SEMANTIC LOCATION EXTRACTION FROM CROWDSOURCED DATA

S. Koswatte [a], K. Mcdougall [a, *], X. Liu [a]

[a] School of Civil Engineering and Surveying, Faculty of Health, Engineering and Sciences, University of Southern Queensland, West Street, QLD 4350, Australia - (Saman.Koswatte, Kevin.McDougall, Xiaoye.Liu)@usq.edu.au

Commission II, ICWG II/IV

KEY WORDS: Geospatial Semantics, SDI, Crowdsourced Data, Ontologies, QLD Floods

ABSTRACT:

Crowdsourced Data (CSD) has recently received increased attention in many application areas including disaster management. Convenience of production and use, data currency and abundancy are some of the key reasons for attracting this high interest. Conversely, quality issues like incompleteness, credibility and relevancy prevent the direct use of such data in important applications like disaster management. Moreover, location information availability of CSD is problematic as it remains very low in many crowd sourced platforms such as Twitter. Also, this recorded location is mostly related to the mobile device or user location and often does not represent the event location. In CSD, event location is discussed descriptively in the comments in addition to the recorded location (which is generated by means of mobile device's GPS or mobile communication network). This study attempts to semantically extract the CSD location information with the help of an ontological Gazetteer and other available resources. 2011 Queensland flood tweets and Ushahidi Crowd Map data were semantically analysed to extract the location information with the support of Queensland Gazetteer which is converted to an ontological gazetteer and a global gazetteer. Some preliminary results show that the use of ontologies and semantics can improve the accuracy of place name identification of CSD and the process of location information extraction.

1. INTRODUCTION

The Crowdsourced Data (CSD) has recently gained increased attention in many fields. Factors like technological development, improvements in mobile communication and availability of sophisticated software in the form of apps are supporting this growth. Moreover, production convenience, ready access, free and openness, currency and abundance of CSD are the key reasons for growing interest. During critical events like disasters, people use social media platforms (twitter, Facebook etc.) to communicate with others as it is the fastest and most convenient way to do so in the modern world. Because of this, the availability of CSD is very high in current disaster events.

This CSD would be a valuable resource for disaster management as the data is current and rich in information. However, there are quality issues such as incompleteness, credibility and relevancy, lack of availability of location information and vagueness of available location. Additionally, inherited problems in structure, documentation and validity of the CSD limit the direct applications of it for scientific and technical analysis (Flanagin and Metzger, 2008, Longueville et al., 2010). Researchers have now understood the value of available CSD and are concentrating their efforts to improve its quality.

Geospatial Information Retrieval (GIR) is important and widely used in many application areas like emergency response, transport planning, hydrology and land-use (Battle and Kolas, 2011). To this end, the most popular approach is to use gazetteers for retrieving GI from the web pages or online contents. Researchers argue that this is not very different from a keyword base search like in search engines (Buscaldi and Rosso, 2009). In recent GIR research, semantics are mainly used along with the gazetteers and other vocabularies. There are number of issues pertaining to GIR and those are discussed in detail in the latter sections of this paper. The scope of this paper is to extract the target geography from the social media communications using a semantic approach.

The objectives of this paper are to semantically recognize, extract and geo-code the content or target location information from the 2011 Queensland Flood CSD (General public Tweets and Ushahidi Crowdmap data) using an ontological gazetteer and other semantic geospatial resources. The paper is structured as follows: section 2 will discuss the background along with important studies conducted in the fields of CSD, GIR from CSD, gazetteers and ontologies. Section 3 introduces the methodology used throughout the study. Next, section 4 provides the preliminary results and discussion. Finally, section 5 will elaborate on the conclusions and future developments of this project.

2. BACKGROUND

2.1 CSD, Twitter and Crisis Mapping platforms

Current, reliable and high quality spatial data are crucial in successful disaster management. During disaster management, available data sources are often not optimally configured to ensure effective data management. Disaster management staff have the option to use government maintained authoritative data or other forms of data like CSD. . CSD provides the opportunity to use other related data that may have higher levels of currency or further depths. However, this data is often problematic due to lack of currency, completeness, access and availability.

* Corresponding author

Conversely, CSD is freely available and mostly contains current information about the event concerned.

