GEOSPATIAL CROWDSOURCED DATA FITNESS ANALYSIS FOR SPATIAL DATA INFRASTRUCTURE BASED DISASTER MANAGEMENT ACTIONS

A Thesis submitted by

Saman Koswatte, BSc, MPhil

For the award of

Doctor of Philosophy

2017

ABSTRACT

The reporting of disasters has changed from official media reports to citizen reporters who are at the disaster scene. This kind of crowd based reporting, related to disasters or any other events, is often identified as “Crowdsourced Data” (CSD). CSD are freely and widely available thanks to the current technological advancements. The quality of CSD is often problematic as it is often created by the citizens of varying skills and backgrounds. CSD is considered unstructured in general, and its quality remains poorly defined. Moreover, the CSD's location availability and the quality of any available locations may be incomplete. The traditional data quality assessment methods and parameters are also often incompatible with the unstructured nature of CSD due to its undocumented nature and missing metadata. Although other research has identified credibility and relevance as possible CSD quality assessment indicators, the available assessment methods for these indicators are still immature.

In the 2011 Australian floods, the citizens and disaster management administrators used the Ushahidi Crowd-mapping platform and the Twitter social media platform to extensively communicate flood related information including hazards, evacuations, help services, road closures and property damage. This research designed a CSD quality assessment framework and tested the quality of the 2011 Australian floods' Ushahidi Crowdmap and Twitter data. In particular, it explored a number of aspects namely, location availability and location quality assessment, semantic extraction of hidden location toponyms and the analysis of the credibility and relevance of reports. This research was conducted based on a Design Science (DS) research method which is often utilised in Information Science (IS) based research.

Location availability of the Ushahidi Crowdmap and the Twitter data assessed the quality of available locations by comparing three different datasets i.e. Google Maps, OpenStreetMap (OSM) and Queensland Department of Natural Resources and Mines' (QDNRM) road data. Missing locations were semantically extracted using Natural Language Processing (NLP) and gazetteer lookup techniques. The Credibility of Ushahidi Crowdmap dataset was assessed using a naïve Bayesian Network (BN) model commonly utilised in spam email detection. CSD relevance was assessed by adapting

Geographic Information Retrieval (GIR) relevance assessment techniques which are also utilised in the IT sector. Thematic and geographic relevance were assessed using Term Frequency – Inverse Document Frequency Vector Space Model (TF-IDF VSM) and NLP based on semantic gazetteers.

Results of the CSD location comparison showed that the combined use of non-authoritative and authoritative data improved location determination. The semantic location analysis results indicated some improvements of the location availability of the tweets and Crowdmap data; however, the quality of new locations was still uncertain. The results of the credibility analysis revealed that the spam email detection approaches are feasible for CSD credibility detection. However, it was critical to train the model in a controlled environment using structured training including modified training samples. The use of GIR techniques for CSD relevance analysis provided promising results. A separate relevance ranked list of the same CSD data was prepared through manual analysis. The results revealed that the two lists generally agreed which indicated the system's potential to analyse relevance in a similar way to humans.

This research showed that the CSD fitness analysis can potentially improve the accuracy, reliability and currency of CSD and may be utilised to fill information gaps available in authoritative sources. The integrated and autonomous CSD qualification framework presented provides a guide for flood disaster first responders and could be adapted to support other forms of emergencies.

CERTIFICATION OF THESIS

This Thesis is entirely the work of Mr Saman Koswatte except where otherwise acknowledged. The work is original and has not previously been submitted for any other award, except where acknowledged.

Principal Supervisor: Professor Kevin McDougall

Associate Supervisor: Dr Xiaoye Liu

Student and supervisors signatures of endorsement are held at the University.

ACKNOWLEDGEMENTS

After four years of my PhD journey, it is time to thank people who supported me in various ways. It is my pleasure to thank all those who supported me to achieve my goals. Firstly, I am sincerely thankful to my principal supervisor, Professor Kevin McDougall for his untiring support, expert advice and guidance throughout the PhD journey. Thank you very much for being a mentor in its complete sense and for showing me the way to learn from mistakes. Secondly, I am indebted to my second supervisor, Dr Xiaoye Liu for giving a wonderful support including expertise, views and corrections whenever it was necessary. Thanks, Xiaoye for being a friend and a supervisor and giving me the strength to walk through this difficult path tirelessly.

I am thankful to academic staff members of the School of Civil Engineering and Surveying of Faculty of Health, Engineering and Science (HES), University of Southern Queensland (USQ) Dr Zhenyu Zhang, Professor Peter Gibbings, Dr Dev Raj Paudyal, Professor Amando Apan, Dr Kithsiri Perera, Mr Shane Simons and Ms Zahra Gharineiat for providing their expertise and various other help. For the administrative arrangements of the research I thank Professor Thiru Aravinthan, Dr Mark Emmerson, Mr Lester Norris and HES Research support team including Mrs Juanita Ryan. My sincere appreciation also goes to Mr Dean Beliveau for accommodation and technical support, Mrs Chris Fogarty for organising PhD meetings and various other support and Ms Robyn Takagaki for operational support. I thank to Faculty librarian Ms Sandra Cochrane for her librarian support. I appreciate the support of Dr Barbara Harmes for English language support and Mr Rod Little for English language editing and proof reading. I am also indebted to Mr Mohan Trada, Mr Adrian Blokland, Mrs Luz Suarez Cadavid, Mr Chris Power and Mr Luke Czaban for the support when working at the surveying store. I also wish to thank USQ ICT and technical support team including Mr Chris de Byl, Miss Sachindra Ranaweera and Mrs Vishaka Wijekoon.

I am thankful to the Australian government and the University of Southern Queensland for providing me with financial support for my PhD studies. My thanks also go to Ms Monique Potts, ABC – Australia for providing the 2011 Australia Floods' Ushahidi Crowdmap data and to Professor Axel Bruns, ARC Centre of Excellence for Creative

Industries and Innovations (ARC-CCI) for providing 2011 QLD Floods public tweets. I must thank to my postgraduate friends at USQ; Sombat Khawprateep, Esmaeil Ahmadinia, Majid Zargar, Madeeha Waseem, Adnan Abed Ahmed Luhaib, Surachai Wongcharee, Susan Alkurdi, Javad Hashempourand and Indika Herath for sharing their time views and other support. Specially, I thank to my friends Gayan Kahandawa, Piumika Ariyadasa, Buddhika Samarawickrama, Nalika Kalpage and their families and all the Toowoomba Sri Lankan community for encouragements and all other support during the stay in Australia.

Thanks, are also due to Sri Lankan government, University Grant Commission, former and current Vice chancellors, academic, non-academic and administrative staff members and students of the Sabaragamua University of Sri Lanka. My special thanks also go to all academic, non-academic, administrative and supportive staff members and students of the Faculty of Geomatics including former and current Deans and Heads of the departments.

My sincere thanks go to my mother Karunawathie and my late father Heenbandara who sacrificed their whole life for the betterment of our future. My father would be the happiest person to see this achievement if he was alive today. My sincere thanks also go to my sister and two brothers Radanika, Dharmasekara and Indika and their families, for caring and love. I am indebted to my late father-in-law Jeewendranath, mother-in-law Sunethra and her sister Mangalika, brother-in-law Kaushalya and his wife Nilushi, grand-father-in-law Ranjith and grand-mother-in-law Ranjanie for their kindness and love. Of course, my very special thanks go to my beautiful wife Sandee and lovely little daughter Thivedhya for their sacrifices in supporting my studies and for patience and understanding during my absence from home. Finally, my thanks go to everyone who supported in various ways that I could not mention here.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

ABC	Australian Broadcasting Cooperation
ADL	Alexandria Digital Library
ANNIE	A Nearly New Information Extraction system
ANZLIC	Australian New Zealand Land Information Council
AP	Average Precision
API	Application Programming Interfaces
ARC-CCI	Australian Research Council - Centre of Excellence for Creative Industries and Innovation
ASDI	Australian Spatial Data Infrastructure
BN	Bayesian Network
CBF	Content Based Filtering
CSD	Crowdsourced Data
CSV	Comma Separated Values
CTM	Conceptual Term Matrix
DS	Design Science
EMERSE	Enhanced Messaging for the Emergency Response Sector
FN	False Negative
FOSS	Free and Open Source Software
FP	False Positive
FST	Finite State Transducer
GATE	General Architecture for Text Engineering
GeoCONAVI	Geographic CONtext Analysis for Volunteered Information
GI	Geographic Information
GIR	Geographic Information Retrieval
GIS	Geographic Information System
GNSS	Global Navigation Satellite System
GPS	Global Positioning Systems
GSM	Global System for Mobile Communications
GSR	Geographic Scope Resolution
GYM	Google, Yahoo and Microsoft

HES	Health, Engineering and Sciences
HTML	Hyper Text Mark-up Language
ICT	Information and Communication Technology
IDE	Integrated Development Environment
IP	Internet Protocol
IR	Information Retrieval
IS	Information Science
ISO	International Organisation of Standards
IT	Information Technology
JAPE	Java Annotation Pattern Engine
JSON	(JavaScript Object Notation)
LR	Learning Resource
MAP	Mean Average Precision
MBR	Minimum Bounding Rectangle
NAVTEQ	Navigation Technologies Corporation
NER	Named Entity Recognition
NLP	Natural Language Processing
NMA	National Mapping Organisation
NS	Natural Science
OOV	Out of Vocabulary
OS	Ordnance Survey
OSM	OpenStreetMap
PGI	Professional Geographic Information
POS	Parts of Speech
PPV	Positive Predictive Value
PR	Processing Resource
QDNRM	Queensland Department of Natural Resources and Management
QLD	Queensland
RDF	Resource Description Framework
SDI	Spatial Data Infrastructure
SMS	Short Message Service
TF-IDF	Term Frequency – Inverse Document Frequency

TGN	Getty Thesaurus of Geographic Names
TN	True Negative
TP	True Positive
TPR	True Positive Rate
UGC	User Generated Content
USQ	University of Southern Queensland
VGI	Volunteered Geographic Information
VPS	Virtual Private Server
VR	Visual Resource
VSM	Vector Space Model
WWW	World Wide Web
WYSWYG	What You See is What You Get

PUBLICATIONS

1. Koswatte, S, McDougall, K & Liu, X 2017, 'VGI and crowdsourced data credibility analysis using spam email detection techniques', *International Journal of Digital Earth*, pp 1-13.
2. Koswatte, S, McDougall, K & Liu, X 2016, 'Semantic Location Extraction from Crowdsourced Data', *ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, pp 543-547.
3. Koswatte, S, McDougall, K & Liu, X 2015, 'SDI and Crowdsourced Spatial Information Management Automation for Disaster Management', *Survey Review*, 47 (344), pp 307-315.
4. Koswatte, S, McDougall, K & Liu, X 2014, 'SDI and Crowdsourced Spatial Information Management Automation for Disaster Management', *FIG Commission 3 Workshop*, 2014 Bologna, Italy.
5. 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.

LIST OF FIGURES

LIST OF TABLES

Chapter 1: **Introduction**

1.1. Background of the research

In recent years, the world has experienced more frequent and increasingly severe disasters. The Nepal earthquake on 25th April, 2015 killed over 8700 people and injured more than 21,000 people. It devastated many historical sites and destroyed over 500,000 houses worth billions of dollars (OCHA 2015). Typhoon Haiyan (locally known as Yolanda), 8th November, 2013 was the deadliest Philippine typhoon recorded, claiming nearly 10,000 lives. Disaster management is a sporadic exercise and it cannot be handled effectively by merely providing more and more resources. It is impossible to predict accurately the occurrence, frequency or severity of these natural disasters. What is required is effective, timely management of the situation to minimise further threats to lives and property damage. The management strategies need to be dynamic as with a major flood there may be more time available however, a cyclone may necessitate rapid management response. The mitigation actions such as carrying out awareness campaigns, strengthening the existing weak structures, preparation of disaster management plans can reduce human and property losses (Khan et al. 2008). Accurate and up-to-date geospatial information is vital in disaster management decision making. However, difficulties in accessing these services or the lack of real-time data of the event can lead to inaccurate or delayed decisions. Therefore, ‘direct input of data from those affected by the emergency, can make a life-saving contribution’ (Jackson et al. 2010).

In modern society, the popularity of social media has significantly changed the reporting and sharing of disaster related information. The ease of access to modern location sensors and readily available free and open source online mapping tools has encouraged the citizens to report these events with the assistance of digital online maps. This kind of citizen reported data, related to disasters or any other events, is identified as Crowdsourced Data (CSD). It can be considered as a 'special case of the more general web phenomenon of user generated content' or simply, Volunteered Geographic Information (VGI) (Goodchild 2007) which is a subset of CSD. Interestingly, CSD is more current and more diverse than conventional geographic information, although quality and credibility issues exist.

In past decades, disaster management has been successfully supported by Spatial Data Infrastructures (SDIs) to provide reliable spatial information (Jackson et al. 2010). The SDIs facilitate the spatial data sharing between organisations to discover, access and use available spatial data (Rajabifard & Williamson 2001; Rajabifard et al. 2002; Tait 2005; Foley 2009). In general, the SDIs will ensure the quality of the spatial data being shared (ESRI 2010). The traditional SDIs are highly institutionally focused and carry ‘a significant organisational inertia… an increasingly complex legislative framework that is difficult to change’ (McDougall 2009). These rigid organisational structures often hinder the benefits and block the direct linkage with CSD which are now widely available in social networks and crowd-mapping platforms.

SDIs are gradually evolving from data centric and process based approaches to user centric models (Sadeghi-Niaraki et al. 2010). Semantic based SDIs can also enable the automatic searching and processing of geospatial data and services. However, to make this a reality, suitable semantic services and user domain ontologies are required (Fernandez & Fernandez 2008). There are various definitions available for ontologies and Gruber (1993) defines ontologies as 'an explicit specification of a conceptualization' which are described formally to 'achieve shared and common understanding of a particular domain of interest' (Janowicz & Keßler 2008). Generally, SDI service providers publish and retrieve spatial information based on the background knowledge. The use of semantics and ontologies can enable the SDI service integration (Sadeghi-Niaraki et al. 2010).

A key challenge of SDIs is to maintain their spatial data currency. This is due to the high cost of accurate spatial data creation and updating, the high level of skills required to maintain the systems and the high costs of the infrastructure needed. There are hundreds of SDIs established throughout the world from local, state, national, regional to global levels (Rajabifard & Williamson 2001). Goodchild (2007), reminds us that ‘the six billion humans constantly moving about the planet collectively possess an incredibly rich store of knowledge about the surface of the earth and its properties’. As a result of the recent developments in internet technology and the growing interest in

providing data through social networks, Budhathoki and Nedovic-Budic (2008) identified the following important questions related to VGI production:

a. Why are SDIs lacking users while millions of people participate in VGI?
b. What factors lead to the differences of VGI participants freely contributing GI (Geospatial Information) and why are SDIs often reluctant to share information?
c. Are SDIs and VGI separate phenomenon or do they have some relationship?
d. Will it be better for the society if VGI is harmonised? If yes, how can it be harmonised?

Therefore, it is important to investigate the possibilities of incorporating freely and widely available CSD along with authoritative data to improve applications such as disaster management which rely on current and reliable spatial data. Moreover, such incorporations may also contribute to the data foundations for future events. The current and ongoing research on CSD is investigating the potential quality improvement through numerous approaches to enabling CSD to be utilised for critical applications such as disaster management. This research focuses on the methods of assessing and improving the quality of CSD and fusing this data with authoritative data such as SDI.

1.2. Research formulation

1.2.1. Statement of research problem

Current, reliable and high quality spatial data are crucial in successful disaster management. During current disaster management, available data sources are often not optimally configured to enable effective data management. Disaster management staff have the options of using government maintained authoritative data or other forms of

data such as CSD in fulfilling their spatial data needs. Authoritative government data has been the main choice for disaster management staff for many years. However, the lack of currency, completeness, access and availability are increasingly key challenges in using such data. 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. It provides a readily available source for real-time and dynamic disaster related information to address the above currency, completeness, access and availability issues pertaining to authoritative data in disaster management. However, the opportunities, challenges and quality issues of such data should be carefully analysed prior to utilising this data in critical applications such as disaster management.

The nature of CSD is often very different from the organised and standardised data available from SDI sources. In many cases, CSD may simply be comments regarding an incident which has occurred and is posted using base data such as Google Map[1] or OpenStreetMap[2] (OSM). CSD creators are generally laypersons or amateurs and hence the end-product may not result in high quality spatial data. Interestingly, the base maps used in crowdsourcing may also be developed by the crowd themselves. CSD are created in an 'informal and ad-hoc' nature and 'does not typically adhere to formal standards of geometric precision or meta-data consistency, neither does it provide consistency in coverage of detail' (Jackson et al. 2010). Basically, there are two views of information namely (1) objective understanding which does not depends on the observer or situation and (2) subjective or situational understanding (Hjørland 2007). On one hand, CSD are often communications or comments over an incident and they can be considered as subjective information. On the other hand, CSD may be based solely on observations rather than measurements and it is difficult to measure its quality by means of objective criteria (Flanagin & Metzger 2008; Longueville et al. 2010). Traditional spatial data quality assessment approaches cannot be readily applied over CSD

[1]https://maps.google.com
[2]https://www.openstreetmap.org

and hence a more sophisticated and knowledge based approach is proposed in which spatial semantics and ontologies are utilized.

The variability of CSD provides a good opportunity to explore if crowdsourced data can be improved to be more authoritative spatial data. To this end, it is argued that disaster management can be more effectively supported by optimising the use of CSD along with authoritative data by incorporating ontologies and geospatial semantics.

1.2.2. Aim, research questions and objectives

In respect of the research problem the main aim of the study was to:

Develop a semantic quality assurance process for crowdsourced data by analysing its quality based on location information availability, credibility and relevance so selected CSD can be fused with authoritative data for improved disaster management decision making.

The study identified the following research questions:

1. What methods could be utilised to improve the quality of CSD to be integrated with authoritative data and what are the important factors to be considered in deciding the CSD quality? (Chapter 2)
2. Can a generic spatial data quality matrix be used in CSD quality assessment? What benchmarks/ standards will guarantee the qualification level of CSD to be integrated with authoritative data? (Chapter 3)
3. Can crowdsourced data (such as tweets) which are missing the intrinsic location data be improved semantically? (Chapters 4, 5)
4. How can we identify the credibility and relevance of CSD? Can the techniques and approaches already available in the domains such as Information Technology (IT) be utilized for data validation? (Chapter 6)
5. How can we improve the quality of CSD to be authoritative data and how can we automate the CSD quality assurance process? (Chapters 7, 8)

To achieve the above main research aim, the specific objectives formulated in this study were to:

1. Review relevant literature to identify the critical dimensions and approaches in assessing the CSD quality and investigate the possibilities to improve CSD (Chapters 2, 3);
2. Develop a process to extract and geo-code the location information of CSD using Gazetteers[3] and ontologies and assess its quality (Chapters 4, 5);
3. Assess the credibility of CSD using appropriate filtering and processing techniques (Chapter 6);
4. Assess the relevance of CSD using Natural Language Processing (NLP) and Geographic Information Retrieval (GIR) techniques (Chapter 6);
5. Propose automation techniques to carry out CSD quality assurance processes and integrate with authoritative data for disaster management activities (Chapters 7, 8).

1.3. Justification for research

The frequency and severity of natural and manmade disasters poses a challenge for accessing and managing reliable and up to date spatial data. In the past, spatial data collection and manipulation have been managed by governments or related mapping authorities and governments that often spend a large percentage of their budgets keeping these spatial data up-to-date and in-line with their related business functions. However, over the last 10-15 years the role of government mapping agencies has changed and a considerable amount of data collection and management is being undertaken by private organisations. The government's exclusive rights of being a spatial data authority has changed in this regard and now the private organisations are slowly becoming significant spatial data providers (McDougall 2009). In the meantime, there is an increasing involvement by the private sector and academic sectors in developing SDIs

[3] Gazetteers are geospatial dictionaries containing place names and related information

(Williamson et al. 2004; Rajabifard et al. 2006). Public authorities have traditionally been reluctant to adapt CSD and are challenged by CSD due to the lack of managerial control and confidence in the quality of the data as the reliability and trust is unknown (Spinsanti & Ostermann 2013). In addition, citizens are more reluctant to voluntarily provide information to government than to private organisations (Economist 2008; Coleman et al. 2009).

The ready access to and availability of substantial amounts of CSD can be an advantage as there is the potential that this data could contain useful information. CSD in general comes from diverse sources and, hence, will be heterogeneous (Spinsanti & Ostermann 2010). Inherit problems in data structure, documenting and validation of the CSD limits the direct application of this data in scientific and technical analysis (Flanagin & Metzger 2008; Longueville et al. 2010). Investigating the ways of integrating subjective information received in the form of CSD with existing SDIs will be a major challenge (Spinsanti & Ostermann 2010). By developing a quality assurance and improvement mechanism of CSD, society may benefit in disaster warning and management. However, a major concern is the quality assessment of crowdsourced data, its actuality, credibility and relevance for a selected context (Bishr & Kuhn 2007; Flanagin & Metzger 2008).

1.4. Research approach

This research follows a Design Science (DS) research approach. The purpose of DS research in information systems (IS) is to achieve ‘knowledge and understanding of a problem domain by building and application of a designed artefact’ (Hevner & Chatterjee 2010). Peffers et al. (2007) describes the DS process in six steps:

1. Problem identification and motivation
2. Definition of the objectives for a solution
3. Design and development

4. Demonstration
5. Evaluation
6. Communication

The research approach (Figure 1.1) was developed using the concepts of the DS research approach.

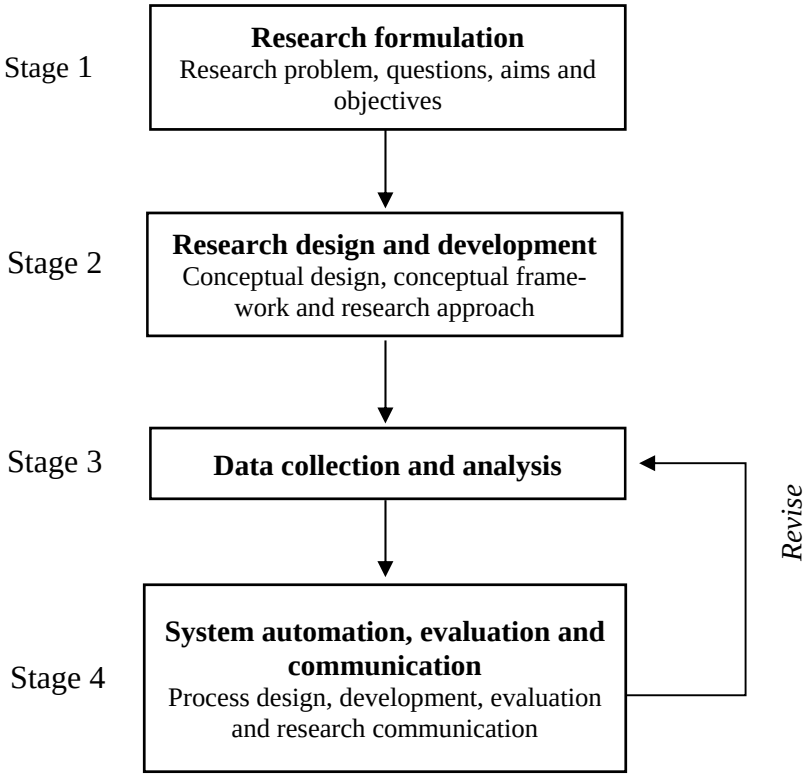


Figure 1.1 Research approach

1.5. Structure of the thesis

The thesis structure (Figure 1.2) includes eight chapters. Chapter one is the introduction and it consists of the background to research, statement of research problem, aim, research questions and objectives, justification for research, research approach, thesis structure and scope and limitations.

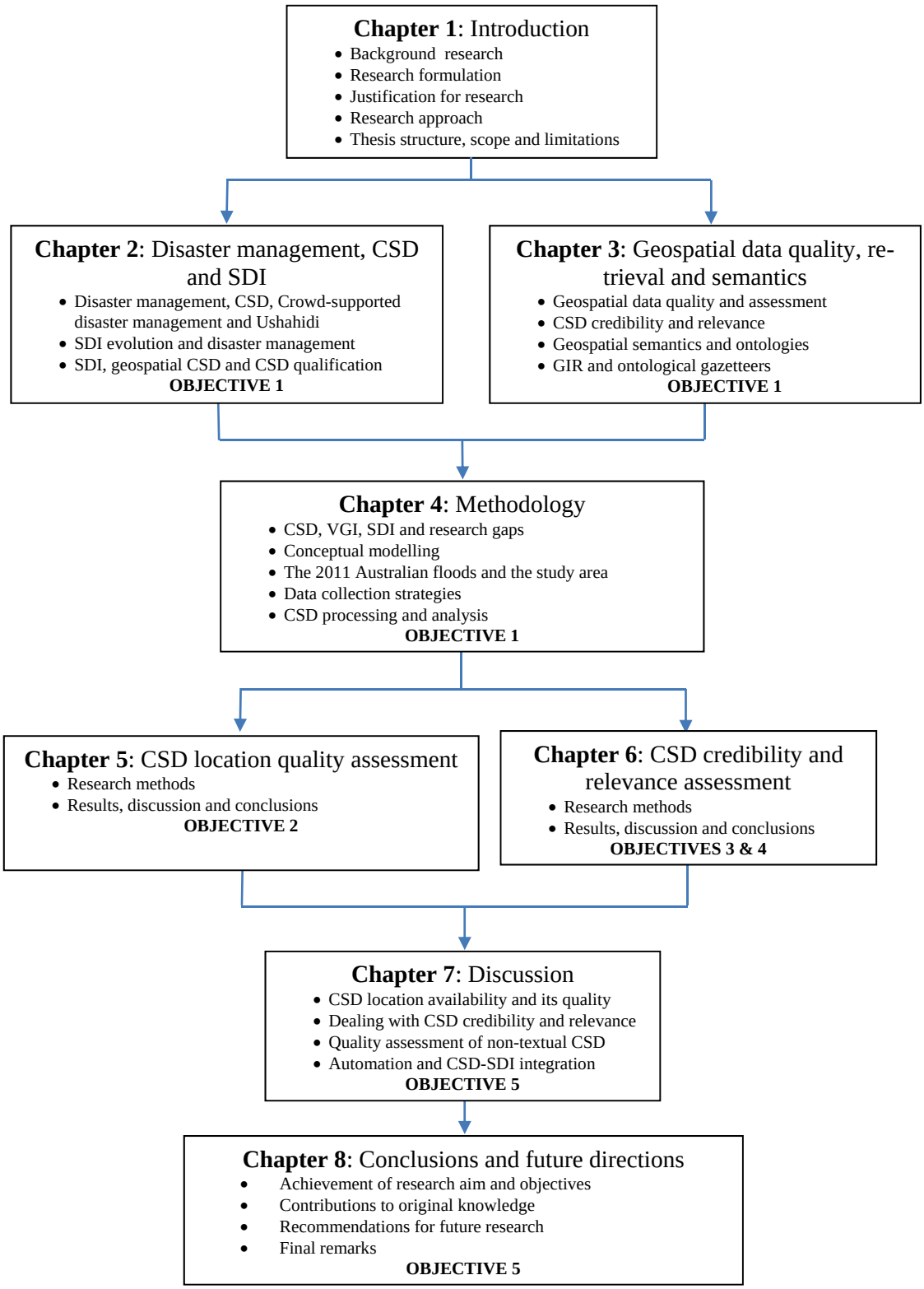


Figure 1.2 Thesis structure

Chapters two and three present the review of literature related to geospatial crowdsourcing, SDIs, disaster management, geospatial data quality, retrieval and geospatial semantics. These two chapters explain how the CSD quality assessment parameters have been identified. Chapter four explains the methodology used in the research along with the conceptual design, study area and a background of the data used. Chapter five explores the CSD location quality assessment and semantic location extraction methods while chapter six develops the CSD credibility and relevance assessment techniques. Chapter seven discusses the research findings, system automation possibilities and CSD integration with authoritative data. Finally, chapter eight describes the conclusions and recommended future work of this research.

1.6. Scope and limitation

The study investigates possible ways of identifying credibility and relevance while improving the quality of CSD to fuse with authoritative data in the context of disaster management actions. Within the possible crisis event domains, this study will be mainly focused on flood disasters. Floods and extreme weather events have resulted in significant disasters in recent times within Australia as well as around the world. The 2011 Australian floods (in South-East Queensland) broke many records, including the area inundated, water level rises, costs for recovery and the utilisation of the social media in crisis communication. More interestingly, involvement of the communities especially through social media such as Twitter[4] were eye opening in the case of 2011 Australian floods (Bruns et al. 2012). This was the motivation to critically assess and analyse possible ways of extracting, value adding and further improving information fed by the community through social media and micro blogging sites.

This research aims to contribute to the body of knowledge by developing a novel CSD quality assurance process to assist critical decision making in disaster management.

[4]https://twitter.com

Furthermore, the research will contribute more broadly to the knowledge and understanding for improving existing CSD quality assurance processes.

1.7. Chapter summary

This chapter has provided the foundation of the research by introducing the research problem along with the aims and objectives. The research and research approach has been explained and justified. The research design and structure of the thesis have been summarised. The research scope and key limitations of the research work have been outlined. The next chapter provides the literature review of CSD, SDI and disaster management research including information requirements and crowd involvements in disaster management.

Chapter 2: **Disaster Management, CSD and SDI**

2.1. Introduction

The purpose of this chapter is to explore the areas of disaster management, CSD and SDIs. The importance, issues and information requirements of disaster management and specifically the flood disaster management are discussed. The chapter also discusses the definitions, supported technologies, data sources, opportunities, applications, issues and challenges of CSD. Crowd supported disaster management and the Ushahidi[5] Crowd-mapping platforms are explored. The evolution of SDIs towards disaster management and the combined use of SDI and CSD are then discussed. Finally, CSD qualification opportunities and approaches of fusion with authoritative data are examined.

2.2. Disaster management

The term 'disaster' originated from French terms 'des' meaning bad and 'aster' meaning star thus in combination refers to 'Bad or Evil star' (Khan et al. 2008). In theory, a disaster is defined as ‘a serious disruption of the functioning of a community or a society causing widespread human, material, economic or environmental losses’ where the demands exceed the coping capacity of the affected community (UNISDR 2009). In natural disasters, the damage is caused by a natural phenomenon or a process. Examples of natural disasters can include floods, earthquakes, typhoons, tornadoes, volcanic eruptions or tsunamis. As it is difficult to predict or influence most natural disasters, the critical intervention is managing the situation and minimising further life threats and property damage.

Disaster management is normally a cyclic process which includes the main phases: mitigation, preparedness, recovery and response (Neal 1997; Poser & Dransch 2010). Figure 2.1 illustrates these phases and examples of associated activities. The mitigation

[5]https://www.ushahidi.com/

refers to the limitation and reduction of any adverse impacts of the disaster. In general, these adverse impacts cannot be prevented in full, however the 'scale or severity can be substantially lessened by various strategies and actions' (UNISDR 2009). The mitigation phase can include the actions of risk identification, risk analysis, risk appraisal and land use planning (Poser & Dransch 2010). Preparedness refers to how to respond to imminent disasters, that is to utilise knowledge models developed by various government and non-government organizations, professional bodies, communities and individuals based on the past experiences to effectively cope with future disaster events (UNISDR 2009). This can include emergency planning and training and establishing and managing early warning systems (Poser & Dransch 2010). During a disaster, the main focus is on search and rescue operations and providing humanitarian aids (Fiedrich et al. 2000; Poser & Dransch 2010) while maintaining the public safety. The objective of the post-disaster recovery phase is to bring activities back to normal after the event. This may include damage assessment, rehabilitation and reconstruction of damaged sites (Poser & Dransch 2010).

Figure 2.1 The phases of disaster management and related activities (Poser & Dransch 2010)

In emergency response, one of the key requirements is to provide shelter and assistance to disaster victims (Rawls & Turnquist 2010). Following a disaster, people will normally search for help and search for their loved ones while disaster first responders' primary concerns are: information integration related to the location and capacities of resource providers, spatial distribution of victims and immediate requirements for operations such as search and rescue (Fiedrich et al. 2000; Rawls & Turnquist 2010). In general, disaster management is a collaborative effort by police, fire, medical and ambulance services, non-governmental organizations and volunteers. The required information management is also a collaborative effort and no single organisation can maintain and keep up to date data (Mansourian et al. 2006). This research work has mainly focussed on the recovery and response phases where the public are mostly interested in sharing information.

2.2.1. Flood disasters and post-flood disaster response

Floods are considered as the most frequent and damaging type of disasters in the world (Jiang et al. 2009) and in Australia caused more loss of life than any other disaster (FitzGerald et al. 2010). Floods are often related with rivers and a common cause of floods is the high level of precipitation and the resulting river overflows. Flash-floods which are the result of extremely heavy rainfall are also common in Australia.

Similar to the other natural disasters, the risk of floods develops quickly and fade away with the time. The disaster management teams may be required to act accordingly in different areas of operations including search and rescue for lifesaving, infrastructure and resource management, resettlements, rehabilitation and communication during and after a flood. Prioritising these activities are essential based on the importance and the level of risk at each stage of the flood. As indicated previously, the focus of this research is the post-flood disaster management or specifically the post-flood disaster response phase.

2.2.2. Disaster management information requirements

Information plays a key role in any type of disaster management and it is important to understand what sort of information is required for the selected type of disaster. Steelman et al. (2012) conducted a study to understand the information exchange during disaster response particularly in wildfires. They identified that during wildfire disaster management there is a high demand for information related to fire status and behaviour, evacuation and road closures, inter/intra-unit communication, resource status and availability, values at risk, information administration, infrastructure and fire potential. In crowd-mapping platforms the type of information provided in each stages of disaster may vary.

Accurate and up-to-date geospatial information is vital in disaster management decision making. However, in cases of difficulties in accessing these services or lack of real-time data about the event, it can lead to inaccurate or delayed decisions. It is of utmost importance to arrive at a clear picture of the situation in any disaster before deciding the actions. The emergency management teams need to access up-to-date spatial data such as road networks, buildings, hospitals, fire stations and medical emergency stations (Mansourian et al. 2006). Ensuring easy access to current and reliable data is a key issue in this situation. The social media supported by Web 2.0, which emphasizes user-generated content and the dynamic form of the web, has made access to current information on a disaster now possible and there is a growing trend to use the social media as an 'important technology' for disaster management (Yates & Paquette 2011).

2.2.3. Post - flood disaster response information requirements and CSD

Disaster management information requirements may vary according to the type, stage and phase of the disaster. Citizens are actively involved in today's disaster communications and communicate information on different stages of disasters. CSD generated

by the citizens based on each phase of a disaster may be useful for effective disaster management. This may include 'information on identification and quantification of hazards, vulnerability parameters such as land use and distribution of assets or information for emergency action planning such as location of hospitals' and highly dynamic information related to the response and recovery phases such as 'observations of water levels and inundated areas, damages to infrastructure or the location of rescue teams' (Poser & Dransch 2010).

In the 2011 Australian floods, communications were occurring in near-real-time about local flooding, road closures and locations of evacuation centres and then changed to locations of bottled water supplies, disposal bin locations, clean up team locations, and lost and found pets when the disaster changed to its recovery phase (McDougall 2012). This is a good example of understanding the very dynamic and varied nature of CSD even within the time span of a single event.

2.3. Crowdsourced Data (CSD)

The Web 2.0 and related communication developments have created an immense repository of information including user generated content which is a variation to collective intelligence (Antoniou 2011). The popularity of crowdsourced data or user generated content has increased among scientific groups since Goodchild (2007) coined the term VGI (Capineri et al. 2016). Sensors such as cameras and Global Navigation Satellite System (GNSS) are now very common and standard features in today's smartphones. Other sensors are also common in today’s crowdsourced data collection for example, mobile weather data sensors which can provide further inputs and real-time data (Sosko & Dalyot 2017). The quality of these sensors is also high and the cost is quite low. Anyone with such a smartphone ‘can collect data and report phenomena more easily and cheaply than through official sources’ (Diaz et al. 2012) and it 'offers to multidisciplinary scientists an unprecedented opportunity to conduct re-

search on a variety of topics at multiple scales' (Capineri et al. 2016). Therefore, researchers are now seeking new approaches for improving and managing the quality of VGI and CSD in order to increase the utilisation of this data.

2.3.1. Definitions

The term 'crowdsourcing' is formed by the words: 'crowd' which refers to the people who participates in the initiative; and 'sourcing' referring to the number of procurement practices aimed at finding, evaluating and engaging suppliers of goods and services (Estellés-Arolas & González-Ladrón-de-Guevara 2012). The data created by crowdsourcing with the help of computer communication and other related technologies is defined as CSD. The crowdsourced geospatial data is identified as Neo-geography (Turner 2006), VGI (Goodchild 2007) or very close variations such as participatory geographic information (Elwood 2008b) or do-it-yourself maps (Jackson 2006). Capineri et al. (2016) identified the following as essential components of VGI:

- A geographical reference which represents the 'where we are' or 'where things are' information in the form of geo-tags, coordinates or toponyms
- The data content which can be transformed to information or knowledge (This may come in the forms of texts, images, symbols, maps, check-ins, photos, videos, drawings, etc.)
- Attributes including accuracy information, information about the users and producers and temporal information

Although, the definition and coining of the term Neo-geography can be debated, the voluntary engagement in creating geographic information is not new (Goodchild 2007). Jackson (2006) pointed out that Di-Ann Eisnor, the co-creator of the Platial[6] website which allows individuals without any programming skills to build personalized maps, first coined the term Neo-geography. Goodchild (2009) identified a defini-

[6]http://www.platial.com

tion for Neo-geography which is 'new geography' (Turner 2006) as ‘the usage of geographical techniques and tools ... for personal and community activities ... by a non-expert group ...’ based on Wikipedia[7] and Turner (2006). Goodchild (2007) defines VGI as 'a special case of the more general Web phenomenon of user-generated content' which is closely related to the crowdsourcing concepts (Capineri et al. 2016). This research uses the terminology VGI and CSD interchangeably as VGI can be considered as a subset of CSD.

2.3.2. Data sources

As indicated previously, CSD comes from diverse sources including social media and microblogging sites. Today's citizens ‘share and learn from their experiences through text (blogs), photos (Flickr[8], Picasa[9], Panoramio[10], …) and maps (Google Maps, Google Earth[11], …) not only seeking but also providing information’ (Spinsanti & Ostermann 2010). In this context people engaged in CSD production range from novices to experts in particular field with data originating from different sources including 'toponyms, GPS tracks, geo-tagged photos, synchronous micro-blogging, social networking applications, blogs, sensor measurements, complete topographic maps etc.' (Antoniu 2016).

