.