Disaster related CSD are usually accessible through both desktop and mobile social media platforms (e.g. twitter[1], Facebook[2], flicker[3], foursquare[4], Ushahidi[5] etc.). It provides a readily available source for real-time and dynamic disaster related information to address the currency, completeness, access and availability issues pertaining to authoritative data in disaster management. Often CSD comprise of comments over an incident which occurred and then posted on top of base data like google maps or open street maps and spatial qualities like location information can be missing. CSD creators are generally laypersons and hence the end product may not result in qualified spatial data. Interestingly, the base maps used in crowdsourcing may also be developed by the crowd. Often the crowdsourced data can be improved to enhance the quality of the final product. To this end, it is argued that disaster management can be improved by optimising the use of CSD along with authentic data incorporating ontology and geospatial semantics.

As indicated previously, CSD can originate from a number of diverse sources, social media like twitter, Facebook, flicker, foursquare etc. or crowd mapping platforms like Ushahidi. Ushahidi (which means 'testimony' in Swahili) is a crowd mapping platform that was developed to easily capture inputs from people by cell phones or emails (Bahree, 2008, Longueville et al., 2010). Even though it's original development goal was to report the election violence in Kenya, over time its usage has changed towards natural disaster crisis mapping. The user can report an incident in various forms including SMS, email or web. The most notable advantage is that the users can report incidents onsite with the help of a mobile device.

Twitter is a very popular social media platform in which the conversations are limited to 140 characters. The users may pass their messages (tweets) very concisely and sometimes using quite different terminology including abbreviations, modified terms or slangs. If the user is skilful and experienced in enabling the location in their mobile device, the messages may include locational data as well. However, in general, the location availability is disabled due to privacy concerns or through the device location settings and therefore care must be taken when considering Twitter as a geospatial data source (Koswatte et al., 2014).

2.2 GIR, Gazetteers and Ontologies

Geospatial Information Retrieval (GIR) is critical in many application domains including emergency response, transport planning, hydrology, land-use and etc. (Battle and Kolas, 2011). Most of the studies attempt to retrieve GI from the web with the help of gazetteers. These approaches have mostly concluded with limited results due to the limitations of clear data definitions. To this end, semantics support the clear specifications of the spatial query. The objective of GIR is to geotag web pages based on its content which involves resolving two types of ambiguities i.e. geo-geo and geo-non-geo (Amitay et al., 2004). A geo-geo ambiguity occurs when two distinct places have the same name (e.g. Rockville in Queensland and Rockville in United States), and geo-non-geo ambiguity occurs when a place name also has a non-geographic meaning (e.g. Forbes is a town in New South Wales and Forbes is a popular magazine in USA).

The geography or the location information in GIR from web contents has identified two main types of location i.e. source and target (Amitay et al., 2004) or reporter and reported location (Koswatte et al., 2014). In this process, the source (or reporter) geography deals with the page/message origin or the server's/mobile device's physical location whilst the target (or reported) geography incorporates the content of the page. The source (reporter) location can also be defined by the provider location and serving location in contrast to web resources (Wang et al., 2005). With regards to a crisis, three types of location has considered in this CSD research i.e. a) reporter location b) incident location and c) content location. The scope of this paper is to extract the target geography (in contrast to GIR from web contents) or the content location (in contrast to GIR from CSD) using a semantic approach.

Gazetteers are geospatial dictionaries containing place names and related information like spatial references and feature types (Janowicz and Keßler, 2008, Machado et al., 2011). Many countries have developed and maintain their own gazetteers. Digital online formats like Alexandria Digital Library Gazetteer[6] (ADL), Getty Thesaurus of Geographic Names[7] (TGN) and GeoNames[8] are available (Machado et al., 2011). Furthermore, integrated semantic geospatial information retrieval systems are also slowly become available. A good example is GeoWordNet[9] (georeferenced version of WordNet[10]) which is an integrated system of GeoNames with WordNet plus the Italian section of MultiWordNet[11] (Giunchiglia et al., 2010, Buscaldi and Rosso, 2009). Gazetteers are widely used in Geospatial Information Retrieval (GIR) research (Borges et al., 2011, Amitay et al., 2004, Hill, 2000, Souza et al., 2005) but it is mostly argued that they are not fully supported in this sense as there are structural limitations and lack of intra-urban place names, no records on spatial relationships among elements other than relying on their proximity based footprints (Machado et al., 2011). Automatic recognition of geographic characteristics from web contents remain challenging and numerous approaches like automatic indexing and geo-referencing (Larson, 1996), ontology-driven approaches (Jones et al., 2001, Fu et al., 2005a), semantic query expansion (Delboni et al., 2007, Fu et al., 2005b) and natural language positioning (Delboni et al., 2007) along with gazetteers and geocoding techniques are proposed (Borges et al., 2011).

