Dimensions of the Use of Volunteered Geographic Information in Mass Crisis Events

Xiannian Chen

Follow this and additional works at: https://researchrepository.wvu.edu/etd

Recommended Citation
https://researchrepository.wvu.edu/etd/5348

This Dissertation is brought to you for free and open access by The Research Repository @ WVU. It has been accepted for inclusion in Graduate Theses, Dissertations, and Problem Reports by an authorized administrator of The Research Repository @ WVU. For more information, please contact ian.harmon@mail.wvu.edu.
Dimensions of the Use of Volunteered Geographic Information in Mass Crisis Events

Xiannian Chen

Dissertation submitted
to the Eberly College of Arts and Sciences
at West Virginia University
in partial fulfillment of the requirements
for the degree
of Doctor of Philosophy in
Geography

Gregory Elmes, Ph.D., Chair
Jamison Conley, Ph.D.,
Robert Duval, Ph.D.
Trevor Harris, Ph.D.
Timothy Warner, Ph.D.

Department of Geology and Geography

Morgantown, West Virginia
2015

Keywords: Volunteered geographic information; Twitter; Tweet; geo-tag; distributed geodatabase; Web GIS; mass evacuation; disaster management

© 2015 Xiannian Chen
ABSTRACT

Dimensions of the Use of Volunteered Geographic Information in Mass Crisis Events

Xiannian Chen

Recent studies have suggested that catastrophic events that trigger mass evacuation require surrounding communities to be well-prepared to act as ingress or pass-through areas for potential evacuees; however surrounding rural communities may have insufficient disaster-related logistical resources. In the response phase of disaster management, officials must be able to deploy resources to demand locations in types and quantities based on real-time requirements. Effective cross-jurisdictional disaster management needs real-time information, which is usually unavailable from official, authoritative sources. Conversely, VGI (volunteered geographic information) has the capability to provide real-time and local information in disaster management. This study investigates the possibility of utilizing real-time or near real-time VGI in mass evacuation scenarios. The study identifies a potential VGI data source, Tweets from Twitter and how to search for, discover and select relevant Tweets. The dissertation proposes research methods for harvesting, managing live Tweets and saving them to a distributed geodatabase for further spatio-temporal analysis and dissemination to users, such as responders and evacuees.

The study implements a Web GIS application, which includes a tweets discovery component, a geo-tagged tweets mapping component, and an online geo-tagged tweets operation component. The major research goals include designing an application programing interface (API) to harvest relevant Tweets and implement a distributed geodatabase system for storage, analysis, and display of the harvested Tweets so that vital information can be distributed in near real-time. Two case studies, based on Super Storm Sandy in 2012 and a shooting at Kent State University in 2014, were used to evaluate the pros and cons of Tweets from Twitter for response in emergency management and offered prototypes for the development of the final on-line Web GIS

Keywords

Volunteered geographic information; Twitter; Tweet; geo-tag; distributed geodatabase; Web GIS; mass evacuation; disaster management
Dedication

I am dedicating this dissertation to my beloved family members who have meant and continue to mean so much to me. First and foremost, to my mother Changxiu He (贺长秀) whose love for me knew no bounds and, who taught me the values of love, contribution, and hard work. Although she is no longer of this world, her memories continue to regulate my life. Mom, I love you, and you live with me forever.

I dedicate this dissertation to my father, Shaoming Chen (陈绍明); my older sister, Lanping Chen (陈兰平), and my brother-in-law, Jiangshan Li (李江山); my younger brother, Xianjiao Chen (陈先交), and my sister-in-law, Juan Li (李涓). Their endless cares, loves, supports and encouragements are most valuable for me to finish this dissertation.

I would also like to dedicate this dissertation to my wife, Fan Yang, and my lovely daughters, Larrisa and Marina for supporting me and loving me. They are the best cheerleaders during my struggles in finishing this dissertation. I love you all, girls.
Acknowledgement

I am a lucky person on my way of pursuing my PhD degree. I would like to thank all the people that helped me on my journey.

My first thank would go to my advisor and the committee chair, Dr. Gregory Elmes. Dr. Elmes spent an incredible amount of time and energies on advising me. He provided me great literature for me to refer to. There were countless meetings and discussions to reach every stage of my study. He helped me to revise my writings. He provided me guidance on how to do my research and how to develop my case studies and the live demo. Dr. Elmes is not only my academic advisor but also a mentor of my personal life. I would always talk to him when I met some tough situations. He encourages me to overcome the difficulties, pushes to move forward, and help me to be successful on my study and research. I would like to say thanks again to Dr. Elmes.

Dr. Trevor Harris opened the door of VGI study to me. I am really obsessed by this research area. I would thank him for his valuable direction on my research and the great suggestions that he has provided me.

I would also like to express my thanks to the other committee members, Dr. Jamison Conley, Dr. Robert Duval, and Dr. Timothy Warner. They helped a lot on my research and dissertations.

Dr. Ge Lin supported my study a lot. He was on my committee before he got a new job at another institute. Dr. Lin is not only a teacher to me but also a friend of mine. I would like to thank him.

Joshua Lohnes of the Department is the first person that I talked with about VGI in the perspective of tweets and Twitter. A special thank would go to him. Barbara Macleannan of the Department keeps updating me on latest disaster related conference/meeting/talk/opportunity. Thank you, Barbara.

WV GIS Technical Center gives a lot of supports on my study, including travels to conferences, the computer hardware and software for my application development. A special thanks will go to Kurt Donaldson, the manager of the Center. He helps in many ways. His help is priceless, and his helps make my life much easier. The other fellow colleagues of the Center, including Evan Fedorko, Eric Hopkins, Kevin Kuhn, Frank LaFone, and Maneesh Sharma, supported my study in different way. I feel so fortunate to be part of the Center.
Abstract .................................................................................................................................................. ii
Dedication ............................................................................................................................................... iii
Acknowledgments .................................................................................................................................... iv
Table of Contents ..................................................................................................................................... v
List of Tables ........................................................................................................................................... ix
List of Figures .......................................................................................................................................... x

1. Introduction .......................................................................................................................................... 1
   1.1 Disaster and disaster management ................................................................................................. 2
   1.2 Mass evacuation .............................................................................................................................. 5
   1.3 Geospatial technologies in mass evacuation .................................................................................... 6
   1.4 Significance of this research ............................................................................................................ 7
   1.5 Research goal and objectives ........................................................................................................ 8
   1.6 Organization of the Dissertation ................................................................................................... 10