2.3.3. Supported technologies

The technological advancements of computing, information systems, positioning and telecommunication has boosted the advent of CSD. Moreover, democratization of mapping and open source initiatives supported this trend towards geospatial CSD production. There are two basic technologies supporting the 'success of crowdsourcing geospatial data; (a) geo-referencing, either using GPS or mobile phone positioning and (b) the Web 2.0 development including broadband communication' (Heipke 2010).

[7] https://wikipedia.org
[8]https://www.flickr.com
[9]https://picasa.google.com
[10]www.panoramio.com
[11]https://earth.google.com

The information infrastructures, mainly Web 2.0 and easily accessible positioning devices (GPS) has enabled ‘users from many differing and diverse backgrounds’ (McDougall 2009).

Turner (2006) identified the following options for figuring out 'where we are?' or 'where things are?'.

GPS or GNSS: GPS is the US built version of Global Navigational Satellite System (GNSS). The key, or the most accurate option for neo-geographers is to use the inbuilt GPS device to capture the location. The inbuilt GPS is a very common sensor in today's mobile devices (e.g. smart phones, tablets etc.) and they are capable to achieving a high accuracy compared to the other available options to the citizens ranging from 3-5 meters horizontally and 10 meters vertically, depending on the environmental conditions. The GPS receiver calculates the global X, Y and Z coordinates by triangulating a minimum of four satellites to locate the user.

Geolocation by IP: An IP (Internet Protocol) address is a unique numerical label assigned to all computers, tablets, phones, printers, etc. when they connect to the internet. Since the IP address is unique it can be used to denote the physical address of the device which is identified as Geo-IP. However, the accuracy of this form of location varies with the form of connection to the internet (i.e. Static IPs, GSM cellular modem or through proxies) which can generally be brought down to city or post code level of accuracy.

Geolocation by cell tower: A cellular tower can be assigned an accurate location coordinate and this location can be shared with the devices connected to the cell tower. Similar to the geolocation by IP, in this instance it uses the wireless base station location as a registered user location. However, this method may use multiple cellular stations and their relative strength to calculate a better location for the device connected. The location accuracy of this method is generally around 30 m and it depends on the location accuracy of the cell tower and the signal strength.

Geolocation by Wi-Fi: Another option is to use the Wi-Fi[12] hotspot location as the user location. A wide range of Wi-Fi hotspots is available in many countries, even in regional locations. The connected device location can be calculated using a similar technique to the cellular towers based on the signal strength. The location accuracy is highly dependent on the signal strength and accuracy of the Wi-Fi hotspot locations and is better than geolocation by IP or cell towers but less reliable than GPS.

Geocoding: Generally, a location is identified as an address (or named location) e.g. '300, Healy Street, South Toowoomba'. Geocoding refers to converting addresses to coordinates (e.g. latitude, longitude) automatically using geo-coding services such as Google geocoder. Reverse geo-coding in which converting geographic coordinates to addresses is also possible through reverse geocoders.

Crowdsourcing applications such as Twitter may opt to use these various mechanisms to enable location in their feeds. However, if users are concerned about their privacy, the locations may be coded as an address or toponym within the text. This research identifies the first four geo-location options as explicit locations and the fifth one as an implicit or semantic location which is still useful for time critical applications including disaster management.

2.3.4. Opportunities and application areas

The crowd was initially considered as data consumers, however new trends and facilities translated the ‘consumers’ into ‘producers’ (Budhathoki & Nedovic-Budic 2008). Turner (2006) suggests new tools and methods to utilise this data created by ‘neo-geographers’ in the field of ‘neo-geography’. Although this geographic content is created by amateurs or non-geographers, the value of the CSD is well understood. The CSD may also create a real opportunity for mapping related organisations to keep their data updated and even enrich their databases (Coleman et al. 2009).

[12] A technology for wireless local area networking

Geospatial CSD, which is sometimes considered as VGI, is important due to its relatively unique characteristics. The knowledge and experience of local producers of such CSD may be better than experts of distant government agencies and may create better content (Goodchild & Li 2012). This argument aligns with Tobler's (1970) first law of geography: 'everything is related to everything else, but near things are more related than distant things' and hence geospatial CSD could be an important asset as a geospatial information source. The other important fact in supporting geospatial CSD is that authoritative data repositories are becoming increasingly out of date. However, new techniques now provide opportunities for collecting more accurate data by citizens (Goodchild & Li 2012).

2.3.5. Issues and research gaps

Researchers are still struggling to explain and understand the reasons for volunteers' motivations to generate geographic related information. Coleman et al. (2009) summarized a list of eight motivations in providing constructive contributions including (1) altruism, (2) professional or personal interest, (3) intellectual stimulation, (4) protection or enhancement of a personal investment, (5) social reward, (6) enhanced personal reputation, (7) outlet for creative & independent self-expression and (8) pride of place.

Data quality is considered as the key issue with respect to the CSD (Flanagin & Metzger 2008; Coleman 2013; Fonte et al. 2015; Senaratne et al. 2016). Variability in quality may emerge when using CSD collected from social media such as Twitter due to spontaneous responses from a heterogeneously interested community (Capineri et al. 2016). Cooper et al. (2011) identified that the biggest challenge of VGI is its inability to be assessed at the time of its creation. Moreover, they suggest that spatial data quality is highly subjective and the VGI producers are unable to assess their contribution as the expected quality is highly depend on the user, purpose and context of the usage. Another issue in CSD is the location ambiguity. With geo-referenced CSD locations, ambiguities such as geo-geo ambiguity may exist (i.e. the same name may

represent multiple locations and same location can have multiple names) or geo-non-geo ambiguity (i.e. the ambiguity of non-locational names and actual locations e.g. Sydney as a name versus a city).

In Twitter, the conversation is limited to 140 characters, so the users need to pass their messages (tweets) very concisely using quite different terminology including abbreviations, modified terms or slang terms. If the user is skilful and experienced in enabling the location in their tweets, their geo-tagged messages may include locational data. However, the location availability is generally disabled due to concerns in regard to privacy or a lack of knowledge of the user on the locational settings. Therefore, care must be taken when considering Twitter as a geospatial data source (Koswatte et al. 2014).

Criscuolo et al. (2016) pointed out three issues pertaining to CSD namely:

1. missing meta information in part or whole (i.e. temporal and geographic references, spatial resolution, quality and validity, constraints, etc.) which allows the precise location in space and time, including the associated usage parameters (i.e. acquisition procedure, measurement accuracy, instrumental precision, time stamp, contact details etc.)
2. difficulties in spatial overlay and thematic integration due to 'different or commonly undefined instrumental precision, reference systems, spatial and temporal granularity, together with the absence of common attributes and conceptual schemas'
3. issues related to trustworthiness of the contributions

CSD generated by the citizens may also be important for managing natural disasters such as floods. However, despite the advantages the use of such information may be challenging due to (1) unpredictable availability, (2) unknown data quality, (3) bias towards severe events, (4) localisation issues and (5) data collection and data structure issues.

In general, crowdsourced geospatial data are mostly unstructured and the accuracy undocumented. Different vocabularies and concepts used in CSD production are often transmitted in different languages. Having a common understanding of these diverse conceptualizations and terminologies is important for critical applications such as disaster management. Misunderstandings could lead to severe consequences, even the loss of life. The conventional way of dealing with conceptual diversity is to use special glossaries and vocabularies developed in the disaster management domain. However, these guides are quite incomplete when handling social media communications.

Researchers have proposed various measures (e.g. based on International Organisation of Standards-ISO[13] principles and guidelines) and indicators (other proxies) (Antoniou & Skopeliti 2015) for VGI and CSD quality assessment. A recent and more detailed review on VGI quality assessment is given in Senaratne et al. (2016). The authors have identified seven measures and ten indicators within the 56 papers they surveyed. The three most prominent measures included credibility (Metzger 2007; Bishr & Mantelas 2008; Van Exel & Dias 2011; Keßler & de Groot 2013), completeness (Koukoletsos et al. 2012) and vagueness (De Longueville et al. 2010).

2.4. Crowd supported disaster management and Ushahidi Crowd-mapping platform

Successful decisions in disaster management are often solely based on the availability of current and reliable data. It is of utmost importance to arrive at a clear picture of the situation in any disaster before deciding the actions. The social media supported by Web 2.0 has made access to current information on a disaster now possible. In the recent times, it can be seen that there is a growing trend to use the social media as a key data source for disaster management information (Gao et al. 2011; Yates &

[13]http://www.iso.org/

Paquette 2011). Examples such as Sahana[14] (Currion et al. 2007), an open source software developed in Sri Lanka to assist with the 2004 Tsunami and later with the 2005 Pakistan earthquake, the 2005 Philippine mud slide, 2007 earthquake in Peru and the 2008 earthquake in China (Bahree 2008) clearly evidence the potential of crowdsourced information in natural disaster mitigation. Moreover, crowdsourced data has contributed greatly 'in fighting wild fires in California, in mapping and monitoring Gulf oil spills and responding to the Haiti Earthquake. Social media has also had a significant impact in riot containment or 'kettling' in London, and the so called 'Arab Spring' in Middle Eastern countries' (Harris & Lafone 2012) and in flood research (Schade et al. 2010).

It can be seen with the recent disasters that 'today's emergencies are typically first seen through the 'eyes' of personal mobile camera phones that transmit in near real-time to international media broadcasting organizations rather than through formal government communication channels' (Jackson et al. 2010). This enables disaster information to potentially reach millions of people within minutes.

Another example of crowdsourced mapping platforms is the Ushahidi platform which was initially developed to easily capture crowd inputs by cell phones or emails (Bahree 2008; Longueville et al. 2010). Ushahidi means 'testimony' in Swahili, and is a platform that was utilised to report on election violence in Kenya. Over time, its popularity increased and the platform has been successfully deployed in a number of disasters around the world. Users can report incidents through various forms including SMS (Short Message Service), email and the internet. The most notable advantage is the convenient use of mobile devices which leads to the onsite incident reporting.

The Uhsahidi platform has been used successfully in a range of disasters including the 2011 Australian floods, the Christchurch New Zealand earth quake and the tsunami in Japan. From December, 2010 to February, 2011 the people of Australia (especially

[14] http://sahanafoundation.org

Queensland), experienced a series of damaging flash floods. In February 2011, Christchurch was hit by a magnitude 6.3 earthquake which caused widespread damage across the city killing 181 people and over NZ $20 billion in property damage. In March 2011, a massive earthquake of magnitude 8.9 estimated at 100 times more powerful than the Christchurch earthquake occurred in Japan causing over 20,000 deaths and more than US $300 billion damage to properties and infrastructure (McDougall 2012). The citizen responses through the established Crowdmaps based on the Ushahidi platform were part of each of these disaster responses. Table 2.1 shows the characteristics of Crowdmaps used in the three disasters.

Table 2.1 Characteristics of Crowdmaps used in three disasters (McDougall 2012)

Characteristic	**Queensland Floods**	**Christchurch Earthquake**	**Japanese Tsunami**
Site establishment time	Approximately 48 hrs	12-24 hrs	6-12 hrs
Utilisation	Alerts, photo, blocked roads, recovery points	Hazards, road closures, drinking water, building damage	Trapped people, dangerous areas that should be avoided, and supplies of food and clean water
Lifecycle	Active for approximately 5 weeks	Active for approximately 3 weeks	Active 8 months after tsunami
Reported quality	99% verified reports	Unknown	6.1% verified
Availability of site	Data currently accessible	Site not available	Active
Number of reports	98,000	>100,000	>12,600

2.5. SDI evolution towards disaster management

2.5.1. Spatial Data Infrastructures (SDIs)

SDIs can be referred to as the collection of technologies, policies and institutional arrangements enabling the ready access to spatial data (Nebert 2004; Ajmar et al. 2008). It provides the basis for facilitation and coordination of spatial data exchange and spatial data sharing among stakeholders from different jurisdictional levels in the spatial data community (Rajabifard & Williamson 2001; McDougall et al. 2009). With reference to the existing definitions for SDI, the main components of SDI are people, access network, policies, standards and data. These components can be classified as the basic components of SDIs. Although a number of the basic components of any SDI will be similar, the content of each component of a SDI might be different from the same component of another SDI (Rajabifard et al. 1999).

Traditionally, SDIs have a top down structure in which organisations govern all the processes and the user generally receives the final product. This is a mismatch with the concepts of Web 2.0 and with the notion of CSD. According to the concepts of Web 2.0, it is identified as a user driven web. In previous forms of the Web, the information/data flow was mainly unidirectional. However, this has recently changed to a bidirectional data flow and user generated content, usability and with interoperability emphasised. To match the concepts and for smooth functional transition of SDIs towards the current form of the web technology, Bishr and Kuhn (2007) suggests the inversion of the process from Top-down to Bottom-up which also clears the path for the next generation SDIs.

In 1990s, the accepted spatial data model was in pyramid style (Figure 2.2) which was based on government data sources; however, in more recent years, this pyramid is increasingly inverted (Harris & Lafone 2012). Related Application Programming Interfaces (APIs) also need to be improved to make them more user-friendly and, therefore

'it is likely that SDIs and data stores will need to be retro-fashioned into API integration systems to ease the integration of past and future data sets' (Harris & Lafone 2012).

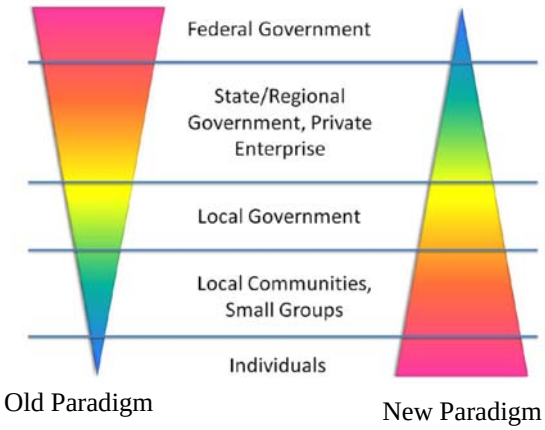


Figure 2.2 The spatial data sources old and new paradigms (Harris & Lafone 2012)

The popularity of location sensor enabled (GPS) mobile devices along with interactive web services such as Google Maps, OSM or Wikimapia[15], have created a user friendly platform for citizens to engage in mapping related activities (Elwood 2008a). This kind of platform will support and encourage crowdsourced geospatial data. Previously, to perform such tasks required highly technical skills and theoretical backgrounds in surveying, mapping and cartography. The citizens' engagement of spatial data generation in this manner has opened a potentially new data source for SDIs.

2.5.2. The role of SDIs in disaster management

SDI facilitates disaster management as it facilitates the collaboration in spatial data collection and sharing among the parties involved in disaster management through web based tools (Mansourian et al. 2006) and in mobile and crowdsource contexts (Laura et al. 2012; Koswatte et al. 2015). In general, SDIs facilitate access to geographically-

[15]http://wikimapia.org

related information using a minimum set of standard practices, protocols, and specifications and commonly delivered electronically via the internet (Ajmar et al. 2008). SDIs can also be used to improve the efficiency in the areas of short-term disaster response capacities, long-term risk reduction, development and environmental protection activities (Ajmar et al. 2008).

As an authoritative data source, SDIs have played a key role in disaster management for a considerable time. Nowadays, it is possible to gather large amounts of spatial data in the form of crowd generated spatial data content. This leads to the question, is it possible to create a link between SDI and CSD to make the disaster management actions more effective? When exploring the evolution of SDIs, it can be seen from its beginning that the majority of SDIs have been led by National Mapping Agencies (NMAs) of various countries (Williamson et al. 2005) and have considered users as passive recipients (Budhathoki & Nedovic-Budic 2008). However, compared with the current trends, it is the users that can potentially contribute towards the development of SDI. In addressing this issue Budhathoki and Nedovic-Budic (2008) suggest we should reconceptualise the role of the ‘user’ of SDIs to be a ‘producer’ and to include crowdsourced spatial data in the SDI-related processes.

2.6. SDI and geospatial CSD: the formal and informal combined

SDIs are generally considered as more formal, being highly institutionalized and having more traditional architectures. In line with the SDI framework, each dataset usually undergoes thorough standardisation procedures and SDI data is generally maintained by skilled and qualified people. Therefore, the cost of creation, management and maintenance is high. SDIs are mainly held by governments and are mostly standards centric as this is important for structuring and communicating data. The standardization is by means of structure (syntax) as well as meaning (semantic) (Hart & Dolbear

2013). Generally, the crowdsourced geospatial data comes from citizens and hence the information is often unstructured, not well documented and loosely coupled with metadata. However, crowd-generated geospatial data is more current and diverse in contrast to SDI data.

Jackson et al. (2010) studied the synergistic use of authoritative government and crowdsourced data for disaster response. They critically compared the clash of two paradigms of CSD and authoritative data as identified in Table 2.2. As can be noted in this table the two forms of data may seem as if they are diametrically opposed. However, the shortfalls of SDIs such as their lack of currency could possibly be more effectively addressed through CSD.

Table 2.2 A comparison of two paradigms: CSD and Authoritative data (Jackson et al. 2010)

CSD	**Authoritative Government Data**
'Simple' consumer driven Web services for data collection and processing.	'Complex' institutional survey and GIS applications.
Near 'real-time' data collection and continuing data input allowing trend analysis.	'Historic' and 'snap-shot' map data.
Free 'un-calibrated' data however, often at high resolution and up-to-the minute.	Quality assured 'expensive' data.
'Unstructured' and mass consumer driven metadata and mashups.	'Structured' and institutional metadata in defined but often rigid ontologies.
Unconstrained capture and distribution of spatial data from 'ubiquitous' mobile devices with high resolution cameras and positioning capabilities.	'Controlled' licensing, access policies and digital rights.
Non-systematic and incomplete coverage.	Systematic and comprehensive coverage.

2.7. CSD qualification to fit with Authoritative data

To use CSD confidently in critical applications, it is important to increase the quality of CSD up to an acceptable level that will match the quality requirements of authoritative data. It has also been identified that the comprehensiveness of coverages is important for effective decision making in disaster management. To this end, it is important to understand the CSD production process i.e. who is producing CSD and what motivates them to do so? Coleman et al. (2009) carried out an interesting study to find out 'the nature of motivation of VGI producers' and identified that the 'pride of place' is an important fact that motivates individuals contributing updates such as centrelines of roads and POIs in their localities in Google Earth, OpenStreetMap and Tele Atlas (now owned by TomTom) or NAVTEQ[16] (Navigation Technologies Corporation) (was acquired by Nokia in 2007/2008). According to Coleman et al. (2009) the majority of VGI contributors are doing it occasionally and mostly are interested amateurs. They also pointed out that there are some intruders contributing at large with negative motivations such as mischief, agenda, malice and/or criminal intent which could have serious consequences by damaging contributions e.g. partially deleting a map. In the case of misinformation, there may be two categories i.e. unintentional in which the provider believes that it is reliable new information however it is not, and deliberate or intentional misinformation based on a conscious agenda (Coleman et al. 2009).

The CSD producers may not be interested in contributing when there are strict specifications (Girres & Touya 2010) as people always value their freedom to act. As they contribute their time and effort in CSD production for free, they see no reason for them to follow any rules or specifications. Girres and Touya (2010) also suggest that the ‘success of VGI lies in the simplicity of contributions’ and needs the ‘ideal balance between specifications and contribution freedom’. Brando and Bucher (2010) proposed to let contributors act freely and check the contributions consistently with specifications. These specifications were expected to assist in three ways by including on

[16] An American Chicago based provider of geographic information system (GIS)

the fly consistency checking, improvement of the quality of CSD through external references and reconciliation of concurrent data editions.

It is obvious that the CSD may be a mix of good and bad geospatial information created by volunteers ranging from complete novices to experts. Filtering the high quality and reliable information out of this mix is very challenging. The geospatial data that comes in the form of CSD may not always be an alternative to the authoritative geospatial data. However, it may be more useful as a tool to fill the information gaps especially currency and incompleteness of the authoritative data. Techniques such as real-time or post spam/rumour detection, source/information credibility and relevance analysis, location quality analysis and quality improvement are some key steps identified for CSD qualification and fusing with authoritative data such as SDIs. The CSD quality assessment parameters, methods and improvement approaches will be discussed in detail in the next chapters.

2.8. Chapter summary

Disaster management, SDI and CSD are interlinked fields in today's connected world. Conventional disaster management mostly relies on authoritative and structured data such as SDIs. With the development of Web 2.0 concepts and inverted web architectures the crowd was empowered and more user centred applications were developed. A new form of free and widely available data has emerged and has created numerous opportunities and challenges. This new form of data was termed as crowdsourced data and it has initiated new research in the context of spatial data quality.

This chapter investigated the backgrounds of disaster management, SDI and CSD. Developments in the disaster management field and its information requirements were discussed with a special focus on flood disasters. The definitions, supported technologies, data sources, opportunities, application areas, issues and research gaps of CSD were explored. Disaster management and how it is supported by new crisis mapping

platforms and data were discussed. The background of SDI, its evolution towards disaster management, the role of SDIs in disaster management and the opportunities of combining SDIs and CSD were explored. Finally, the CSD and authoritative data such as SDI combination opportunities were discussed.

Although the CSD has emerged as a potential data source for many applications, existing disaster management agencies have not fully utilised or recognised the value of CSD. Therefore, further research is required to ensure the quality of CSD and its value in a disaster context.

The next chapter will discuss CSD quality aspects, retrieval methods and geospatial semantics.

Chapter 3: **Geospatial data Quality, Retrieval and Semantics**

3.1. Introduction

The previous chapter examined the concepts, developments and related literature about disaster management, CSD and SDIs. The challenges, opportunities and research gaps regarding the utilisation of CSD were also discussed. The purpose of this chapter is to understand the geospatial data quality, retrieval and semantics in relation to CSD. The spatial data quality aspects, quality elements and assessment approaches including the history are discussed and the CSD credibility, relevance and related research approaches are explored. The use of the naïve Bayes theorem-based spam email detection approach for CSD credibility detection is explored. The methods and relevance assessment used in Geographic Information Retrieval (GIR) for assessing CSD relevance for post-flood disaster management are explained. Moreover, GIR concepts, the use of Natural Language Processing (NLP) techniques, semantics, ontologies and gazetteers for GIR are examined. Finally, a development procedure for an ontological gazetteer and the General Architecture for Text Engineering (GATE[17]) software and its components are also discussed.

3.2. Geospatial data quality aspects and CSD quality assessment approaches

Spatial data quality has been a key issue for researchers since the advent of Geographic Information Systems (GIS) technology (Chrisman 1984; Oort 2006; Criscuolo et al. 2016) with many researchers and organisations identifying spatial data quality issues (Devillers et al. 2010). In 2002 ‘quality aspects of geographic information was enshrined in the ISO codes 19113 (quality principles) and 19114 (quality evaluation procedure) (Haklay 2010). Similarly, the quality of CSD is also a complex issue which

[17] https://gate.ac.uk

includes characteristics of data and its producers and the context of the application (Criscuolo et al. 2016).

From a consumer's perspective, high-quality spatial data should be 'intrinsically good' and 'contextually appropriate' to the task in hand (Wang & Strong 1996). Measuring the intrinsic quality is more relevant to conventional geospatial data and judgment of extrinsic quality is more useful to crowdsourced type data. Generally, the refresh rate of CSD (i.e. the update frequency) is believed to be higher than conventional authoritative data, particularly when they are associated with dynamic and cyclical events. Therefore, CSD might be considered as more up-to-date than conventional data repositories. As uncertainty is a key issue related to CSD, the quality assessment is vital before this data is utilised in applications such as disaster management (Koswatte et al. 2015).

CSD and VGI quality assessment research has suggested various metrics (Antoniou & Skopeliti 2015). Detailed reviews on the recent CSD and VGI quality assessment approaches can be found on Wiggins et al. (2011), Bordogna et al. (2014), Antoniou and Skopeliti (2015), Senaratne et al. (2016) and Criscuolo et al. (2016).

CSD and VGI quality control research has identified two quality assessment approaches, namely quality by measures (quality as accuracy) or quality by indicators (quality as credibility) (Flanagin & Metzger 2008; Antoniou & Skopeliti 2015). However, the quality as accuracy (quality by measures) may not be considered to be the best approach as VGI and CSD quality is mostly undocumented (Antoniou 2011). Hence, the general spatial data accuracy assessment parameters including completeness, logical consistency, positional accuracy, temporal accuracy, thematic accuracy, purpose, usage, lineage (Girres & Touya 2010; Haklay 2010; Goodchild & Li 2012), attribute accuracy (Girres & Touya 2010), semantic accuracy (Goodchild & Li 2012), definition, coverage, legitimacy and accessibility (Kim 2013) are still questionable in VGI/CSD quality assessment based on measures. A number of researchers (Flanagin & Metzger 2008; Grira et al. 2010; Ostermann & Spinsanti 2011; Craglia et al. 2012;

Kim 2013; Spinsanti & Ostermann 2013; Antoniou & Skopeliti 2015; Criscuolo et al. 2016; Hung et al. 2016) have suggested credibility and relevance are more pragmatic CSD quality indicators. However, the existing methods in this regime are still immature.

Many parameters related to geospatial CSD production such as the level of user expertise, their motivation, methods of data creation and expected maximum accuracy level are mostly uncertain (Flanagin & Metzger 2008). In solving the quality issues pertaining to CSD for disaster management applications, it is useful to understand the role of disaster responders and the specific requirements of CSD in respect to their spatial data needs. On one hand, relevant and credible data of the event concerned are crucial in their decision making. On the other hand, recent studies suggest the need to further analyse the quality of CSD by assessing credibility and relevance. Reinforcing the control over the data production chain such as using standardised data creation methods, using well trained volunteers (Lee 1994), crowd quality control (Bishr & Mantelas 2008), and filtering good quality information from abundant data (Krumm & Mummidi 2008) are some of the identified credibility resolving strategies. However, the definitions and semantic meanings of relevance and credibility are still not clear.

Recent research conducted by Hung et al. (2016) identified the possibility of using statistical methods to assess the credibility of VGI. They used the 2011 Australian flood VGI data set as the training data and the 2013 Brisbane floods data as the testing data set. Their approach (Figure 3.1) was to use binary logistic regression modelling to achieve an overall accuracy 90.5% for a training model and 80.4% accuracy for the testing data set. They highlighted the potential of using statistical approaches for efficiently analysing the CSD credibility and for rapid decision making in the disaster management sector even without real-time or near real-time information.

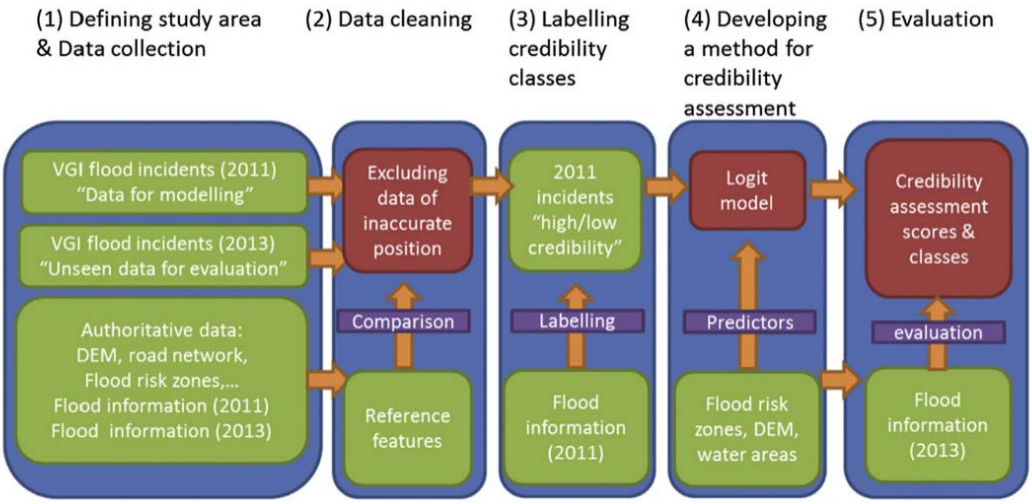


Figure 3.1 A flood related VGI credibility assessment method (Hung et al. 2016)

Longueville et al. (2010) proposed a generic workflow which used prior information about the phenomenon and reasoning techniques to improve the reliability of VGI. Ostermann and Spinsanti (2010) extended this work to develop a method to validate and assess the relevance of VGI in the context of forest fires. Figure 3.2 shows the workflow they used in this scenario. They used four elements including source reputation, spatial and temporal validity of the source's profile, tagging VGI with keywords and VGI cross-referencing.

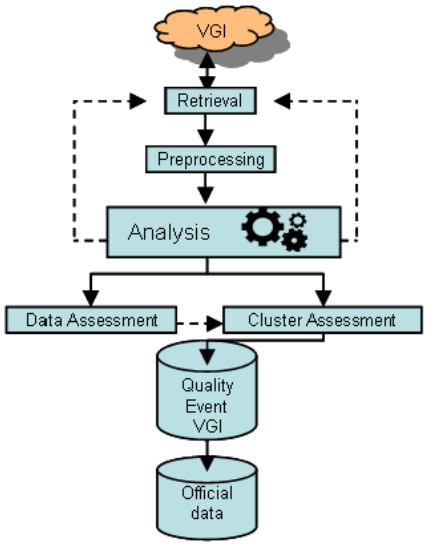


Figure 3.2 The VGI quality assessment workflow (Spinsanti & Ostermann 2010)

Osterman and Spinsanti (2011) again came up with a conceptual framework to automatically assess the quality of VGI in disaster events. Their approach simplistically analysed CSD based on keywords which could arise in the context of forest fires. They identified three main challenges pertaining to CSD: the sheer volume, unclear structure and poor-quality control. However, what was lacking in their model was to semantically handle the different meanings in the keywords. CSD credibility can be manipulated by considering people's perception of information source credibility (Flanagin & Metzger 2008).

Brando and Bucher (2010) proposed an approach for improving CSD quality based on formal specifications and external reference data. Their focus was to improve quality in both production and provision of metadata for the users. Girres and Touya (2010) assessed the quality of the French OpenStreetMap (OSM) dataset using the elements of geometric, attribute, semantic and temporal accuracy, logical consistency, completeness, lineage and usage. They suggested that the improvement of geospatial CSD requires finding the ideal balance between specifications and contribution freedom.

Bishr and Kuhn (2007) developed a method to use trust as a function of the temporal and spatial dimensions. According to the authors, three important challenges of CSD that are similar to VGI are 'how to filter correct data, how to provide metadata for VGI and how to explicate semantics of VGI'. Collaborators with a high reputation should be able to share data with 'more reliability and quality' (Bishr & Kuhn 2007). Bakri and Fairbairn (2011) conducted research to study the use of semantic similarity measures to analyse the semantic heterogeneity between CSD and formal data and concluded that semantic similarity alone will not solve the issue.

3.3. CSD credibility and relevance

3.3.1. CSD credibility

Hovland et al. (1953) defined credibility as 'the believability of a source or message' which comprises primarily of two dimensions, trustworthiness and expertise. However, as identified by Flanagin and Metzger (2008), the dimensions of trust and expertise can also be considered as being subjectively perceived, as the study of credibility is highly interdisciplinary and the definition of credibility varies according to the field of study. While the scientific community view credibility as an objective property of information quality, the communication and social psychology researchers treat credibility more as a perceptual variable (Fogg & Tseng 1999; Flanagin & Metzger 2008). According to Fogg and Tseng (1999) credibility is defined as 'a perceived quality made up of multiple dimensions such as trustworthiness and expertise' or simply as believability.

In recent times, there has been an increased interest in the use of CSD for both research and commercial applications. VGI production and use have also become simpler than ever before with technological developments in mobile communication, positioning technologies, smart phone applications and other infrastructure developments which support easy to use mobile applications. However, data quality issues such as credibility, relevance, reliability, data structures, incomplete location information, missing metadata and validity continue to limit its usage and potential benefits (Flanagin & Metzger 2008; De Longueville et al. 2010; Koswatte et al. 2016).

VGI quality can be described in terms of quality measures and quality indicators (Antoniou & Skopeliti 2015). The quality measures of spatial data have largely focused on quantitative measures such as completeness, logical consistency, positional accuracy, temporal accuracy and thematic accuracy whilst the quality indicators are often more difficult to measure and refer to areas such as purpose, usage, trustworthiness, content quality, credibility and relevance (Senaratne et al. 2016). However, in

CSD it may not always be appropriate to trust the information provided by the volunteers as their experience and expertise varies dramatically and assessing the credibility of the provider may be impractical. In particular, the volunteers in a disaster situation are often extremely heterogeneous and their input only occurs during a short period. Hence, it is difficult to profile these contributors, unlike many users of Twitter which may have a long history of activity. Therefore, a key challenge is to assess the credibility of the provided data in order to utilise it for future decision making.

A popular approach to assess credibility in spam email detection is to numerically estimate the 'degree on belief' (Robinson 2003) by analysing the email content using natural language processing and machine learning techniques. Natural language processing is a commonly used term to describe the use of computing techniques to analyse and understand natural language and speech. These approaches have been successfully applied to the detection of spam in Twitter messages (Wang 2010).

Credibility analysis approaches and the methods will vary depending on the context. Studies conducted by Bishr and Kuhn (2007), Noy et al. (2008), Janowicz et al. (2010), Sadeghi-Niaraki et al. (2010), and Shvaiko and Euzenat (2013) have identified the importance and usefulness of spatial semantics and ontologies in assessing the quality of CSD. Most approaches tackle CSD quality by qualifying contributors and contributions (Brando & Bucher 2010). The quality based on contributions has mostly been validated using rating systems (Elwood 2008a; Brando & Bucher 2010) and using a reference data set (Haklay 2010; Goodchild & Li 2012). The flood disaster related CSD is different in the sense of its timespan and contributors. They are collected in a very short period of time and the contributors will also vary with the event. Credibility analysis through source reputation analysis will be highly challenging in this context, so a more feasible option is the analysis of information credibility.

Given the variability of contributors of CSD during a disaster event, and the complexities in qualifying the expertise or experience of contributors, a content analysis approach may provide the greatest likelihood of success for this research.

3.3.1.1. Statistical approaches for CSD credibility detection in disaster management

Disaster related CSD is quite different in the sense of its lifetime and contributors. Data are often collected over a very short period of time with many different contributors during the event. Kim (2013) developed a framework to assess the credibility of a VGI dataset from the 2010 Haiti earthquake based on a Bayesian Network model. The outcomes of this earthquake damage assessment study were compared with the results from official sources. The author reported that 'the experiments have not only demonstrated microscopic effects on the individual data, but also showed the macroscopic variations of the overall damage patterns by the credibility model'. The model was identified as being more suitable for post disaster management purposes.

In filter based classification processes, it is important to simplify the message content using transformations including tokenizing (extracting words), stemming (removing derivational affixes) and lemmatizing (remove inflectional endings and returning the base or dictionary form of the word) (Figure 3.3) which may improve the classification accuracy and performance (Guzella & Caminhas 2009).

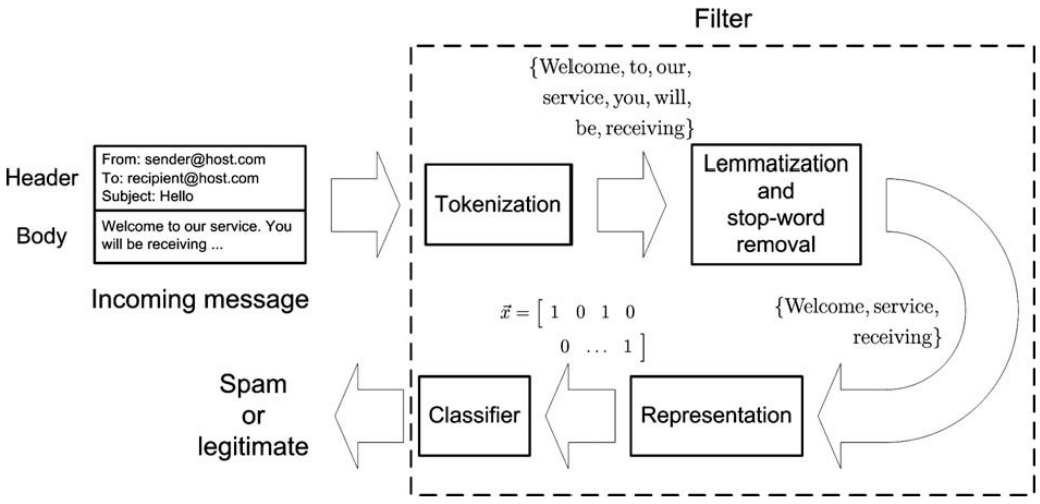


Figure 3.3 Main steps involved in filter based email classification (Guzella & Caminhas 2009)

Credibility can be calculated and rated into different levels which may be useful for disaster management staff. However, in critical events such as disaster management, a binary form of credibility representation would be simpler and less confusing for the general public (Ostermann & Spinsanti 2011).

Why use spam email detection as an approach for CSD credibility detection

Spam email is considered as 'unsolicited bulk email' in its shortest definition (Blanzieri & Bryl 2008). Spam emails cost industries billions of dollars annually through the misuse of computing resources and the additional time required by users to sort emails. Spam emails can often carry computer viruses and also violate users' privacy (Blanzieri & Bryl 2008). Compared to the spam emails, CSD has some similarities and differences. Firstly, CSD also has a mixture of content that varies in credibility and the CSD events often generate large volumes of data. Emails, including spam emails, often have a specified structure (sender, body text and header), however, CSD often lacks structure. Finally, the aim of the filtering data to identify legitimate or credible content is similar in both cases.

Spam email detection (Cranor & LaMacchia 1998; Pantel & Lin 1998; Robinson 2003; Metsis et al. 2006; Lopes et al. 2011), junk-email detection (Sahami et al. 1998) or anti-spam filtering (Androutsopoulos et al. 2000; Schneider 2003) research has a long history which grew from the commercialization of the internet in mid 1990s (Cranor & LaMacchia 1998). Researchers have explored various approaches with Content Based Filters (CBF) or Bayesian filters being the most popular anti-spam systems (Lopes et al. 2011). Wang (2010) tested a Bayesian classifier for spam detection in Twitter and confirmed that Bayesian classifiers performed highly in terms of weighted recall and precision, and outperformed the decision tree, neural network, support vector machines, and k-nearest neighbour's classifications.

Castillo et al. (2011) analysed the news worthiness of tweets using a supervised classifier whilst Kang et al. (2012) analysed the 'credible individual tweets or users' based on three models (social model, content model and hybrid model) using Bayesian and

other classifiers. These studies support the use of a modified Bayesian approach for assessing the credibility of crowdsourced data.