Ontologies are explicit specifications of shared conceptualizations and are key to establishing shared formal vocabularies (Du et al., 2013, Gruber, 1993). They are fundamentally important when dealing with heterogeneous systems and considered as a main pillar in so called semantic web. When considering the geo-spatial system manipulations it should be specifically conceptualized considering special geographic properties like inherited location and spatial integrity. Geo-spatial ontologies include geo-spatial entities,

[1] https://twitter.com
[2] https://www.facebook.com
[3] https://www.flickr.com
[4] https://foursquare.com
[5] https://www.ushahidi.com

[6] http://legacy.alexandria.ucsb.edu
[7] http://www.getty.edu/research/tools/vocabularies/tgn
[8] http://www.geonames.org
[9] https://datahub.io/dataset/geowordnet
[10] https://wordnet.princeton.edu
[11] http://multiwordnet.fbk.eu/

geographic classes and topological relations (Giunchiglia et al., 2010) and describe conceptual hierarchies and terminological interrelations of geospatial domain, and facts about spatial individuals along with location and geometry information (Du et al., 2013).

3. METHODOLOGY

3.1 The 2011 floods in Queensland and the study area

In January 2011, the state of Queensland Australia experienced one of the largest disaster events in its history. Nearly, 70 towns and 200,000 people were affected by severe flooding, claiming 35 lives and costing over $10 billion. This study will analyse citizen involvement in this natural disaster through the data that was collected via the #QLDFloods hashtag based Twitter communications and Ushahidi based crisis mapping platform. The study area (Figure 1) covers an area approximately 4000 km^2 where the majority of tweets and Ushahdi posts originated.

3.2 2011 QLD Flood CSD

The two months, December and January, 2011 were a very critical period for the Queenslanders with a series of flood events due to a La Nina event. With all of the flooding the social media, including Twitter and ABC's[12] Ushahidi based QLD Flood Crisis Map were busy with communications including severe weather alerts.

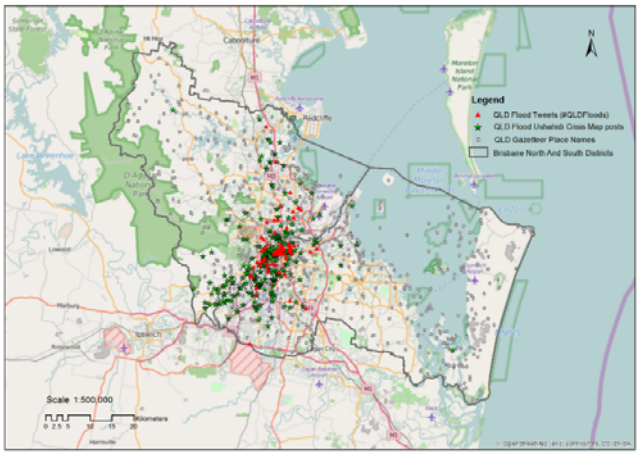

Figure 1. Study area and 2011 QLD Flood CSD

2011 QLD Flood Tweets: Through a special project carried out by ARC Centre of Excellence for Creative Industries and Innovation[13] (CCI), all the 2011 QLD Flood related tweets have been recorded using the open source tool yourTwapperkeeper[14] which is based on Twitter API and developed using PHP[15] and MySQL[16]. More than 35,000 tweets (based on the #qldfloods hash tag) were sent during 10-16 January, 2011 while more than 11,600 of them on 12th January alone. Moreover, there were more than 15,500 Twitter users participated using #qldfloods hash tag. During this period, leading accounts included the Queensland Police Service Media Unit (@QPSMedia), ABC News (@abcnews), and the Courier-Mail (@couriermail). @QPSMedia, (Bruns et al., 2012).

According to the findings of Koswatte et al., (2014) it was identified that the location availability of the 2011 QLD Flood tweets were only 1%.

ABC's Ushahidi based QLD Flood Crisis Map: During the early stages of the flood event, the Australian Broadcasting Corporation (ABC) maintained an interactive map based on the Ushahidi crowdmap platform to gather information related to the Queensland floods 2011. The public's uptake of the site was quite remarkable and more than 230,000 site visits over a 24 day period. According to the ABC's statistics, approximately 1,500 reports were received on the site and nearly 500 of these were from the public whilst another 1000 were generated by ABC moderators. Most reports were made through the online interface, however a small percentage of reports were made via emails, twitter and through SMS. The floodmap was most commonly browsed using the Internet Explorer browser (77%) via Windows operating systems (81%). Surprisingly, browsing using mobile devices was less than 5% of total visits (Potts et al., 2011). For mobile users, Ushahidi iPhone and Android apps were available.