2. Literature review ................................................................................................................................ 11
   2.1 Mass Evacuation ............................................................................................................................. 11
      2.1.1 Modeling and simulation of emergency related resource planning ........................................ 12
      2.1.2 Studies of human responses in mass evacuation .................................................................... 13
      2.1.3 Findings on mass evacuation .................................................................................................. 16
   2.2 Distributed GIS ............................................................................................................................... 16
      2.2.1 Challenges of using GIS in disaster management ................................................................. 17
      2.2.2 Distributed GIS Architecture ................................................................................................ 18
      2.2.3 Two types of distributed GIS systems ..................................................................................... 20
      2.2.4 Distributed GIS in mass evacuation ....................................................................................... 21
      2.2.5 Distributed GIS: Summary ..................................................................................................... 26
   2.3 Volunteered Geographic Information .............................................................................................. 27
6.2.1 How to use the time management tool ................................................................. 97
6.3 Map operation ........................................................................................................ 98
6.4 Summary .................................................................................................................. 99

7. Summary and Conclusion ......................................................................................... 100
  7.1 The contributions of this research ....................................................................... 101
  7.2 Discussion ............................................................................................................. 103
  7.3 Possible future research ....................................................................................... 107
  7.4 Conclusion ............................................................................................................ 109

References .................................................................................................................... 110

Online resources .......................................................................................................... 124
List of Tables

Table 5-1: Summary of collected Tweets for event of 2014 Kent State Shooting ..................... 71
Table 5-2: Tweet symbology design ......................................................................................... 77
Table 5-3: Summary for the sample VGI dataset ...................................................................... 82
List of Figures

Figure 1.1: Cycles in Emergency Management: adapted from National Research Council 2007 . 4
Figure 1.2: General map of an evacuation area, from Goldblatt and Weinisch 2005 ................. 5
Figure 2.1: Three-Tier Client/Server Architecture ................................................................. 19
Figure 2.2: Web-based system to facilitate disaster management using an SDI environment (Mansourian et al. 2006 p. 1576) ................................................................. 22
Figure 2.3: Types of Error ......................................................................................................... 40
Figure 3.1: Conceptual framework emphasizing the importance of VGI at the response phase of the disaster management cycle ................................................................................. 44
Figure 4.1: REST API Tweets request and response diagram. Source: Twitter.com .......... 52
Figure 4.2: Streaming API Tweets request and response diagram. Source: Twitter.com .... 53
Figure 4.3: the distribution of search centers and radii in the Baltimore-Washington D.C. Metropolitan study area .............................................................................................. 56
Figure 4.4: the flow diagram of Tweets search by using the Twitter Search API ................. 58
Figure 4.5: System architecture for live tweet storage ............................................................ 63
Figure 4.6: Implementation of live tweet saving ........................................................................ 64
Figure 5.1: Location and radius for Tweets collection of the 2014 Kent State University campus shooting .............................................................................................................. 70
Figure 5.2: Kent State Campus Shooting Daily Tweet Collection ......................................... 72
Figure 5.3: Kent State Campus Shooting Hourly Tweet Collection ........................................ 73
Figure 5.4: Study Area of Case Study Two– Washington-Baltimore Metropolitan Area and the possible ingress area of West Virginia’s Eastern Panhandle ......................................................................... 74
Figure 5.5: Two major components of case study two: Tweets collection component and Tweets saving and VGI mapping component ............................................................................. 75
Figure 5.6: the Tweets collection sub-region distribution for 2012 Hurricane Sandy at the Washington-Baltimore Metropolitan Area ........................................................................... 76
Figure 5.7: The system interface ............................................................................................... 79
Figure 5.8: Heat map based on VGI in the medical category .................................................... 83
Figure 5.9: Real time data collection and live data access ........................................................ 85
Figure 5.10: An algorithm to remove possible duplicate Tweets ............................................. 87
Figure 5.11: Flowchart used to address two possible VGI errors ............................................ 90
Figure 6.1: Overview of the online geo-Tweets collection system .......................................................... 94
Figure 6.2: Classified geo-Tweets with heat map ...................................................................................... 94
Figure 6.3: specify Tweets search parameters .......................................................................................... 95
Figure 6.4: Status before click on button ................................................................................................. 96
Figure 6.5: Status after click on button .................................................................................................... 96
Figure 6.6: Geo-Tweets management interface ......................................................................................... 97
Figure 6.7: Time management tool ........................................................................................................ 97
Figure 6.8: time slider with time interval in hour and time frame in 8 days ............................................. 98
Figure 6.9: time slider with time interval in hour and time frame in 2 days ............................................. 98
Figure 6.10: List of available basemaps .................................................................................................. 99
Chapter 1: Introduction

1. Introduction

In August and September 2005 the tragedy of Hurricane Katrina captured the attention of the nation and the world. Hurricane Katrina, with more than 1836 victims and $148.8 billion in economic losses was, at the time, the largest, most severe and expensive disaster in the nation’s history (Burby 2006, Lott et al 2013). Cutter and Emrich (2005) stated that the trend of significant natural disasters is upward; that is to say, the number of presidential declared disasters in the 1990s (467) doubled the number declared in the 1980s (242), which was, in turn, fewer than the number in the 2000s (581) (Federal Emergency Management Agency, FEMA). The damaging hurricane seasons of 2004 and 2005 attracted attention to the nation’s capabilities of responding to emergencies and evacuating citizens in a safe and timely manner (National Research Council, 2008). In fact, disasters of large scale have the capability to overwhelm the response capabilities of local and state governments, and to strain capacity at the national level. In 2005, hurricanes Katrina and Rita demonstrated that the mass evacuation of large numbers of people carries considerable risk for the affected population in the disaster areas (Nigg et al, 2006). However, disasters resulting in mass evacuation are not limited to hurricanes or other natural disasters; man-made disasters such as a ‘dirty bomb’ may also call for mass evacuation (Gerber et al, 2006).

Research conducted by Gerber et al. (2006) and Meit et al. (2007, 2008) indicates that rural preparedness planning is an important consideration in mass evacuation, which is necessary if a large number of urban area residents evacuate to rural communities. Rural communities could be quickly overwhelmed by the surge in population because they have insufficient resources to handle the influx. Research has found that man-made disasters, such as a ‘dirty bomb’ incident, are not predictable and generally would cause comparatively little direct loss of life, but such disasters definitely would cause high potential for spontaneous and organized evacuation of urban residents and the subsequent chaos in unprepared receiving areas (Gerber et al 2006; Meit et al 2007, 2008).