3.3.1.2. A naïve Bayesian Network based model for CSD credibility detection

The Bayesian Networks (BNs) were initially identified as powerful tools for knowledge representations and inference. With the advent of naïve Bayesian Networks, which are simple BNs that assume all attributes are independent, the classification power of BNs were expanded (Cheng & Greiner 1999).

A credibility detection function can be defined as,

$$f(m,\theta) = \begin{cases} t_{credible} \\ t_{unreliable} \end{cases} \quad \begin{array}{l} \text{if} f(m,\theta) > T \text{ message is credible} \\ \text{Otherwise message classified as unreliable} \end{array}$$

Where m is a message to be classified, θ is a vector of parameters, and $t_{credible}$ and $t_{unreliable}$ are tags to be assigned based on the threshold *T* to the messages.

The vector of parameters θ is the result of training the classifier on a pre-collected dataset:

$$\theta = \Theta(M)$$

$$M = \{(m_1, l_1), (m_2, l_2), \dots (m_n, l_n)\}, l_i \in \{t_{credible}, t_{unreliable}\}$$

Where $m_1, m_2 \dots m_n$ are previously collected messages, $l_1, l_2 \dots l_n$ are the corresponding labels, and Θ is the training function.

As Guzella and Caminhas (2009) defined; if a given message is represented by $\vec{x} = [x_1, x_2, \dots x_n]$ which belongs to class $c \in (s: spam, l: legitimate)$, the probability $\Pr(c|\vec{x})$ that a message is classified as c and represented by $\vec{x}$ can be written as,

$$Pr(c|\vec{x}) = \frac{Pr(\vec{x}|c)\cdot Pr(c)}{Pr(\vec{x})} = \frac{Pr(\vec{x}|c)\cdot Pr(c)}{Pr(\vec{x}|s)\cdot Pr(s)+Pr(\vec{x}|l)\cdot Pr(l)} \quad \ldots(1)$$

Where;

$\Pr(c)$is overall probability that any given message is classified as c

$\Pr(\vec{x})$is the a priori probability of a random message represented by $\vec{x}$

$\Pr(\vec{x}|s)$ and $\Pr(\vec{x}|l)$are the probabilities that a message is classified as spam or legitimate respectively

$\Pr(s)$ and $\Pr(l)$ are overall probabilities that any given message is classified as spam or legitimate respectively.

The naïve classifier assumes that all features in $\vec{x}$ are conditionally independent to every other feature and the probability $Pr(\vec{x}|c)$ can be defined considering *N* number of messages as,

$$Pr(\vec{x}|c) = \prod_{i=1}^{N} Pr(x_i|c)$$

So, the equation (1) becomes,

$$Pr(\vec{x}|c) = \frac{\prod_{i=1}^{N} Pr(x_i|\mathrm{c})\,.\,Pr(c)}{\prod_{i=1}^{N} Pr(x_i|s)\,.\,Pr(s) + \prod_{i=1}^{N} Pr(x_i|l)\,.\,Pr(l)}$$

with $Pr(x_i|c)$,$c \in [s,\ l]$ given by,

$$Pr(x\_i \,|\, c) = Pr(X\_i = x\_i \,|\, c) = f(Pr(t_i|c, \mathbb{D}_{tr})\,, x_i)$$

Where function f depends on the representation of the message. The probability $Pr(t_i|c, \mathbb{D}_{tr})$ is determined based on the occurrence of term t_i in the training dataset $\mathbb{D}_{tr}$.

3.3.2. CSD relevance

Relevance is naturally cognitive and 'the greater the cognitive effects the greater the relevance and the smaller the processing efforts to derive these effects, the greater the relevance' (White 2011). It is highly dependent on the end user's requirements regardless of being a product or information. The context of relevance has long been studied in diverse fields including but not limited to philosophy, communication, logic, psychology, artificial intelligence, natural language processing, documentation, information science and information retrieval (Saracevic 1996). Raper (2007) showed that the geographic relevance is an important factor in dealing with mobile information seeking. Saracevic (1996) identified five types of relevance based on the relevance literature, namely (1) topical or cognitive relevance, (2) algorithmic relevance, (3) pertinence or intellectual relevance, (4) situational relevance and (5) motivational or affective relevance. This research proposes to focus on the situational relevance which can be defined as 'usefulness of the viewed and assessed information' towards the task in hand and information needs of the user (Andrade & Silva 2006) and is more appropriate to assessing the CSD relevance for post-flood disaster management context.

Many approaches have been tested to measure the relevance of CSD with different data types and different purposes. Spinsanti and Ostermann (2013) designed a system which is termed as GeoCONAVI (Geographic CONtext Analysis for Volunteered Information) to test the validity of CSD related to forest fires gathered from social media networks. Caragea et al. (2011) developed a framework called Enhanced Messaging for the Emergency Response Sector (EMERSE) to automatically classify microblogging reposts related to the 2007 Haitian earthquake in order to support emergency responders. They classified the data gathered from the Haiti Ushahidi Crowdmap which was based on the Ushahidi crowd-mapping platform. Parker et al. (2011) suggested that VGI has potential value and usability benefits to the end-users over the Professional Geographic Information (PGI).

Cowan (2013) argued that the cognitive knowledge and user relevance feedback which are used in the Information Retrieval (IR) field to improve the search engine results can be used to identify the 'most relevant content in VGI and relevant data'. Geographic Information Retrieval (GIR) extends the IR's thematic based textual relevance assessment of documents by incorporating the geographic context. This also builds on Tobler's (1970) first law of geography indicated in the section 2.3.4.

GIR seeks to retrieve geographically relevant documents (Larson 1996; Jones et al. 2001; Wang et al. 2005; Andrade & Silva 2006; Jones & Purves 2008; De Sabbata & Reichenbacher 2010; Janowicz et al. 2011; Kumar 2011) or identify unambiguous geographic associations (Amitay et al. 2004) based on the user's requirements. Simple word or toponym matching is not adequate for geographic information retrieval purposes (Jones et al. 2001). Therefore, toponym matching based on semantic similarity measures may be the most appropriate approach.

GIR and CSD relevance

Geographic relevance is applied in many of today's human information seeking activities e.g. in search engines. It can be defined as 'a relation between a geographic information need and the spatio-temporal expression of the geographic information objects needed to satisfy it' (Raper 2007). The fields of IR and modern web based GIS systems have now matured to provide professional outputs for their own information requests. These developments suggest that the combined use of GIS and IR systems to handle the requests on geo-textual information are more effective (Cai 2002).

GIR is considered as a special case of IR which uses geographic indexing and geographic retrieval (Zaila & Montesi 2015). The key objective of GIR is to identify the place names or toponyms within a corpus (a large structured set of text, e.g. web sites, documents or social media posts) and their corresponding geographic location that is 'concept@location' (Andrade & Silva 2006). On the one hand, it is a process that manages the imprecision and ambiguity as geographic names are often ambiguous (Zaila & Montesi 2015). On the other hand, it is a process of ranking the relevance in two

dimensions namely thematic and geographic (Andrade & Silva 2006) with the assumption that they are independent from each other (Kumar 2011).

Although, the GIR field is relatively new (Kumar 2011), numerous mechanisms have been proposed such as weighted geo-textual similarity measures (Andrade & Silva 2006), extended vector space model (Cai 2002), probabilistic models (De Sabbata & Reichenbacher 2010), dynamic assessment of the specificity of the users' search context (Yu & Cai 2007), and semantic and ontology based models (Martins et al. 2006). Similarly, techniques can be applied along with NLP techniques to detect relevance of data with very low signal-to-noise ratios such as social media data (Stowe et al. 2016), and even in a near-real-time context (Monteiro et al. 2016). De Sabbata and Reichenbacher (2010) suggest that GIR concepts can be utilized to estimate the relevance of geographic objects which are based on user context by converting geographic distances into similarity scores.

The CSD relevance for post-flood disaster management

During a crisis three temporal stages are evident namely pre-crisis, crisis and post crisis (Lettieri et al. 2009). Disaster management can be divided into four main processes namely disaster mitigation, preparedness, disaster response and recovery (Lettieri et al. 2009; Poser & Dransch 2010). The goal of the post-disaster recovery stage is to bring the living conditions of the victims back to normal. CSD relevance may be utilised in each of these stages; however, a key focus of this research is to analyse CSD relevance for flood disaster management in the post-disaster management response context.

3.3.2.1. Adapting GIR process for CSD relevance analysis

Monterio et al. (2016) highlighted four techniques associated with the various stages of GIR based search engine pipelines, namely, (1) geographic indexing, (2) query expansion, (3) recognition and use of place names and (4) geographic ranking. A number of key challenges lie in the area of analysing and processing sets of documents and queries, textual-geographical indexing and ranking the documents using the relevance

criteria (Kumar 2011). Linguistics and cognitive science research has identified spatial representation and conceptual representation as key cognitive facilities while the real challenge in GIR is to find the right balance between conceptual (thematic) and spatial (geographic) approaches (Cai 2002).

Managing the thematic scope

The presence of relevant terms in a document provides an indication of relevance of the document for a selected task. From an information analysis perspective, the terms can be weighted based on the importance of the task in hand. A commonly used weighting method is Term Frequency-Inverse Document Frequency (TF-IDF) model. In this model, higher weights are assigned for specific terms appearing more frequently in a document. This is based on the premise that, the more frequently a given term appears, the more likely that document is relevant to the search. Conversely, a low weight will be assigned to more commonly available terms in the whole document set.

TF-IDF is a very popular weighting function used in information retrieval algorithms where the importance of a term or word to a document is statistically estimated.

$$TF(t) = \frac{\text{Number of times the term } t \text{ occurs in a message}}{\text{Total number of terms in the message}}$$

$$IDF(t) = \log_e \left[\frac{\text{Total number of messages}}{\text{Total number of messages the terms } t \text{ exits}}\right]$$

Therefore, the *TF-IDF* weight for term *t* in message *m* can be denoted as:

$$TF - IDF_{t,m} = TF_{t,m} * IDF_{t,m} \quad \ldots(2)$$

And the thematic similarity score Sim_T for a message *m* for the term *t* in query *q* can be calculated by:

$$Sim_{T\,(q,t)} = \sum_{t \epsilon q} TF - IDF_{t,m} \quad \ldots(3)$$

Vector Space Model (VSM)

When the TF-IDF values of document terms are calculated, it can represent the document in a vector space model which is an algebraic model for representing text documents. In here, each document is represented by vectors of identifiers i.e. index terms weighted based on their importance using a model such as the TF-IDF model. The axes of the vector space are denoted by the terms of the document. If two vectors are identical the angle between the vectors will be zero and produce maximum similarity (Salton et al. 1975). .. The query terms can also be represented in a VSM. Any form of vector mathematics can be applied to this kind of system to identify relationships (e.g. document similarity) including the analysis of CSD thematic relevance using the open source Lucene[18] IR system. The Lucene software is a high-performance, fully featured text search engine library written entirely in Java which performs term weighting based on a TF-IDF model. Recently, the popular Okapi BM25 probabilistic model was introduced to the Lucene system. The Okapi BM25 is considered as a probabilistic implementation of TF-IDF model.

VSM cosine similarity

In the VSM model, if the angle between two vectors is zero (i.e. cosine similarity) it is considered that the two messages are identical. The similarity between two messages m_1 and m_2 or cosine of the angles between two vectors can be calculated by:

$$Sim(m_1, m_2) = \frac{\vec{V}(m_1) \cdot \vec{V}(m_2)}{|\vec{V}(m_1)||\vec{V}(m_2)|} \quad \ldots(4)$$

Where: $\vec{V}(m_1)$ and $\vec{V}(m_1)$ are the vector representations of messages m_1 and m_2 and the same can be used to calculate the similarity between a query q and a message m.

The documents are normalised for the variable lengths which is an advantage of using cosine similarity function.

[18] http://lucene.apache.org

Managing the geographic scope

Managing the geographic scope or discovering and disambiguating toponyms that exist in the text document has been identified as Geographic Scope Resolution (GSR) (Alexopoulos et al. 2013; Monteiro et al. 2016). Generally, GSR consists of three tasks namely (1) geo-parsing (identifying toponyms), (2) reference resolution (toponym resolution) and (3) grounding reference (mapping toponyms to a footprint) (Monteiro et al. 2016). Common geo-parsing methods are gazetteer lookup based (searching and tested the location terms against a Gazetteer), rule based (identifying location terms based on pre-defined rules) and machine learning based methods (trained to detect location terms based on correlation measures with gold data i.e. training corpus) (Leidner & Lieberman 2011). The reference resolution which is mapping toponyms is mandatory when ambiguities are available (Monteiro et al. 2016). This work is generally supported by external resources such as gazetteers or spatial databases. In grounding reference (geocoding/ geo-referencing or geotagging are common synonyms) a set of coordinates (latitude, longitude or grid coordinates) is assigned to the identified toponyms. This is mostly supported by reference datasets such as gazetteers and geocoding algorithms.

This research suggests that the natural language processing based gazetteer lookup approaches are viable to semantically extract location information from CSD. The geographic information retrieval can be performed by the natural language processing software such as GATE. This type of work may be supported by an ontological gazetteer for both toponym identification and ambiguity resolution. The grounding references (geo-tagging) can also be assisted by the ontological gazetteer.

Usually after the GSR process, there is a need to calculate the geographic focus of a message. In the geographic focus detection stage, an ordered list is prepared based on the importance or relevance to the user query (Lieberman et al. 2007). Different approaches are available for geographic focus detection such as measuring the geographic similarity and relevance ranking. The geographic similarity measures can be

calculated based on region overlaps (Frontiera et al. 2008) or calculating a non-linear normalised distance between the scopes of the document and query (Andrade & Silva 2006; Lieberman et al. 2007; Zaila & Montesi 2015). Andrade and Silva (2006) explored a model which combined the ontological geographic relevance calculations whilst Zaila and Montesi (2015) proposed a model based on topological relations, metric proximity calculations and ontological geographic similarity calculations.

The similarity Sim_G between the geographic scope of the query (S_q) and geographic scope of the message (S_m) based on the ontology information can be represented by:

$$Sim_G(S_q, S_m) = K \times \{Insd(S_q, S_m) + Proxm(S_q, S_m)\} + (1 - K) \times Sib(S_q, S_m) \quad \ldots (5)$$

Where: $0 \leq K \leq 1$ so that the final value lies between 0 and 1.

Inside (*Insd*): If S_m is inside S_q the weight based on the number of decedents in the ontology,

$$Insd(S_q, S_m) = \frac{NumberOfDecendents(S_m)+1}{NumberOfDecendents(S_q)+1}$$ and 0 otherwise.

Proximity (*Proxm*): Based on the inverse distance where the distance normalized by the diagonal of Minimum Bounding Rectangle (MBR) of the query scope.

$$Proxm(S_q, S_m) = \frac{1}{1+\frac{Dist(S_q,S_m)}{Diagonal(S_q)}} \quad \ldots (6)$$

Siblings (*Sib*): Tests whether S_m and S_q are siblings,

$Sib(S_q, S_m)$=1 if S_m and S_q are siblings in the ontology, 0 otherwise.

3.3.2.2. Indexing, Relevance Ranking and Merging thematic and geographic scopes

Various algorithms and methods have been proposed within GIR research for indexing, relevance ranking and merging thematic and geographic scopes. It is often advantageous to consider the specificity of query scope in assessing the CSD thematic relevance as suggested by Yu and Cai (2007). They also reported that Dempster-Shafer's method of evidence combination shows superior results in their experimental study which is also very close to human judgments in many cases. However, the weighted sum method for relevance fusion is more common in GIR research (Martins et al. 2005; Andrade & Silva 2006; Yu & Cai 2007; Zaila & Montesi 2015).

As Yu and Cai (2007) defined,

The thematic specificity Spc_T of query $q = \{t_1, t_2, \dots, t_n\}$ is;

$$Spc_T = -\sum_{t \in q} \omega_t * CTM(t) \log\left(\frac{N_t+1}{N}\right) \quad \dots(7)$$

Where: $t_k$be the k^{th} term of the query q,

ω_t is the weight for each term,

$CTM(t)$ is the Conceptual Term Matrix of term *t* from the WordNet[19] ontology,

N_t is the number of messages containing term *t* and N is the total number of messages in the dataset.

The conceptual term matrix $CTM(t)$ is calculated by (1) extracting conceptual information representatives of term *t* (i.e. number of senses, number of synonyms, level number and number of children) from the WordNet ontology in the form of integer

[19]https://wordnet.princeton.edu/

values in CTM, (2) weighting to transform the values into weights based on the importance of different information types and (3) combining weighted values to calculate the final single score in $CTM(t)$.

The geographic specificity Spc_G of geo-referenced query q can be calculated by:

$$Spc_G = -\log\left(\frac{Area(G_q)}{Area(G_M)}\right) \quad \ldots(8)$$

Where: $G_q$be the geometry representative of the associated geographic scope of query q,

$Area(G_q)$ is the area of the geographic scope of q,

$Area(G_M)$ is the area of the coverage of all messages in the dataset.

The final rank as a weighted sum of individual scores can be represented by:

$$Rel(q,m) = \omega_T * Sim_T(q,m) + \omega_G * Sim_G(q,m) \quad \ldots (9)$$

Where q is a query, m is a message, Sim_T and Sim_G are thematic and geographic relevance functions, and ω_T and ω_G are weights of the two relevance scores.

The normalized weights of relevance scores ω_T and ω_G are calculated by:

$$\omega_T = \frac{1}{\ln(e+Spc_T)} \ , \ \omega_G = \frac{1}{\ln(e+Spc_G)} \quad \ldots (10)$$

Where: Spc_T and Spc_G are the thematic and geographic specificities as defined above.

3.4. Geospatial semantics and ontologies

The term ‘ontology’ comes directly from philosophy and it goes back to Aristotle. Although, its definition may vary according to the phenomena, in general, ontologies

are explicit specifications of shared conceptualizations and are key to establishing shared formal vocabularies (Gruber 1993; Studer et al. 1998; Du et al. 2013). Furthermore, ontologies provide a vocabulary of terms and relations of a domain of concerns. Domain ontologies express the knowledge valid to a particular domain. Ontologies provide a means of ensuring semantic interoperability in dynamic environments. They can be developed in the scope of global (top level) using top-down approaches and local (domain specific) from the bottom up approaches. Although there is no one accepted way or methodology for developing ontologies, they can be developed by starting without clearly known requirements i.e. quite vague objectives (Brusa et al. 2006).

Ontologies are fundamentally important when dealing with heterogeneous systems and considered as a main pillar in the so called semantic web. When considering the geospatial system management, it might be specifically conceptualized considering special geographic properties such as inherited location and spatial integrity. Geo-spatial ontologies include geo-spatial entities, 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).

3.4.1. Ontology development for semantic gazetteers

The fundamental rules of ontology development are (1) there is no correct way to model ontology and there will be other alternatives, (2) ontology development is necessarily an iterative process, (3) ontology concepts should be close to the objects (physical or logical) and relationships (nouns or verbs that describe the domain) in the domain of interest (Noy & McGuinness 2001). Figure 3.4 shows the ontology development workflow proposed by Scheuer et al.(2013).

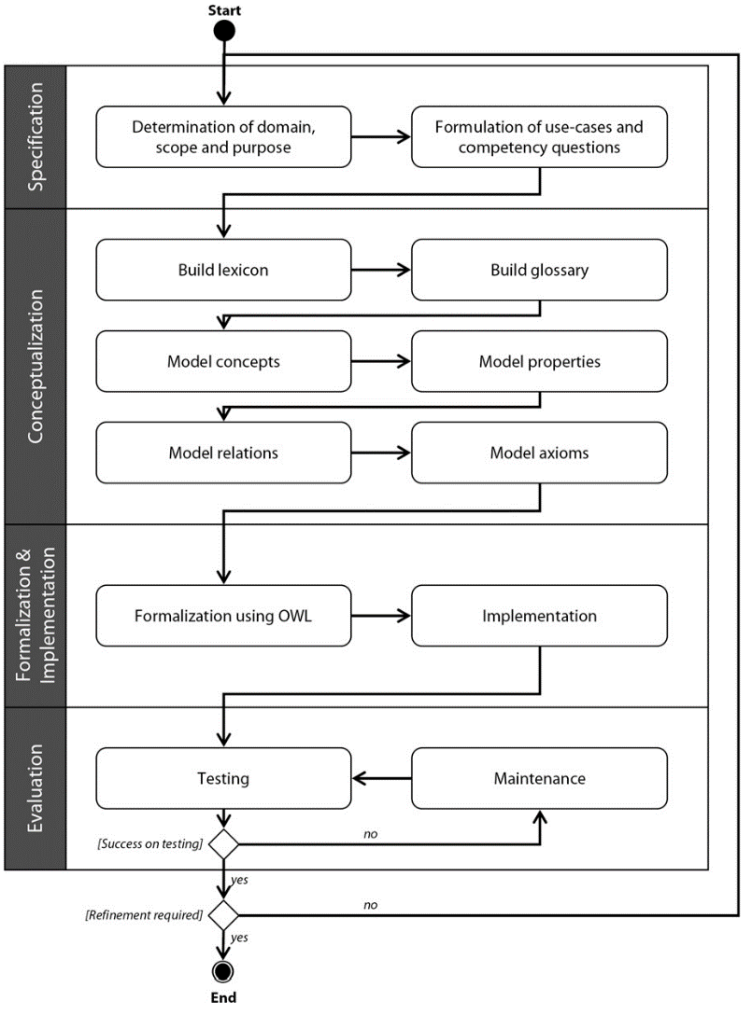


Figure 3.4 Ontology development workflow (Scheuer et al. 2013)

3.5. Geospatial Information Retrieval (GIR)

GIR is important and widely used in many application areas including disaster management, transport planning, hydrology and land-use (Battle & Kolas 2011). To this end, the most popular approach is to use gazetteers for retrieving geographic information from the web pages or online contents. Researchers argue that this is not very different from a keyword base search such as in search engines (Buscaldi & 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 thesis.

3.5.1. GIR and ontological gazetteers

The objective of GIR is to geotag web pages based on their content and involves resolving two types of ambiguities i.e. geo-geo and geo-non-geo (Amitay et al. 2004). The geography or the location information in GIR from web content has 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 have been considered in this research i.e. (a) reporter location (b) incident location and (c) content location.

Gazetteers are geospatial dictionaries containing place names and related information describing spatial references and feature types (Janowicz & Keßler 2008; Machado et al. 2011). Many countries have developed and maintain their own gazetteers. Digital online formats such as Alexandria Digital Library Gazetteer[20] (ADL), Getty Thesaurus of Geographic Names[21] (TGN) and GeoNames[22] are available (Machado et al. 2011). Furthermore, integrated semantic geospatial information retrieval systems are also slowly become available. A good example is GeoWordNet[23] (georeferenced version of WordNet) which is an integrated system of GeoNames with WordNet plus the Italian section of MultiWordNet[24] (Buscaldi & Rosso 2009; Giunchiglia et al. 2010). Gazetteers are widely used in GIR research (Hill 2000; Amitay et al. 2004; Souza et al.

[20]http://legacy.alexandria.ucsb.edu
[21]http://www.getty.edu/research/tools/vocabularies/tgn
[22] http://www.geonames.org
[23]https://datahub.io/dataset/geowordnet
[24]http://multiwordnet.fbk.eu/

2005; Borges et al. 2011). However, it is mostly argued that they are not fully supported as there are structural limitations including lack of intra-urban place names and no records available 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 including automatic indexing and geo-referencing (Larson 1996), ontology-driven approaches (Jones et al. 2001; Fu et al. 2005b), semantic query expansion (Fu et al. 2005a; Delboni et al. 2007) and natural language processing (Delboni et al. 2007) along with gazetteers and geocoding techniques are proposed (Borges et al. 2011).

3.5.2. GATE semantic GIR/NLP tools

The use of NLP tools in information retrieval and information extraction work is popular. The GATE software (Figure 3.5) is a robust and scalable open-source java based tool (Cunningham et al. 2002) developed by the University of Sheffield, United Kingdom for text processing including semantic processing. This tool's main development focus is for Named Entity Recognition (NER) where the texts are grouped into predefined categories such as persons, organisations and locations. The GATE's system components are termed as "resources". The main three elements of the GATE system are Language Resources (LRs), Processing Resources (PRs) and Visual Resources (VRs). LRs include the entities lexicons, corpora or ontologies. Generally, in linguistics, a corpus (corpora in plural) is defined as a large and structured set of text or documents. PRs are parsers, generators or modellers and VRs represent visualisation and editing components (Cunningham et al. 2002).

GATE's A Nearly New Information Extraction (ANNIE) module consists of the following set of processing` resources: tokenizer, sentence splitter, POS tagger, gazetteer, finite state transduction grammar and ortho-matcher. The resources communicate via GATE's annotation API.

The *tokeniser* splits text into simple tokens, such as numbers, punctuation, symbols, and words of different types (e.g. with an initial capital, all upper case, etc.).

The *sentence splitter* segments the text into sentences. This is required for the tagger to process and needs to run prior to the *tagger*. Both the splitter and tagger are generally domain and application-independent.

The *POS tagger* adds part-of-speech tags as a feature to each Token annotation. The splitter and tagger are not mandatory parts of the system.

The *gazetteer* consists of lists such as cities, organisations, days of the week, etc. It contains some entities, but also names of useful key words, such as company designators (e.g. "Ltd."), titles (e.g. "Dr"), etc.

The *semantic tagger* (or JAPE transducer) consists of hand-crafted rules written in the JAPE (Java Annotation Pattern Engine) language, which describe patterns to be matched and annotations to be created.

The ortho-matcher performs co-referencing, or entity tracking, by recognising relations between entities. It also has a secondary role of improving NER by assigning annotations to previously unclassified names, based on relations with existing entities (Maynard et al. 2008).

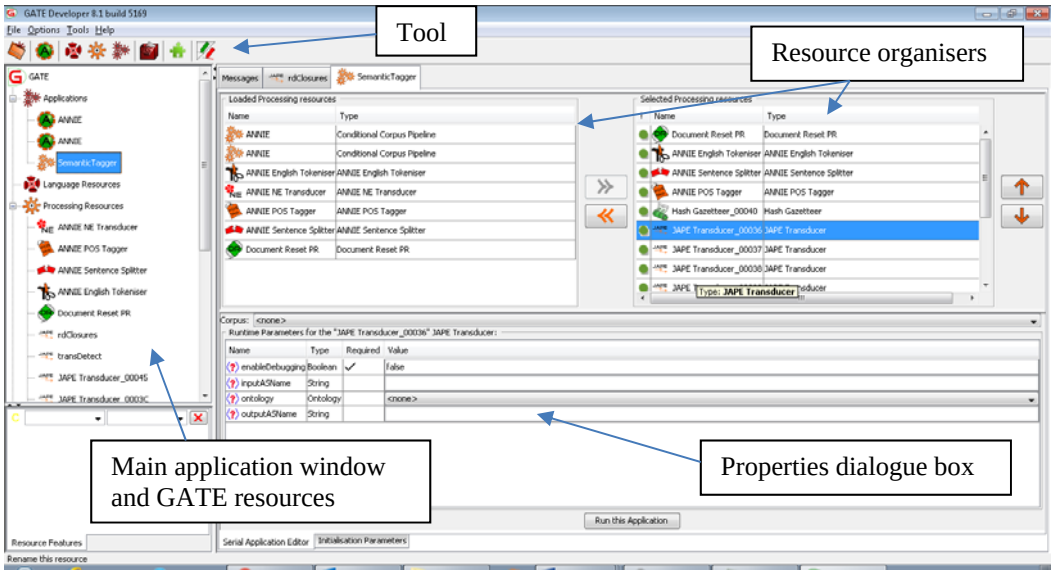


Figure 3.5 Interface of the GATE GIR/NLP system

3.6. Chapter summary

The quality of geospatial data is critical to deciding the quality of the outcomes of any project or application concerned. Spatial data quality assessment techniques and parameters have long been considered among mapping/GIS professionals, academics and researchers. However, general spatial data quality matrices are not applicable in measuring the quality of CSD due to structural, procedural and technological discrepancies as well as missing metadata. As a solution, various quality indicators have been proposed such as relevance and credibility. Credibility issues arise as CSD come from heterogeneous sources and are captured both by professionals and amateurs. The relevance or fitness for the purpose is highly cognitive and depends on the task at hand.

This chapter presented CSD credibility and relevance assessment approaches. A credibility assessment method using a naïve Bayesian Network based model which is commonly used in spam email detection systems was explored. This chapter also investigated a relevance assessment approach by adapting relevance assessment techniques available in the GIR domain. The thematic and geographic relevance assessment methods using the TF-IDF VSM, NLP based semantic gazetteers lookup were discussed along with the use of thematic and geographic specificity of the queries for relevance ranking. This chapter also explained the GIR concepts, use of NLP techniques, semantics, ontologies and gazetteers for GIR. The utilisation of an ontological gazetteer, GATE software and its components were discussed.

Chapter two identified the limitations available in current disaster management processes and the information currently being utilised. It also identified the importance of considering alternative data sources such as CSD to address these limitations. However, the quality aspects of CSD including location, credibility and relevance are often problematic. The compatibility of existing quality assessment techniques and assessment parameters for CSD are still not well developed. Therefore, more work is required to identify appropriate CSD quality assessment parameters, techniques and ap-

proaches through further research. This chapter identified these research gaps and investigated the possibility of incorporating various approaches and techniques that are utilised in other fields, including Information Technology.

The next chapter explains in detail the research approach used in this dissertation.

Chapter 4: **Research Approach**

4.1. Introduction

The previous chapters of the thesis investigated the theoretical background and the research potential of the selected field of study. Chapter one set the background for the research and chapters two and three were dedicated to identifying the research trends in the fields of CSD, disaster management and the advantages and fusion possibilities of CSD with authoritative data such as from SDIs. This chapter explains the research approach including an overview of research design and methods used to achieve the objectives of the research. It identifies the research gaps based on the understandings of CSD, VGI and SDI data. It explains and justifies the research methods used in this study and describes the conceptual model which includes CSD in crowd-supported disaster management and CSD quality control. The study area and the data collection procedures are firstly discussed, followed by an overview of the 2011 Australia flood tweets and Ushahidi data. Finally, it briefly describes the CSD processing and analysis methods conducted in this research.

4.2. Understanding CSD, VGI and SDI data and identifying research gaps

Within the scope of this study, it was identified that a deeper understanding of the similarities and differences among CSD, SDI and VGI will support the improved data modelling and linkages. SDIs are generally considered as more formal structures being highly institutionalized and having more traditional architectures. In line with the SDI framework, each dataset usually undergoes thorough standardization procedures. SDIs are mainly developed by governments and are highly standards centric, which is important for structuring and communicating data. CSD and VGI often come from laypersons and are unstructured, poorly documented and loosely coupled with metadata. VGI can be considered as a subset of CSD and both are generally more current and diverse in contrast to SDI. In general, VGI can be considered as semi-

structured data as many sources of VGI are semi-structured (Lemmens et al. 2016). Figure 4.1 depicts a comparison of VGI, SDI and next generation SDIs (i.e. the form of SDIs which are more semantic and accept new spatial data formats such as VGI) in terms of data quality, standards, currency and breadth of data.

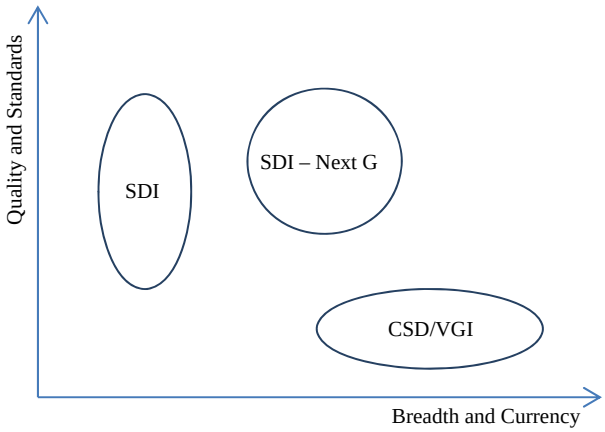


Figure 4.1 SDI, CSD, VGI and next generation SDIs (SDI-Next G)

As identified throughout this research there are number of issues as well as important opportunities pertaining to the CSD. Credibility and relevance are key issues and the currency and information abundance are the main advantages identified. Anyone can easily create CSD with a mobile phone by simply sending an SMS. This simplicity in creation leads to the issues and opportunities mentioned above. The unstructured form leads to accuracy and validity issues while the simplicity of creation encourages people to send more and more information in the form of CSD. This research examines the breadth of information required to resolve accuracy issues. Selection, filtering and integration are utilized to enhance the accuracy of the location component of CSD.

Current disaster management largely relies on authoritative data sources which are often considered as very reliable. However, the limited availability and access issues of authoritative data slows the disaster management process. Conversely, CSD is timely and more available, and has increased potential to be used as an alternative data source in disaster management. Therefore, current disaster management approaches

and strategies need to be reviewed to consider alternative data sources such as CSD which are highly available and up-to-date. However, the quality of CSD in the sense of credibility and reliability is still unclear. Therefore, further research is required to understand the quality of CSD to be utilised efficiently in current and future disaster management.

4.3. Conceptual modelling

4.3.1. Research methods

Research refers in general to search for knowledge and can be defined as a 'scientific and systematic search for pertinent information on a specific topic' (Kothari 2004). Research methods refer to the 'behaviour and instruments used in selecting and constructing a research technique' and may vary with the type of the research concerned i.e. descriptive/analytical, applied/fundamental, qualitative/quantitative or other types which are variations of these types (Kothari 2004). There are two key paradigms distinctively applied in information system implementations and its research, namely, Natural Science (NS) (or behavioural science) and Design Science (DS) research paradigms (Hevner et al. 2004). The Natural Science (NS) research attempts to understand the reality whereas Design Science (DS) research tries to create things that serve for human purposes (March & Smith 1995). Design science in information systems is identified also as a problem-solving process. March and Smith (1995) argue that both of NS (build, evaluate, theorise and justify) and DS (representational constructs, models, methods and instantiations) activities are required to ensure the information technology research is relevant and effective. Moreover, some people argue that those paradigms are practically inseparable and 'are two sides of the same coin and that scientific research should be evaluated in light of its practical implications' (Hevner et al. 2004).

Natural Science (NS) research method

Natural Science (NS) research methods intend to ‘develop and justify theories (i.e., principles and laws) that explain or predict organizational and human phenomena surrounding the analysis, design, implementation, management, and use of information systems’ (Hevner et al. 2004). March and Smith (1995) reported that NS research tries to conceptualise and characterise phenomena using 'higher order constructions, laws, models, and theories that make claims about the nature of reality'. They pointed out that it consists of two activities namely 'discovery' which is the process of generating or proposing scientific claims (e.g., theories, laws)' and justification which includes 'activities by which such claims are tested for validity'.

Design Science (DS) research method

'Design Science (DS), as the other side of the information science research cycle, creates and evaluates Information Technology (IT) artefacts intended to solve identified organizational problems' (Hevner et al. 2004). Hevner and Chatterjee (2010) introduced the following seven guidelines for conducting DS research i.e. (1) design as an artefact (creation of an innovative, purposeful artefact) (2) problem relevance (design for a specified problem domain) (3) design evaluation (through evaluation of artefact) (4) research contributions (solving unknown problem or know problem more effectively) (5) research rigor (rigorously defined, formally represented, coherent, and internally consistent) (6) design as a search process (search process to find an effective solution) (7) communication of research (effective communication).

This research falls under the applied research type and the approach taken consists of four stages based on the DS research methods broadly addressing Peffers et al.'s (2007) six steps mentioned in the section 1.4.

The first Stage, the research formulation, includes the aims and objective formulation, literature review, research question determination and especially the identification of the research approach and research methods.

Stage two consists of research design and development and includes further literature review to identify CSD quality assessment elements. CSD quality assessment element identification is an important part of this research. Spatial data quality assessment research has a long history; however, CSD quality assessment is still considered immature. There are no set parameters for CSD quality assessment and hence the need for careful analysis and to identify the most appropriate elements for CSD quality assessment. CSD is unique by its nature of production and the diversity of the producers. Therefore, the quality assessment and parameter design will be very challenging.

Stage three of the research involves the design and development with an emphasis on data collection and ontology development. CSD quality assessment and improvement is also considered in this stage.

Finally, Stage four entails the system integration and designing of an automated system for CSD quality assessment and designing a framework for fusing qualified and improved CSD with SDIs.

4.3.2. Research approach and conceptual framework

Chapters one to three of this thesis identified the key research question and the background of the problem. In the research formulation stage, the research problem, aim and objectives were defined. Figure 4.2 depicts the modified research approach based on DS research approach and Figure 4.3 illustrates the relationship between the research problem, research questions, aim and objectives, research approach and the outcomes. The research approach, outcomes and the quality assessment will be modified and refined according to the expected level of quality of the outcomes.