Within the ABC's Queensland Flood Crisis Map dataset, there were approximately 700 reports during the period of 9-15th of January, 2011, which included the location information where it originated. These records were filtered and extracted for further analysis.

Selected samples from both 2011 Queensland Flood Tweets and Crisis Map data which fell inside the North and South Districts (Figure 1) of Brisbane City, Queensland, Australia were used as input CSD in this study. The study area was selected based on the high density of crisis communications which occurred. The sample contains 89 Tweets, 268 Ushahidi posts and 800 Queensland Gazetteer place name entries which are all provider location enabled.

3.3 The Research Approach

The Figure 2 illustrates the overall research approach. The study used Natural Language Processing (NLP) and annotation techniques incorporating additional resources like gazetteers. The GATE[17] (General Architecture for Text Engineering) software which is a robust and scalable open-source java based tool (Cunningham et al., 2002) developed by the University of Sheffield, United Kingdom was used for text processing including the semantic processing. GATE system components are termed as resources. The main three elements are Language Resources (LRs), Processing Resources (PRs) and Visual Resources (VRs). LRs are the entities like lexicons, corpora or ontologies. PRs are parsers, generators or modellers and VRs represent visualisation and editing components that participate in Graphical User Interfaces (Cunningham et al., 2002).

The first step of this research was to design and develop an ontology set for the Queensland Gazetteer. An ontology schema (Figure 3) was designed for Queensland Gazetteer based on OMT-G Gazetteer conceptual schema (Souza et al., 2005) and OnLocus simplified schema (Borges et al., 2011).

[12] http://www.abc.net.au
[13] http://www.cci.edu.au/
[14] https://github.com/540co/yourTwapperKeeper
[15] https://www.php.net
[16] https://www.mysql.com
[17] https://gate.ac.uk

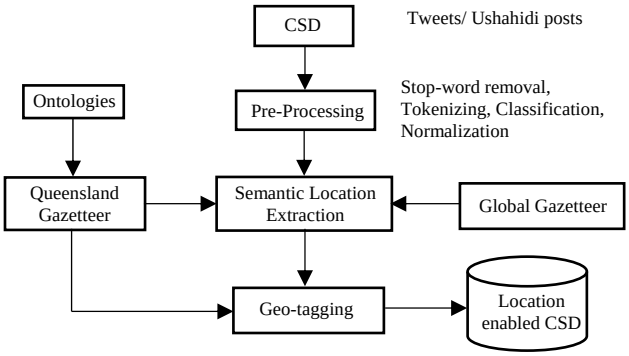

Figure 2. Semantic CSD Location extraction and Geo-tagging

Ontology design and development: Ontologies are key in semantic information processing. An ontology set was developed to convert the general Queensland Gazetteer to an ontological Gazetteer. That was to enable the semantic processing of the selected CSD in this study. Noy and McGuinness (2001) Ontology Development 101 guide was followed for developing the ontology which includes;

- ontology class definition
- class arrangement in a taxonomic hierarchy
- slot definition and value description
- value feeding for slot instances

The designed ontology was constructed using the GATE's ontology tools which provide the ontology viewing/editing facilities.

Processing and analysis using GATE software: The two datasets were analysed separately using the GATE. Processing Resources (PRs) of the GATE software; ANNIE's (A Nearly New Information Extraction system) English Tokenizer, Sentence Splitter, POS tagger, Transducer, and GATE's morphological analyser, and Document reset were used along with Queensland Place Name Gazetteer for non-semantic analysing. In the semantic analysis phase, ANNI OntoRootGazetteer along with Flexible Gazetteer were used along with the above processing resources.

The PRs were organized in the order of Document reset, Tokenizer, POS tagger, Morphological analyser, Gazetteer and then the Transducer for more effective processing and better results.