This dissertation focuses on the uses of VGI in resource management (principally resource dispatch) in mass evacuation. The major use is to help emergency managers to better understand the resources requirements for sound resource dispatches so that potential evacuation ingress areas,
especially rural communities, can receive real-time resource supplies to avoid the shortage on resources due to evacuee influx during a possible mass evacuation event.

1.1 Disaster and disaster management

As opposed to an emergency, which is defined as “a sudden, unpredictable event that poses a substantial threat to life or property” (NRC 2007, p.12), a disaster is “a calamitous event that overwhelms the impacted community’s ability to respond effectively” (NRC 2007, p.13). This research study differentiates disaster from emergency because disasters, e.g. terrorist attacks, floods or hurricanes, “necessitate a significant evacuation from the central business district or other locations in an urbanized area and in which meeting the associated surge demands and coordination requirements is likely to strain the capacity of a single jurisdiction or transit agency” (NRC 2008, p.21). An emergency does not imply the evacuation is necessary.

Based on Waugh’s (2003) and Kapucu’s (2005) studies, Hossain and Kuti (2010) define a disaster as a “destructive event creating negative social and economic conditions that interrupt the day-to-day activities of a society” (p.755). From both definitions (National Research Council 2007; Hossain and Kuti 2010) disasters present sudden negative impacts to the populations of the affected communities, including “physical injury, loss of life, property damage, physical and emotional hardship, destruction of physical infrastructure, and failure of administrative and operation systems” (NRC 2007, p.14). Disasters can be classified into two categories: natural disasters - events that are the results of natural processes, e.g. earthquake, hurricane, flood etc.; and manmade disasters - events that directly result from human activities, or human developments, e.g. chemical spills and releases, nuclear releases and dirty bombs, terrorism, or war (Alexander, 2000; Burton et al., 1993; Cutter, 2001). It is arguable however whether any disaster is purely ‘natural’ as human activities frequently combine to aggravate an initial natural phenomenon and may even initiate them.

Significant examples of major natural disasters include the Indonesian Tsunami (2004, which killed about 227,898 people around the shores of the Indian Ocean) (US Geological Survey, or USGS), Hurricane Katrina (2005, which killed at least 1,836 people, CNN 2014); and the Wenchuan Earthquake, China of May 12th 2008, which killed about 70,000 people and left about
5 million people homeless, Zhang and Ho 2011). The World Health Organization reported that high population mobility was a contributing factor to the severity and length of the Ebola outbreak in West Africa (WHO 2015; http://www.who.int/csr/disease/ebola/one-year-report/factors/en/). In each of these cases a natural event was exacerbated by human activity, whether that was the concentration of population in risk-prone areas, or the deficiencies of structural features such as levees or buildings. Manmade disasters on the other hand may be the outcome of accidental or deliberate acts. One possible action is the misuse or abuse of technology, including technological failures, such as the Chernobyl Nuclear Disaster (Ukraine, then Ukraine SSR, part of the Soviet Union in 1986). Another possibility for anthropogenic events is terrorism, such as the 9/11 terrorist attacks on the World Trade Center and the Pentagon (USA, 2001). Another type of manmade disaster is a politically-driven social event, such as the Darfur Conflict in Sudan in Africa, which began in 2003 and is still on-going; or the Syrian civil war (2014 – 15, 10 + million displaced persons or refugees) and the nearly 1 million Ukrainians displaced internally by war. An example of a hybrid natural / manmade disaster is the Japanese tsunami of March 2011 followed by the melt-down of the Fukushima Daiichi Nuclear Power Plants.

According to International Federation of Red Cross and Red Crescent Societies’ *World Disaster Report* (International Federation of Red Cross and Red Crescent Societies, or IFRC, 2013), globally 1,149,920 people were killed and 2,167,404,000 people reported affected in disasters between 2003 and 2012 and the total damage to property and infrastructure amounted to more than US$1,571 billion (2012 prices). In the United States alone, there have been 151 weather/climate related disasters from 1980 to 2013, and each disaster caused more than $1 billion of damages/costs. The total losses of the 151 events exceeded US$ 1 trillion (Smith and Katz 2013).

In a response to natural and anthropogenic disasters emergency management is developed and implemented. Emergency management is “the organization and management of resources and responsibilities for dealing with all aspects of emergencies” (NRC 2007, p.14). There are four commonly recognized phases in emergency management: preparedness, response, recovery, and mitigation (Figure 1.1). Those four phases are not distinctly separate from each other, indeed they overlap each other. The activities in the preparedness phase aim to help the community to prepare for possible disaster events, for example, by identifying possible hazards, developing plans for quick response to and rapid recovery from certain types of disaster, and conducting training and
drills. The response phase immediately follows a disaster, during which activities are undertaken straightaway to provide emergency assistance. Response includes such activities as to rescue people that are in danger, to reduce and stabilize life-threatening conditions, to provide life-sustaining aid, and to prevent additional cost to human life, property, and infrastructures. Recovery phase activities are designed to return the affected community to a semblance of normality, or ideally to its pre-disaster or better conditions. The activities include short-term activities, which temporarily restore community services for better response, and long term activities, which aim at rebuilding or reconstructing the community and infrastructure of the affected area. Mitigation phase activities are long-term, including preventing a possible disaster, minimizing damages resulting from disasters, identifying possible hazards and minimizing community vulnerability. While the present study targets response phase activities, it is developed during the preparedness phase and implemented during the response phase.

Figure 1.1: Cycles in Emergency Management: adapted from National Research Council 2007

This study focuses on evacuations that may happen during the response phase, in response to large scale disasters, in which people always need to be relocated from the disaster affected area. The geographic regions involved in relocations could be as small as the directly affected areas, e.g. from buildings to neighborhoods or communities, however under some circumstances, the scale could extend beyond the local scene of a focused disaster to a large geographic region, e.g. multi-community, cross-jurisdiction, regional, or even multi-state). The following section will discuss mass evacuation - involving a large number of human beings relocating within a large geographic region.
1.2 Mass evacuation

Mass evacuation is inevitable as a response to a catastrophic incident that results in “extraordinary levels of mass casualties, damage, or disruption severely affecting the population, infrastructure, environment, economy, national morale, and/or government functions” (FEMA, p.1). For example, 80 percent of the citizens of New Orleans, Louisiana were forced to evacuate as a response to Hurricane Katrina (2005) (FEMA) and 2,400,000 people were evacuated from Houston, Texas in response to Hurricane Rita (Litman 2006). Even though those examples are hurricane-related natural disasters, industrial disasters may also trigger mass evacuation for example the Fukushima Daiichi nuclear disaster (NRC, 2008; Gerber et al 2006; Meit et al 2007).