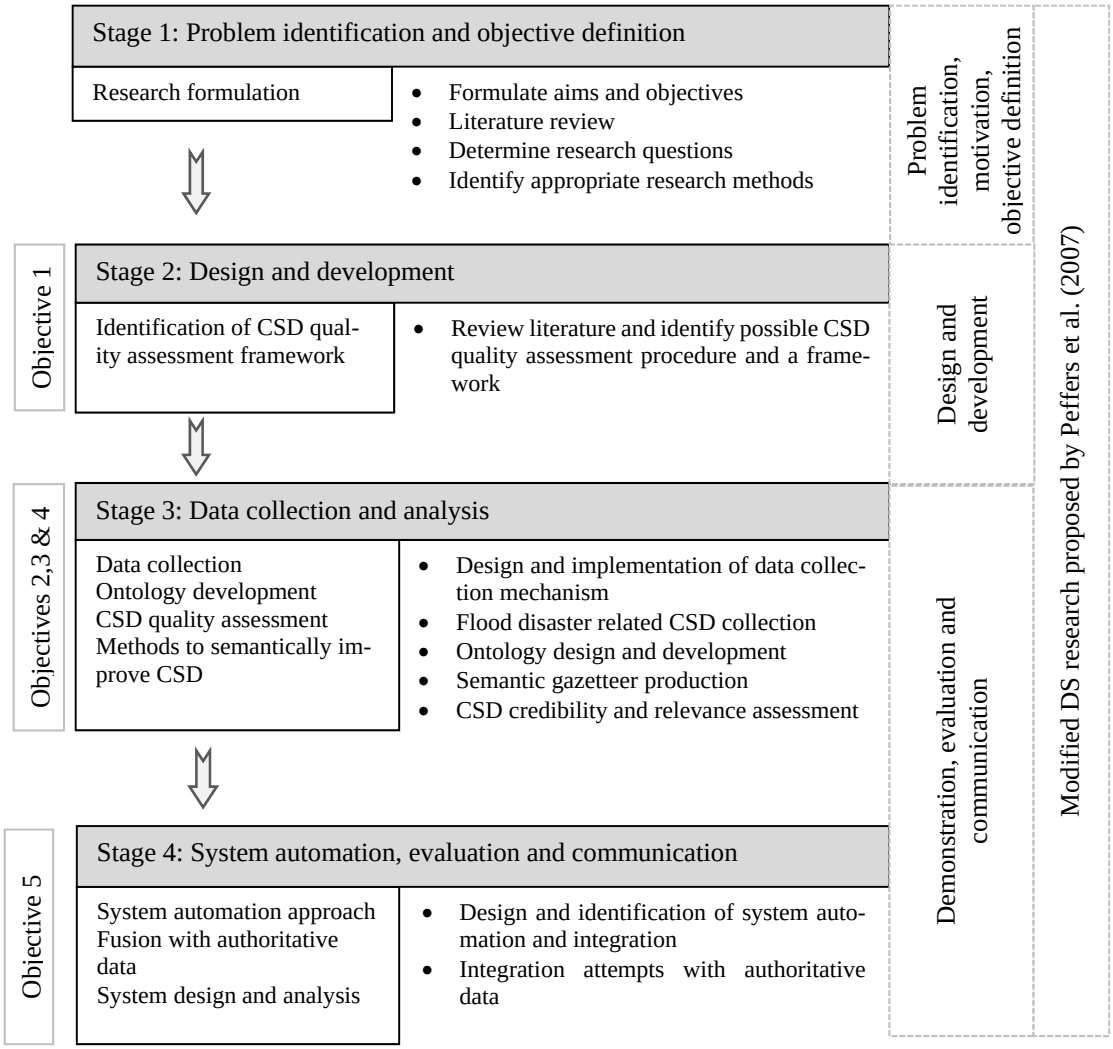


Figure 4.2 Modified research approach based on DS research proposed by (Peffers et al. 2007)

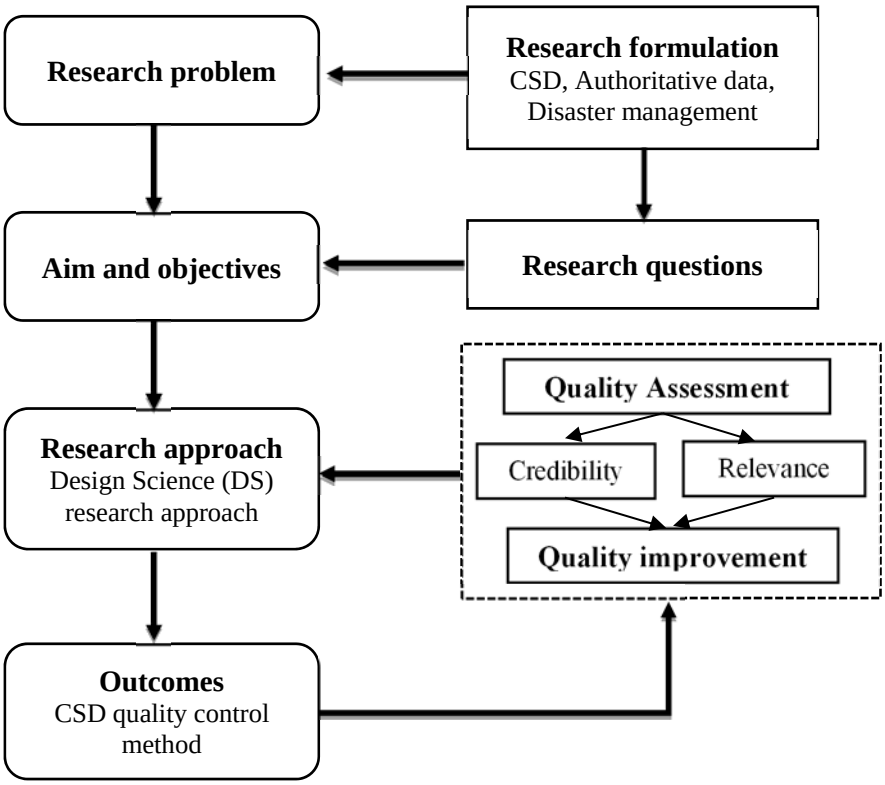


Figure 4.3 Conceptual research workflow

4.3.3. CSD in crowd-supported disaster management

In general, CSD comes from diverse sources and the quality is undocumented and often indeterminable as indicated in previous sections. Although, CSD are often un-structured, there can be some form of structure when the data comes from specifically designed applications such as Crowd-mapping platforms. However, these CSD are often not complete or as consistent as authoritative data. Quality control and improving the CSD towards authoritative data is very challenging and CSD quality should be carefully assessed prior to its use in critical applications such as disaster management.

In a broader context, the key objective of this research is the quality control of CSD. The review of literature has identified that CSD quality improvement is challenging and the available approaches are still immature. Figure 4.4 shows a very high-level

view of the CSD quality control approach used in this study. The CSD quality improvement will include CSD collection, improvement and re-use mechanisms which may also include many other sub-processes. In an ideal situation, the unstructured CSD will end up as highly structured authoritative data. An important advantage of this approach is that existing or system generated authoritative data can be used/re-used within the system to improve new CSD.

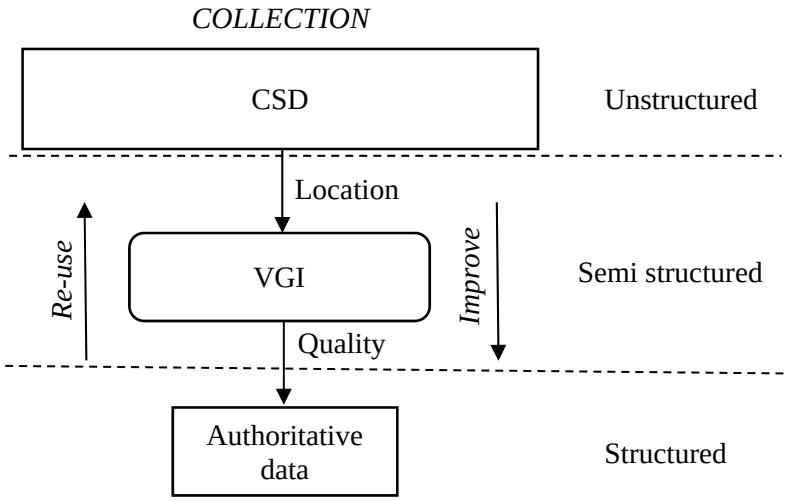


Figure 4.4 CSD quality control conceptual overview

Figure 4.5 depicts the crowdsourced spatial data life-cycle. Data collection mechanisms can be automated due to the rapidly increasing amounts of data. Social media flatforms including Twitter provide their own APIs for data collection purposes or third party tools such as yourTwapperKeeper are also available. The next step in this cycle is to extract the required data from the massive collections of data. Extracted data should undergo improvements prior to use in applications.

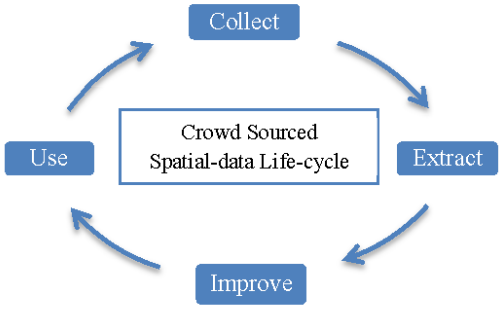


Figure 4.5 Crowd-supported spatial data life-cycle

In today's disasters, there is more and more citizen involvement in the reporting and updating of disaster related information. Figure 4.6 illustrates the crowd-supported disaster management operations. In particular, citizens are largely involved in observing an event and reporting their observations as CSD. The emergency responders can access these CSD and perform actions towards the disaster event management. However, the use of raw CSD which is produced by the citizens can be problematic. Therefore, an important and critical operation identified in this research is to introduce the CSD quality control.

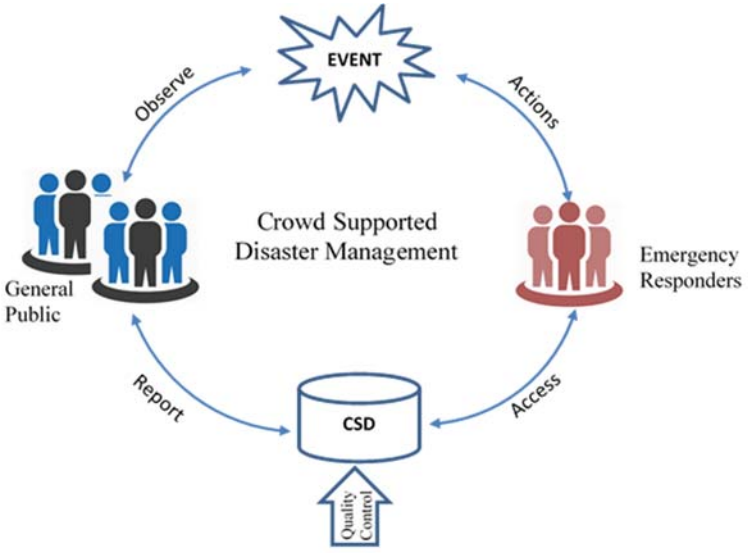


Figure 4.6 Crowd-supported disaster management

4.3.4. CSD quality control in crowd-supported disaster management

This research models the CSD quality assessment by considering a range of dimensions. Figure 4.7 depicts the data flow and different steps to be conducted in disaster management operation planning. Various activities such as filtering may be required to perform during the CSD qualification process. This may include selecting/discarding data at different levels such as messages/sentences or terms based on their importance. Therefore, it identifies the change in the amount of data when the quality of data is considered. As more critical decisions are required, only the most important data is utilised.

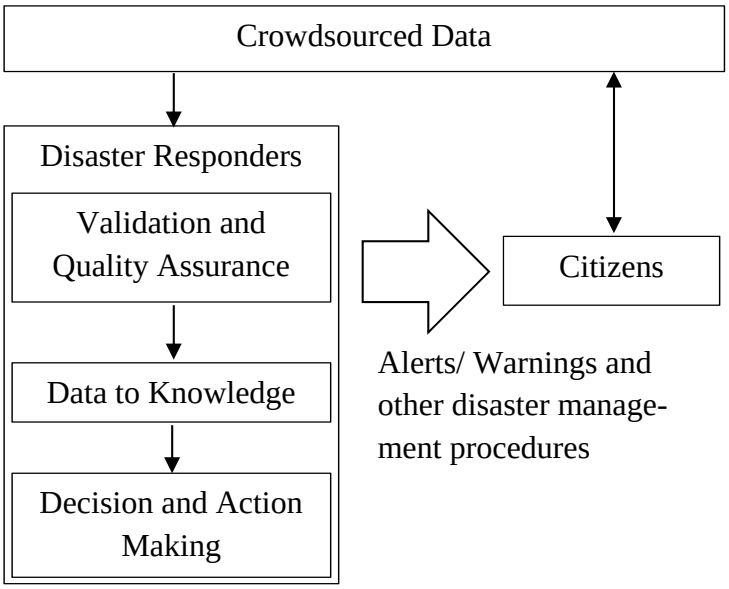


Figure 4.7 Data flow and different steps in crowd-supported disaster management

As identified above, CSD quality assessment is critical in crowd-supported disaster management and so defining CSD quality assessment parameters is challenging. The use of general geospatial data quality assessment parameters is not practical due to CSD's highly unstructured format, its varying quality and undocumented nature.

Within this study, a CSD quality assessment approach was developed by understanding the nature of flood disaster related CSD and the specific requirements of emergency management staff.

4.3.5. CSD quality control workflow

Figure 4.8 describes the semantic official data generation workflow designed in this research. Initially, CSD (in this study related to the flood disasters) from the sources including Twitter and/or Ushahidi were used as the base data for analysis. This CSD required pre-processing to select the required data based on the application. This included processes such as natural language processing and semantic processing and utilised resources such as ontologies and gazetteers. This type of processing is also useful in generating/ updating existing ontology and authoritative data such as gazetteers. The data then undergo quality control processing including relevancy assessments, credibility assessments, semantic location assessments and data improvements. The output data from this processing can generate more appropriate semantic data and be used in applications including disaster management. Moreover, the output data may be useful to process/ improve new CSD.

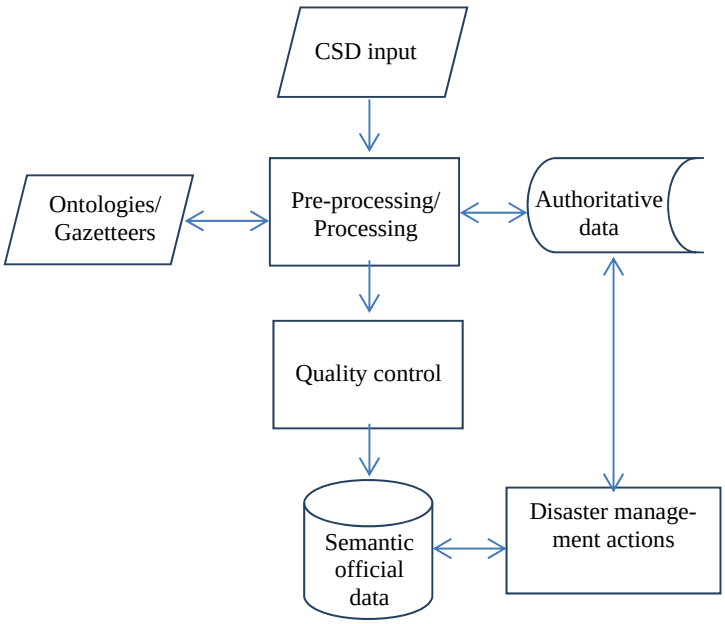


Figure 4.8 Semantic official data generation workflow

4.4. The 2011 Australian floods and the study area

In January 2011, the state of Queensland, Australia, 'experienced its second biggest flood since the beginning of the 20th Century' (Honert et al. 2011). Nearly, 90 towns and 200 000 people were affected by severe flooding, claiming 38 lives and costing over AUD$ 2 billion (https://en.wikipedia.org/wiki/2010-11_Queen sland_floods#cite_note-BBC2010-12-31-2). This research analysed citizen involvement in this natural disaster through the data that was collected via the Ushahidi based Crowd-mapping platform and Twitter. The study-area for the entire research (Figure 4.9) covers an area of approximately 1,355,000 km^2 with the majority of crowdsourced data originating from around the South-East of the state including the capital city, Brisbane.

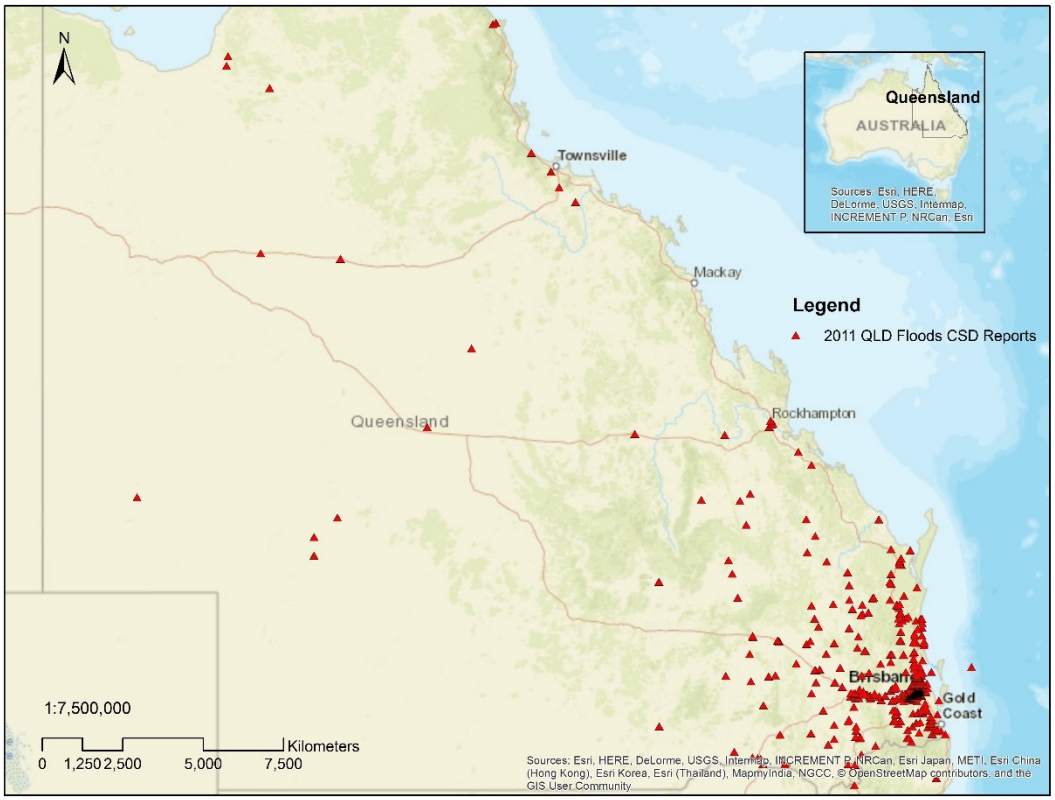


Figure 4.9 Study area and 2011 Australian floods CSD reports

4.5. Data collection strategies

The research investigated two online data collection mechanisms namely social media and crowd-mapping scenarios. Disaster event information was collected through Twitter social media using yourTwapperKeeper[25] tweet capture tool and Ushahidi posts using Ushahidi crowd-mapping implementation.

4.5.1. The yourTwapperKeeper public tweet archival tool

Twitter is an important source of social media data. It is a wonderful platform to access very large amounts of time based public opinions and views on real-world events. The Twitter API provides structured access to communication data in standard formats such as JSON (JavaScript Object Notation), CSV (Comma Separated Values) and Microsoft® Excel using minimal programming effort (Burgess & Bruns 2012). There are various commercial and free tools including TAGS[26], Tweet Archivist[27], TWchat[28] and yourTwapperKeeper developed based on the Twitter search/streaming API. This research selected the free and open source yourTwapperKeeper tweet archiving tool to capture CSD on flood disaster events. The yourTwapperKeeper is the open version of TwapperKeeper tool which was later fully integrated with HootSuite[29] social media management tool. Figure 4.10 shows a resulting window of yourTwapperKeeper tweet capture tool utilised in this research.

[25]https://github.com/540co/yourTwapperKeeper
[26]https://tags.hawksey.info/
[27]https://www.tweetarchivist.com/
[28]http://twchat.com/
[29]https://hootsuite.com/

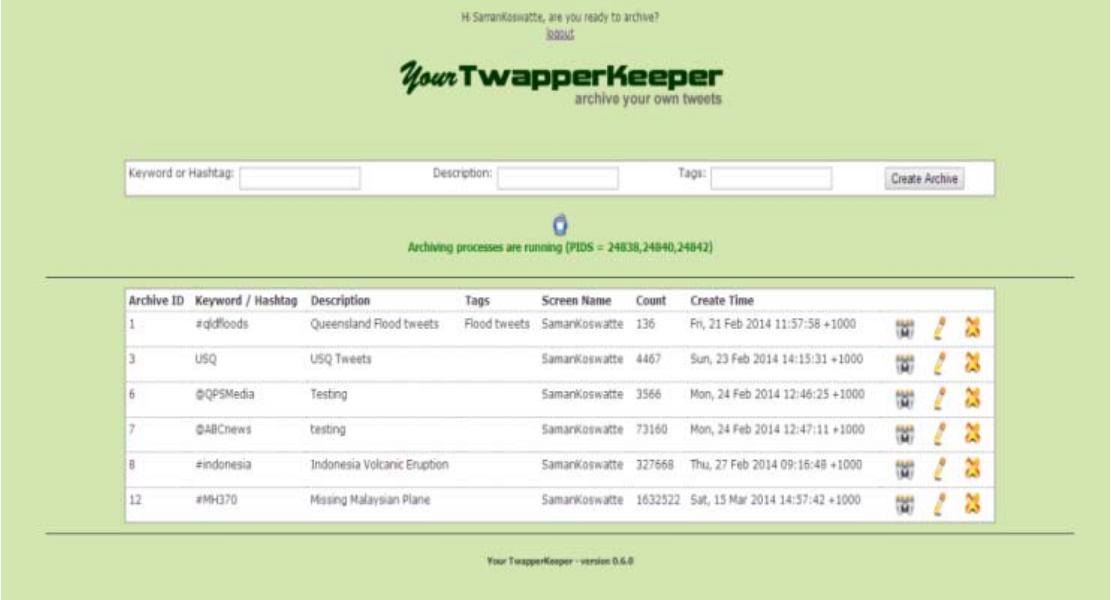


Figure 4.10 The yourTwapperKeeper public tweet collector

4.5.2. 2011 Australian floods' tweets

The two months, December 2010 and January, 2011 were very critical periods for Queensland with a series of flood events due to a La Nina event across the state. With all of the flooding, the social media, including Twitter, was busy with communications including severe weather alerts.

Through a special project carried out by ARC Centre of Excellence for Creative Industries and Innovation (ARC-CCI[30]), the 2011 Australian floods related tweets were recorded using the tool yourTwapperkeeper. More than 35,000 tweets (based on the #qldfloods hash tag) were sent during 10-16 January 2011, with more than 11,600 of them on 12th January alone. Only a very small portion of the Twitter reports included locations i.e. less than 0.5% when analysing the full dataset for location availability. However, it was found that close to 1% of tweets were explicitly location enabled (Figure 4.11) when the pre-processing of these tweets was carried out after discarding the re-tweets. Moreover, there were more than 15,500 Twitter users who participated using the #qldfloods hash-tag and peaked during 11th and 12th January 2011 (Figure

[30]http://www.cci.edu.au/

4.12a) with a variety of different tweet types (Figure 4.12b). During this period, leading accounts included the Queensland Police Service Media Unit (@QPSMedia), ABC News (@abcnews), and the Courier-Mail (@couriermail) (Bruns et al. 2012).

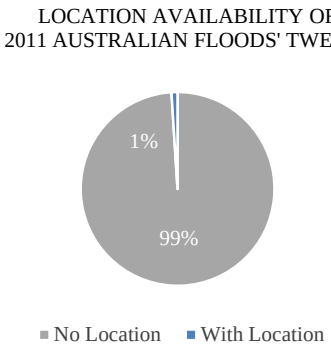
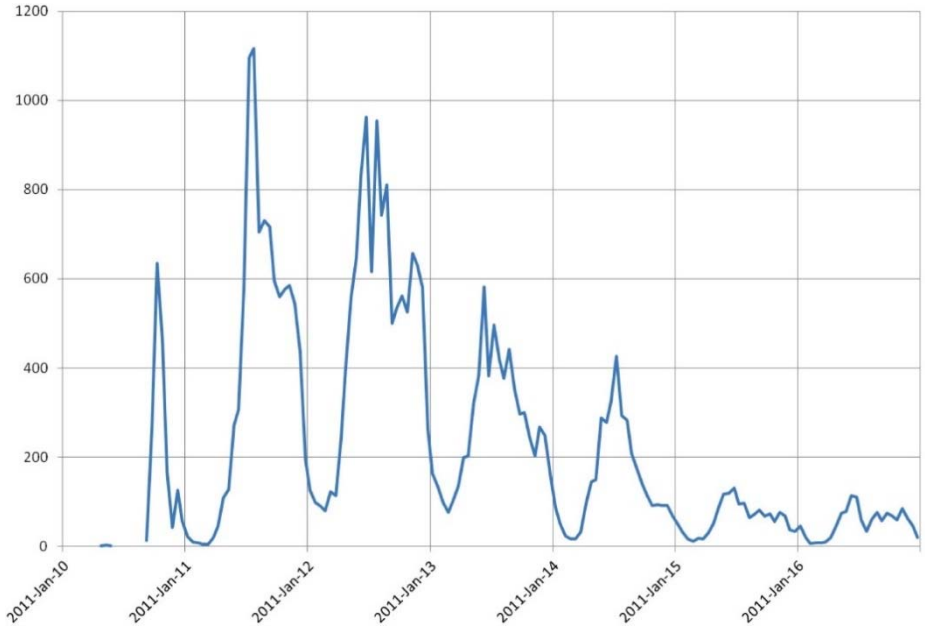
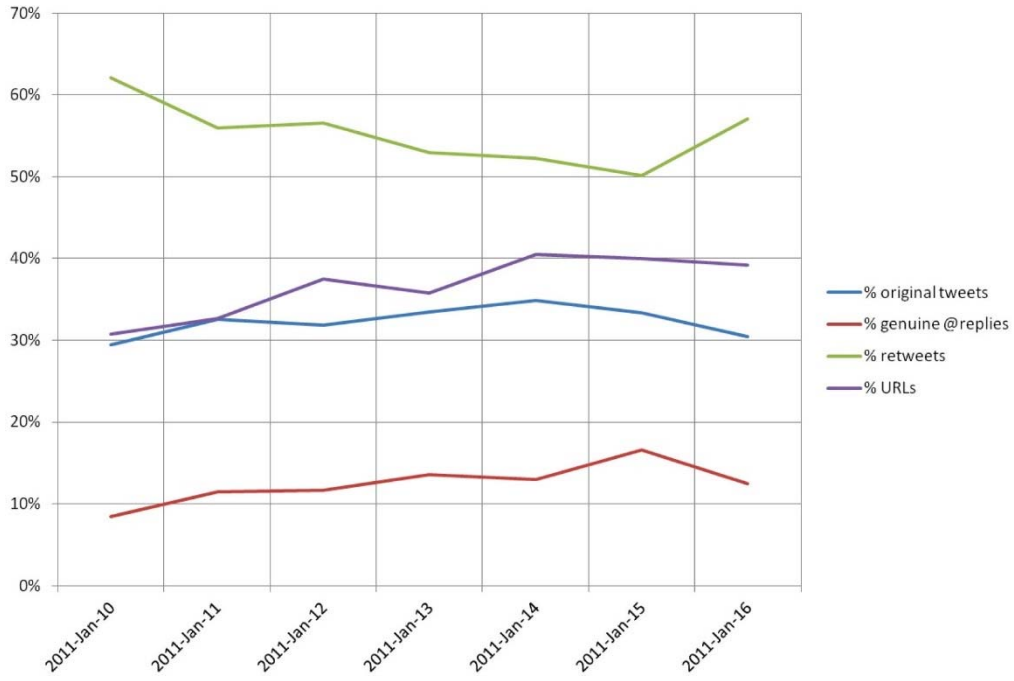


Figure 4.11 Location availability of the 2011 Australian floods tweets



a) #qldfloods tweets per hour, 10-16 Jan. 2011



b) Tweet types

Figure 4.12 Tweet receiving frequency and tweet types (Bruns et al. 2012)

4.5.3. Ushahidi Crowd-mapping platform

Crowd mapping has become popular in many fields including but not limited to the fields of scientific research. It has become a useful platform for both professionals and non-professionals to report their views/opinions on a map with a reasonable location accuracy. As indicated in section 2.4, the Ushahidi crowd-mapping platform was originally created to report election violence in Kenya. Now, its utilisation is wider and include curating local resources i.e. managing local knowledge in business and mapping crisis information. The ease of customisation and free and openness have attracted the interest of many users. It is possible to easily create a mapping instance in the Ushahidi system server or to set up one in a locally configured map server and database service.

4.5.4. ABC's Ushahidi based flood crisis-map (Crowdmap)

During the early stages of the flood event, the Australian Broadcasting Corporation (ABC) maintained an interactive map (Figure 4.13) based on the Ushahidi Crowd-mapping platform (similar to interface shown in Figure 4.14) to gather information related to the 2011 Australian floods. The public uptake of the site was quite remarkable with 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 flood-map 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.

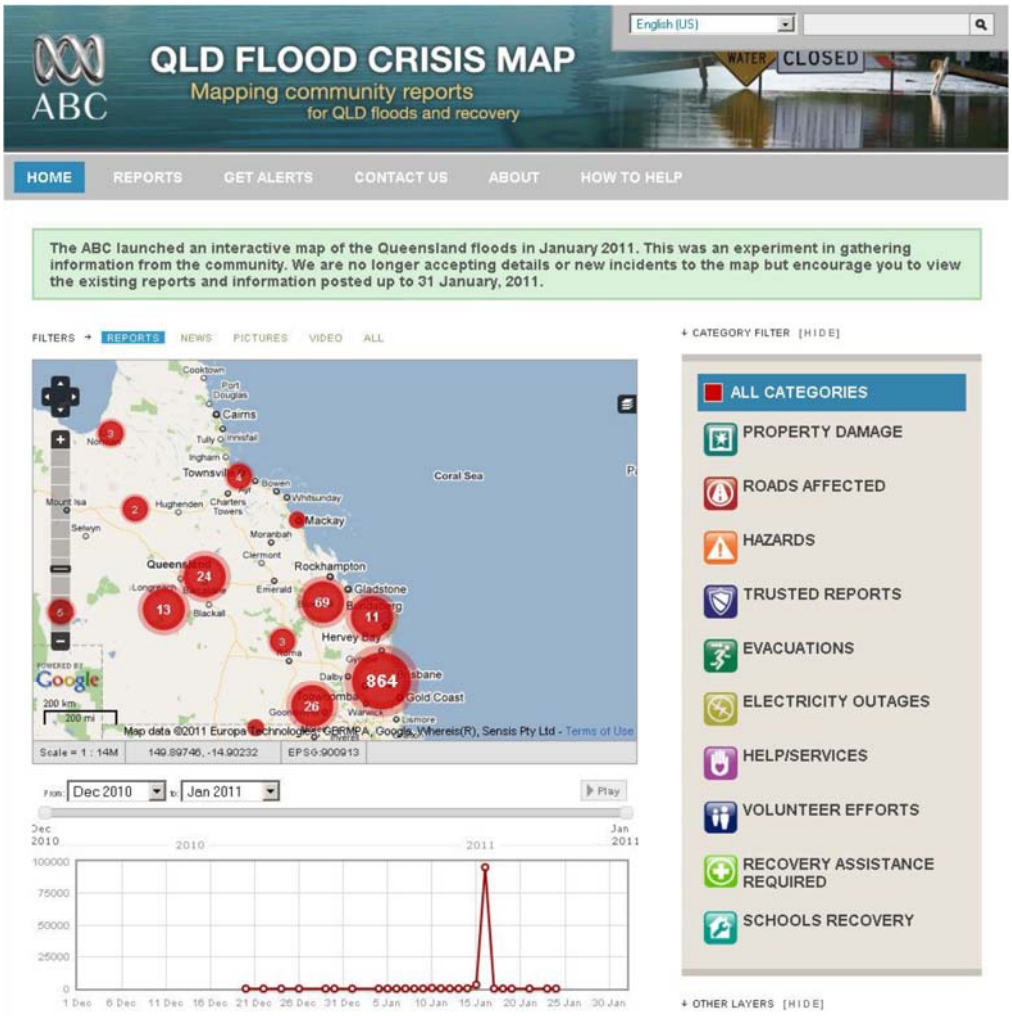


Figure 4.13 ABC's Australian floods Ushahidi Crowdmap (Potts et al. 2011)

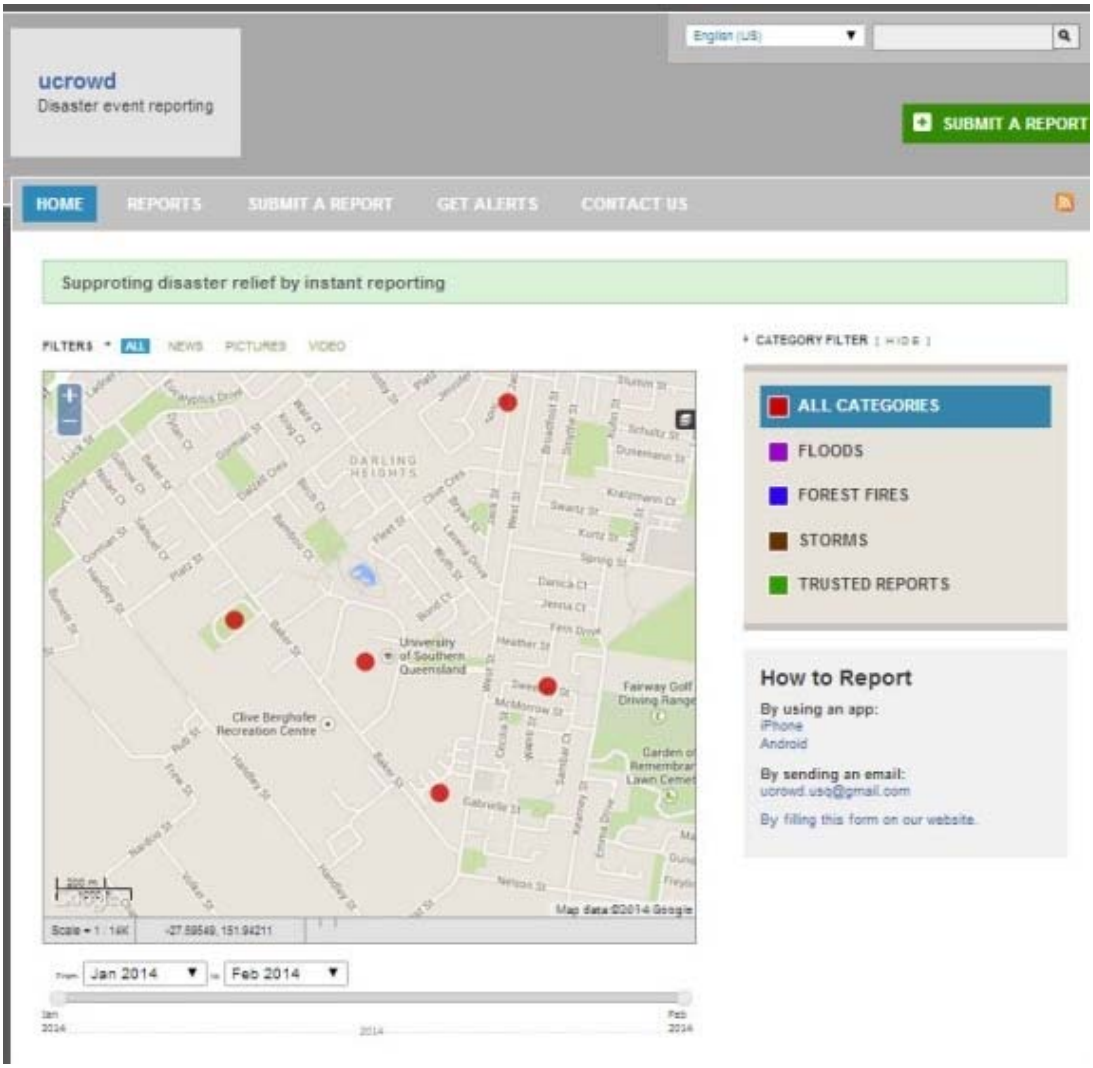


Figure 4.14 Ushahidi Crowd-mapping platform interface

Within the ABC's Australian floods' Ushahidi Crowdmap dataset, there were approximately 700 reports during the period of 9th to 15th of January, 2011, which included the location information where it originated. These records were filtered and extracted for further analysis.

4.6. CSD processing and analysis

4.6.1. CSD location quality assessment

The selected CSD were initially pre-processed to test the location availability. Within the reports submitted during the period of 9th to 15th of January 2011, through the ABC's QLD flood crisis map, 33% (i.e. 704 out of 2136) were location enabled while the remainder of the reports provided other useful data but no location (Figure 4.15a). However, the dataset included many duplicate reports such as re-tweets, which were forwarded messages of the original posts. After removing the duplicates there were only 663 unique reports and 391 reports were location was enabled. Therefore, the location availability of Crowdmap reports can be considered as 59% of the data (Figure 4.15b). Location enablement was possible through mobile devices with GPS, or alternatively, if the user reported using the crisis-map, it was possible to mark the location geographically on the Crowdmap. In both cases, the planimetric location was encoded in latitude and longitude in decimal degrees. The base map of Crowdmap utilised the Google Maps Engine.

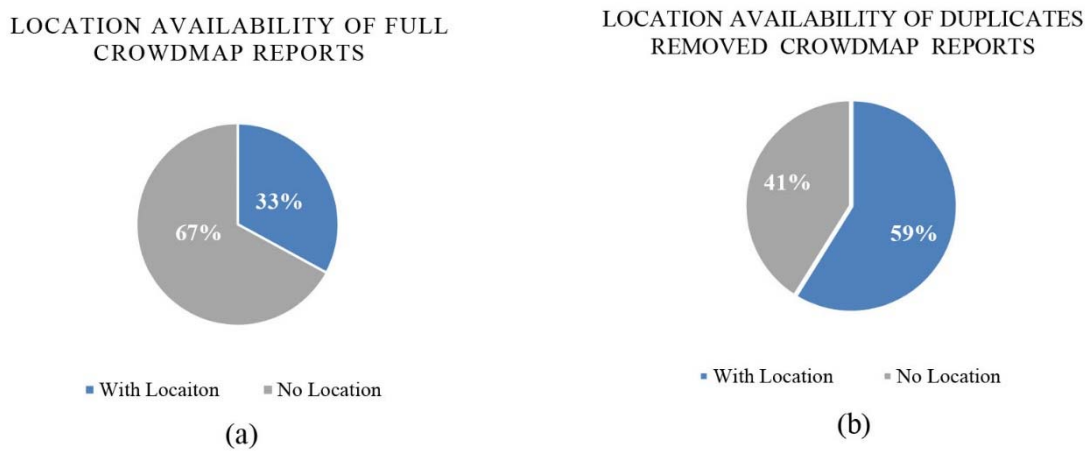


Figure 4.15 Location availability of the Crowdmap reports

The available locations of the CSD content were selected to be analysed by comparing different spatial data sets such as free and open source data, proprietary and closed

type data and authoritative data. It is important to identify the missing location information of the CSD and improve the location quality of CSD and then to fuse this with high quality data such as authoritative data. This research identified the possibilities of using semantic concepts for identifying the missing location data and improving the location quality of CSD. The CSD location quality assessment and semantic location extraction methods along with results and discussion are presented in chapter five of this thesis.

4.6.2. CSD credibility and relevance analysis

The quality of geospatial data is important for obtaining high quality outputs in geospatial applications. The CSD are quite different to traditional geospatial data and requires special forms of quality assessment techniques. This research identified the importance of assessing CSD against credibility and relevance dimensions. The Crowdmap CSD content was analysed against relevance and credibility aspects to assess its quality. Credibility in general determines the trustworthiness or believability of a dataset and relevance determines the fitness for the purpose. This research examined the applicability of spam email detection approaches for CSD credibility detection and GIR techniques for relevance assessment. A naïve Bayesian Network based spam email legitimacy was used to check a sample of the Crowdmap dataset while the relevance of it was assessed using an adapted geo-thematic relevance ranking techniques. Chapter six examines the methods used and presents the results achieved.