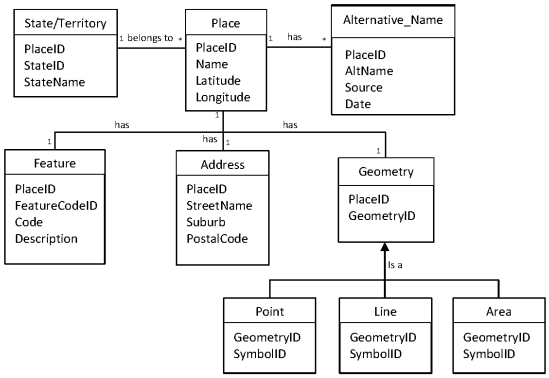

Figure 3. Simplified schema used in ontological Queensland place name Gazetteer

4. PRELIMINARY RESULTS AND DISCUSSION

This research work is still progressing and only some preliminary analysis has been undertaken thus far. The Table 1 shows the annotation accuracy matrix. As explained in GATE's Annotation Diff tool, precision (P) is a measure of number of correctly identified items as a percentage of the number of items identified, recall (R) measures the number of correctly identified items as a percentage of the total number of correct items and f-measure (F) is the weighted average of those two. ANNIE Gazetteer is a global gazetteer used in GATE as the default gazetteer. QLDGazetteer is Queensland's official place name gazetteer while QLDGazOnto is its ontological version developed in this study. It was developed with a main focus on the Ushahidi dataset and the results were dominant in tagging the Ushahidi dataset based on the ontological gazetteer. In future developments, it is planned to generalize the ontology set by considering a control dataset.

The results in Table 1 indicate that it is more advantageous to use local gazetteers in place name extraction. The use of the ANNIE gazetteer which is a global gazetteer provides the poorest results of all. The use of QLDGazetter alone dramatically improves the recall (R) factor but other measures are still better with the combined use of ANNIE and QLDGazetteer. Even though the combined use of global and local gazetteers shows some improvements, care needs to be taken not to introduce more geo-geo ambiguities. These ambiguities will be analysed in the future versions of this study. It can be seen some indications that the use of semantics would improve the place name extraction of CSD. The results of the use of ontological Gazetteer QLDGazOnto clearly improves the detection accuracy of Ushahidi posts. In case of Twitter posts, the semantic local Gazetteer outperforms the global ANNIE Gazetteer. No significant differences for the use of combined local and global gazetteers was detected.

Future research is planned to further improve the designed ontologies along with the ontological gazetteer. It is expected more stable and improved results through these modifications.

It is recognised that the results indicate a bias to the Ushahidi annotation accuracy as the ontology was developed on the same dataset. However, the annotation accuracy results of the Twitter dataset is encouraging as it is independent of the ontology development.

Table 1: Comparison of gazetteer success for Twitter and Ushahidi

Composition of Gazetteers	Ushahidi			Twitter		
	P	R	F	P	R	F
ANNIE Gazetteer	0.14	0.28	0.19	0.21	0.54	0.30
ANNIE+QLDGazetteer	0.32	0.41	0.36	0.37	0.64	0.47
QLDGazetteer	0.19	0.64	0.29	0.34	0.62	0.44
QLDGazOnto	0.96	0.90	0.93	0.36	0.55	0.44

5. CONCLUSION AND FUTURE WORK

In this paper we investigated how to extract the missing location of CSD. In this context we extracted location information from two different information sources from

Ushahidi and Twitter social media platforms. An ontology set was designed and developed to make the general Queensland Place Name Gazetteer a semantic Gazetteer which was termed QLDGazOnto in this study. The study is still progressing and the initial results were encouraging and open for further improvements.

In future, the ontology development will be more generalized and controlled. Furthermore, it is planned to examine and resolve the ambiguities of the identified location. The identified place names will be converted to a true location by hierarchical analysis and considering the adjacent place names through a selected span. The study also plans to apply the identified process to larger datasets.

ACKNOWLEDGEMENTS

Authors wishes to acknowledge Monique Potts, ABC – Australia for providing the QLD Crisis Map data and CCI and the researchers Bruns et al. (2012) for providing the Queensland 2011 tweet dataset.

REFERENCES

Amitay, E., Har' El, N., Sivan, R. & Soffer, A., 2004. Web-a-where: geotagging web content. Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, 2004 Sheffield, United Kingdom. 1009040: ACM, pp. 273-280.

Bahree, M., 2008. Citizen Voices. Forbes Magazine, pp. 182-183.

Battle, R. & Kolas, D., 2011. Linking geospatial data with GeoSPARQL. Semantic Web J Interoperability, http://www.semantic-web-journal.net/sites/default/files/swj176.pdf (Accessed: 25th November, 2015).