Goldblatt and Weinisch (2005) illustrate the public evacuation response to a man-made disaster (e.g. a chemical spill or a radiologic release) from a geographical region perspective (Figure 1.2). The central circular area is determined by the location of the release and the wedge by the downwind direction; people in this region need to be evacuated mandatorily. The region immediately surrounding the core and wedge is a region of voluntary evacuation, and the outermost ring depicts a shadow evacuation region, where evacuations may happen sporadically and informally depending on the risk perceived by residents. Voluntary evacuation and shadow evacuation must be incorporated in the evacuation plan as they complicate mass evacuation.

![General map of an evacuation area, from Goldblatt and Weinisch 2005](image)

Figure 1.2: General map of an evacuation area, from Goldblatt and Weinisch 2005

Mass evacuations initially involve a local response. When a catastrophic event strikes, local communities experience it first. Depending on the location of the event (which indicates the size
of population involved) and the type of event, such as a large scale or extremely powerful catastrophic event, local evacuations could quickly evolve to encompass a regional, state, or even national level (Gerber et al., 2006). According to recent studies, more than 50 percent of the evacuees are likely to self-evacuate and could drive more than 150 miles away, meaning that evacuees will not stay in the affected town or shelter-in-place, but will look for shelter in another area (Gerber et al, 2006; Meit et al, 2007; Hess and Gotham, 2007).

Mass evacuations that have cross-jurisdiction impacts need the potential evacuation ingress area to be well planned and prepared for a possible short-term, high-volume population surge. This research project focuses on how to provide real time information to help disaster management professionals dispatch mass evacuation-related resources in the destination areas.

1.3 Geospatial technologies in mass evacuation

GIS and other geospatial technologies have been widely used in all aspects of disaster management (National Research Council 2007), including mass evacuation. The use of GIS could be different during the four phases of disaster management, but mainly include identifying hazard events, monitoring the development and progression of a disaster, providing location information of property or victims, assessing the costs/damages of a disaster, guiding activities, e.g. land uses, evacuation routes, and delineating hazard zones to prepare for potential disasters. The contributions of geospatial technologies include saving lives, and reducing damages to society and infrastructure. The emergency response community relies heavily on the ability to discover and use accurate up-to-date spatial information in order to respond to disasters and other emergency events.

Although geospatial technologies greatly contribute to disaster management, geospatial technologies face some limitations. The National Research Council states that “in this era of heightened requirements for prompt and effective response, rapid access to disparate geospatial information sources is essential” (2007, p.87). Especially in the response phase of emergency management, the limitations of geospatial technologies in disaster management include, but are not limited to, accessing accurate and essential data, disparate and inadequate data quality, slow or incomplete data acquisition and integration and the complete lack of real-time geospatial data (Cutter 2003, ESRI 2000, Farris et al 2006, NRC 2007)).
In emergency management, especially during the response and recovery phases, traditional geospatial technologies face challenges, including that professional geospatial technology is too complicated for non-professionals; and that the geospatial system is not available to emergency personnel during the event. In response to those challenges, distributed geospatial technologies can play key roles in disaster management. Distributed geospatial data and technologies, such as GIS, and other geospatial technologies have “been built and deployed using the standards and software of the Internet” (Tait 2005). The benefits of using distributed geospatial data and technologies arise as a result that many standalone geospatial systems and data sources “can be linked and accessed as a single virtual system” (Tait 2005).

Since the term volunteered geographical information (VGI) was coined (Goodchild 2007), its importance in disaster management, including mass evacuation, has been gradually recognized. Disaster management agencies do not have sufficient staff and resources to acquire and synthesize rapidly changing geographical information in a real-time or near real-time manner. On the other hand, average citizens are frequently equipped with powers of observation coupled with a means (e.g. smart phone, GPS) to record and measure the rapid changes during a catastrophic event, have the abilities to geo-register their observations and to transmit and synthesize their observations into readily understood maps and status reports, and have the capabilities to post their findings to social media websites (Goodchild and Glennon 2010; NRC 2013; Congress Hearing 2013).

The values of volunteered geographical information are made known in various perspectives (Goodchild 2007, Goodchild and Glennon 2010): the potential to be a significant source of scientist’s understanding of the surface of the Earth; the ability to collect information about local activities in various geographic locations that may go reported by the world’s media, and about life at a local level; rich in context, to capture changes in real-time or near real-time, and to be produced more quickly, and in increasing quality.

1.4 Significance of this research

Human population growth and the increasing spread of human activity over many vulnerable areas; as a result, natural disasters have growing effects on lives and property. Although technology enriches human activities, technical disasters and terrorist attacks become more probable through its increased availability. In disaster management, mass evacuation requires human beings to move from disaster-affected areas to shelter in other urban areas or rural communities and egress
from the disaster area tends to receive most attention from researchers. Studies range in focus from building evacuation in a fire emergency (Lo et al 2006), evacuation from small region (such as neighborhoods or towns) in a forest wildfire disaster (Cova et al 2005), and evacuation from a wide region (such as counties or large cities) in a Katrina-type hurricane disaster (Chen 2006, Dixit 2008, Tanaka et al 2007). In those studies, attention was focused on identifying the best evacuation routes, optimizing the evacuation routes, minimizing evacuation time and locating shelter. Such research focuses on the evacuation of evacuees from the disaster scene and placing them in nearby shelters. As revealed by the surveys conducted by Gerber et al. (2006) and Meit et al. (2007) focus is on the condition of an evacuation route or on extending the evacuation route to a greater distance.

In mass evacuation, rural communities along the evacuation routes are likely to play key roles, either as evacuation destinations, as evacuation drive-through nodes, or as both. Those rural communities that are affected by flows of evacuees should have good plans for allocation and distribution of emergency-related resources, including shelter, food and water, gas, health care, volunteers and law enforcement. Emergency managers at every level should be able to access resource-related information easily. Most importantly, they should be able to access real-time resource data, to allocate resources in the dispatch centers to locations of demand; to send requests to high level authorities for extra resources; to dispatch mobile medical resources for emergency; and to dispatch mobile gas tankers to stranded automobiles along the evacuation route. Under most situations, rural communities along the evacuation route have limited everyday resources and stocks in range and volume. Rural communities need the power to command adequate resources to support entering and drive-through evacuees.

As the possibility of mass evacuation becomes greater, rural communities need to be well-prepared for such incidents, but currently there are few effective mechanisms to integrate and deploy efficiently the available resources during disaster planning and response (Kapucu et al, 2007).