4.7. Chapter summary

The design of the research, identification of the research gaps and finding the best approach is very important in scientific research. This chapter detailed the research design and approach used in this research. It has detailed an understanding of CSD, VGI and SDI along with identification of research gaps. This chapter selected the Design Science as the viable research approach to address the research questions. It also

explained and justified the research method used in this study. The conceptual modelling which included CSD in crowd-supported disaster management and CSD quality control were detailed. The study area selected for this study and the data collection procedures along with the methods utilised were discussed. The CSD processing approaches identified in this research were discussed. The next chapter will describe the location quality analysis of CSD and compare the different sources of spatial data such as free and open source, proprietary and authoritative spatial data.

Chapter 5: **CSD Location Quality Assessment**

5.1. Introduction

The previous chapter introduced the research approach along with the methods used to achieve the research objectives. CSD often comes with variable quality information including the location data. This chapter discusses the location data quality assessment of CSD compared to the other forms of spatial data including free and open source, proprietary and authoritative spatial data. The detailed research methods utilised to assess the Ushahidi Crowdmap public reports are presented. Firstly, an analysis and comparison of street names was conducted to identify the variability in both the submitted report data and the data that existed through the various authoritative and third party sources. Secondly, semantic CSD location information retrieval and semantic gazetteer creation were undertaken to extract location data. This included ontology development for converting the Queensland gazetteer into a semantic gazetteer. The results from each of these analyses are then discussed with respect to the CSD location analysis and semantic location extraction.

5.2. Research methods

5.2.1. 2011 Australian floods' CSD reports street name comparison

The location quality of the CSD is important for its further utilisation in any application, particularly where it may impact on other people. Often, the CSD location availability is restricted due to privacy or other reasons. Moreover, the location data that are available may also be problematic as it is often produced by variously skilled volunteers. This section of the research examines the quality of available locational dimensions of CSD by comparing a range of other datasets including authoritative data such as Queensland Department of Natural Resources and Mines (QDNRM) and more open data such as OSM data and proprietary data such as Google Maps.

The QDNRM street and road network data is held by the State Government of Queensland and has been compiled through authoritative data collection of data at both local and state government levels. The data is used extensively within local and state government agencies to support their day to day business operations. OpenStreetMap (OSM) is considered as a crowdsourced product (Haklay & Weber 2008) and its 'Australia and New Zealand street' datasets have been compiled through a range of volunteered data and other available data. These data layers were downloaded and consisted of street name columns which were used to undertake a comparison with ABC's Ushahidi Crowdmap report locational data. The workflow for the comparison of crisis map report locations is illustrated in Figure 5.1. In this comparison, extracted location data available from the Crowdmap was spatially related with the Queensland QDNRM roads and OSM streets layers separately (i.e. to test the location quality against authoritative data and more open type data) using the ESRI's ArcGIS[31] spatial analysis tools.

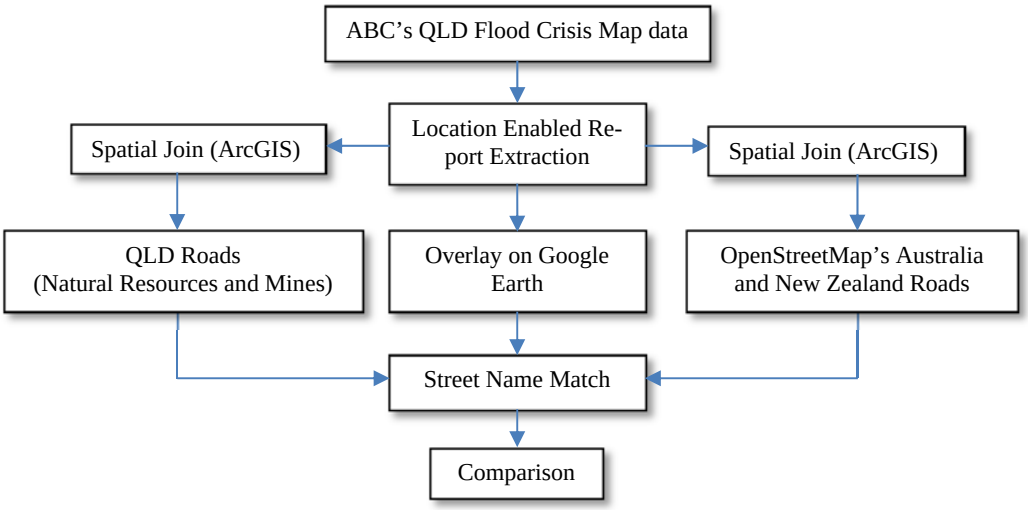


Figure 5.1 Workflow of comparison

Firstly, the reports with location available information were extracted from the ABC's Ushahidi Crowdmap using Microsoft Excel software. As identified in section 4.6, only

[31]http://www.esri.com/arcgis/

33% of reports from the full Crowdmap reports had location data, but once the duplicates were removed from the Crowdmap reports it resulted in approximately 59% of data (Figure 4.15). The street names from the QDNRM roads were then imported to a point attribute table using the ArcGIS software. Distances from these points to the closest available street on the QDNRM roads layer were then computed (Figure 5.2) using the spatial join tool in the overlay analysis toolset of the software. The tool joins attributes from one feature to another based on a spatial relationship. The process matches rows from the Join Features (Crowdmap reports) to the Target Features (DNRM roads) based on their relative spatial locations and the condition where match option is selected as 'CLOSEST'. The distance to the closest feature is then calculated and stored in the attribute table. If it is required to find the second, third, or N^{th} closest feature, the 'Generate Near Table' tool from the Proximity toolset can return a table with this data. However, this study did not select to generate other secondary closest features.

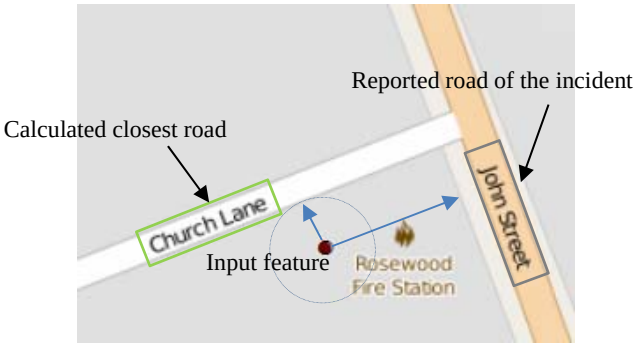


Figure 5.2 ArcGIS spatial join and closest feature identification

To perform a one to one match of the incident location reported through the crisis map, a comparison was then undertaken with the QDNRM and OSM Street names. The same location data from the ABC's Ushahidi Crowdmap reports were also overlayed on Google Earth for a further high-level analysis. This was undertaken to further understand why the distance to the closest street from the points identified were not

matching and why the street name or the reported location mismatched? In this process, various possible reasons for the data mismatches were identified and are discussed in the results and discussion sections of this chapter.

5.2.2. Semantic location information retrieval from CSD

5.2.2.1. The study area for semantic location extraction

For this analysis, subsets of the Ushahidi Crowdmap data and the 2011 Australian floods' tweets data were selected. The dataset included the public's social media interaction during the 2011 Australian floods disaster using the #QLDFloods hashtag via Twitter and the Ushahidi based crisis mapping reports. The study area selected for the semantic location extraction process (Figure 5.3) covers an area approximately 4000 km^2 where the majority of tweets and Ushahdi posts originated.

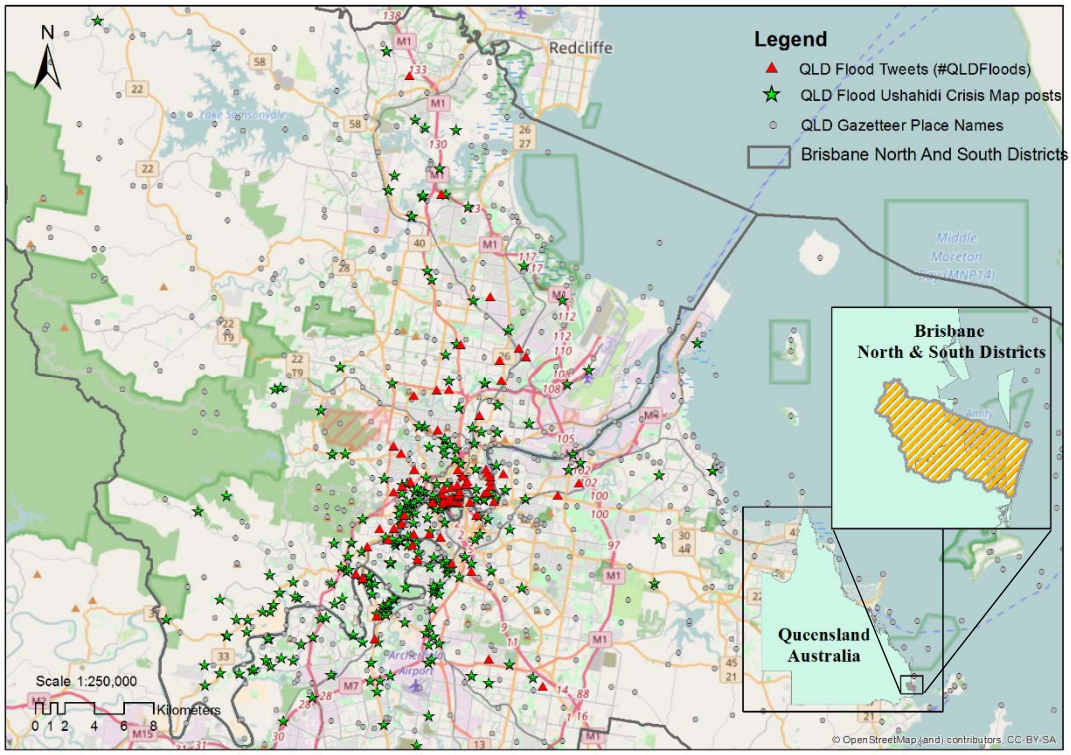


Figure 5.3 Study area and 2011 Australian floods CSD

Selected samples from both the 2011 Australian floods tweets and the Crowdmap data which fell inside the North and South Districts of Brisbane City, Queensland, Australia (Figure 5.3) were used as input CSD in this analysis. The analysis area was selected based on the high density of crisis communications which occurred in this area. The sample contained 89 tweets, 268 Ushahidi posts and 800 Queensland Gazetteer place name entries which were all provider location enabled.

5.2.2.2. Semantic CSD location extraction approach

Figure 5.4 illustrates the overall approach to the semantic CSD location extraction. The first step of this process was to design and develop an ontology set for the Queensland Gazetteer to convert it to a semantic gazetteer. The ontology development process to convert the Queensland Gazetteer into a semantic gazetteer is discussed in the next section (section 5.2.2.3).

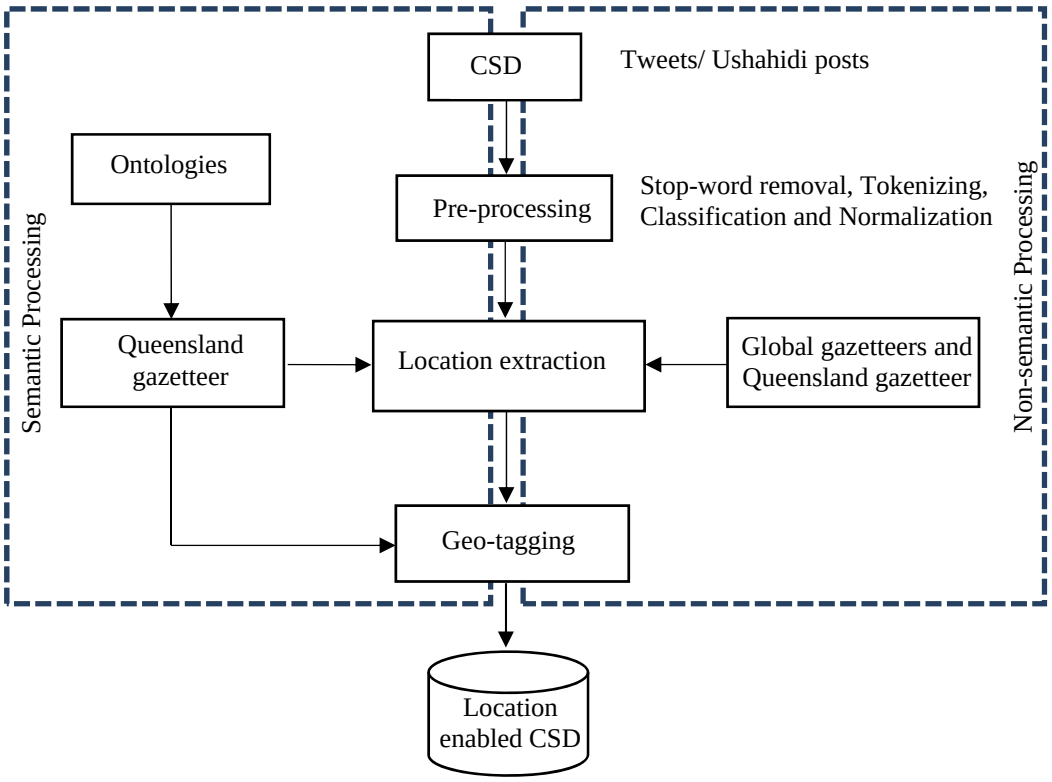


Figure 5.4 Semantic CSD location extraction and geo-tagging

The selected CSD dataset which contained Twitter and Ushahidi Crowdmap data on the 2011 Australian floods was initially pre-processed using stop-word removal, tokenizing, classification and normalization steps. The stop-word removal process identified and removed terms such as 'the, it, as' which are usually very common words in the English language. The removal of these words reduces the unnecessary words in the text. The tokenizing refers to the breaking up of sentences into elements (i.e. words, keywords, phrases, symbols etc.) called tokens. These tokens can then be further processed.

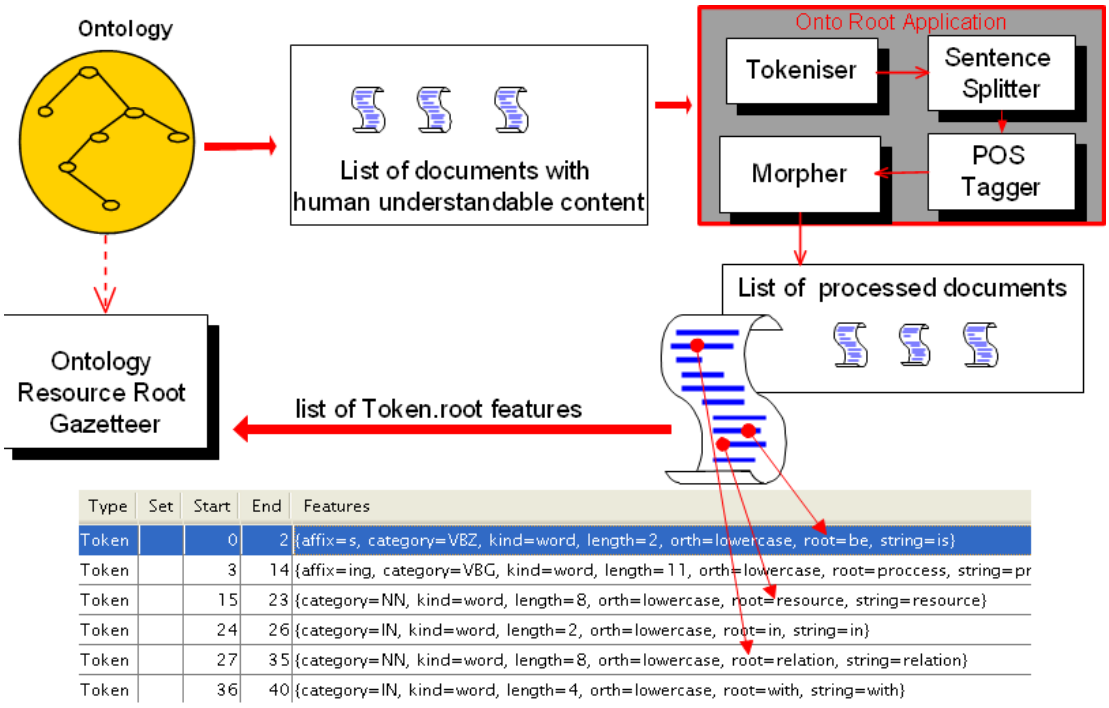
Tweets often contain some unusual words including slang (e.g. Aussie – Australian, Ambo – Ambulance Driver, barbie – barbecue) or abbreviations (e.g. RT – Re-Tweet, pls – please, rmv – remove) which must be carefully analysed and removed as required. Therefore, as part of the pre-processing, these terms were also identified and classified manually using a random sample (150 tweets) of the 2011 Australian floods' tweets to enable further processing. In the normalization phase of the pre-processing, other identified special terms were converted into meaningful content using GATE's 'Tweet Normalizer' tool. The 'Tweet Normaliser' was useful to identify spelling corrections and to expand common abbreviations and Twitter specific terms.

The approach used for location extraction in this analysis was to use the NLP based gazetteer look-up technique. Using this approach, each of the terms of the message content were matched with a list of gazetteer terms to identify the possible location terms i.e. toponyms. The location extraction was conducted in two ways namely, semantic location extraction and non-semantic location extraction. The non-semantic location extraction used two global gazetteers and a local non-semantic gazetteer. The semantic location extraction was conducted using the semantic Queensland Gazetteer developed in this research and is further explained in the section 5.2.2.3.

The GATE software consists of many useful modules such as the ANNIE system. The ANNIE system includes a tokenizer, sentence splitter, Parts of Speech (POS) tagger, gazetteer, Finite State Transducer (FST) and ortho-matcher. Generally, the tokenizer

splits a sentence into tokens such as words, keywords, phrases and symbols. These tokens are useful for further processing steps. The sentence splitter then assembles the tokenized text into complete sentences. The POS tagger assigns Part-of-speech tags (usually in a selected language) to the words. In general, a gazetteer is a geographical naming dictionary however, the gazetteer used in GATE software consists of a list of names of entities such as cities, organisations, days of the week and other terms. The FST (Finite State Transducer) is used to generate new relationships or different outputs using rewrite rules by performing mathematical operations over the existing annotations. The GATE software uses the Java Annotation Pattern Engine (JAPE) language for developing the FST rules. JAPE rules also generate co-references (for example to identify the variants of proper nouns e.g. Peter Smith and Mr. Smith) which is also referred to as orthographic co-reference identification. This can be done by the orthomatcher module of the GATE software using JAPE rules. Please refer to section 3.5.2 of this dissertation for more information on the GATE software and its Processing Resources (PRs).

The two datasets were then analysed separately using the PRs of the GATE software and GATE's morphological analyser which considered the tokens and their POS tag to identify the lemma (i.e. base or dictionary form of a word) and affix (end words i.e. s, ss, ies etc.). The Document Reset process was used to reset the previous annotations in the document along with Queensland Place Name Gazetteer for non-semantic analysing. In the semantic analysis phase, ANNIE OntoRootGazetteer and Flexible Gazetteer were used along with the above processing resources. The OntoRootGazetteer is normally a dynamically created gazetteer which is capable of producing semantic annotations. The Flexible Gazetteer provides flexibility to users to choose their own customized input. Figure 5.5 describes the OntoRootGazetteer development process from the ontology. The human understandable documents are initially processed with the GATE's PRs using the ontology and the root tokens are added to the OntoRootGazetteer list.



Type	Set	Start	End	Features
Token		0	2	{affix=s, category=VBZ, kind=word, length=2, orth=lowercase, root=be, string=is}
Token		3	14	{affix=ing, category=VBG, kind=word, length=11, orth=lowercase, root=proccess, string=pr
Token		15	23	{category=NN, kind=word, length=8, orth=lowercase, root=resource, string=resource}
Token		24	26	{category=IN, kind=word, length=2, orth=lowercase, root=in, string=in}
Token		27	35	{category=NN, kind=word, length=8, orth=lowercase, root=relation, string=relation}
Token		36	40	{category=IN, kind=word, length=4, orth=lowercase, root=with, string=with}

Figure 5.5 Building OntoRootGazetteer from the ontology (Source: https://gate.ac.uk/sale/tao/splitch13.html)

The following PRs were selected and organised in the order below to perform the semantic geo-tagging of the selected CSD.

1. Document Reset: To reset the previous annotations and bring the document to its original state if it has to iterate through the PRs or redo the work.
2. ANNIE English Tokenizer: To tokenize the content to very simple tokens such as numbers, punctuation and words based on the English language.
3. ANNIE Sentence Splitter: To split the text in to new sentences.
4. ANNIE POS tagger: To detect and annotate part-of-speech tags.
5. GATE Morphological Analyser: To analyse the morphological variations and identify the lemma (i.e. the dictionary form of the word) and affix of the terms considering tokens and POS tags.

6. OntoRootGazetteer and Flexible Gazetteer: To use as a reference list of toponyms for the semantic processing. An ontology is required which can be developed using the GATE's 'Gaze Ontology Gazetteer Editor' ontology design and editing tool.
7. JAPE Transducer: JAPE (JAVA Annotation Patterns Engine) establishes rules that can be used to combine annotations and build complex annotations to draw more semantic contexts of the contents.

The annotations of messages were filtered manually and assigned the best location annotation for each message of Twitter and Ushahidi data where the annotations were available. Each of these messages were then manually geocoded for further processing and analysis.

5.2.2.3. Ontology development for QLD Placename Gazetteer

As discussed previously, ontologies are key to semantic information processing. An ontology set was developed to convert the general Queensland Gazetteer into an ontological gazetteer. This was to enable the semantic processing of the selected CSD in this study. This research followed Noy and McGuinness' (2001) Ontology Development 101 guide and the ontology development workflow proposed by Scheuer et al. (2013) for developing the QLDGazOnto ontology set.

Answers were defined to the ontology development questions suggester by Noy and McGuinness (2001). Table 5.1 shows the questions and initial answers defined in this process.

Table 5.1 Ontology development questions and initially defined answers

Question	Answer initially defined
What is the domain that ontology will cover?	Flood disaster management.
For what we are going to use the ontology?	For analysing the quality of Flood related CSD.
For what type of questions the information in the ontology should answer?	Toponym identification, hierarchical placenames, spatial semantics, ambiguity resolutions etc.
Who is going to use and maintain the ontology?	Flood disaster management staff

Reusing existing ontologies is useful for saving development time and efforts. This research identified the GeoNames and WordNet ontologies as useful for the selected task. Although they did not fully cover the domain and scope of the study, modified versions were useful for the ontology development process. A comprehensive list of flood disaster management related terms was then created. The relevant terms from the selected existing ontologies and Queensland local gazetteer were utilised for this task.

A schema was designed for the ontological gazetteer (QLDGazOnto) during the conceptualisation process and the GeoNames ontology adapted and reused. Each element of schema represented the heirachical classess in the ontology set designed and their semantic combinations (i.e. a place can have one or more alternative names). There were different options for selecting ontology development software and tools including GATE ontology API, Protégé[32], Fluent editor[33], Neon-Toolkit[34] etc. The ontology development process in this study was carried out using the GATE's ontology tools which provided the ontology viewing/editing facilities.

[32]http://protege.stanford.edu/
[33]http://www.cognitum.eu/semantics/FluentEditor/
[34]http://neon-toolkit.org

Figure 5.6 shows a view of the ontology set developed for QLDGazOnto by adapting the GeoNames ontology using the GATE software's ontology editor.

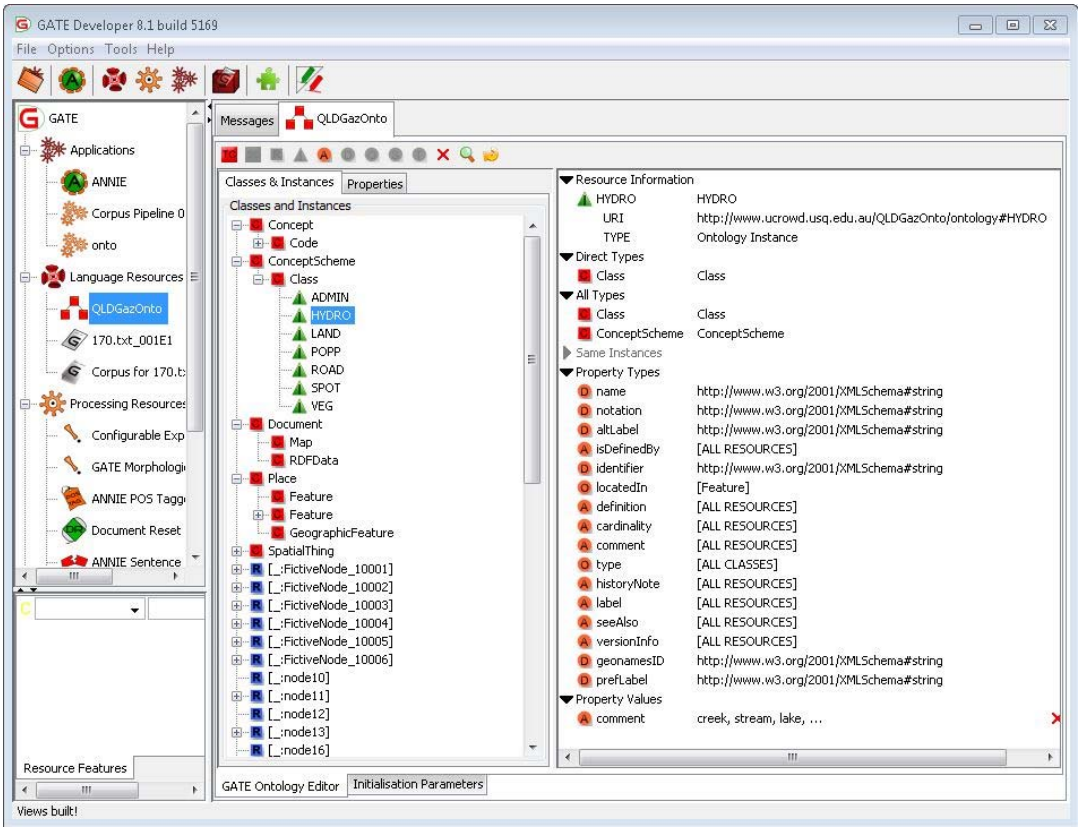


Figure 5.6 QLDGazOnto development using GATE's ontology editor

5.3. Results and Discussion

5.3.1. CSD Location Analysis Results and Discussion

Google Maps versus QDNRM data

The users of the Crowdmap site mostly reported the location of incidents based on the Google Maps data unless they used the GPS location of a mobile device. These reported locations were compared to the closest street name in the QDNRM roads database that was identified using the closest road or street feature calculation by spatial

join explained in the section 5.2.1. The identified road names from the database were matched with the street name of the reported location. The results indicated that approximately 38% of the incident locations' street names reported though the Crowdmap agreed with the QDNRM street names (Figure 5.7). This result raises an important question for the disaster responders. As official disaster responders rely on authoritative street name data, will they trust information based on other spatial data sources such Google Maps, as information verification is critical? However, this analysis does not provide sufficient depth of understanding as to the most accurate spatial data sources currently available and provides no evidence about the validity of information on the other dimensions such as credibility or relevance.

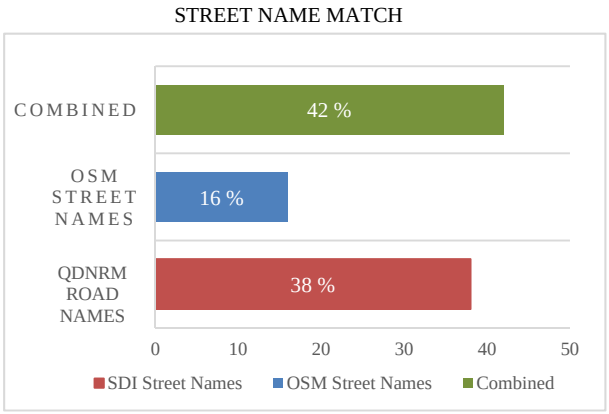


Figure 5.7 Street name matching results

Google Maps versus OSM roads

The use of OSM is becoming increasingly popular throughout the world. However, the lack of data coverage (Haklay et al. 2010) for many parts of the world, including Australia (Pullar & Hayes 2017), is limiting its wider utilisation. Within the study, the Ushahidi Crowdmap report locations were also compared with OSM Queensland street names. The matching results were quite low in comparison to the QDNRM street comparison results. Only around 16% of the reported incident locations were matched with OSM street names within the study area (Figure 5.7), even though both Google Maps and OSMs were conceptually similar, the data availability, content and quality

were obviously different. The key differences were that OSM is an open data project while Google Maps is closed at the level of its raw data capture and is designed for commercial operations. As google maps is a commercial product it is natural for the company to focus on geographical areas that may have commercial return. Therefore, some geographic areas are not as up-to-date as other areas. OSM is an open platform and anyone can freely contribute to the system or can use its services. Due to this fact, the content of the OSM is growing faster and sometimes can maintain high granularity even in remote and less developed areas. This can happen due to the interest over an event or a personal interest over an area concerned. A good example is the significant coverage change of the OSM after the Haiti earthquake in 2010 (Zook et al. 2010).

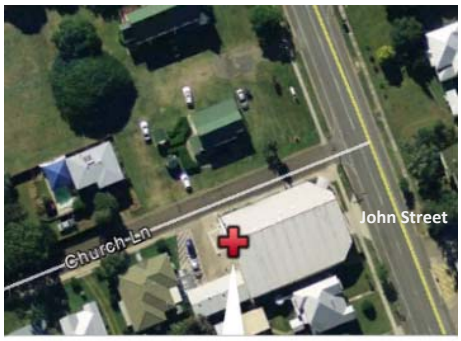
Issues of QDNRM and OSM street data

In understanding the issues related to the spatial data integration and potential automation of location data improvement, the street name matching results from this study (Figure 5.7) are very important. QDNRM data produced by Department of Natural Resources and Mines is the authoritative dataset for the state and outperformed the OSM data according to the analysis of the results. However, improvement in the QDNRM data is potentially possible as the study identified potential incompleteness of the street names. By combining the two datasets, an improved data matching was achieved with up to 42% of the records matching (Figure 5.7) and therefore identified the potential for the next generation of spatial data users to be part of the data improvement through fusing authoritative and crowdsourced data for data improvement. Throughout the analysis, a number of issues were identified, where the street name did not match although the distance was less than 15 m from the road centreline. Some possible reasons for these issues are detailed as follows.

Information provider decisions on importance

Figure 5.8 shows a view of a report received and overlayed on Google Earth. According to the sender, the location of the Emergency Centre was in John Street. However, according to the spatial relationship analysis, it identified the nearest street as Church

Lane. In this case, the location data that was provided located the incident closer to Church Lane rather than John Street, which is the main road. In this case, the official address and the reported address have been confused. If the spatial join had opted to select the 2nd or 3rd closest match it could have found the reported street name correctly. However, this research only selected the 1st closest match and the user reported street name was identified as incorrect in this scenario.



Emergency Centre: Rosewood Emergency Services

Emergency Centre: Rosewood Emergency Services	
FID	1327
F1	1453
INCIDENT_T	Emergency Centre: Rosewood Emergency Services
INCIDENT_D	
LOCATION	Rosewood Fire Station 70 John Street Rosewood QLD 4340
DESCRIPTIO	At 23:00 11-Jan-2011 Sylvia from Ipswich showgrounds evacuation centre stated that this centre was open. No further sources identified.
CATEGORY	Evacuations Relief Centres/Emergency accomodation
LATITUDE	-27.637828
LONGITUDE	152.591277
APPROVED	YES
VERIFIED	NO

Figure 5.8 Effect of personal choice for reported location

Information provider's local knowledge

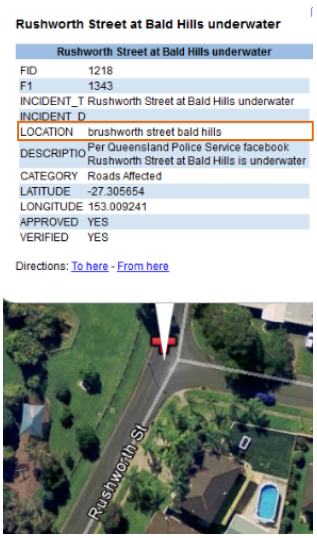
Figure 5.9 depicts two views of the report locations on Google Earth. The reporter has provided the location as Railway Street, however actually it is Victoria Street according to Google Earth. When conducting a vicinity search for Railway Street, there was a connection between Victoria Street and Railway Street, which led the provider's decision to provide the detailed location as Railway Street. After exploring cases of mismatch, it was identified that there were a number of issues related to the datasets being compared including, information incorrectly provided, incomplete road or street information and lack of information detail.

Forest Hill Evacuation Centre

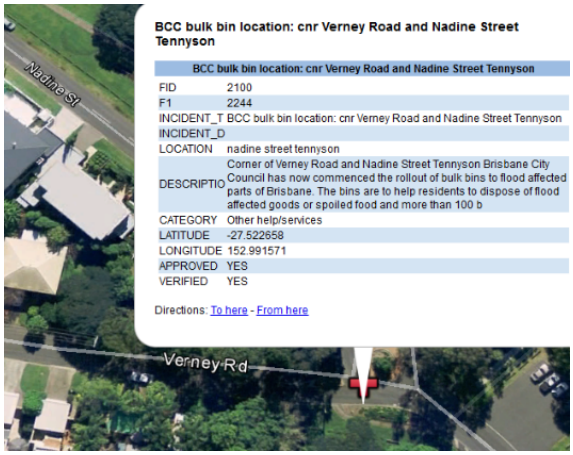
Forest Hill Evacuation Centre	
FID	584
F1	702
INCIDENT_T	Forest Hill Evacuation Centre
INCIDENT_D	
LOCATION	School of Arts Hall Railway St Forest Hill Queensland
DESCRIPTIO	School of Arts Hall Railway Street Forest Hill. For more information go to Queensland Police Facebook page. http://www.facebook.com/QueenslandPolice#!/QueenslandPolice?v=wall
CATEGORY	Relief Centres/Emergency accomodation Trusted Reports
LATITUDE	-27.590635
LONGITUDE	152.356509
APPROVED	YES
VERIFIED	YES

Figure 5.9 Effect of local knowledge for reported location

There were numerous cases such as incorrectly supplied information, incomplete data and sometimes lack of data. The incorrect data may be the result of the typographical errors of the reporter (Figure 5.10a) or incomplete road segment information (Figure 5.10b). According to the Google Earth street information, the report is accurate. However, when it was analysed along with the QDNRM roads, the street name appeared as 'a roundabout' only and resulted in a mismatch. Furthermore, the OSM street data were mostly incomplete and lacking details for the reported locations. The matching percentage was less than 20% for the entire dataset (Figure 5.7).



(a) Typing errors



(b) Incomplete road information

Figure 5.10 Results of incomplete data

Differences in location of the incident and the location of the information provider

As Figure 5.11 illustrates, there is a considerable difference between the actual location that the reporter described and the encoded location of the report. A possible reason for this could be that the information provider was on the move or there were obstacles blocking the reporter such as flood waters. The reporter realised that in Figure 5.11 where it was identified that Melton Road was closed, but they may have reported the incident when they reached a convenient and a safer location to do so. Other possible scenarios are the reporter can be in a moving vehicle, reporter is different from the observer, more than one observers and one reporter etc.

Figure 5.11 Differences of report and reporter locations

5.3.2. Semantic CSD location extraction results and discussion

As reported in this analysis, the CSD's explicit location availability is low and the quality of the available location can often be problematic. In response to this limitation the next stage of the analysis semantically analysed the CSD message contents to identify location information from the message's textual descriptions. In the message descriptions, the users often included the type of incident and a brief description of the incident location. Therefore, the analysis tested the possibility of semantically extracting these textual locations as toponyms and then geocoding this data as real locations (i.e. latitude and longitude or grid coordinates). The approach utilised was explained in section 5.2.2.2 and used a semantic gazetteer lookup and Natural Language Processing techniques.

The selected CSD messages (i.e. 89 tweets and 268 Ushahidi posts) were semantically annotated and Table 5.2 lists the annotation accuracy matrix. The GATE's 'Annotation Diff' tool defines:

- **Precision** as the measure of the number of correctly identified items as a percentage of the number of items identified
- **Recall** as the measure of the number of correctly identified items as a percentage of the total number of correct items and

- **F-Measure** as the weighted average of those two.

The ANNIE Gazetteer is a global gazetteer used in GATE as the default gazetteer. The QLDGazetteer is Queensland's official place name gazetteer while QLDGazOnto was its ontological version developed in this study. It was developed with the main focus on the Ushahidi dataset and the results were dominant in tagging the Ushahidi dataset based on the Ontological Gazetteer. Generally, the ontology development is a cyclic process and often needs revision. In the analysis undertaken, the ontology was designed and tested using a sample (i.e. 150 Ushahidi posts) of the CSD data which does provide some limitations.

Table 5.2 Comparison of gazetteer success for Twitter and Ushahidi

Composition of Gazetteers	**Ushahidi**			**Twitter**		
	Precision	**Recall**	**F-Measure**	**Precision**	**Recall**	**F-measure**
ANNIE Gazetteer	14	28	19	21	54	30
ANNIE+QLDGazetteer	32	41	36	37	64	47
QLDGazetteer	19	64	29	34	62	44
QLDGazOnto	96	90	93	36	55	44

The results in Table 5.2 indicate that the all annotation quality indicators (i.e. Precision, Recall and F-Measure as percentages where higher values indicate better results) were low when using the GATE's global ANNIE gazetteer. For the Twitter data annotation, it was comparatively higher and the recall value was over 50%. The combined use of global and local gazetteers i.e. Queensland Place name gazetteer (QLDGazetteer) and ANNIE global gazetteer provided higher quality output than the global gazetteer alone for both Twitter and Ushahidi data. The use of a local gazetteer provided higher annotation quality than the global gazetteer. However, quality of that option was lower than the local and global gazetteer combination. Interestingly, the recall remained high as over 60% in both datasets for the local gazetteer based annotation.

Annotation quality measures of the semantic local gazetteer (QLDGazOnto) use were comparatively higher for almost all indicators for the Ushahidi dataset and indicated slightly higher precision and F-Measure with a slightly lower recall factor for the Twitter data annotations. With the annotation quality results, it was more advantageous to use local gazetteers for place name extraction. Although, the combined use of global and local gazetteers showed some improvements, care needs to be taken not to introduce more geo-geo ambiguities. The use of semantics has indicated some improvements in the CSD place name extraction. However, further studies are required to validate this trend.

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 were encouraging as it is independent of the ontology development.