Borges, K. A., Davis Jr, C. A., Laender, A. H. & Medeiros, C. B., 2011. Ontology-driven discovery of geospatial evidence in web pages. *GeoInformatica*, 15, pp. 609-631.

Bruns, A., Burgess, J. E., Crawford, K. & Shaw, F., 2012. # qldfloods and@ QPSMedia: Crisis communication on Twitter in the 2011 south east Queensland floods.

Buscaldi, D. & Rosso, P., 2009. Using geowordnet for geographical information retrieval. *Evaluating Systems for Multilingual and Multimodal Information Access*. Berlin Heidelberg, Springer. pp. 863-866.

Delboni, T. M., Borges, K. A., Laender, A. H. & Davis, C. A., 2007. Semantic expansion of geographic web queries based on natural language positioning expressions. *Transactions in GIS*, 11, pp. 377-397.

Cunningham, H., Maynard, D., Bontcheva, K. & Tablan, V., 2002. A framework and graphical development environment for robust NLP tools and applications. ACL, 2002. 168-175.

Du, H., Alechina, N., Jackson, M. & Hart, G., 2013. Matching Formal and Informal Geospatial Ontologies. *Geographic Information Science at the Heart of Europe*. Springer. pp. 155-171.

Flanagin, A. J. & Metzger, M. J., 2008. The credibility of volunteered geographic information. *GeoJournal*, 72, pp. 137-148.

Fu, G., Jones, C. B. & Abdelmoty, A. I., 2005a. Building a Geographical Ontology for Intelligent Spatial Search on the Web. *Databases and Applications*. pp. 167-172.

Fu, G., Jones, C. B. & Abdelmoty, A. I., 2005b. Ontology-based spatial query expansion in information retrieval. On the move to meaningful internet systems 2005: CoopIS, DOA, and ODBASE. Springer. pp. 1466-1482

Giunchiglia, F., Maltese, V., Farazi, F. & Dutta, B., 2010. GeoWordNet: a resource for geo-spatial applications. *The Semantic Web: Research and Applications*. Springer. pp. 121-136.

Gruber, T. R., 1993. A translation approach to portable ontology specifications. *Knowledge acquisition*. 5, pp. 199-220.

Hill, L. L., 2000. Core elements of digital gazetteers: placenames, categories, and footprints. *Research and advanced technology for digital libraries*. Springer.

Janowicz, K. & Keßler, C., 2008. The role of ontology in improving gazetteer interaction. *International Journal of Geographical Information Science*, 22, pp. 1129-1157.

Jones, C. B., Alani, H. & Tudhope, D., 2001. Geographical information retrieval with ontologies of place. *Spatial information theory*. Springer. pp. 322-335.

Koswatte, S., McDougall, K. & Liu, X., 2014. Ontology driven VGI filtering to empower next generation SDIs for disaster management. In: Winter, S. & Rizos, C., eds. Research at Locate 14, 07-09 April 2014 Canberra, Australia.

Larson, R. R., 1996. Geographic information retrieval and spatial browsing. Geographic information systems and libraries: patrons, maps, and spatial information, Clinic on Library Applications of Data Processing, April 10-12, 1995.

Longueville, B., Luraschi, G., Smits, P., Peedell, S. & Groeve, T., 2010. Citizens as sensors for natural hazards: A VGI integration workflow. *Geomatica*, 64, pp. 41-59.

Machado, I. M. R., de Alencar, R. O., de Oliveira Campos Jr, R. & Davis Jr, C. A., 2011. An ontological gazetteer and its application for place name disambiguation in text. *Journal of the Brazilian Computer Society*, 17, pp. 267-279.

Noy, N. F. & McGuinness, D. L., 2001. Ontology development 101: A guide to creating your first ontology. Stanford knowledge systems laboratory technical report KSL-01-05 and Stanford medical informatics technical report SMI-2001-0880.

Potts, M., Lo, P. & McGuinness, R., 2011. Ushahidi Queensland Floods Trial Evaluation Paper: A collaboration between ABC Innovation and ABC Radio.

Souza, L., Davis, C., Borges, K. A., Delboni, T. M. & Laender, A. H., 2005. The role of gazetteers in geographic knowledge discovery on the web. Web Congress, LA-WEB 2005. Third Latin American. IEEE-2005.

Wang, C., Xie, X., Wang, L., Lu, Y. & Ma, W.-Y., 2005. Detecting geographic locations from web resources. Proceedings of the workshop on Geographic information retrieval, 2005. ACM, pp. 17-24.