1.5 Research goal and objectives

Because emergency managers have to remain abreast of a rapidly changing situation, reliable real-time information is critical in mass evacuation and other disaster/crisis events. Live updates are hard to get in the chaotic situations during a catastrophic event and are necessary to help decision-makers select appropriate choices. Volunteered geographic information (VGI), especially social media-based VGI, is real-time information produced by volunteers, who are in the event area and
have knowledge of their surroundings which can directly benefit disaster management. The overall goal of this research is to improve the ability of users, including managers, responders and average citizens, to make informed decisions, in otherwise chaotic situations during the response phase of a disaster.

To attain this ultimate goal, several research objectives need to be fulfilled. The first objective is to identify the challenges facing the implementation of distributed geospatial technologies in crises and mass evacuation; including the application of real-time information that could benefit disaster/crisis events but that are not always available from official sources. Finding solutions to these challenges is the second objective. As web 2.0-based, user-generated-content, VGI or crowdsourced data are produced by non-specialists from a local perspective in real-time and provide the latest updates of an event. In this study, Tweets from Twitter are chosen for investigation. This objective also includes identifying the tasks that are required for an emergency management model to collect near real-time data. The tasks include identifying the proper API (Application Programming Interface), understanding the Twitter Search API, from several different available types of APIs, to produce the correct design of the Twitter Search API parameters.

After the ideal data and methods used for acquiring the data have been identified, the next objective is to design and implement a web platform for online VGI collection and mapping. The web interface provides the capabilities for users to specify the search parameters, such as search keywords (subject-based), search location, time frame, and others. It provides the capabilities for online mapping the VGI with other traditional GIS data. It also provides the capabilities for online management of the collected VGI, including changing the display, to summarizing spatial and temporal trends, generating heat-maps and other visualizations, and to display the information in different time frames.

Another major objective is to design a server-side system for data saving and distribution. A distributed geodatabase and a web feature service are used for online live data saving; a web service realizes live data dissemination. Additionally, the web service includes a template of the map design.

In order to substantiate the proposed methods, the next objective is to carry out two case studies to evaluate the corresponding features, including the capability to integrate VGI with other
traditional GIS data, the capability to access real-time data, the discovery of related knowledge, the extraction of event and location related data, the symbolization of Tweets, and the evaluation of two significant types of errors that challenge the reliability of VGI.

1.6 Organization of the Dissertation

The dissertation begins with this chapter (chapter one), which introduces the research background, including, mass evacuation to the ingress area with a focus on rural communities; the role of distributed geospatial technology and the emergence of VGI.

The relevant literature is presented and reviewed in three sections in the second chapter: mass evacuation, distributed geospatial technologies, and volunteered geographic information. Beginning with an overview of mass evacuation, the related literature is reviewed from two perspectives: modeling and simulation of emergency-related resource planning and studies of human behavior in mass evacuation. The next segment of the literature review addresses distributed geospatial technologies, and their roles and limitations in mass evacuation and other disaster/crisis events. The emergence of VGI is documented, focusing on the enabling technologies (Web 2.0, social media, neogeography, and crowdsourcing), and their value in mass evacuation and other disaster/crisis events. Additionally, the problems and limitations of using VGI are set out.

The third chapter details the research objectives in depth; setting out the conceptual framework and the research goals. The fourth chapter introduces the research methods for real-time VGI collection and VGI dissemination. The fifth chapter illustrates the potential and value of incorporating VGI in disaster/crisis events using two case-studies and evaluates the advantages and disadvantages of using VGI. Finally, the sixth chapter summarizes the research achievements and gives particulars of future directions uncovered by the study.
Chapter 2: Literature review

2. Literature review

The study of disasters is an extremely broad field involving many disciplines and specialties. The literature underpinning and informing the topics of this dissertation is addressed and evaluated in the following three sections: mass evacuation, distributed GIS, and VGI.

2.1 Mass Evacuation

We live in an age of mass movement. Some level of mass evacuation is inevitable in response to large scale crisis events, such as natural disasters, man-made disasters, technical failures, and other sources of social chaos. Uncontrolled migrations such as from Africa and the Middle East to surrounding regions and Europe dominated the headlines in spring 2015. This study however focuses on more controllable events. In any circumstances, mass evacuation is a complex task and usually needs to be completed in a short time frame relative to the time needed for recovery from the event; a period often measured in hours rather than days. The fundamental question of mass evacuation is how should the finite resources of a transportation system, a shelter system, a food support system, and other necessary services within a community or communities be used to enable the maximum number of people to reach safety and to sustain them once they get there (Bongol, 2005). Efficient and effective evacuation is a complex task that primarily relies on the coordination and utilization of roadway capacity, traffic management equipment, mobile fuel support, shelter, food supply, and available emergency response resources.

Chaotic evacuations from hurricanes in New Orleans and Houston in 2005 prompted local officials across the country to review their plans for emptying their cities in response after a large-scale natural disaster or a terrorist attack (The New York Times, 2005). A number of researchers have studied the process of mass evacuation in different situations, e.g., Wolshon et al. (2006) studied the evacuation plan for Hurricane Katrina; Gerber et al. (2006) studied the possibility of uncontrolled mass evacuation following a terrorist attack in DC area; Meit et al. (2007) studied possible rural preparedness and population surge issues in response to a dirty bomb or pandemic influenza. The following sections will first review several perspectives on mass evacuation, including the modeling and simulation of human behavior and of emergency management-related
resources. The second section discusses findings from the reviews. The third section reviews the roles and the challenges that geospatial technologies face in disaster management.

2.1.1 Modeling and simulation of emergency related resource planning

Technological advances have greatly improved emergency management operations, including mass evacuation. For example, Tufekci (1995), ESRI (2000), Church and Cova (2000), Pine (2004), Farell (2005), Farris et al. (2006) have studied such enhancements in their work. The following section discusses those works, separately and chronologically. After Hurricane Andrew hit Florida in 1992, Tufekci (1995) investigated integrating GIS to build a personal computer-based emergency hurricane evacuation planning module, the Regional Evacuation Modeling System (REMS), which uses simulation and network optimization models to estimate the evacuation time and the traffic flow on a given transportation network.

ESRI discusses the uses of GIS in every aspect of disasters and emergency management, including the preparedness, response, recovery, and mitigation phases. It is possible that evacuation demand may exceed the capacity provided for evacuation at any given time. To handle this problem, Church and Cova (2000) presented an optimization model, named the critical cluster model, which integrates GIS to produce maps of evacuation risk or vulnerability, and to identify small areas or neighborhoods that have a high ratio of population at risk to exit capacity.