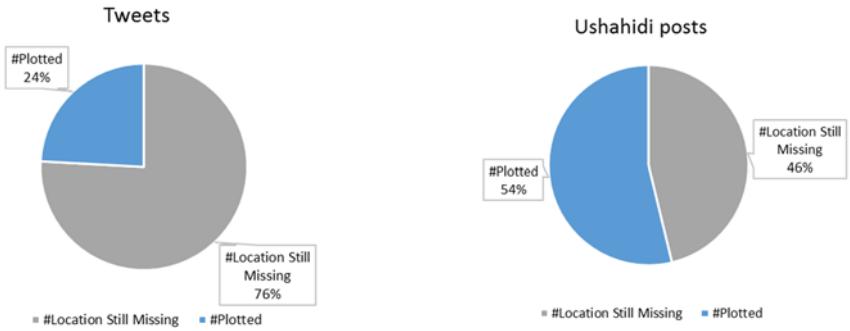


Figure 5.12 Location availability of semantically processed CSD

In the geocoding phase, 24% of the Twitter messages (Figure 5.12) in the study area were able to be semantically geocoded through the research approach. For Ushahidi this was 54% which is very close to the location availability of original Ushahidi Crowdmap reports when duplicates were removed. Figure 5.13 shows the new semantically detected locations of the selected Ushahidi and Twitter CSD samples.

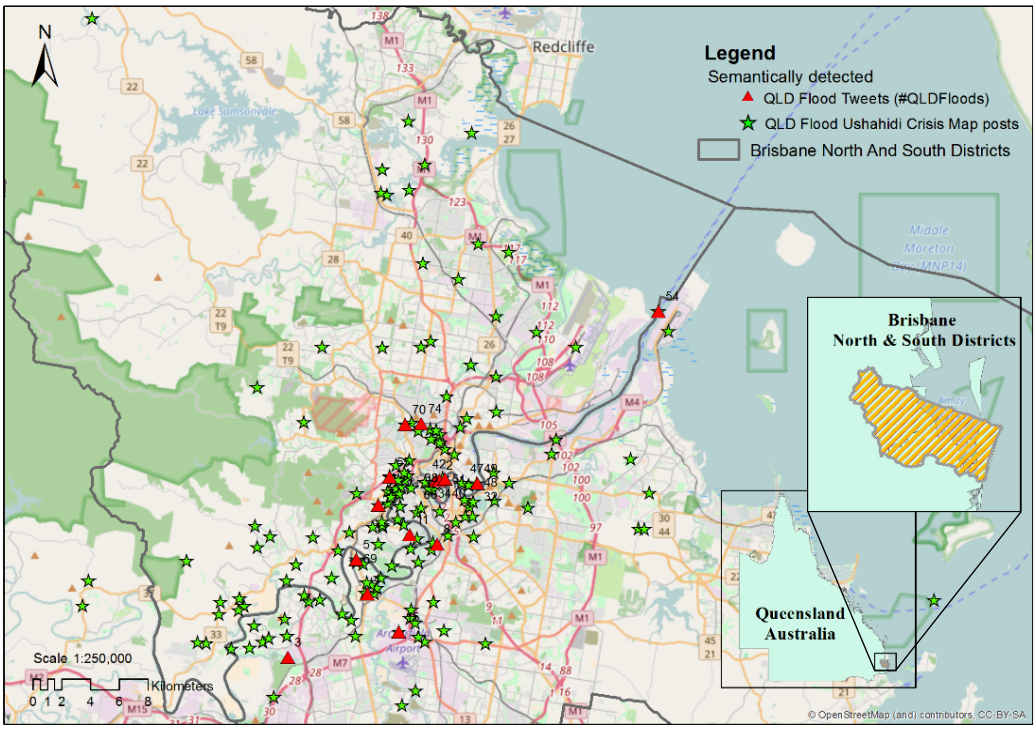


Figure 5.13 Semantically detected new CSD locations

5.4. Chapter summary

Modern spatial data such as CSD may consist of locations in the form of explicit (i.e. locations which are derived from the location sensors attached to the smart phone used) or implicitly (i.e. explained textually in the CSD descriptions). This analysis found that the CSD's implicit location extraction is challenging but possible. This chapter explored CSD location quality analysis methods and compared the available location information of CSD with three different forms of base data namely, free and open-source, proprietary and closed and authoritative data. Moreover, it presented an approach to semantically extract the implicit location hidden in CSD textual descriptions using semantic gazetteer lookup and Natural Language Processing techniques. Results of the semantic location extraction were then discussed along with some identified

issues. The next chapter will describe the CSD credibility and relevance analysis based on approaches used in other fields of information communication and technology.

Chapter 6: **CSD Credibility and Relevance Assessment**

6.1. Introduction

The previous chapter presented the research outcomes on the Ushahidi Crowdmap public reports street name comparison and the semantic location information retrieval from CSD. The first part of this chapter explores the research methods applied to CSD credibility and relevance analysis. The CSD credibility assessment methods are described along with the model design, system training and testing procedures. Next, the CSD relevance detection methods are described and finally, the results, discussions and conclusions of the CSD credibility and relevance analysis are detailed.

6.2. Research methods

6.2.1. CSD credibility analysis

The CSD credibility detection approach consisted of two distinct phases including a system training phase and a credibility detection phase. As indicated previously, during the 2011 Australian floods, the Australian Broadcasting Corporation (ABC) developed a customised version of the Ushahidi Crowdmap to report/map disaster communications. This data was comprised primarily of text based content that was submitted by volunteers during the flood event. The data included input from a heterogeneous range of volunteers who submitted reports during a relatively short period of time (approximately 7 days) via various channels including a mobile app, a website, SMS messages, emails, phone calls and Twitter. A part of that dataset was used in this analysis to train the CSD detection system and then to test the credibility of remainder of the dataset. Figure 6.1 shows a simplified view of the credibility detection process.

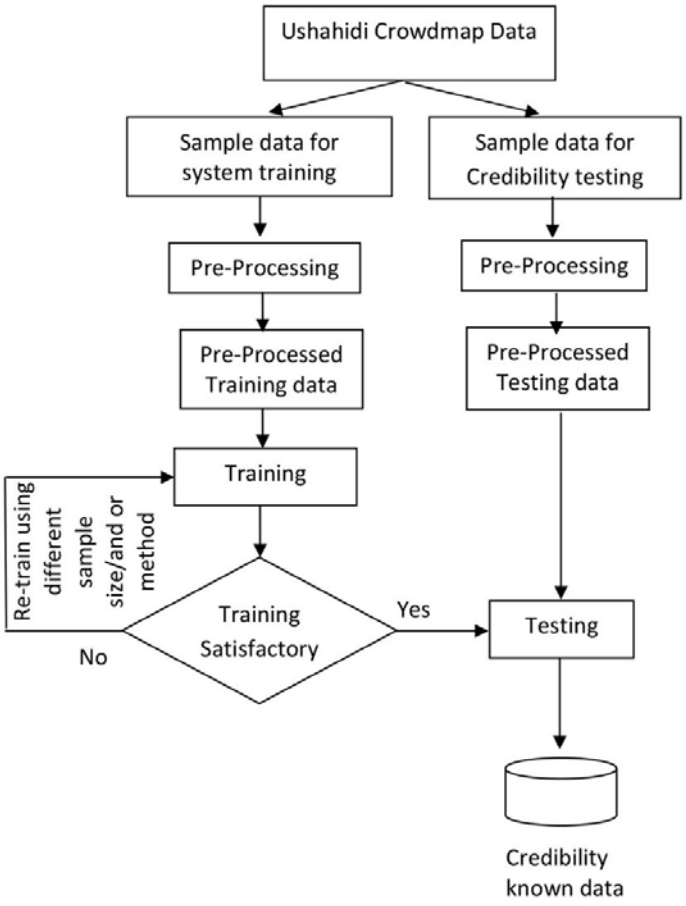


Figure 6.1 Simplified credibility analysis process

6.2.1.1. CSD credibility detection algorithm based on spam email detection approach

An algorithm for the CSD credibility detection based on the naïve Bayesian Network was developed for the analysis. The Java[35] programming language was used for coding the system within the NetBeans[36] Integrated Development Environment (IDE). The pseudo code of the algorithm consisted of two phases including training and testing, and is listed below.

[35]https://java.com

[36]https://netbeans.org/

Phase 1: Start training

Select Classifier and Training Data set

for each Message m_i in Training Dataset D_{tr} **do**

for each Word in the Corpus **do**

Calculate the *Credible* and *Unreliable* Probabilities and store in the Hash Table

end for

end for

End training

Phase 2: Start classification

Select Classifier, Testing Dataset and Hash Table

for each Message m_i in the Training Dataset D_{tr} **do**

for each Word in the Corpus **do**

Calculate the Word Probability for being *Credible and Unreliable*

Update Hash Table

end for

Calculate combined Probability for the Message

if combined Probability > Threshold

Label Message as *Credible*

else

Label Message as *Unreliable*

end if

end for

End classification

The probability threshold was determined after the initial testing and was set at the 0.9 probability level (Threshold = 0.9).

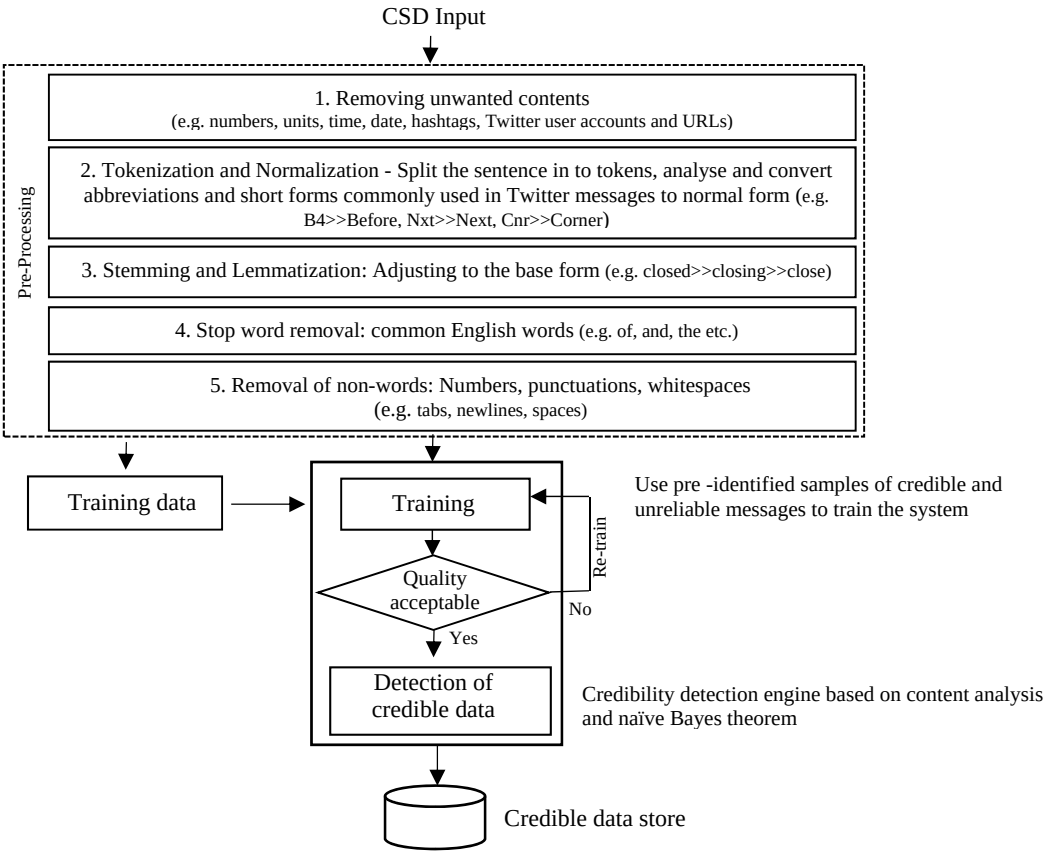


Figure 6.2 CSD Credibility detection workflow

Figure 6.2 illustrates the key steps in CSD credibility detection approach based on the naïve Bayesian Network and the classical "bag of words" model popular in spam email detection.

The ABC's 2011 Australian Floods Crisis Map dataset (Ushahidi Crowdmap) was used as the input CSD. The dataset was initially pre-processed using the steps explained in Figure 6.2. After the data pre-processing, the system was trained using a training sample dataset. Within the ABC's Ushahidi Crowdmap, there were approximately 700 reports during the period of 9^{th}-15^{th} of January 2011 which often included information about the location where the report had originated. After the initial duplicates were removed, there were 663 unique Ushahidi Crowdmap reports remaining. The duplicates of the dataset were removed using the *'Remove duplicates'* tool of the Microsoft Excel software.

For training and testing purposes, approximately 20% of the total reports (143 reports) were randomly selected from this Ushahidi Crowdmap dataset. Eighty percent of these reports (110 reports) were then selected as training data and the remaining 20% selected as the testing data (33 reports). The remainder of the full dataset (520 reports) was later used for the credibility detection analysis.

The whole dataset was initially pre-processed to prepare for the training, testing and credibility detection. The training data set was classified through a manual decision process which identified messages that where either *credible* or *unreliable* based on the credibility of terms within the message. The classification was undertaken by a reviewer who had local and expert knowledge of the disaster area. The system was then trained and tested using the testing data set under two different environments namely, unforced and forced conditions, to test the accuracy and performance improvements.

In the unforced training, the data processing of the test data followed the normal pre-processing steps and was then used directly for refining the training of the system. The results of this unforced training provided a report on the level of possible false positives in the classification. A high level of false positives is indicative of a possible bias in the classification process and is often referred to as *Bayesian poisoning* (Graham-Cumming 2006). The purpose of the forced training was then to review the false positives and other classified data to improve the quality of the classification process and hence re-train the system. The forced training required human intervention to improve the training of the system and therefore some terms which had artificially increased the credibility of the messages were identified and removed. This enabled the training of the system to be further refined and to more effectively distinguish the credible or unreliable messages.

The forced training process consisted of the following stages:

- The location terms were removed/disabled from both the credible and unreliable messages

- Highly credible terms such as *evacuation centre, road close, police, hospital* etc. were removed from unreliable messages to give more weight to similar terms in the credible messages and to avoid Bayesian poisoning
- Removing messages which could cause a high false positive rate and therefore avoid Bayesian poisoning (i.e. removing whole messages such as "*Lots of road closures due to flooding, our back fence was partially pulled down by the flooded Coonowrin creek overnight.*")

When location terms appeared frequently in messages, these terms tended to increase the probability of the message being credible when in reality this was not the case. This impacted both the credible and unreliable messages. This impact was reduced by removing all the location terms in both credible and unreliable training sample messages. The Queensland Place Names Gazetteer was used as the basis for removing location terms as it provided a list of registered geographic locations and places. All incoming message terms were cross checked against the gazetteer list and discarded if found. Due to the large range and complexity of local or vernacular place names, these were not identified and would therefore be ignored by the gazetteer.

The full message structure from the Ushahidi reports included information on *message number, incident title, incident date, location, description, category, latitude* and *longitude*. For example:

"101, Road closure due to flooding, 9/01/2011 20:00, Esk-kilcoy Rd, Fast running water over the road at the bottom of the decent[sic] below lookout, Roads Affected, -27.060215, 152.553593".

Some of the message descriptions were very brief in the Ushahidi Crowdmap data. The content of these messages was further reduced when some of the pre-processing activities were undertaken including the removal of numbers, units, time, dates, hashtags, Twitter user accounts and URLs. If the number of characters of these messages were less than 30 characters, the data columns "*Incident Titl*e" and "*Description*"

were manually combined (see Table 6.1) to make the descriptions more comprehensive and meaningful.

Table 6.1 Example of the combination results of the *Incident title* and *Description* of the Ushahidi Crowdmap message fields

Incident title	Description	Combined message
Road Closed-Manly Rd between new Cleveland Rd and Castlerea St, Manly	Road closed due to flooding	Road Closed-Manly Rd between new Cleveland Rd and Castlerea St, Manly road closed due to flooding

In some cases, this combination did not provide a meaningful result and did not satisfy the above condition. Therefore, the "*Location*" column was also combined in these situations (see Table 6.2) to improve the message meaning. However, a small number of messages had to be discarded as they failed in any of the above operations.

Table 6.2 Example of the combination result of the *Incident title*, *Description* and *Location* of the Ushahidi Crowdmap message fields

Incident title	Description	Location	Combined message
Roads Affected	Not passable	Gailey Rd, St Lucia	Roads Affected Not passable Gailey Rd, St Lucia

The following example shows how the original Ushahidi Crowdmap message was processed after tokenisation, stemming, lemmatisation and stop-word removal before being used for training, testing and credibility detection.

Original Ushahidi Crowdmap message:

'Access to Stanthorpe town is severely restricted and all residents along Quart Pot Creek have been ordered to evacuate'.

Tokenized, stemmed and lemmatized message:

'access to Stanthorpe town be severely restrict and all resident along Quart Pot Creek

have be order to evacuate'.

Stop-word removed message:
'access stanthorpe town severely restrict resident along quart pot creek order evacuate'.

6.2.2. CSD relevance analysis

Previous research showed that CSD relevance analysis had been investigated using a variety of methods. This research provided a solid theoretical foundation to utilise Geographic Information Retrieval techniques to assess the CSD relevance. However, the suitability of each approach depends on the data and the task in hand. To test the geographic information retrieval a test was implemented using a Java framework, Lucene IR software and the GATE natural language processing software. The Ushahidi Crowdmap dataset of 2011 Australian floods was used as the testing dataset. From the Crowdmap reports, 200 random messages were selected for this analysis for easy manipulation and to better understand the system's behaviour. After the pre-possessing explained below, 182 reports remained for the thematic and geographic scope analysis.

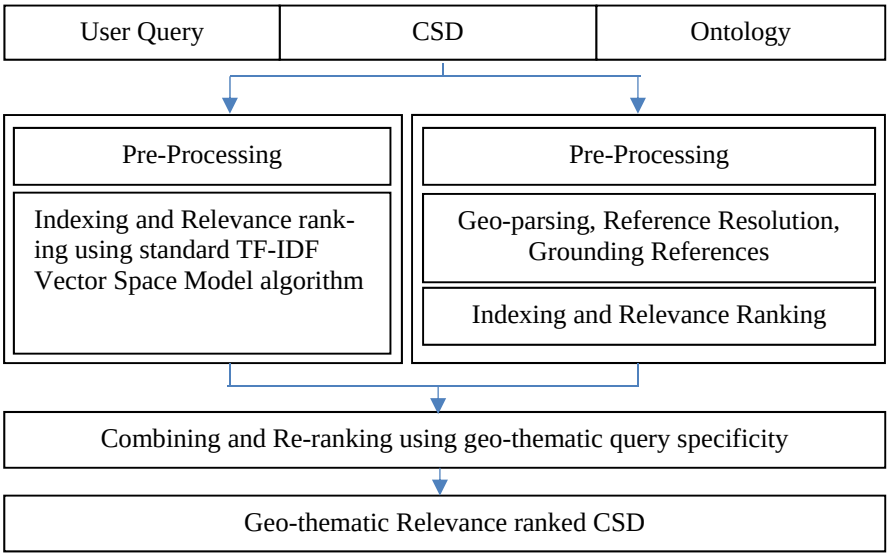


Figure 6.3 CSD relevance detection approach adapted from Zaila and Montesi's (2015) GIR architecture

Figure 6.3 depicts the overall CSD relevance analysis approach adopted in this research. The CSD was analysed based on two key scopes, the thematic scope and the geographic scope. These are explained in more detail below. In each instance, the pre-processing of CSD was carried out to prepare the unstructured raw dataset for further processing. This included actions such as duplicate removal, tokenizing, stop-word removing (i.e. removing common terms similar to prepositions etc., stemming and lemmatization (i.e. bringing the word to its base form such as 'flooding' to 'flood') and removing non-words such as numbers, white spaces etc. Other steps in each of the scope analyses are explained in the sections below.

6.2.2.1. Thematic scope analysis

The thematic scope relevance analysis used the Lucene open-source keyword matching information retrieval system. Lucene is based on standard Term Frequency – Inverse Document Frequency Vector Space Model (TF-IDF VSM) model explained in the section 3.3.2.1.

The thematic scope analysis was conducted using two Java programs which were constructed based on Lucene 6.0 API and its standard analyser. The first Java program was developed for indexing the dataset and the second program was used to perform the searching.

6.2.2.2. Geographic scope analysis

The geographic scope analysis differs from the thematic analysis. As identified in the section 3.3.2.1, the geographic scope analysis tasks including geo-parsing, reference resolution and grounding references can be performed using a natural language processing based gazetteer lookup approach. These tasks were carried out using the GATE software. The selected sample of the CSD dataset had to first undergo pre-processing in order to filter inappropriate content such as duplicates. However, tokenizing, stemming and lemmatizing pre-processing tasks which were used in the thematic scope analysis were not performed during the pre-processing of geographic scope analysis

as they were undertaken within the GATE software through a morphological analysis. The geo-parsing, reference resolution and grounding reference tasks that were performed are detailed below.

Generally, geo-parsing is expected to identify and tag toponyms and the geographic reference resolution to identify the best (i.e. most appropriate) toponym for the CSD report. The reference resolution is more challenging when ambiguities such as geo-geo or geo-non-geo occur. For both of the above tasks, it was required to identify possible toponyms in the message content by searching a reference list such as a gazetteer. The semantic geo-parsing (or semantic location extraction) process was explained in detail in the section 5.2.2.2.

The process of geographic reference resolution includes mapping toponyms and identifying the most appropriate toponym for the content. This is quite important when there are possible ambiguities. Mostly, these situations consist of relationships terms such as 'near' and contain contextually important information that can be resolved using an algorithm proposed by Martins et al. (2006) in association with context based semantic processing. Their approach was to split the queries into triples to form <what, relation, where> relations by concatenating the individual tokens. The relation terms were identified using a list of possible values such as 'near', 'between', 'crossing' and 'south of' etc.

Java Annotation Pattern Engine (JAPE) transducers were more useful in the process of geographic reference resolution to identify appropriate location according to the relationships described above. A number of JAPE rules were developed to resolve the ambiguities and to tag the messages with the most appropriate toponym with the help of QLDGazOnto ontological gazetteer. The final step of geographic scope resolution analysis was to geo-tag by assigning geographic coordinates to the CSD messages. Once the geographic reference resolution was completed all the messages were tagged

with a toponym using JAPE rules (see Figure 6.4). For the geo-tagging task, the toponyms were searched from the QLDGazOnto semantic gazetteer and then assigned the relevant geo-coordinates.

```
Phase: OntoMatching // phase name
Input: Lookup
Options: control = applet // control type
Rule: GeoTag   // rule name
({Lookup.class == Place}
) //search for place names in the semantic gazetteer
:place-->
:place.Mention = {class = :place.Lookup.class, inst = :place.Lookup.inst}
//match and tag with toponym
```

Figure 6.4 Example JAPE rule used for semantic geo-tagging

After completing the Geographic Scope Resolution (GSR) process, the next task was to calculate the geographic similarity measures.

The geographic similarities were calculated using the equation (5) below (also listed in Section 3.3.2.1) by considering the geographic scope of the query and the geographic scope of each CSD report using the QLDGazOnto ontology information.

The value for the variable K in equation (5) below was identified as 0.8 after manual testing.

$$Sim_G(S_q, S_m) = K \times \{Insd(S_q, S_m) + Proxm(S_q, S_m)\} + (1 - K) \times Sib(S_q, S_m) \quad \ldots (5)$$

Example for geographic similarity calculation:

Query (q): "Road closed due to flood in Toowoomba"

Processed Message (m): "resident dalby ask evacuate home likely inundate by western downs disaster coordination centre dalby"

Scope of the S_q = Toowoomba

Scope of the Message S_m= Dalby

$Insd(S_q, S_m) = 0$ as Dalby is not Inside Toowoomba

$$Dist(S_q, S_m) \approx 80\ km$$

$$Diagonal(S_q) = 220\ km$$

By substituting the above values in equation (6)

$$Proxm(S_q, S_m) = \frac{1}{1 + \frac{80}{220}} = 0.74$$

$Sib(S_q, S_m) = 1$ as Dalby and Toowoomba are siblings in the ontology set.

Therefore, by substituting above values in the equation (5);

$$Sim_G(S_q, S_m) = 0.8 \times \{0 + 0.74\} + (1 - 0.8) \times 1 = 0.79$$

Finally, the geographic and thematic relevance lists were merged to create the final geo-thematic relevance ranked list. The final ranked list was calculated using the weighted sum method equation (9) and (10) (see Section 3.3.2.2) proposed by Yu and Cai (2007) which considered the thematic and geographic specificities of the query.

Example for geographic specificity calculation:

Area of the geographic scope of query (Convex hull): $Area(G_q) = 16535\text{km}^2$

Area of the coverage of all messages in the dataset: $Area(G_M) = 996865\ \text{km}^2$

Substituting these values in the equation (8) (see Section 3.3.2.2) the geographic specificity Spc_G(i.e. how specific the geographic scope of the query is) of the above query (q) was calculated as 1.78.

Example for thematic specificity calculation:

The Conceptual Term Matrix (CTM) for each term of the query (q) was calculated using the WordNet ontology as below.

Step 1: Extraction of CTM for terms in the query

Conceptual values for each of the query terms were extracted (Table 6.3) using information of the WordNet ontology.

Table 6.3 Extracted CTM for each term of the query (q)

Query Term	**Parts of Speech**	**# Senses C1**	**# Synonyms C2**	**# Level C3**	**# Children C4**
Road	R1 (Noun)	2	1	2	20
	R2 (Verb)	0	0	0	0
	R3 (Adjective)	0	0	0	0
Closed	R1 (Noun)	0	0	0	0
	R2 (Verb)	17	8	15	12
	R3 (Adjective)	9	15	0	0
Flood	R1 (Noun)	6	3	6	4
	R2 (Verb)	4	3	4	3
	R3 (Adjective)	0	0	0	0

Step 2: Weighing

The purpose of the weighting was to convert the extracted CTM integer values into weights in the range of 0 and 1. Twelve membership functions were developed using the Parts of Speech statistics (POS) (Min, Max, AVG) (Table 6.4) of the WordNet ontology and general weighting functions (Figure 6.5 (a) and (b)).

$$f(x) = \begin{cases} 0 & ,x \geq Max \\ 0.5 & ,x = Avg \\ 1 & ,x = Min \\ f(x-\Delta x) - \frac{0.5 * \Delta x}{Avg - Min} & ,Min < x < Avg \\ f(x-\Delta x) - \frac{0.5 * \Delta x}{Max - Avg} & ,Max > x > Avg \end{cases} \qquad f(x) = \begin{cases} 0 & ,x = Min \\ 0.5 & ,x = Avg \\ 1 & ,x \geq Max \\ f(x-\Delta x) + \frac{0.5 * \Delta x}{Avg - Min} & ,Min < x < Avg \\ f(x-\Delta x) + \frac{0.5 * \Delta x}{Max - Avg} & ,Max > x > Avg \end{cases}$$

(a) (b)

Figure 6.5 General weighting function for (a) Nouns, verbs, adjectives – senses, synonyms, and children (b) Nouns and verbs levels (Sakre et al. 2009)

The weighted CTM matrix (see Table 6.5 for example) for each term of the query was computed using the weighting functions formulated using the general weighting functions listed in the Figure 6.5 and the POS statistics (Table 6.4) of the WordNet ontology.

Table 6.4 Min, Max, AVG statistics for Parts of Speech of WordNet (Sakre et al. 2009)

Part of Speech	**Conceptual Type**	**[MIN, MAX]**	**AVG**
Noun	Senses	[1,7]	2.76
	Synonyms	[0,7]	1.58
	Levels	[1,16]	7.5
	Children	[0,77]	31
Verb	Senses	[1,7]	3.54
	Synonyms	[0,7]	1.96
	Levels	[1,8]	3.64
	Children	[0,29]	10.8
Adjective/Adverb	Senses	[1,7]	2.79
	Synonyms	[0,7]	1.7
	Levels	N/A	N/A
	Children	N/A	N/A

Table 6.5 Example weighted CTM of the term 'Flood'

	C_1 (Senses)	**C_2 (Synonyms)**	**C_2 (Level)**	**C_2 (Children)**
R_1 (Noun)	0.100	0.350	0.375	0.875
R_2 (Verb)	0.450	0.400	0.500	0.850
R_3 (Adjective)	0	0	0	0

Step 3: Fusing

After computing weighted CTM for each query term the weights were then combined using the row and column combining equations listed below and the weight fusing matrix (Figure 6.6) where the value set to 0.5.

Firstly, each column of the weighted CTM were fused using the equation;

$$C_n = \frac{\sum_m V_{mn} \times W_{mn}}{\sum_m W_{mn}}$$

Where C_n is the concept number and *m*, *n* are the rows and columns respectively.

This resulted in a row vector *R*= {0.1833, 0.2500, 0.2917, 0.5750} which was then fused using the row averaging equation below to calculate the final CTM for each term of the query;

$$CTM_q = \frac{\sum_n C_n \times W_n}{\sum_n W_n}$$

Therefore, the final conceptual weighted CTM value for the term 'Flood' was calculated as 0.3251 and 0.1709, 0.1209 for the terms 'Road' and 'Closed' respectively. The term 'Toowoomba' was discarded as there was no results for the any concepts within the ontology.

	C_1	C_2	C_3	C_4
R_1	0.100	0.350	0.375	0.875
R_2	0.450	0.400	0.500	0.850
R_3	0	0	0	0

Conceptual Term Matrix

x

C_1	C_2	C_3	C_4
0.5	0.5	0.5	0.5
0.5	0.5	0.5	0.5
0.5	0.5	0.5	0.5

Weight Fusing Matrix

Figure 6.6 Fusing CTM weights using the Weight Fusing Matrix

Finally, the thematic specificity Spc_T (i.e. how specific the thematic scope of the query) of the above query (q) was calculated as 0.541 using the equation (7) in the section 3.3.2.2.

6.3. Results and discussion

6.3.1. Results and discussion of the CSD credibility analysis

The CSD credibility analysis was performed using the Ushahidi Crowdmap dataset and a naïve Bayesian Network model trained using a selected sample of the dataset. The results of the analysis are explained below.

6.3.1.1. Results of the initial training and testing using different sized training data

The system developed for CSD credibility detection was initially trained using two different sized training data sets to assess any variations in the outcomes based on the size of the training data set used. The first training data set from the Ushahidi Crowdmap data consisted of 35 messages of which there were 25 credible messages and 10 messages identified as unreliable. The second training set was a larger training

sample and consisted of 77 messages with 53 credible messages and 24 messages identified as unreliable.

A test dataset of 33 messages was then tested using both the smaller and larger training data sets to train the system under both forced and unforced conditions. The test dataset was also manually pre-classified to identify credible messages and unreliable messages in order to confirm the accuracy and performance during the testing. Table 6.6 shows example messages of correctly and misclassified results. A possible reason for misclassifications may be the appearance of more credible types of terms in unreliable reports and the vice versa. Tables 6.7 to 6.10 show the classification results for the four test environments. Test 1 utilised the smaller training data set (35 messages) to train the system and then used 33 test messages under unforced training conditions.

Table 6.6 Examples of correctly and incorrectly classified messages.

Correctly Classified		**Misclassified**	
Credible	**Unreliable**	**Actually Credible – Classified Unreliable**	**Actually Unreliable –Classified Credible**
Gold Coast - Springbrook Road is closed between Belmont Park Dv and Pine Creek Rd due to flooding. Also closed at the Austineville Rd turn off.	Fast running water over the road at the bottom of the decent below lookout	The Ipswich Show Grounds has been declared the evacuation centre. For more information go to the Queensland Police Service Facebook page.	Lots of road closures due to flooding, our back fence was partially pulled down by the flooded Coonowrin creek overnight.
An emergency evacuation centre was set up at the Gatton Shire Hall in North Street Gatton. Residents who were evacuated from their homes in Grantham, Helidon and the low-lying parts of Gatton.	Evacuation - Toogoolawah from river risingg My sis has been told to evacuate by the SES as the water has reached her fence	Flash flooding at Goodna. Picture of Leslie Park off Bertha Street Goodna where the floodwater is rising.	I live in Dubai and own a house in Ellena Street Paddington BNE, can you please advise if this house was affected?
Queensland Police Service: The D'Aguilar Highway at Kilcoy is now closed in both directions. Police remind motorists not to attempt to cross flooded roads or causeways. Police remind motorists not to attempt to cross flooded roads or causeways.	Thanks local baker keep spirit keep bake provide bread other side town picture nothing	There are two evacuation centres in Dalby. South State School, Bunya and Owen Steets, in Dalby. You can also evacuate to the Dalby Agriculture College in Nicholson Street.	Creek (now river) still a long way over and flowing fast - no traffic passable.

Test 2 utilised the same training data set (35 messages) and the 33 test messages but this time under forced training conditions. Test 3 utilised the larger training data set (77 messages) to train the system and then the 33 test messages under unforced training

conditions. Finally, Test 4 utilised the larger training data set (77 messages) with the 33 test messages but again this time under forced training conditions.

The terms True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) were used to compare the results of the classification. A True Positive result correctly predicts a “Credible” outcome when it is “Credible”, a True Negative result correctly predicts an “Unreliable” outcome when it is “Unreliable”, a False Positive result falsely predicts a “Credible” outcome when it should be “Unreliable”, and finally, a False Negative result falsely predicts an “Unreliable” outcome when it should be “Credible”.

Table 6.7 Test 1 - Unforced training using the small training sample (35 messages) and 33 test messages.

	Classified as Credible	**Classified as Unreliable**	**Total**
Actually Credible	24 (TP)	1 (FN)	25
Actually Unreliable	7 (FP)	1 (TN)	8
Total	31	2	33

Table 6.7 results indicates that the system correctly classified 24 out of 25 credible messages during unforced training, but only one out of the eight messages identified as unreliable was correctly classified. This outcome resulted in a high number of False Positives for the unforced training which indicated that further training was required.

When the system utilised the same training data set but ran under forced training conditions the results as expected varied (Table 6.8). Of the 25 credible messages 23 messages were correctly classified and only two messages incorrectly classified. These results only varied slightly from the unforced training outcomes in regard to detecting credible messages correctly. However, there was a significant improvement in the correct detection of unreliable messages with all messages being correctly classified during this test. Overall, the results were considered acceptable with a high classification

accuracy for both the credible and unreliable messages classification and hence validated the forced training conditions.

Table 6.8 Test 2 - Forced training using small training sample (35 messages) and 33 test messages.

	Classified as Credible	**Classified as Unreliable**	**Total**
Actually Credible	23 (TP)	2 (FN)	25
Actually Unreliable	0 (FP)	8 (TN)	8
Total	23	10	33

Next, the size of the training sample was increased from 35 messages to 77 messages and then the unforced and forced training was repeated on the same test data set. The results of the unforced training are shown in Table 6.9 and identify that for the credible message classification, 21 out of 25 messages were correctly classified which was a small decrease in accuracy compared to the previous result (Table 6.7). However, the classification accuracy of unreliable messages improved from one correctly classified message to five correctly classified messages out of the eight to be classified.

Table 6.9 Testing 3 – Unforced training using the larger training sample (77 messages) and 33 test messages.

	Classified as Credible	**Classified as Unreliable**	**Total**
Actually Credible	21 (TP)	4 (FN)	25
Actually Unreliable	3 (FP)	5 (TN)	8
Total	24	9	33

Finally, Table 6.10 shows the results of the classification using the larger training data set under forced training conditions. The results of the testing are identical to the forced training using the smaller training data set with 23 out of 25 credible messages correctly classified and all eight unreliable messages were also correctly classified. This

indicated that the forced training conditions were consistent and were not impacted by the changed training sample size.

Table 6.10 Test 4 - Forced training using the larger training sample (77 messages) and 33 test messages.

	Classified as Credible	**Classified as Unreliable**	**Total**
Actually Credible	23 (TP)	2 (FN)	25
Actually Unreliable	0 (FP)	8 (TN)	8
Total	23	10	33

A number of measures such as accuracy, precision, sensitivity and the F1 score provided an indication of each classification's effectiveness. The accuracy, which is the ratio of correctly predicted observations, was calculated by the formula (TP+TN)/(TP+TN+FP+FN). The precision or Positive Predictive Value (PPV) is the ratio of correct positive observations and was calculated by TP/(TP + FP). The F1 score (F1) is used to measure classification performance using the weighted recall and precision, where the recall is the percentage of relevant instances that are retrieved and was calculated using formula 2*TP / (2*TP + FP + FN). The sensitivity or True Positive Rate (TPR) was calculated by TP / (TP + FN).

The classification quality for the four tests are summarised in Table 6.11. The accuracy and precision was higher for the forced training outcomes for both training sample sizes and indicates the impact of the forced training. It can also be seen that the classification accuracy and precision increased slightly for the unforced training outcomes when the larger training sample size was utilised. However, the precision and accuracy outcomes for the forced training were similar and indicate that there may be a lesser dependency on the size of the training data set when forced training is utilised. The F1-Score did not change with the sample size but the measures indicate that the forced training again performed better than the unforced training scenarios. Finally, the classification sensitivity remained constant for the forced training for both training sample sizes but dropped slightly with the larger training sample size for the unforced training

test outcomes.

Table 6.11 Quality of the CSD Classification

Test Scenario	Accuracy	Precision	F1-Score	Sensitivity
Test – 1 Unforced Using the small training sample (35 messages) and 33 test messages	76	77	86	96
Test -2 Forced Using the larger training sample (77 messages) and 33 test messages	94	100	96	92
Test – 3 Unforced Using the small training sample (35 messages) and 33 test messages	79	88	86	84
Test – 4 Forced Using the larger training sample (77 messages) and 33 test messages	94	100	96	92

6.3.1.2. Results of the full Ushahidi Crowdmap data CSD credibility analysis

After the training testing of the system was completed to an acceptable classification quality, the full Ushahidi Crowdmap sample of remaining 433 messages was analysed for credibility. As the Figure 6.7 (a) indicates, 54% (234 out of 433) of the messages were identified as credible using an unforced training classification. However, when the system was run under forced conditions, 77% (334 out of 433) of the messages were identified as credible (Figure 6.7 (b)). This was a more confident value than the previous result as the accuracy and precision of the credibility detection was higher.