Pine (2004) outlined research needs and operational issues associated with the application of technology to emergency management systems. He pointed out that emergency management partners, including public, private, and non-profit agencies, should integrate new technologies, which allow broader approaches to address emergency-related issues in emergency management operations. These technologies include powerful portable computers (mobile and tablet), satellite imaging and sensors, global positioning system (GPS), Geographic Information System (GIS), wired and/or wireless network, and a sensor web.

Based on his study of the coastal area of South Carolina, Farell (2005) proposed 13 alternative solutions to the low capacity of South Carolina’s roadway infrastructure for evacuation; most of the proposed solutions are heavily dependent on computer and information technologies. These thirteen solutions can be categorized into seven groups: expand the use of technology to provide public information; improve communication; establish planned evacuation options; public shelters
and safe houses; disaster reserve units and volunteers; infrastructure improvement; and partnerships. Solutions in the group of expanded use of technology to provide the public with effective information include generating and providing emergency-related information to the public and specialists alike. Those technologies include: (1) computer modeling; (2) intelligent information systems and smart communication devices, which are computer networks, combined with sensors and cameras, to provide real-time information and monitoring (Farell, 2005).

During the response phase of emergency incident decision-making quick access to the population at risk, critical infrastructure, and response resources is increasingly important. Farris et al. (2006) argued that post-9/11 conditions make this requirement more urgent; for example, the need for quick access to an array of geocoded information about properties and population at risk and the distribution and availability of response resources. By combining the needs of geocoded information that are critical in the emergency response phase, Farris et al proposed and created a GIS, which can help organizations and communities prepare for 21st century disaster management.

2.1.2 Studies of human responses in mass evacuation

The anticipation and prediction of human responses has a great influence on mass evacuation. Understanding, predicting, and managing human responses to disaster are essential to any successful evacuation. There are several surveys on human response to catastrophic events, such as terrorist attacks and pandemic influenza. Such surveys include two policy briefs related to rural preparedness and population surge questions from the National Opinion Research Center (NORC) Rural Health Analyses (Meit 2007; Meit et al. 2008), and from Gerber et al (2006). These surveys indicate the importance of the in-depth prior study of candidate ingress areas. Led by Meit, NORC published a brief, “Urban to Rural Evacuation: Planning for Rural Population Surge” (Meit 2008). The article assesses the likelihood of urban evacuation to rural communities and provides recommendations for rural planning and response. This study includes a comprehensive literature review, qualitative analyses based on 17 preparedness experts (6 from the national level, 5 from urban communities, and 6 from rural communities). Also included is a quantitative analyses based on a national survey of 1505 urban residents about their intended actions following a dirty bomb or pandemic influenza scenario and their likely travel distance and destinations. The study found that more than fifty-five percent of the survey respondents explicitly expressed their preference to travel to rural communities should they evacuate. For a given rural community of a population
about 10,000, the number of evacuees, say 2,000, which may reflect only a small percentage (1 percent) of the evacuating urban community of more than 200 thousand population, presents a significant population surge (20 percent). The evacuees that choose to travel to another urban area or rural destination may potentially exhaust resources, such as water, food, and gas, from the communities along their way. Rural communities should consider the potential for depletion of supplies and resources by such transient evacuees. Four of the nine recommendations from the study are listed here: (1) rural communities in close proximity to urban centers should consider the question of urban to rural evacuation among preparedness concerns; (2) rural communities should engage in regional planning to pool resources in preparation for a potential urban evacuation; (3) rural communities should consider the potential consequences not only as an evacuee destination, but also as evacuees travel through their communities on their way to other urban and rural destinations; and (4) tools, training, and technical assistance should be developed to support rural preparedness planning to address issues of population surge following urban evacuation.

The NORC Walsh Center for Rural Health Analyses report, “Spontaneous Evacuation Following a Dirty Bomb or Pandemic Influenza: Highlights from a National Survey of Urban Residents’ Intended Behavior”, summarizes five major findings from a nationally representative survey of urban residents regarding their evacuation intentions (Meit, 2007). The major findings include: (1) while between 77% and 91% of urban residents intend to evacuate by following government suggestions or orders following a dirty bomb or pandemic influenza, close to 40% of urban residents intend to evacuate despite government advice to shelter in place, i.e. not to evacuate; (2) without outside influence from government or other sources, 65% of the urban residents will evacuate following a dirty bomb scenario and 40% of the urban residents will evacuate during a pandemic influenza scenario; (3) in both the dirty bomb scenario and pandemic influenza scenarios, 55% of respondents indicate that they are likely to travel to a rural destination; (4) based on these first three observations, communication and coordination are needed between urban and rural preparedness planners to assure to meet the needs from the evacuees who defy government orders and seek shelter in rural communities; and (5) for effective preparedness planning, estimates of the numbers, travel directions, and characteristics of evacuees are needed.

The research study, “The Potential for an Uncontrolled Mass Evacuation of the DC Metro Area Following a Terrorist Attack” which was led by Gerber (2006) is among the first studies of how
an urban public, would be likely respond to a catastrophic event. While this research is focused on a hypothetical terrorist disaster, the results are applicable in part to natural disasters. Several findings in Gerber’s study indicated that planning for catastrophic events, as part of homeland security preparedness activities by state and local governments in the region’s surrounding urban areas, is both warranted and advisable. Such findings include: (1) a large percent of the survey respondents are very likely or somewhat likely to evacuate; (2) a large percent of the survey respondents state that a catastrophic disaster that can cause mass evacuation is likely to happen; (3) while the possible evacuation destinations are nationally-dispersed, 40 percent of the respondents are likely to evacuate from the disaster event without having a destination in mind.

From the above three studies of mass evacuation, the findings of the human responses in the urban area in responding to catastrophic events indicated that shelter-in-place is not an option for at least 50 percent of the urban residents once an urban area is threatened by a catastrophic event. The evacuees are likely to seek shelter, at least temporarily, in rural communities. Further research on potential ingress areas, especially rural communities, is critical in mass evacuation. Based on the review of human responses of urban residents in responding to catastrophic events, studies have directed attention to further research on the resources and resilience of rural communities.

Human response modeling and behavioral simulation are significant aspects to be considered in mass evacuation planning. Lindell and Prater (2007) have studied human behavior by estimating evacuation-time and Pires (2005) assessed human cognition in an evacuation model. Lindell and Prater (2007) highlighted the computation of evacuation-time estimates (ETEs) for communities threatened by hurricanes. Instead of building an evacuation model from a purely transportation engineering perspective, Lindell and Prater address some fundamental components (e.g. in categories such as, departure timing, destination / evacuation route choice, trip generation) that affect hurricane evacuation time estimations (ETEs), reveal the relationships between these variables and management models, and provide recommendations for future research. Chiu et al. (2007) presented a network transformation and demand specification approach to model no-notice mass evacuation, which is likely to occur right after a large-scale and unexpected disaster event strikes. The approach formulates the joint evacuation destination-route-flow-departure (JEDRFD) problem that occurs during a no-notice mass evacuation into an optimizing, dynamic traffic assignment model. By combining demand modeling techniques, Chui’s model can solve problems
by simultaneously optimizing the location and number of evacuation destinations, traffic assignment, and evacuation departure schedules.