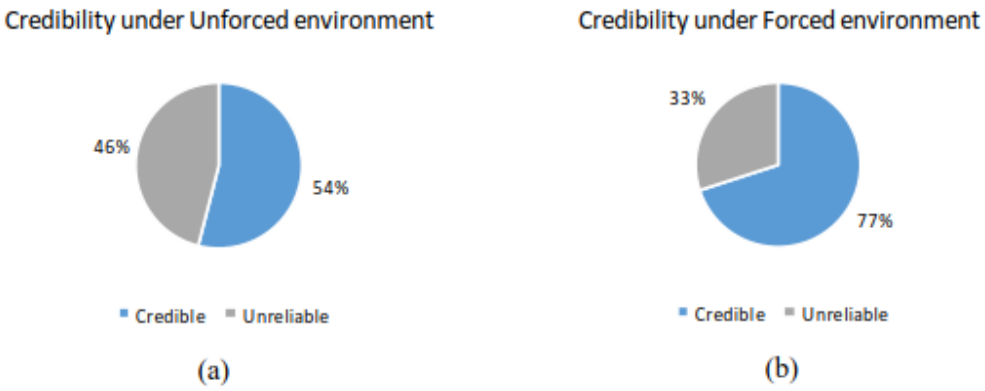


Figure 6.7 Assessed Credibility of 2011 Australian floods Ushahidi Crowdmap data

6.3.2. Results and discussion of the CSD relevance analysis

In the CSD relevance analysis, 182 Ushahidi Crowdmap messages were selected for the geo-thematic relevance analysis after the initial pre-processing. There are various quality metrics to test the performance and quality of the results from this analysis. Measures such as recall and precision are popular measures in these classification systems. However, precision is often regarded as a more important measure than recall in rank based IR systems if the user does not intended to retrieve all relevant records (Inkpen 2007). In relation to the information retrieval, precision refers to the fraction of retrieved documents that are relevant to the query, whilst recall is a representation of the fraction of documents that are relevant to the query that are successfully retrieved (https://wikipedia.org). The Precision can be calculated by:

$$Precision = \frac{|\{relevant\ documents\} \cap \{retrieved\ documents\}|}{|\{retrieved\ documents\}|}$$

Other measures including Average Precision (AP), Mean Average Precision (MAP) and Precision at K are the measures often used in modern web based information retrieval systems as the recall may not represent a meaningful measure where thousands of relevant documents are present in such systems. Average Precision refers to the precision averaged across all values of recall between 0 and 1.

The Average Precision AP can be calculated by:

$$AP = \frac{\sum_{k=1}^{n}(P(k) \times (R(k))}{number\ of\ relevant\ documents}$$

$$\text{and} \qquad MAP = \frac{\sum_{q=1}^{Q} AP(q)}{Q}$$

Where k is the rank in the retrieved message list and $P(k)$ is the precision at cut off k in the list and $R(k)$ is an indicator function which provides 1 if the message at position k is relevant and 0 otherwise. The AP for a query q refers to the average precision for each relevant message retrieved and finally, the MAP is the mean average precision of all Q queries. The measure Precision at K (P@K) reports the fraction of messages ranked in the top k results marked as relevant.

Thematic scope analysis results

The quality of thematic scope analysis used the Lucene benchmark quality assessment package. In this analysis two configuration files were constructed, one containing the queries and the other containing the manually classified test reference collection. The test reference collection consisted of relevant and non-relevant sets of documents for each query. These configuration files were used for the quality analysis along with the indexed file of CSD messages.

Table 6.12 shows the performance test results of the thematic scope analysis using the Lucene software. This research selected the AP, MAP and P@K metrics which are accepted information retrieval quality benchmark metrics (Agichtein et al. 2006). The table 6.12 also shows the number of hits (i.e. the number of messages identified relevant to the each query) along with the average precision, precision at level 5 (P@5) and precision at level 10 (P@10) of the analysis. According to the Lucene benchmark quality package results, the average precision of the relevance of the message retrieval to the queries were generally above or close to 0.6 which indicates the system performed well. The P@5 was generally above 0.4 and the minimum value was 0.3 which

means the system better identified relevant documents at the top levels. The MAP of the quality assessment was calculated as 0.792 which is a good indication of systems performance for relevance assessment as the value 1 indicates the best performance.

Table 6.12 Quality assessment results of thematic scope analysis

No.	Query	# hits	Average Precision	P@5	P@10
1	Road closed due to flood in Toowoomba	120	0.655	0.600	0.300
2	Highway closed	69	0.897	0.800	0.600
3	Evacuation centre open	21	0.595	0.400	0.300
4	Heavy rainfall Toowoomba	45	0.911	0.800	0.600
5	Flash flooding Toowoomba	55	0.903	0.800	0.500

Results of the geographic scope analysis

The grounding references of the geographic relevance assessment was performed using a Java program based on Google geo-coding API. The location availability of CSD messages were close to 90% (i.e. 163 out of 182 messages) after the Geographic Scope Resolution (GSR) process. The geographic similarities were calculated using equation (5) (see Section 3.3.2.1) with the value of K set to 0.8. The geographic scope of the queries was selected as Toowoomba local government area which was a polygon feature. In the case of a polygon feature, it can use the Minimum Bounding Rectangle (MBR) or convex hull as the feature representing the geographic scope. Both the MBR and convex hull options (Figure 6.8) were tested in calculating locations inside and within proximity using equation (5) The results identified that if the MBR was used instead of the convex hull, there was a 46% increase in area and therefore the selection of 12 additional points which did not belong to the Toowoomba local government area.

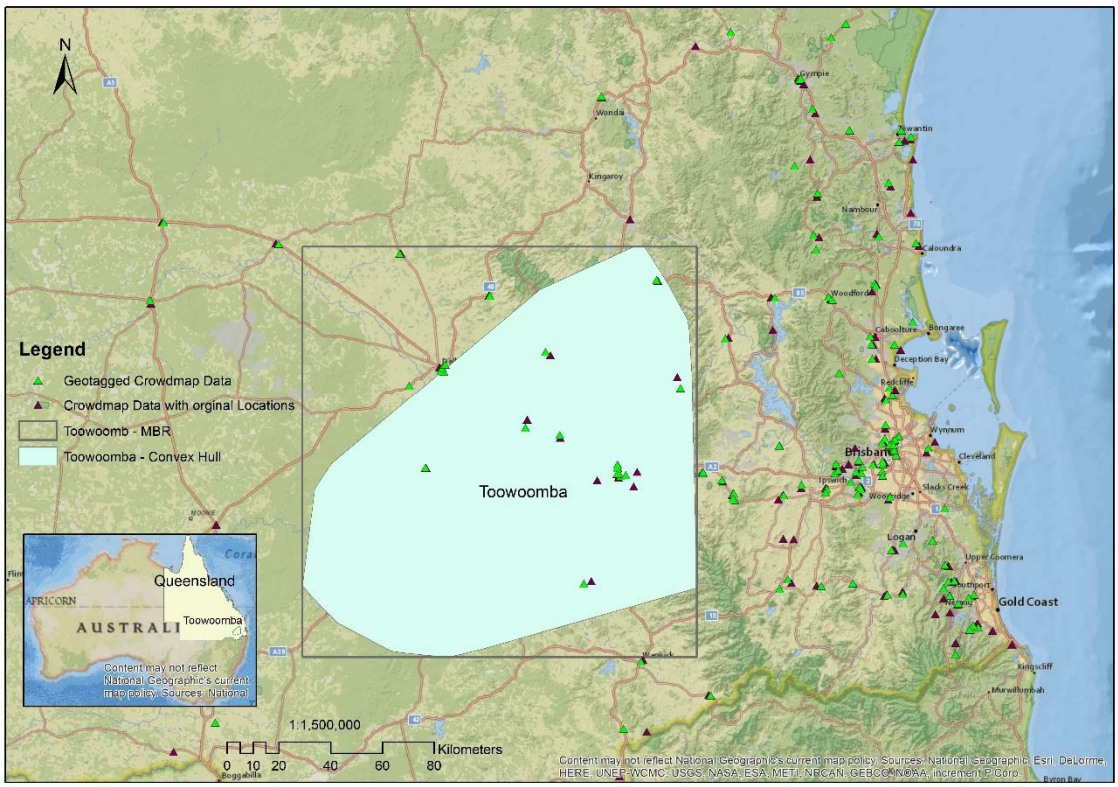


Figure 6.8 Crowdmap data and Toowoomba local government area using MBR and Convex-hull

Results of the final geo-thematic relevance ranking

The thematic and geographic specificities of the queries were calculated using equation (7) and (8) (see Section 3.3.2.2) using the georeferenced queries and geocoded CSD messages. The geographic specificity and thematic specificity were 0.44 and 0.67 respectively. The values indicate that the queries used were less thematically specific and more geographically specific as value 1 indicates the highest specificity. The final rankings were performed using equations (9) and (10) (see Section 3.3.2.2). The results were then compared with a human ranked list to see the system's ability to analyse the relevance compared to a human. Spearman's rho, which is a commonly used statistical test to compare agreement between two rankings where the value 1 indicates perfect match and -1 indicates complete inverse ranking (Yu & Cai 2007), was calculated. The Spearman's rho was 0.62 which indicates the two lists closely agree to each other and confirms the validity of the approach for CSD relevance assessments.

6.4. Conclusions of the CSD credibility and relevance analysis

Conclusions of the CSD credibility analysis

The CSD message credibility detection is a challenging task due to the high degree of variability of the data, the lack of a consistent data structure, the variability of the data providers and the limited metadata available. This study identified that Bayesian spam email detection approaches can be applied successfully to the challenge of classifying the credibility of CSD. However, the training approaches and the size of the training data set can influence the quality and performance of the training outcomes.

Due to the variability of the data, it is recommended that forced training is undertaken to achieve the highest accuracy and performance. In particular, the forced training provided a higher level of confidence in eliminating the number of False Positive (FP) outcomes which were the incorrect classification of messages. The size of the training data set was found to be less critical when a forced training approach was utilised with the results of the classification outcomes being similar for both the smaller and larger training data sets. However, if the system training was unforced, a larger training data set is recommended.

Conclusions of the CSD relevance analysis

The known spatial data quality is a distinct advantage for the confidence in quality outputs of any spatial data dependent project. This research tested a relevant Crowdsourced Data retrieval method for disaster management activities. In disasters such as floods, speedy identification of relevant and credible spatial information is important and is required to support victims and save lives.

This research analysed CSD based on thematic and geographic relevance using GIR techniques. For the thematic relevance assessments, it used the Term Frequency and Inversed Document Frequency Vector Space Model (TF-IDF VSM) based on the pop-

ular Lucene full featured text search engine library. CSD in general is curated by different people with different experiences and different knowledge levels using heterogeneous devices. In the Crowdmap content, it can be seen that people communicated similar incidents in various ways i.e. the intended meaning of road closures reported such as 'road closed, no go zone, water over the road, road under water, road flooded, road impassable, highway cut, water across road, etc.'. Identifying similar stories using a keyword based search is challenging. This research suggests the use of semantic based thematic relevance assessment for highly unstructured and heterogeneous data such as CSD.

The research used the Natural Language Processing (NLP) based gazetteer lookup for Geographic Scope Resolution (GSR) in the thematic relevance assessment. It applied stop-word and common-word filters to minimise the effect of frequently occurring terms. However, this research identified drawbacks in the application of filters i.e. removal of terms such as 'Can' in toponyms such as 'Tin Can Bay'. It is recommended therefore the need to further understand similar effects and to identify precautions to prevent the removal of important terms. For geo-tagging purposes, the research used the Google geo-coding service with the support of the local semantic gazetteer (QLD-GazOnto) for ambiguity resolution. This was very useful to resolve geo-geo and geo non-geo ambiguities (e.g. Killarney in Ireland and Killarney in Australia, John Krebs is a personal name and there is a bridge called John Krebs Bridge in Murgon, Queensland, Australia). It is recommended that it is important to use local gazetteers for ambiguity resolutions in the GSR process.

The research considered query specificity for final geo-thematic relevance joining and ranking. This was useful in identifying contextually more relevant CSD messages for flood disaster managers and other stakeholders. It is suggested that further work be completed to test and compare the performance and usefulness of other available geo-thematic relevance combination approaches.

It is also noted that the GIR field is a fast-growing research area and new techniques are introduced quickly. This research suggests the need to test innovative and more stable approaches used in GIR to validate the applicability of similar approaches for CSD relevance studies.

6.5. Chapter summary

This chapter explored CSD credibility and relevance assessment methods along with the results of the analysis and discussed the implications of the results. The first part of the chapter discussed the CSD credibility and relevance assessments. The CSD credibility detection used a naïve Bayesian Network based model which had been utilised successfully for spam email detection. This chapter detailed the model design, system training and testing procedures followed for CSD credibility assessment. The CSD relevance detection methods were then explained and the use of GIR approaches along with Natural Language Processing (NLP) techniques for CSD relevance assessment were discussed. Finally, the results and their importance with respect to CSD credibility and relevance were discussed. The next chapter will detail the potential implications of the outcomes achieved in chapters five and six and look to identify possible integration and system automation.

Chapter 7: **Discussion**

7.1. Introduction

Chapters five and six detailed location quality analysis, semantic location extraction, credibility and relevance analysis of CSD. The purpose of this chapter is to discuss the findings of the research, in particular, the opportunities to integrate the research outcomes with authoritative data and to identify possible system automation methods. The CSD quality indicators, assessment and improvement methods were explored in previous chapters. However, an automated mechanism for CSD quality management may also be required to make the data useful for stakeholders such as disaster first responders. Therefore, in this chapter a framework is presented to automate the CSD quality management and to enable the information to be effectively integrated with authoritative data. SDIs are also rapidly evolving with the technological changes happening in this fast-moving world. These developments include changes in the technical, architectural, policy and other dimensions of SDIs which are critical for maintaining the highest data quality, improved access and the smooth functioning of SDIs. Due to this dynamic environment, SDI-CSD coupling will be challenging. This chapter explores the opportunities, challenges and issues pertaining to authoritative data and CSD integration.

7.2. CSD location availability and its quality

CSD location determination can be problematic due to its limited availability or the quality of the available location data. Privacy issues, default location settings and the reluctance for sharing the location information in smartphones are some of the possible reasons for the lack of availability of location information. People are wary of enabling location on their smartphones with their CSD feeds due to the privacy issues. Although they may be willing to provide the location, users may opt to use the network location which is not as accurate as GPS. GPS sensors in the modern smart devices are improving dramatically and most smartphones provide street level accuracy that can easily be

used for navigation applications. However, they are still not achieving surveying grade accuracy i.e. centimetre level or sub-meter accuracies and the maximum accuracy we can expect from the CSD produced by citizens may be at the street level accuracy. This research found that the missing location data could often be determined through street address or other textual locations (i.e. toponyms) and a method was presented to extract locations within the CSD textual descriptions. However, the location enabled CSD produced in this manner can have limitations if it is to be used in the applications which require higher locational accuracies.

The results of chapter five provided a clearer understanding of the quality of the CSD location by comparing three different datasets; Google Maps (used by the crowd-mapping platform), OpenStreetMap (OSM) and the QDNRM roads data. The nature of the three datasets were different with OSM being free and open, Google Maps being a commercially developed web mapping service, and QDNRM being a government data set available for restricted use. In general, OSM and Google maps are conceptually similar to each other but different to QDNRM data. QDNRM's accuracy is higher and OSM and Google Maps are at a lower spatial accuracy but often more current than QDNRM.

The CSD location quality analysis identified that the QDNRM dataset had better location agreement with CSD locations than the OSM data. It also confirmed that the Google Maps data for the study area was spatially more accurate when compared to the OSM data. However, in terms of information currency, official government databases often lag behind the more dynamic and flexible open data sources. The results of the combined analysis supported this proposition. It was also identified that data integration needs to be undertaken with care as it can lead to the introduction of duplicates and errors such as geo-geo ambiguities. The issue of geo-geo ambiguities can be resolved by using semantic local gazetteers and so the combined use of CSD and authoritative data along with semantic gazetteers may be advantageous.

This research initially analysed CSD from two different sources i.e. feeds from the 2011 Australian floods' Twitter and the Ushahidi Crowdmap data. It was found that the Twitter data was unstructured in comparison to the other spatial data, while the Crowdmap data was found to be partially structured. Twitter users report information using the 140 characters allowed for a Tweet. On the other hand, Ushahidi users follow a number of steps to submit a report including selecting the report category, location, description etc. Although this tends to create more structured data, Ushahidi can also accept reports from Twitter, emails, SMS and other sources.

During the research, it was identified that the location enabled feeds available from Twitter data were very low (i.e. approximately 1%) whilst the Crowdmap data location availability was approximately 60%. The use of a semantic location extraction method was able to determine a location of up to 25% of the Twitter data sample. However, this figure can vary dramatically depending on the type of Twitter data, users and context and is considered low from a spatial data capture point of view. This finding highlighted a number of limitations when considering Twitter as a spatial data source and therefore, the remainder of the quality assessment steps (i.e. credibility and relevance) were conducted using only the Crowdmap data. It also highlighted the importance of educating the people to use Twitter in a meaningful manner i.e. enabling the location sensors on their devices and create complete reports especially if they are seeking to contribute to disaster event reporting.

7.3. Dealing with CSD credibility and relevance

Prior to using CSD for any application, the credibility and relevance should be assessed to provide a higher level of confidence for the end user. This research assessed the credibility using a spam email detection approach and the relevance was assessed by using a GIR technique. Both approaches have their origins in the IT domain. The credibility and relevance assessment results showed that, although the location quality of

CSD remains comparatively low, it can be considered as a credible and relevant product. However, it is still not clear the order in which each of these elements should be assessed. For example, is it better to determine the location quality or the credibility first or can better location also compliment the credibility?

CSD credibility

Credibility, in general, denotes trust or believability whilst relevance indicates the relationship of the information to a particular purpose. Both factors are important in deciding whether a dataset can be used confidently for a chosen purpose. As CSD is often communicated during an event, such as a natural disaster, the credibility and relevance is closely linked to the situation and may provide a degree of confidence in respect to its use. Information credibility is a critical consideration in the modern connected world. People are now very connected with others through social media or online services for business, cultural, administrative, social or other purposes. The trust of shared information may be the most critical factor determining the strength of the bond, regardless of the purpose or connection mode. Humans can easily decide the degree of credibility or relevance of related information but, in the semantic web this decision is expected to be made by the software. Integrated or automated mechanisms of credibility and relevance assessment will be important in extracting relevant information in a timely manner.

Information credibility is often required to be assessed across application areas such as communication, social media, health data, academic information, finance, business and management. This research explored an approach which had been previously utilised for identifying the credibility of emails, more commonly termed as spam email detection. Email credibility is assessed on a number of email components including the header information, credibility of the sender, email title and the body text. However, in the case of CSD, there is often limited information on the volunteers who submitted the information, so assessing the credibility based on the sender was not able to be investigated further.

Most of the Ushahidi Crowdmap reports were generated from the Crowdmap interface whilst other sources originated from emails and SMS messages. From the Crowdmap reports which were sent during 9th to 15th of January 2011, approximately 74% of them were tagged as being initially "verified" by the Crowdmap administrators and therefore, may be considered as having some level of credibility. However, approximately 99% of the reports were also finally approved by the administrators to be published in the Crowdmap. From this data, it is difficult to ascertain if the verification assessment or publishing by the administrators was anything more than a simple check to ensure the content supplied by the volunteer was complete and appropriate to the disaster event.

The initial training of the naïve Bayesian Network based credibility detection system developed in this research was challenging due to a large number of false positives during the initial testing. It was suspected that this was largely due to the heterogeneous nature of the CSD. The training of the system was then undertaken using a forceful (or manual) training process with modified training data which resulted in more successful results. This suggested that the credibility of unstructured and heterogeneous data such as CSD can be classified using modified training samples and a rigorously trained naïve Bayesian Network based model. However, this needs to be tested for generic applications and context independent CSD.

CSD relevance

Relevance or the fitness for purpose was another important factor identified by this research for determining the CSD quality. CSD is often generated during a natural disaster event such as a flood. So logically, a set of CSD which is based on a flood event should contain information more relevant to flood disasters and so it should be expected that certain terms will have higher relevance if the data is assessed based on a similar disaster profile. However, CSD is largely heterogeneous and understanding volunteer responses and terminology cannot be easily determined. In other words, the relevance is highly dependent on personal circumstances and individual responses.

Therefore, this research identified it was appropriate to rank the data based on relevance rather than splitting the content further.

Geographic context analysis is popular in modern information retrieval systems as many of the search queries include place names or other forms of location. In Geographic Information Retrieval relevance is often considered in two assessment dimensions, namely thematic relevance and geographic relevance. This research used these dimensions to assess the relevance of the CSD however, the parameters were refined to identify CSD relevance specifically for post-flood disaster management. Another important relevance dimension emerging in the IR field is the time or temporal dimension. CSD is generally more current than the authoritative data, however, CSD related to a particular event may vary from a future event and therefore generate new and different content. In such cases, it would be important to assess the temporal relevance of the CSD along with the thematic and geographic relevancies.

7.4. Quality assessment of non-textual CSD

CSD can be available in different forms including texts, photos, and maps and sometimes may be a combination of all. These forms of CSD may be related to different events such as disasters, social or political events. Twitter and other microblogging platforms often generate textual CSD however, they may also include other content such as images. Flickr and Panoramio are popular photo sharing platforms which are used to share travel experiences by sharing geo-tagged images. Crowdsourced maps including OpenStreetMaps or Ushahidi Crowdmaps often contain GPS points or GPS tracks captured by citizens during their travel or in specific events such as disasters. Social media platforms such as Facebook often include mixed content of images and text.

As the content varies, the complexity also varies from a spatial data quality point of view. Therefore, different approaches are required to assess the quality including the

assessment of credibility and relevance of CSD content. Although, the pictures are generally self-descriptive, the usefulness or value may be varying with the application considered. For example, in the case of disaster management, an image of a road sign or water depth indicator may be critical. However, an image of muddy water may be less informative. Metadata attached to the images, such as coordinates of the captured location, information about the person who captured the data, date, time and information about the event may be useful for assessing the credibility. Current smartphones with cameras can add this data which are useful for detecting the relevance and or credibility of photo/image based CSD. However, quality assessment of photo/ image based CSD has not been extensively investigated. This research analysed text based CSD of a flood disaster event but this approach may not be appropriate to assess other forms of CSD.

7.5. Towards system automation

An automated system of CSD quality assessment will be important for timely decision making during disaster management. The selected application area for testing the identified theories in this research was post-flood disaster management. Similar to other disaster management, post-flood disaster management is time critical and dependent on spatial data. Therefore, the quality of spatial data used, the time taken for assessing the quality of data and the currency of the spatial data selected are critical for effective decision-making. Figure 7.1 depicts an automated system architecture for CSD quality assessment which consists of four important stages including (1) input and pre-processing, (2) location testing, (3) location extraction and (4) quality assessment and output. All these stages have been individually tested in this research and were explained in chapters five and six. The software, tools and programs were developed using the Java programming language and therefore, the proposed architecture could be implemented using Java APIs and other related tools.

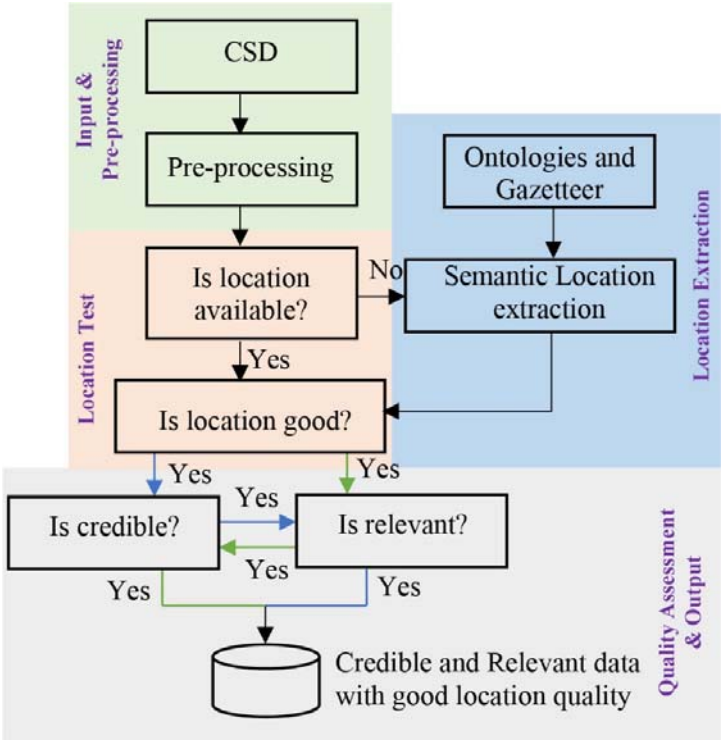


Figure 7.1 Automated CSD quality assessment architecture

Input and pre-processing

In the proposed architecture, the input CSD could originate from many forms of crowd generated data such as social media data (i.e. Twitter, Facebook etc.) or crowd-mapping data such as Ushahidi posts which contain textual data. Generally, CSD in its raw format is unstructured and often contains many abbreviations or slang terms. Therefore, it requires pre-processing steps such as Twitter text normalisation. Pre-processing may also require removal of duplicate information, tokenizing, stop-word removal, stemming and lemmatization as explained in sections 3.3.1.1, 6.2.1.1 and Figure 6.2.

Location test

The location quality assessment methods conducted in this research were explained in chapter five. The location test included two components namely, a location availability check and location quality test. The explicit location availability check was a simple task and was manually tested using the Microsoft Excel software for this research.

However, in the automatic workflow this can be done using a simple Java program by reading the relevant location sections of the CSV file. If the location is identified as missing, such records can be flagged which would then initiate the location extraction process. In this research, the quality of the locations was tested using ArcGIS spatial analysis tools. This could also be implemented in an automated workflow using ArcGIS Runtime SDK for Java.

Location extraction

The location extraction was tested by using semantic and non-semantic location extraction techniques. It was found that the semantic technique outperformed the non-semantic location extraction technique and so the use of the semantic approach for the automated processing is recommended. A gazetteer lookup-based approach in conjunction with geo-spatial semantics and natural langue processing techniques could be utilised using the GATE software and its processing resources. The semantic location extraction procedure was explained in section 5.2.2.2 and could be easily be automated using GATE API, Java programming and batch processing methods.

Quality assessment and output

The quality assessment and output stage consists of three phases namely credibility analysis, relevance analysis and the storing of the quality assessed data. The credibility and relevance assessment methods were detailed in chapter six and utilised a naïve Bayesian Network detection approach. The system was fully coded using the Java programming language and the automation of these programs is possible. Similarly, the relevance analysis was based on a GIR approach using a natural language processing based gazetteer lookup with the GATE software for resolving the geographic scope. The thematic scope was resolved using the TF-IDF VSM analysis with the Lucene textual IR system. All these tools and software are Java based and can directly be used for the automated system. The location quality assessed data and credibility and geo-thematic relevance assessed CSD can then be stored in a spatial database such as a spatially enabled PostgreSQL server with PostGIS spatial extensions.

7.6. Integration framework for quality assessed CSD with authoritative data

This dissertation has identified the importance of utilising CSD as an authoritative data source. It is critical to understand the similarities and differences between CSD and authoritative data. Authoritative data are often well structured, organised and managed by trained professionals with tight control over the data capture, editing and maintenance. On the other hand, CSD are often unstructured, not well planned and often produced by untrained citizens with few rules or controls. It is obvious that connecting these two paradigms is challenging and needs a clear understanding about the issues, opportunities and challenges.

Issues, challenges and opportunities of SDI-CSD integration

Authoritative data, similar to SDIs, have clear policies, standards, technology and user requirements which are defined at the very early stages of its development workflow. Therefore, at the data collection phase the collection methods are generally pre-determined, the required accuracy levels are set and standards are defined. The CSD generation is usually very flexible and unconstrained which is an advantage from the data collection point of view however, this creates ongoing issues with respect to data quality.

The metadata and other documented information such as the details of the data collection device, the data collection accuracy, the collector's profile, environmental factors, processing and storage devices and data structures are considered as being critical in authoritative data collection. This generally extends the data collection time and hence the cost of the data collection. However, this becomes an advantage in the later stages of the spatial data management chain. The CSD can be created rapidly with little or no cost and is often just a matter of the touch of a button or a few mouse clicks. In this case, the volunteer is not expected to collect metadata. If a user logs onto a system such as Twitter with a public account, their profile information may be attached to the

CSD reports. However, this information, which was generated at the time the account was created, does not necessarily reflect the real background of the user and therefore limits the opportunities for further source credibility analysis.

There can be thousands of CSD records generated during an event, often over a large area and within a very short period of time. This makes the information currency of the CSD very high and is identified as a key advantage of this form of data. Therefore, identifying at least a small portion of this data as quality data could be extremely useful for contributing to the post disaster management of a disaster event.

Other critical issues related to CSD were the limited location data available and the quality issues of the location data when it was available. As a solution, a location quality assessment method and missing location extraction method were examined in this research. CSD credibility and relevance assessment methods were also investigated. This research suggests that the quality assessed CSD could be suitable for supplementing information gaps and updating the authoritative data sets without the need to utilise an authoritative collection processes. This can also work in reverse, with authoritative data being beneficial in assessing CSD quality. Therefore, CSD and authoritative data management could be considered an interconnected process as indicated in Figure 7.2.

Figure 7.2 CSD and authoritative data integration process

As Figure 7.2 identifies, a CSD qualification step may consist of a number of sub-processes such as location quality assessment, credibility and relevance assessments, may be required before the data can be considered as qualified.

Other issues also need to be considered for the successful integration of data into other repositories. These include the consideration of policy, economic, legal, cultural and

technological impacts. Hence, the progress in the use of non-authoritative data to supplement authoritative data sets has been slow. However, with the increased level of user interaction and the willingness of users to provide feedback and corrections, authorities are now implementing user feedback to correct data errors. Governance and policy frameworks for SDIs have been largely conservative by nature and the institutional inertia of large government organisations has resulted in the limited uptake of CSD for business purposes.

From the CSD perspective, the use of authoritative data to improve the quality of CSD outcomes should be a relatively straightforward process. However, a key issue that is faced by volunteer and crowd sourcing groups is the limited access that is often provided to base data layers by the authoritative data custodians. The open data movement is now changing these access arrangements with governments now required to make their data accessible to the citizens. Some organisations have facilitated access to their data through web services which provides a ready link for mobile applications.

7.7. Chapter summary

CSD is identified as a useful source of data for many applications including disaster management. It can also be considered as a spatial data input source to other established data repositories however, rigorous quality assessments are required to ensure the quality of this spatial data is appropriate. This chapter discussed the findings of this research in the areas of CSD location availability and its quality, credibility and relevance. A framework for CSD quality assessment system automation was presented and discussed. An integration framework was presented for quality assessed CSD with authoritative data. The issues, challenges and opportunities of operationalising data integration were also discussed. The next chapter will discuss the conclusions and recommendations of the research.

Chapter 8: **Conclusions and recommendations for future directions**

8.1. Introduction

This research investigated the CSD quality control methods and assessed their fitness for the SDI based disaster management activities. It explored the CSD quality assessment methods including location, credibility and relevance analysis. The CSD's location information was compared with freely available spatial data from OpenStreetMaps and proprietary spatial data from Google Maps and authoritative spatial data from QDNRM. The credibility and relevance of CSD were assessed using credible and relevant information identification techniques used in the Information Technology (IT) industry. Finally, it identified an automated CSD quality control process and the integration possibilities with SDIs.

This chapter discusses the outcomes achieved during this research and highlights the significance of the research to theory and practice. It reflects on the original research problem and suggests directions for future research.

8.2. Achievement of research aim and objectives

As indicated in the first chapter, the central aim of this research was to:

"Develop a semantic quality assurance process for crowdsource data by analysing its quality based on location information availability, credibility and relevance so selected CSD can be fused with authoritative data for improved disaster management decision making".

To achieve this aim of the research, chapter four identified Design Science (DS) as the most appropriate research approach. This approach was successfully utilized to achieve the main aim of the research and was tested in chapters five, six and seven. Utilizing the DS research approach, chapter five reported on the location quality assessment, semantic location identification and the process of improving the missing

location information of CSD. Chapter six explored the methods used to assess the CSD credibility and relevance. The credibility of CSD was assessed using a naïve Bayesian Network based spam detection model. The relevance of CSD was analysed using geo-thematic information relevance assessment methods utilised in the Geographic Information Retrieval domain. Finally, chapter seven outlined the methods used to automate the CSD credibility and relevance detection and presented a quality improved CSD and authoritative data integration framework.

8.2.1. Objective 1: Review relevant literature to identify the critical dimensions and approaches in assessing the CSD quality and investigate the possibilities to improve CSD

This research reviewed the relevant literature and found that the general spatial data quality assessment approaches were not appropriate and often invalid for the CSD quality assessment. It identified that the CSD quality should be carefully analysed as they are often created by people with varied experience and knowledge using devices of differing accuracy. It has also identified the need to critically analyse the CSD location and found that credibility and relevance as the most appropriate CSD quality assessment indicators. The information credibility and relevance are often assessed in the IT sector to identify legitimate and the most appropriate information such as the legitimacy of emails received or searching for quality products or services through the online search engines. Modern online information searches often include location in the queries while modern search engines use GIR techniques to identify the relevant information from millions of information sources available in today's connected world. The literature review of this research has identified the similarities between assessing the legitimacy of emails and assessing the credibility of CSD. It also identified the possibility of using GIR approaches for assessing relevant CSD for tasks such as post-flood disaster management.

This research utilised the spam email detection and GIR techniques for CSD credibility and relevance assessment along with critical assessment of the CSD location quality for flood disaster management actions.

8.2.2. Objective 2: Develop a process to extract and geocode the location information of CSD using Gazetteers and ontologies and assess its quality

Chapter five presented methods for the CSD location quality assessment and semantically extracted the implicit location information from CSD message descriptions. The availability of high quality location information in spatial data may be useful for obtaining quality outputs. A popular approach for assessing the quality of spatial data location information is to compare this data with the authoritative spatial data or any other forms of spatial data with acceptable quality. This research compared the CSD location quality with three different sources of spatial data with acceptable location quality and completeness.

As indicated previously, the location availability cannot always be guaranteed in modern CSD spatial data and the quality of the CSD location information may often vary. This research identified a viable option to extract the hidden implicit location information in textual descriptions of the CSD reports. Natural Language Processing (NLP) based text processing is a popular method utilised to identify the useful information in textual data. The use of semantics in the text processing is useful to detect important information which are often not easy to detect using the general NLP approaches. This research used semantics and NLP techniques to successfully process CSD along with local and global gazetteers to identify hidden toponyms and to geocode the CSD messages.

8.2.3. Objective 3: Assess the credibility of CSD using appropriate filtering and processing techniques

The results of the research showed that the use of information credibility assessment techniques used in the IT sector for identifying legitimate emails can be successfully implemented to assess the credibility of CSD. The naïve Bayesian Networks are proven to be successful in filtering spam emails by analysing various properties of the email messages including the contents. Similarly, this research assessed the CSD message contents using a modified naïve Bayesian Network based model with forced training to detect credible reports from CSD. As indicated previously, CSD is often a mix of credible and unreliable messages. Therefore, this study trained the system to detect credible information by carefully modifying the messages in a training sample. The results showed that a naïve Bayesian system with forced training can accurately detect credible information from CSD.

8.2.4. Objective 4: Assess the relevance of CSD using Natural Language Processing (NLP) and Geographic Information Retrieval (GIR) techniques

The research presented a successful CSD relevance analysis method based on relevance assessment techniques used for geographic information retrieval. The method was based on NLP and GIR techniques. Identifying credible as well as relevant information is important to the particular task in hand. CSD can contain highly relevant information as well as information that may be irrelevant. This research identified that identifying relevant information from CSD is challenging however, utilising the information relevance detection approaches used in the IT domain can be used for CSD relevance detection. It was found that GIR assesses the relevance of data and ranks the identified documents based on both the thematic and geographic relevancies. This research illustrated that assessing CSD based on the thematic and geographic relevancies can also generate a relevance ranked list of CSD for specific tasks such as post-flood disaster management.

8.2.5. Objective 5: Propose automation techniques to carry out CSD quality assurance processes and integrate with authoritative data for disaster management activities

Chapter seven discussed the findings of this research and presented an approach for CSD quality assessment automation along with a framework to integrate the quality assessed CSD with authoritative data. Chapter five presented CSD location quality assessment and semantic location extraction methods and chapter six presented CSD credibility and relevance assessment methods. Individual processing and assessment of the CSD location quality, semantic location extraction, credibility and relevance would not be practical for real life applications of CSD such as disaster management. Therefore, an integrated and automated mechanism to perform all above tasks within a single framework should be employed. Chapter seven presented a framework for an integrated system for applications such as disaster management. The proposed automation architecture is based on Java APIs and Java programming language which could easily integrate most of the modules developed throughout this research.

8.3. Contributions to original knowledge

The CSD quality assessment and its research are still relatively immature however, the areas are growing rapidly. This research contributes to the field of CSD quality assessment in theory, practice and methods.

This research and previous research have identified that the CSD quality assessment is challenging due to the CSD's unique nature and characteristics. Researchers are actively working to address the CSD quality issues and to fill the gaps pertaining to CSD and related applications. The research identified the importance of, and proposed novel approaches to, successfully assessing CSD quality including location, credibility and relevance.

A number of researchers have reported that the location availability of CSD is very low and the location quality of CSD is often vague. This research also has identified the same issues and assessed the CSD's location quality against different datasets with variable quality. Attempts to semantically extract the CSD location and assess the location quality have been potentially successful. This research has contributed to the knowledge by presenting a useful method to assess the quality by semantically detecting, extracting and geo-tagging the CSD location.

Previous research has proposed and tested various CSD quality assessment approaches. However, few researchers have identified the opportunities available in the other domains such as IT, for adapting successful credibility and relevance assessment techniques for CSD credibility and relevance assessment. This research also found similarities with the legitimate email detection systems in the IT domain and credibility detection of CSD. A method was described to adapt the naïve Bayesian Network for the CSD credibility detection. Although, CSD is highly unstructured and often a mix of high and low-quality information, creating a challenging environment for identifying the credible information, this research tackled the issue by carefully selecting and managing the training sample to rigorously train the model.

Although, it is considered as a growing field, the CSD relevance assessment approaches are still immature. This research utilised an adapted information relevance assessment model from the GIR for the CSD relevance assessment.

Most of the CSD credibility and relevance approaches proposed by previous research have assessed only one aspect at a time (i.e. independently). This research has identified the importance of assessing both of the aspects together and proposed a model to assess credibility and relevance simultaneously. This may be an important initiative for applications similar to the disaster management where it is required to determine the quality of information within a single quality assessment framework.

8.4. Recommendations for further research

This research has identified various avenues for the further research based on the findings. The possible areas for the further research are described below.