2.1.3 Findings on mass evacuation

From the foregoing review, it would appear that most research study disaster affected areas and their evacuation, whereas studies specifically on ingress areas are still few in number. The necessity for the identification of the candidate ingress areas for mass evacuation has been gradually recognized by more scholars and government entities. In the response to a large-scale mass evacuation triggered by a catastrophic event, the neighboring regions (counties, states) need access to a geospatial technology-based resource management system that focuses on variables associated with evacuee ingress. To summarize, the findings from the literature review include that instead of sheltering in place, most affected people will evacuate, and though final evacuation destinations may be dispersed nationally, a neighboring rural community could become the new destination, albeit temporarily.

2.2 Distributed GIS

Emergency events, especially disasters, have long presented tragic disruptions to human lives, social infrastructure, economy, capital investment and development processes (Mansourian et al. 2006). The fundamental requirement for emergency (disaster) management is reliable, valid, accurate and updated information, and such information must have a spatial component (Amdahel 2001; Cutter 2003). Relative to the size and resources of a rural community, a disaster comprises a large set of rapid-onset incidents. Thus, the occurrence and the outcomes of disasters are uncertain. Also, disasters are dynamic events making the management of disaster-related occurrences and facts not only an organizational challenge, but also a highly technical endeavor (NRC 2007).

In the response-phase of disaster management, decision-makers need to be updated on the latest situation. Additionally, emergency response is time-sensitive, with little allowance for delay in decision-making and response operations, therefore any problem or delay in data collection, access, usage and dissemination has negative impacts on the quality of decision-making and hence the quality of disaster response. With this in mind, it is necessary to utilize appropriate frameworks
2.2.1 Challenges of using GIS in disaster management

As spatial databases become more widely available and the cost of software goes down with the introduction of open systems, more emergency managers are able to acquire GIS expertise and the roles of GIS in disaster management are expanded. GIS-based emergency management has become a feature of state and local government’s emergency management procedures (Zerger and Smith 2002). Despite the advantages of the utilization of GIS in emergency management, there are often problems associated with the use of GIS technologies with respect to data collection, dissemination, access and usage especially during the emergency response phase, which is dynamic and time-sensitive (Jain and McLean 2003; Mansourian et al. 2006). Therefore decision-makers need the system to reflect the real-time situation for them to be able to make prompt and correct responses.

Traditional GIS systems were developed for use on a stand-alone platform, such as a desktop computer. Such architecture limits the ability of a system on one computer to communicate with the same system located on another computer; consequently users at different locations find it difficult to share information with one another. Data updates on one system are not simultaneously synchronized to the other data users. As for an emergency management GIS system/model, a traditional, pc-based GIS system/model is not equipped to accept dynamic input from in-field, first-time responders. Thus in the absence of real-time, up-to-date information the manager in the emergency management control center cannot make appropriate decisions. For example, during an emergency situation, the needs for the logistics of food, fuel, medical aid and shelter, are dynamic. Emergency managers need the dynamic changes of the situation to be updated in the emergency management system and reflected on a map simultaneously for them to dispatch the right responders to the correct location, or to dispatch the appropriate logistic resource to the right demand location. Such situations require the model to be able to accept dynamic input from a distributed user interface, and to generate real-time responses to the in-field responders, or generate concurrent reports for in-house decision-makers to make correct decisions.

Distributed geospatial technology is the combination of geospatial technologies, such as GIS, distributed computing technologies, and the standards of the Internet (Cai, 2005; Walle et al. 2010).
Distributed GIS are defined as “geographic information services provided though the Internet (both wired and wireless networks) and allows people to access geographic information, spatial analytical tools, and GIS-based web services without owning a GIS and data” (Peng and Tsou, 2003, p.1). As computer networks are essential in supporting computing infrastructure and meeting the needs of today’s enterprise operations, marrying GIS with computer networks creates various types of distributed GIS computing system, such as Internet GIS (or Web GIS), wireless GIS, mobile GIS and distributed GIServices.

The primary functions of distributed GIS are the sharing of geographical data and GIS processing tools across organizations and among developers, owners and users. Three levels of sharing from low to high can be identified: (1) Online archive, search, and download, such as data clearinghouse, and data portals (Cai, 2005). A web user can easily search and download the needed data from this kind of Internet GIS application. (2) Interactive map services and mobile navigation services commonly include display, zooming in and out, and querying of spatial information. With these services, web users can interact with many thematic data layers via the provided functions, such as navigation functions (zoom in, zoom out, pan, zoom back, etc.), query functions (query by attributes, located by location, etc.) and data downloading (download the vector data via the service). (3) Online GIS modeling and spatial analysis are also required. Internet GIS sharing is most valuable in helping users have functions that he/she wants without having to own complicated and expensive GIS software. Also, the emergence of geospatial portal technologies has made the access to such kind of distributed geographical information services much simpler through service discovery tools (Maguire and Longley 2005, Tait 2005).

2.2.2 Distributed GIS Architecture

Distributed GIS are constructed using different architectures. Three-tier client-server architecture is most common (Figure 2.1). The first tier is called "the client tier," through which the user makes requests and views geospatial data. The second tier is named the middleware tier, and includes the Web Server and the Server Connectors to bridge the communication between clients and the map servers. The third tier is the data storage tier, which includes the map server and the database server. The three-tier software architecture of web-based GIS provides customizable functions for different mapping applications and scalable implementations for different hardware systems.
The advent of web services as a new paradigm for distributed application development, future generations of distributed GIS will allow users to create task-oriented clients dynamically utilizing interoperable services (Albrecht, 1997). Application developers will have the flexibility to select services based on the requirements of the end-user by choosing the service implementations that are best suited to the task at hand. Additionally, developers of applications for non-GIS users can select the required set of functionality without the need to implement a full GIS capability. As mentioned above, a dynamic collaborative model - “Middleware” - is basic for distributed GIS computing. Distributed GIService plays an important role in the context of distributed GIS computing. Similarly to a general web service, a GIService is “a software application that can be discovered, described and accessed based on a specific description language and protocol, and does not depend on the context or state of other services or supporting platforms” (Daconta, Obrst and Smith 2003).