8.4.1. Semantics and CSD location

The CSD's location is available in different forms such as real coordinates (i.e. GPS coordinates) or hidden in the text in the form of addresses (or toponyms). This research has shown that the hidden location can be extracted using semantics and gazetteer lookup techniques. Semantics are beneficial for identifying complex combinations of toponyms within the text where the direct word matching often fails to do so. These were also useful for resolving geo to geo and geo to non-geo ambiguities. This research utilised a semantic gazetteer for both extracting toponyms and ambiguity resolutions. However, other forms of non-spatial ontologies e.g. medical or environmental ontologies could be used to filter out non-relevant terms along with semantic gazetteers.

Current developments in the social media and related communications are relatively new to the semantics and ontology concepts. This research found limitations with available ontologies in dealing with the social media data. It was also identified that the available disaster management ontologies were incomplete with regards to the current social media terminology. Therefore, further work is suggested on semantic concepts and methods to design the required ontologies which are suitable for social media and related communication.

There were additional locations derived from the CSD reports when the semantic location extraction was performed. This created multiple locations and therefore generated ambiguities. Careful analyses of the multiple locations were required to determine the most appropriate location in these situations. However, the decision with respect to the best location was not an easy decision unless contextual analysis was undertaken to understand the intended location that the producer expected to report. The most accurate position may not always be the best answer as the user may be at one location

however, was intending to generate a report about an incident that occurred at a different location. Therefore, this research recommends further research to semantically analyse the CSD reports' text to accurately identify the intended contextual location.

In GIS, the related attribute information of a location can be useful for further geo processing purposes. This research extracted many attribute information types during the toponym identification process of the CSD location extraction. However, much information was discarded and filtered to identify the required toponyms accurately. If the system was modified to collect such appropriate attribute information this would be useful in the value adding of the extracted location information. This study recommends testing possible methods to identify useful attribute information along with the location extraction process.

The findings of this research suggest that further work can be conducted on the usability of semantics for identifying hidden toponyms by semantically combining text patterns. It may also be important to study how the wording used to explain geographic locations in CSD which can be unique to a local group or to a language in particular. Natural language processing, linguistics and semantics will be useful for this further work.

8.4.2. Web 3.0 and CSD, the future of quality

The Web 3.0 or the semantic web explores connective intelligence by connecting data, concepts, applications and ultimately people. This research recommends the use of semantics for analysing and improving the CSD quality. In the semantic web, semantic services should be widely available in different forms including quality assessment and improvement services. This research suggests the development of semantic CSD quality management tools in the form of ubiquitous services that are ready to use in the future smartphones to efficiently surf the semantic web. Another possible initiative is to test the quality of generated CSD reports at the time of their creation. In this scenario, the above CSD quality management tools can be used to test the CSD reports

at the time of creation and tag them with a quality level. This will save time, effort and resources and improve the value of CSD.

8.5. Final remarks

The location information requirements of the rapidly changing world cannot rely on traditional spatial data curation and quality management mechanisms as they can take a considerable amount of time. More timely data creation and quality management approaches are required in most modern applications which rely on spatial data. In future, there will be increased demand for faster data capture, processes and procedures. However, existing authoritative spatial data frameworks such as the SDIs cannot be neglected as many applications still require accuracy rather than currency of the data. The most appropriate solution identified by this research is to run both systems in parallel and fill the gaps of each of the systems through data integration. The research presented a CSD quality assessment and integration framework with authoritative data such as that available through SDIs. The findings of the research will be beneficial to the government, public, community and businesses seeking timely, accurate and relevant spatial data.

References

Agichtein, E, Brill, E & Dumais, S 2006, 'Improving web search ranking by incorporating user behavior information', in *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval* ACM, Seattle, Washington, USA, pp. 19-26.

Ajmar, A, Perez, F & Terzo, O 2008, 'WFP spatial data infrastructure (SDI) implementation in support of emergency management', *International Archives of Photogrammetry and Remote Sensing*, vol. 37, pp. 1097-104.

Alexopoulos, P, Ruiz, C & Villazon-terrazas, B 2013, 'KLocator: An Ontology-Based Framework for Scenario-Driven Geographical Scope Resolution', *International Journal on Advances in Intelligent Systems*, vol. 6, no. 3 & 4, pp. 177-87.

Amitay, E, Har'El, N, Sivan, R & Soffer, A 2004, 'Web-a-where: geotagging web content', in *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval* ACM, Sheffield, United Kingdom, pp. 273-80.

Andrade, L & Silva, MJ 2006, 'Relevance Ranking for Geographic IR', in *Proceedings of Workshop on Geographic Information Retrieval - SIGIR '06,* Seattle, USA.

Androutsopoulos, I, Koutsias, J, Chandrinos, KV & Spyropoulos, CD 2000, 'An experimental comparison of naive Bayesian and keyword-based anti-spam filtering with personal e-mail messages', in *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, Athens, Greece, pp. 160-7.

Antoniou, V 2011, 'User generated spatial content: an analysis of the phenomenon and its challenges for mapping agencies', PhD thesis, UCL (University College London).

Antoniou, V & Skopeliti, A 2015, 'Measures and Indicators of Vgi Quality: AN Overview', *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. II-3/W5, pp. 345-51.

Antoniu, V 2016, 'Volunteered Geographic Informaiton Measuring Quality, Understnading the Value', *GEOmedia*, vol. 20, no. 1.

Bahree, M 2008, 'Citizen Voices', *Forbes Magazine*, vol. 182, no. 12, p. 83.

Bakri, A & Fairbairn, D 2011, 'User Generated Content and Formal Data Sources for Integrating Geospatial Data', in *Proceedings of the 25th International Cartographic Conference*, Paris, France.

Battle, R & Kolas, D 2011, 'Linking geospatial data with GeoSPARQL', *Semantic Web Journal - Interoperability, Usability, Applicability*, vol. 24, pp. 1-11.

Bishr, M & Kuhn, W 2007, 'Geospatial information bottom-up: A matter of trust and semantics', in S Fabrikant & M Wachowicz (eds), *The European information society: Leading the way with geo-information*, Springer, Berlin, pp. 365-87.

Bishr, M & Mantelas, L 2008, 'A trust and reputation model for filtering and classifying knowledge about urban growth', *GeoJournal*, vol. 72, no. 3-4, pp. 229-37.

Blanzieri, E & Bryl, A 2008, 'A survey of learning-based techniques of email spam filtering', *Artificial intelligence review*, vol. 29, no. 1, pp. 63-92.

Bordogna, G, Carrara, P, Criscuolo, L, Pepe, M & Rampini, A 2014, 'On predicting and improving the quality of Volunteer Geographic Information projects', *International Journal of Digital Earth*, vol. 9, no. 2, pp. 134-55.

Borges, KA, Davis Jr, CA, Laender, AH & Medeiros, CB 2011, 'Ontology-driven discovery of geospatial evidence in web pages', *GeoInformatica*, vol. 15, no. 4, pp. 609-31.

Brando, C & Bucher, B 2010, 'Quality in user generated spatial content: A matter of specifications', in *Proceedings of the 13th AGILE International Conference on Geographic Information Science*, Springer, Guimarães, Portugal, pp. 11-4.

Bruns, A, Burgess, JE, Crawford, K & Shaw, F 2012, *#qldfloods and @QPSMedia: Crisis communication on Twitter in the 2011 south east Queensland floods*, ARC Centre of Excellence for Creative Industries and Innovation - Brisbane, Queensland, viewed 15.04.2015, <https://eprints.qut.edu.au/48241/1/floodsreport.pdf>.

Brusa, G, Caliusco, ML & Chiotti, O 2006, 'A process for building a domain ontology: an experience in developing a government budgetary ontology', in *Proceedings of the second Australasian workshop on Advances in ontologies (AOW 2006)*, Australian Computer Society, Inc., Hobart, Australia, pp. 7-15.

Budhathoki, NR & Nedovic-Budic, Z 2008, 'Reconceptualizing the role of the user of spatial data infrastructure', *GeoJournal*, vol. 72, no. 3-4, pp. 149-60.

Burgess, J & Bruns, A 2012, 'Twitter archives and the challenges of "Big Social Data" for media and communication research', *M/C Journal*, vol. 15, no. 5.

Buscaldi, D & Rosso, P 2009, 'Using geowordnet for geographical information retrieval', in *Evaluating Systems for Multilingual and Multimodal Information Access*, Springer, pp. 863-6.

Cai, G 2002, 'GeoVSM: An integrated retrieval model for geographic information', in *Proceedings of International Conference on Geographic Information Science (GIScience 2002)*, Springer, Boulder, CO, USA, pp. 65-79.

Capineri, C, Haklay, M, Huang, H, Antoniou, V, Kettunen, J, Ostermann, F & Purves, R 2016, *European Handbook of Crowdsourced Geographic Information*, Ubiquity Press: London, UK.

Caragea, C, McNeese, N, Jaiswal, A, Traylor, G, Kim, HW, Mitra, P, Wu, D, Tapia, AH, Giles, L & Jansen, BJ 2011, 'Classifying text messages for the haiti earthquake', in *Proceedings of the 8th international conference on information systems for crisis response and management (ISCRAM2011)*, Lisbon, Portugal.

Castillo, C, Mendoza, M & Poblete, B 2011, 'Information credibility on twitter', in *Proceedings of the 20th international conference on World wide web*, ACM, Hyderabad, India, pp. 675-84.

Cheng, J & Greiner, R 1999, 'Comparing Bayesian network classifiers', in *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence* Morgan Kaufmann Publishers Inc., Stockholm, Sweden, pp. 101-8.

Chrisman, NR 1984, 'Part 2: issues and problems relating to cartographic data use, exchange and transfer: the role of quality information in the long-term functioning of a geographic information system', *Cartographica: The International Journal for Geographic Information and Geovisualization*, vol. 21, no. 2-3, pp. 79-88.

Coleman, DJ 2013, 'Potential contributions and challenges of VGI for conventional topographic base-mapping programs', in *Crowdsourcing Geographic Knowledge*, Springer, pp. 245-63.

Coleman, DJ, Georgiadou, Y & Labonte, J 2009, 'Volunteered geographic information: The nature and motivation of produsers', *International Journal of Spatial Data Infrastructures Research*, vol. 4, no. 1, pp. 332-58.

Cooper, AK, Coetzee, S, Kaczmarek, I, Kourie, DG, Iwaniak, A & Kubik, T 2011, 'Challenges for quality in volunteered geographical information', in *Proceedings of AfricaGEO 2011 Conference*, Cape Town, South Africa.

Cowan, T 2013, 'A Framework for Investigating Volunteered Geographic Information Relevance in Planning', Master thesis, University of Waterloo.

Craglia, M, Ostermann, F & Spinsanti, L 2012, 'Digital Earth from vision to practice: making sense of citizen-generated content', *International Journal of Digital Earth*, vol. 5, no. 5, pp. 398-416.

Cranor, LF & LaMacchia, BA 1998, 'Spam!', *Communications of the ACM*, vol. 41, no. 8, pp. 74-83.

Criscuolo, L, Carrara, P, Bordogna, G, Pepe, M, Zucca, F, Seppi, R, Oggioni, A & Rampini, A 2016, 'Handling quality in crowdsourced geographic information', in C Capineri, et al. (eds), *European Handbook of Crowdsourced Geographic Information*, ch 5, pp. 57-74.

Cunningham, H, Maynard, D, Bontcheva, K & Tablan, V 2002, 'A framework and graphical development environment for robust NLP tools and applications', in *Proceedings of Annual Meeting of the Association of Computational Linguistics*, Philadelphia, Pennsylvania, pp. 168-75.

Currion, P, Silva, Cd & Van de Walle, B 2007, 'Open source software for disaster management', *Communications of the ACM*, vol. 50, no. 3, pp. 61-5.

De Longueville, B, Ostlander, N & Keskitalo, C 2010, 'Addressing vagueness in Volunteered Geographic Information (VGI)–A case study', *International Journal of Spatial Data Infrastructures Research*, vol. 5, pp. 1725-0463.

De Sabbata, S & Reichenbacher, T 2010, 'A probabilistic model of geographic relevance', in *Proceedings of the 6th Workshop on Geographic Information Retrieval (GIR-10)*, ACM, Zurich, Switzerland, p. 23.

Delboni, TM, Borges, KA, Laender, AH & Davis, CA 2007, 'Semantic expansion of geographic web queries based on natural language positioning expressions', *Transactions in GIS*, vol. 11, no. 3, pp. 377-97.

Devillers, R, Stein, A, Bédard, Y, Chrisman, N, Fisher, P & Shi, W 2010, 'Thirty years of research on spatial data quality: achievements, failures, and opportunities', *Transactions in GIS*, vol. 14, no. 4, pp. 387-400.

Diaz, L, Remke, A, Kauppinien, T, Degbelo, A, Foerster, T, Stasch, C, Rieke, M, Baranski, B, Broring, A & Wytzisk, A 2012, 'Future SDI – Impulses from Geoinformatics Research and IT Trends', *International Journal of Spatial Data Infrastructures Research*, vol. 07, pp. 378-410.

Du, H, Alechina, N, Jackson, M & Hart, G 2013, 'Matching Formal and Informal Geospatial Ontologies', in *Geographic Information Science at the Heart of Europe*, Springer, pp. 155-71.

Economist 2008, *Identity Parade*, The Economist Special Report Technology and Government, viewed 20.08.2016, <http://www.economist.com/node/10638196>.

Elwood, S 2008a, 'Volunteered geographic information: key questions, concepts and methods to guide emerging research and practice', *GeoJournal*, vol. 72, no. 3, pp. 133-5.

Elwood, S 2008b, 'Volunteered geographic information: future research directions motivated by critical, participatory, and feminist GIS', *GeoJournal*, vol. 72, no. 3-4, pp. 173-83.

ESRI 2010, *Spatial Data Infrastructure: A Collabarative Network*, ESRI, <http://www.esri.com/library/brochures/pdfs/spatial-data-infrastructure.pdf>.

Estellés-Arolas, E & González-Ladrón-de-Guevara, F 2012, 'Towards an integrated crowdsourcing definition', *Journal of Information science*, vol. 38, no. 2, pp. 189-200.

Fernandez, TD & Fernandez, JLC 2008, 'Towards Semantic Spatial Data Infrastructures: A framework for sustainable development', in *Proceedings of the 10th GSDI Conference*, Port of Spain, Trinidad and Tobago.

Fiedrich, F, Gehbauer, F & Rickers, U 2000, 'Optimized resource allocation for emergency response after earthquake disasters', *Safety Science*, vol. 35, no. 1–3, pp. 41-57.

FitzGerald, G, Du, W, Jamal, A, Clark, M & Hou, XY 2010, 'Flood fatalities in contemporary Australia (1997–2008)', *Emergency Medicine Australasia*, vol. 22, no. 2, pp. 180-6.

Flanagin, AJ & Metzger, MJ 2008, 'The credibility of volunteered geographic information', *GeoJournal*, vol. 72, no. 3-4, pp. 137-48.

Fogg, B & Tseng, H 1999, 'The elements of computer credibility', in *Proceedings of SIGCHI conference on Human Factors in Computing Systems*, ACM, Pittsburgh, PA, USA, pp. 80-7.

Foley, R 2009, 'Integrated Spatial Data Infrastructures', in *The International Encyclopaedia of Human Geography*, Elsevier, London, vol. 5, pp. 507-11.

Fonte, C, Bastin, L, Foody, G, Kellenberger, T, Kerle, N, Mooney, P, Olteanu-Raimond, AM & See, L 2015, 'VGI quality control', *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 2, no. 3, pp. 317-24.

Frontiera, P, Larson, R & Radke, J 2008, 'A comparison of geometric approaches to assessing spatial similarity for GIR', *International Journal of Geographical Information Science*, vol. 22, no. 3, pp. 337-60.

Fu, G, Jones, CB & Abdelmoty, AI 2005a, 'Ontology-based spatial query expansion in information retrieval', in *Proceedings of OTM Confederated International Conferences "On the Move to Meaningful Internet Systems"*, Springer, Berlin, Heidelberg, pp. 1466-82.

Fu, G, Jones, CB & Abdelmoty, AI 2005b, 'Building a Geographical Ontology for Intelligent Spatial Search on the Web', in *Proceedings of Databases and Applications*, Innsbruck, Austria, pp. 167-72.

Gao, H, Barbier, G, Goolsby, R & Zeng, D 2011, 'Harnessing the crowdsourcing power of social media for disaster relief', *IEEE Intelligent Systems*, vol. 26, no. 3, pp. 10-4.

Girres, JF & Touya, G 2010, 'Quality assessment of the French OpenStreetMap dataset', *Transactions in GIS*, vol. 14, no. 4, pp. 435-59.

Giunchiglia, F, Maltese, V, Farazi, F & Dutta, B 2010, 'GeoWordNet: a resource for geo-spatial applications', in *Proceedings of Extended Semantic Web Conference*, Springer, Berlin, Heidelberg, pp. 121-36.

Goodchild, MF 2007, 'Citizens as sensors: the world of volunteered geography', *GeoJournal*, vol. 69, no. 4, pp. 211-21.

Goodchild, MF 2009, 'NeoGeography and the nature of geographic expertise', *Journal of Location Based Services*, vol. 3, no. 2, pp. 82-96.

Goodchild, MF & Li, L 2012, 'Assuring the quality of volunteered geographic information', *Spatial Statistics*, vol. 1, pp. 110-20.

Graham-Cumming, J 2006, 'Does Bayesian poisoning exist', *Spam Bulletin*, vol. 2, p. 69.

Grira, J, Bédard, Y & Roche, S 2010, 'Spatial data uncertainty in the VGI world: Going from consumer to producer', *Geomatica*, vol. 64, no. 1, pp. 61-72.

Gruber, TR 1993, 'A translation approach to portable ontology specifications', *Knowledge acquisition*, vol. 5, no. 2, pp. 199-220.

Guzella, TS & Caminhas, WM 2009, 'A review of machine learning approaches to Spam filtering', *Expert Systems with Applications*, vol. 36, no. 7, pp. 10206-22.

Haklay, M 2010, 'How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets', *Environment and planning. B, Planning & design*, vol. 37, no. 4, pp. 682-703.

Haklay, M & Weber, P 2008, 'Openstreetmap: User-generated street maps', *IEEE Pervasive Computing*, vol. 7, no. 4, pp. 12-8.

Haklay, M, Basiouka, S, Antoniou, V & Ather, A 2010, 'How many volunteers does it take to map an area well? The validity of Linus' law to volunteered geographic information', *The Cartographic Journal*, vol. 47, no. 4, pp. 315-22.

Harris, TM & Lafone, HF 2012, 'Toward an informal Spatial Data Infrastructure: Voluntary Geographic Information, Neogeography, and the role of citizen sensors', in K Cerbova & O Cerba (eds), *SDI, Communities and Social Media*, Czech Centre for Science and Society, Prague, Czech Republic, pp. 8-21.

Hart, G & Dolbear, C 2013, *Linked Data: A Geographic Perspective*, CRC Press, NW, USA.

Heipke, C 2010, 'Crowdsourcing geospatial data', *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 65, no. 6, pp. 550-7.

Hevner, A & Chatterjee, S 2010, 'Design science research in information systems', in *Design research in information systems*, Springer, USA, pp. 9-22.

Hevner, VAR, T, MS, Jinsoo, P & Sudha, R 2004, 'Design science in information systems research', *MIS quarterly*, vol. 28, no. 1, pp. 75-105.

Hill, LL 2000, 'Core elements of digital gazetteers: placenames, categories, and footprints', in *Research and advanced technology for digital libraries*, Springer, pp. 280-90.

Hjørland, B 2007, 'Information: objective or subjective/situational?', *Journal of the American Society for information Science and Technology*, vol. 58, no. 10, pp. 1448-56.

Honert, Vd, Robin, C & McAneney, J 2011, 'The 2011 Brisbane Floods: Causes, Impacts and Implications', *Water*, vol. 3, no. 4, pp. 1149-73.

Hovland, CI, Janis, IL & Kelley, HH 1953, *Communication and persuasion; psychological studies of opinion change*, Yale University Press, New Haven, Connecticut, USA.

Hung, KC, Kalantari, M & Rajabifard, A 2016, 'Methods for assessing the credibility of volunteered geographic information in flood response: A case study in Brisbane, Australia', *Applied Geography*, vol. 68, pp. 37-47.

Inkpen, D 2007, 'Information retrieval on the internet', viewed 05.12.2015,<http://www.site.uottawa.ca/~diana/csi4107/IR_draft.pdf>.

Jackson, J 2006, 'Neogeography" blends blogs with online maps', *National Geographic News*, vol. 6, p. 2008.

Jackson, M, Rahemtulla, H & Morley, J 2010, 'The synergistic use of authenticated and crowd-sourced data for emergency response', in *Proceedings of the 2nd International Workshop on Validation of Geo-Information Products for Crisis Management (VAL-gEO)*, Ispra, Italy.

Janowicz, K & Keßler, C 2008, 'The role of ontology in improving gazetteer interaction', *International Journal of Geographical Information Science*, vol. 22, no. 10, pp. 1129-57.

Janowicz, K, Raubal, M & Kuhn, W 2011, 'The semantics of similarity in geographic information retrieval', *Journal of Spatial Information Science*, vol. 2011, no. 2, pp. 29-57.

Janowicz, K, Schade, S, Broring, A, Kebler, C, Maue, P & Stasch, C 2010, 'Semantic enablement for spatial data infrastructures', *Transactions in GIS*, vol. 14, no. 2, pp. 111-29.

Jiang, W, Deng, L, Chen, L, Wu, J & Li, J 2009, 'Risk assessment and validation of flood disaster based on fuzzy mathematics', *Progress in Natural Science*, vol. 19, no. 10, pp. 1419-25.

Jones, CB & Purves, RS 2008, 'Geographical information retrieval', *International Journal of Geographical Information Science*, vol. 22, no. 3, pp. 219-28.

Jones, CB, Alani, H & Tudhope, D 2001, 'Geographical information retrieval with ontologies of place', in *Spatial information theory*, Springer, pp. 322-35.

Kang, B, O'Donovan, J & Höllerer, T 2012, 'Modeling topic specific credibility on twitter', in *Proceedings of the 17th International Conference on Intelligent User Interfaces (IUI-12)*, ACM, Lisbon, Portugal, pp. 179-88.

Keßler, C & de Groot, RTA 2013, 'Trust as a proxy measure for the quality of volunteered geographic information in the case of OpenStreetMap', in *Geographic information science at the heart of Europe*, Springer, pp. 21-37.

Khan, H, Vasilescu, LG & Khan, A 2008, 'Disaster Management Cycle - A Theoretical Approach', *Journal of Management and Marketing*, vol. 6, no. 1, pp. 43-55.

Kim, H 2013, 'Credibility assessment of volunteered geographic information for emergency management: a Bayesian network modeling approach', MSc thesis, University of Illinois.

Koswatte, S, McDougall, K & Liu, X 2014, 'Ontology driven VGI filtering to empower next generation SDIs for disaster management', in *Research at Locate 14*, S Winter & C Rizos (eds.), Canberra, Australia.

Koswatte, S, McDougall, K & Liu, X 2015, 'SDI and crowdsourced spatial information management automation for disaster management', *Survey Review*, vol. 47, no. 344, pp. 307-15.

Koswatte, S, McDougall, K & Liu, X 2016, 'Semantic Location Extraction from Crowdsourced Data', *ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, pp. 543-7.

Kothari, CR 2004, *Research methodology: Methods and techniques*, New Age International, Delhi, India.

Koukoletsos, T, Haklay, M & Ellul, C 2012, 'Assessing data completeness of VGI through an automated matching procedure for linear data', *Transactions in GIS*, vol. 16, no. 4, pp. 477-98.

Krumm, JC & Mummidi, LN 2008, 'Discovering points of interest from users map annotations', *GeoJournal*, vol. 72, no. 3-4, pp. 215-27.

Kumar, C 2011, 'Relevance and ranking in geographic information retrieval', in *Proceedings of the Fourth BCS-IRSG conference on Future Directions in Information Access*, British Computer Society, pp. 2-7.

Larson, RR 1996, 'Geographic information retrieval and spatial browsing', in *Proceedings of Clinic on Library Applications of Data Processing - 1995*, Illinois, USA.

Laura, D, Albert, R, Tomi, K, Auriol, D, Theodor, F, Christoph, S, Matthes, R, Bastian, S, Bastian , B, Arne, B & Andreas, W 2012, 'Future SDI – Impulses from Geoinformatics Research and IT Trends', *International Journal of Spatial Data Infrastructures Research*, vol. 07.

Lee, V 1994, 'Volunteer monitoring: a brief history', *The Volunteer Monitor*, vol. 6, no. 1, pp. 29-33.

Leidner, JL & Lieberman, MD 2011, 'Detecting geographical references in the form of place names and associated spatial natural language', *SIGSPATIAL Special*, vol. 3, no. 2, pp. 5-11.

Lemmens, R, Falquet, G, De Sabbata, S, Jiang, B & Bucher, B 2016, 'Querying VGI by semantic enrichment', *European Handbook of Crowdsourced Geographic Information*, p. 185.

Lettieri, E, Masella, C & Radaelli, G 2009, 'Disaster management: findings from a systematic review', *Disaster Prevention and Management: An International Journal*, vol. 18, no. 2, pp. 117-36.

Lieberman, MD, Samet, H, Sankaranarayanan, J & Sperling, J 2007, 'STEWARD: architecture of a spatio-textual search engine', in *Proceedings of the 15th annual ACM international symposium on Advances in geographic information systems* ACM, Seattle, Washington, p. 25.

Longueville, BD, Luraschi, G, Smits, P, Peedell, S & Groeve, TD 2010, 'Citizens as sensors for natural hazards: A VGI integration workflow', *Geomatica*, vol. 64, no. 1, pp. 41-59.

Lopes, C, Cortez, P, Sousa, P, Rocha, M & Rio, M 2011, 'Symbiotic filtering for spam email detection', *Expert Systems with Applications*, vol. 38, no. 8, pp. 9365-72.

Machado, IMR, de Alencar, RO, de Oliveira Campos Jr, R & Davis Jr, CA 2011, 'An ontological gazetteer and its application for place name disambiguation in text', *Journal of the Brazilian Computer Society*, vol. 17, no. 4, pp. 267-79.

Mansourian, A, Rajabifard, A, Valadan Zoej, M & Williamson, I 2006, 'Using SDI and web-based system to facilitate disaster management', *Computers & Geosciences*, vol. 32, no. 3, pp. 303-15.

March, ST & Smith, GF 1995, 'Design and natural science research on information technology', *Decision Support Systems*, vol. 15, no. 4, pp. 251-66.

Martins, B, Silva, MJ & Andrade, L 2005, 'Indexing and ranking in Geo-IR systems', in *Proceedings of Workshop on Geographic information retrieval* ACM, Bremen, Germany, pp. 31-4.

Martins, B, Silva, MJ, Freitas, S & Afonso, AP 2006, 'Handling Locations in Search Engine Queries', *GIR*, vol. 6, pp. 1-6.

Maynard, D, Li, Y & Peters, W 2008, 'NLP techniques for term extraction and ontology population', in P Buitelaar & P Cimiano (eds), *Ontology Learning and Population: Bridging the Gap between Text and Knowledge*, IOS Press, vol. 167, pp. 107-27.

McDougall, K 2009, 'Volunteered geographic information for building SDI', in *Proceedings of Surveying and Spatial Sciences Institute Biennial International Conference (SSC 2009),* Adelaide, Australia, pp. 645-53.

McDougall, K 2012, 'An assessment of the contribution of volunteered geographic information during recent natural disasters', in *Proceedings of GSDI-2012*, GSDI Association Press, Quebec City, Canada, pp. 201-14.

McDougall, K, Rajabifard, A & Williamson, I 2009, 'Local government and SDI: understanding their capacity to share data', in *Proceedings of GSDI-11*, Rotterdam, Netherlands.

Metsis, V, Androutsopoulos, I & Paliouras, G 2006, 'Spam filtering with naive bayes-which naive bayes?', in *Proceedings of the Third Conference on Email and Anti-Spam (CEAS 2006)*, California, USA, pp. 27-8.

Metzger, MJ 2007, 'Making sense of credibility on the Web: Models for evaluating online information and recommendations for future research', *Journal of the American Society for information Science and Technology*, vol. 58, no. 13, pp. 2078-91.

Monteiro, BR, Davis, CA & Fonseca, F 2016, 'A survey on the geographic scope of textual documents', *Computers & Geosciences*, vol. 96, pp. 23-34.

Neal, DM 1997, 'Reconsidering the Phases of Disasters', *International journal of mass emergencies and disasters*, vol. 15, no. 2, pp. 239-64.

Nebert, DD 2004, *The SDI Cookbook: Developing Spatial Data Infrastructures, version 2*, Reston: GSDI.

Noy, NF & McGuinness, DL 2001, *Ontology development 101: A guide to creating your first ontology*, Stanford knowledge systems laboratory technical report and Stanford medical informatics technical report, Knowledge Systems Laboratory, Stanford University, Stanford, CA.

Noy, NF, Griffith, N & Musen, MA 2008, 'Collecting community-based mappings in an ontology repository', in *Proceedings of the 7th International Semantic Web Conference (ISWC 2008)*, Springer, Karlsruhe, Germany, pp. 371-86.

OCHA 2015, *Nepal: Earthquake 2015 Situation Report No. 20 (as of 3 June 2015)*, viewed 08.01.2016, <http://reliefweb.int/sites/reliefweb.int/files/resources/OCHANepalEarthquakeSituationReportNo.20%283June2015%29_Final.pdf>.

Oort, PV 2006, 'Spatial data quality: from description to application', PhD thesis, Wageningen University, Netherlands.

Ostermann, FO & Spinsanti, L 2011, 'A conceptual workflow for automatically assessing the quality of volunteered geographic information for crisis management', in *Proceedings of AGILE 2011*, University of Utrecht, Utrecht.

Pantel, P & Lin, D 1998, 'Spamcop: A spam classification & organization program', in *Proceedings of AAAI-98 Workshop on Learning for Text Categorization*, Madison, Wisconsin, pp. 95-8.

Parker, CJ, May, A & Mitchell, V 2011, 'Relevance of volunteered geographic information in a real world context', in *Proceedings of GISRUK 2011 conference*, Portsmouth, UK.

Peffers, K, Tuunanen, T, Rothenberger, MA & Chatterjee, S 2007, 'A design science research methodology for information systems research', *Journal of management information systems*, vol. 24, no. 3, pp. 45-77.

Poser, K & Dransch, D 2010, 'Volunteered geographic information for disaster management with application to rapid flood damage estimation', *Geomatica*, vol. 64, no. 1, pp. 89-98.

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

Pullar, D & Hayes, S 2017, 'Will the future maps for Australia be published by ‘nobodies’', *Journal of Spatial Science*, pp. 1-8.

Rajabifard, A & Williamson, IP 2001, 'Spatial data infrastructures: concept, SDI hierarchy and future directions', in *Proceedings of GEOMATICS'80 conference*, Tehran, Iran.

Rajabifard, A, Chan, TO & Williamson, IP 1999, 'The nature of regional spatial data infrastructures', in *Proceedings of AURISA99 Conference*, Blue Mountains, Australia, pp. 22-6.

Rajabifard, A, Feeney, MEF & Williamson, IP 2002, 'Future directions for SDI development', *International Journal of Applied Earth Observation and Geoinformation*, vol. 4, no. 1, pp. 11-22.

Rajabifard, A, Binns, A, Masser, I & Williamson, I 2006, 'The role of sub-national government and the private sector in future spatial data infrastructures', *International Journal of Geographical Information Science*, vol. 20, no. 7, pp. 727-41.

Raper, J 2007, 'Geographic relevance', *Journal of Documentation*, vol. 63, no. 6, pp. 836-52.

Rawls, CG & Turnquist, MA 2010, 'Pre-positioning of emergency supplies for disaster response', *Transportation Research Part B: Methodological*, vol. 44, no. 4, pp. 521-34.

Robinson, G 2003, 'A statistical approach to the spam problem', *Linux journal*, vol. 2003, no. 107, p. 3.

Sadeghi-Niaraki, A, Rajabifard, A, Kim, K & Seo, J 2010, 'Ontology Based SDI to Facilitate Spatially Enabled Society', in *Proceedings of GSDI 12 World Conference*, Singapore, pp. 19-22.

Sahami, M, Dumais, S, Heckerman, D & Horvitz, E 1998, 'A Bayesian approach to filtering junk e-mail', in *Proceedings of Learning for Text Categorization (ICML/AAAI-98) Conference*, Madison, Wisconsin, pp. 98-105.

Sakre, MM, Kouta, MM & Allam, AM 2009, 'Weighting query terms using wordnet ontology', *International Journal of Computer Science and Network Security*, vol. 9, no. 4, pp. 349-58.

Salton, G, Wong, A & Yang, CS 1975, 'A vector space model for automatic indexing', *Communications of the ACM*, vol. 18, no. 11, pp. 613-20.

Saracevic, T 1996, 'Relevance reconsidered', in *Proceedings of the Second conference on conceptions of library and information science (CoLIS 2)*, Copenhagen, pp. 201-18.

Schade, S, Luraschi, G, Longueville, BD, Cox, S & Díaz, L 2010, 'Citizens as Sensors for Crisis Events: Sensor Web Enablement for Volunteered Geographic Information', in *Proceedings of WebMGS 2010*, Como, Italy.

Scheuer, S, Haase, D & Meyer, V 2013, 'Towards a flood risk assessment ontology – Knowledge integration into a multi-criteria risk assessment approach', *Computers, Environment and Urban Systems*, vol. 37, pp. 82-94.

Schneider, KM 2003, 'A comparison of event models for Naive Bayes anti-spam e-mail filtering', in *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics (EACL '03)*, Association for Computational Linguistics, Budapest, Hungary, pp. 307-14.

Senaratne, H, Mobasheri, A, Ali, AL, Capineri, C & Haklay, M 2016, 'A review of volunteered geographic information quality assessment methods', *International Journal of Geographical Information Science*, pp. 1-29.

Shvaiko, P & Euzenat, J 2013, 'Ontology Matching: State of the Art and Future Challenges', *IEEE Transactions on Knowledge & Data Engineering*, vol. 25, no. 1, pp. 158-76.

Sosko, S & Dalyot, S 2017, 'Crowdsourcing User-Generated Mobile Sensor Weather Data for Densifying Static Geosensor Networks', *ISPRS International Journal of Geo-Information*, vol. 6, no. 3, p. 61.

Souza, L, Davis, C, Borges, KA, Delboni, TM & Laender, AH 2005, 'The role of gazetteers in geographic knowledge discovery on the web', in *Proceedings of the Third Latin American Web Congress (LA-WEB 2005)* IEEE, Washington, DC, USA, p. 9.

Spinsanti, L & Ostermann, F 2010, 'Validation and relevance assessment of volunteered geographic information in the case of forest fires', in *Proceedings of Validation of geo-information products for crisis management workshop (ValGeo 2010)*, JRC Ispra.

Spinsanti, L & Ostermann, F 2013, 'Automated geographic context analysis for volunteered information', *Applied Geography*, vol. 43, pp. 36-44.

Steelman, TA, Nowell, B, Bayoumi, D & McCaffrey, S 2012, 'Understanding Information Exchange During Disaster Response: Methodological Insights From Infocentric Analysis', *Administration & Society*, vol. 46, no. 6, pp. 707-43.

Stowe, K, Paul, M, Palmer, M, Palen, L & Anderson, K 2016, 'Identifying and Categorizing Disaster-Related Tweets', in *Proceedings of conference on Empirical Methods in Natural Language Processing*, Austin, Texas, USA, pp. 1-6.

Studer, R, Benjamins, VR & Fensel, D 1998, 'Knowledge engineering: principles and methods', *Data & knowledge engineering*, vol. 25, no. 1, pp. 161-97.

Tait, MG 2005, 'Implementing geoportals: applications of distributed GIS', *Computers, Environment and Urban Systems*, vol. 29, no. 1, pp. 33-47.

Tobler, WR 1970, 'A computer movie simulating urban growth in the Detroit region', *Economic geography*, pp. 234-40.

Turner, A 2006, *Introduction to neogeography*, O'Reilly Media, Sebastopol, CA.

UNISDR 2009, *UNISDR Terminology on Disaster Risk Reduction*, viewed 16.01.2016, <http://www.unisdr.org/files/7817_UNISDRTerminologyEnglish.pdf>.

Van Exel, M & Dias, E 2011, 'Towards a methodology for trust stratification in VGI', in *Proceedings of VGI Pre-Conference at AAG*, Seattle, Washington, USA.

Wang, AH 2010, 'Don't follow me: Spam detection in Twitter', in *Proceedings of International Conference on Security and Cryptography (SECRYPT 2010)*, Athens, Greece, pp. 1-10.

Wang, C, Xie, X, Wang, L, Lu, Y & Ma, WY 2005, 'Detecting geographic locations from web resources', in *Proceedings of Workshop on Geographic information retrieval at conference on Information and Knowledge Management (CIKM '05)*, ACM, Bremen, Germany, pp. 17-24.

Wang, RY & Strong, DM 1996, 'Beyond accuracy: What data quality means to data consumers', *Journal of management information systems*, vol. 12, no. 4, pp. 5-33.

White, HD 2011, 'Relevance theory and citations', *Journal of Pragmatics*, vol. 43, no. 14, pp. 3345-61.

Wiggins, A, Newman, G, Stevenson, RD & Crowston, K 2011, 'Mechanisms for Data Quality and Validation in Citizen Science', in *Proceedings of 2011 IEEE Seventh International Conference on e-Science Workshops (eScienceW)*, IEEE, Stockholm, Sweden, pp. 14-9.

Williamson, I, Grant, D & Rajabifard, A 2005, 'Land Administration and Spatial Data Infrastructures', in *Proceedings of GSDI-8* Cario, Egypt, pp. 1-13.

Williamson, IP, Rajabifard, A & Feeney, MEF 2004, *Developing spatial data infrastructures: from concept to reality*, CRC Press.

Yates, D & Paquette, S 2011, 'Emergency knowledge management and social media technologies: A case study of the 2010 Haitian earthquake', *International journal of information management*, vol. 31, no. 1, pp. 6-13.

Yu, B & Cai, G 2007, 'A query-aware document ranking method for geographic information retrieval', in *Proceedings of the 4th ACM workshop on Geographical information retrieval* ACM, Lisbon, Portugal, pp. 49-54.

Zaila, YL & Montesi, D 2015, 'Geographic information extraction, disambiguation and ranking techniques', in *Proceedings of the 9th Workshop on Geographic Information Retrieval* ACM, Paris, France, pp. 1-7.

Zook, M, Graham, M, Shelton, T & Gorman, S 2010, 'Volunteered geographic information and crowdsourcing disaster relief: a case study of the Haitian earthquake', *World Medical & Health Policy*, vol. 2, no. 2, pp. 7-33.