According to Yang and Tao (2006) distributed Geospatial Information Service (Distributed GIService) refers to an emerging paradigm for offering geospatial data and processing services by using distributed computing technologies. Through on-line access and integration of distributed data and geo-processing services, distributed GIServices provide reduced technology risk, better ability to leverage the value of legacy data and systems, and more efficient information sharing. Development of distributed GIService is largely driven by: 1) the increasing use of service-oriented architecture (SOA); 2) the adoption of interoperable standards for sharing geospatial information resources; and 3) the fast development of enabling distributed computing technologies, (e.g., agent, grid, and peer-to-peer P2P) for transparent and reliable access to computing infrastructure (Yang and Tao, 2006).
2.2.3 Two types of distributed GIS systems

Distributed GIS systems include (but are not limited to) Internet or Web-based GIS, and mobile or field-based GIS. There have been many successful applications of distributed GIS systems. For example, USGS’s National Map (nationalmap.gov) is a good example of an online map viewer (http://nationalmap.gov/viewer.html) providing “public access to high-quality, geospatial data and information from multiple partners to help support decision making by resource managers and the public” (The National Map). ESRI’s ArcPad is an example application of a mobile GIS application. ArcPad is “software for mobile GIS and field mapping applications using handheld and mobile devices. ArcPad provides field-based personnel with the ability to capture, analyze, and display geographic information, without the use of costly and outdated paper field map books” (http://www.esri.com/software/arcgis/arcpad last accessed 4/22/2015).

Mobile GIS is the extension of a Geographical Information System from the office into the field. Mobile GIS may be based on smart phones, toughened laptops or tablet PCs and have some or all of the functions of a desktop GIS, and is heavily reliant upon wireless technologies to access remotely stored data. There are smaller PDA-based systems, which have a more limited and specific set of functions. These PDA-base systems allow the capture, manipulation and storage of data in remote locations, which has the advantage of real time updating of central databases, thus removing the need for storage of multiple copies of data and lessening data replication. Mobile GIS has been used in field data collection. These systems use wireless communication systems, mobile computers, and positioning systems to achieve the ability to access, process, and display geospatial information in the field. ESRI developed ArcPad for mobile GIS systems. ArcPad can be installed on a GPS receiver or other Windows supported PDA (see http://www.esri.com/software/arcgis/arcpad for system requirements). ArcPad supports GPS, and a digital camera integrated for geo-tagged field navigation and documentation. For example, using ArcPad, the North Carolina Division of Public Health developed a mobile GIS application to speed disaster relief (ESRI, 2007). This application demonstrated that a mobile GIS has the ability to collect information and mobilize aid following natural disasters. USDA has developed an ArcPad-based field data recorder software, PedonCE, for describing soils. PedonCE provides the field soil scientist with a tool to record quality soil descriptions electronically in the field (USDA, 2005).
Internet GIS is the principal distributed GIS computing system, and is expected to change the level of participation by the public in geospatial related activities and information. With the availability of Internet GIS, the public has easy access to GIS via smart phones and tablets to learn about the location of activities and information. Internet GIS has been widely used in a variety of different applications, and allows the public to interact with GIS data and maps over the Internet without owning GIS software programs, which are too professional, complicated, and expensive for general users. For example, the WV GIS Technical Center developed an Internet GIS application, the WV Flood Hazard Determination Map, to help “floodplain managers, insurance agents, developers, real estate agents, local planners and citizens attempting make informed decisions about the degree of flood risk faced by particular pieces of property”. The major functionality of such application is to help the public to find how their location relates to flood hazards (http://www.mapwv.gov/flood/ last accessed 04/22/2015). Huang (2003) developed an Internet-based dynamic and interactive visualization application to view DEM data and a sequence of rainfall and potential evapotranspiration data, and to predicate the pattern of soil water deficit, as well as the resulting stream discharges. Tsou (2004) integrated image processing tools into web-based GIS for environmental monitoring and natural resource management. Abdalla and Tao (2005) introduced an Internet GIS approach for earthquake disaster modeling and visualization. Immediately following the 5/12 Wenchuan Earthquake (2008), ESRI China developed a series of web-based GIS applications for the general public to better view the situation of the earthquake, for the decision-makers to better monitor the damages and to send out the help teams (ESRI 2008).

2.2.4 Distributed GIS in mass evacuation

An Internet GIS (or web GIS) can be a good tool for facilitating disaster management due to the need for a high degree of interaction between decision-makers in the disaster management community, particularly during disaster response. Therefore, an Internet GIS using spatial data infrastructure (SDI) can facilitate disaster management by providing a better way of spatial data collection, access, management and usage. Mansourian, et al. (2006) demonstrate an overall picture of a web-based system to facilitate disaster management using an SDI environment (see Figure 2.2).
As Figure 2.2 shows, each of the organizations involved in responding to an emergency is responsible for updating one or more datasets within the Emergency Operation Center (EOC) database before, during and after a disaster. Each organization can then utilize required datasets from the EOC database for their own use. This example demonstrates the important concept of partnerships in producing and updating datasets, as well as the concept of sharing datasets, which allows each organization to work on a common database. All of the organizations are connected to the EOC database through internet or intranet, which is the network that organizations use for data sharing and access their own required datasets. Field workers of each and every organization access emergency-related spatial data or non-spatial data via the client-side of the distributed GIS system. The field workers update the current situation via wireless GIS, such as PDAs, so the dynamic situation of the emergency is synchronized in the distributed system simultaneously. In the EOC the decision-makers receive real-time updates from the server side of the Internet GIS system. As mentioned above, ESRI developed an Internet GIS application for the 5/12 Wenchuan Earthquake (2008). With ArcInfo, ArcGIS Spatial Analyst, ArcGIS Tracking Analyst, ArcIMS, ArcSDE, and ArcGIS Server, this application helped to track follow-up incidents and their extent. Staff used GPS-enabled phones and devices wirelessly to transmit data tied to locations back to
the primary epidemic tracking system. Officers in the EOC receive the real-time information from
the in-field workers. Users open up their desktop computers, wireless laptop computers, web-
enabled phones, or other mobile devices to view spatial data as it is being collected and published
by the China Earthquake Administration (ESRI 2008).

2.2.4.1 Distributed GIS-based data sharing in disaster management

Many different government agencies, non-profit organizations and private sectors are involved in
emergency management and they use a diverse range of spatial data. Distributed GIS provides
many benefits for emergency management, especially in data sharing (Snoeren et al, 2007). For
effective emergency management, these data should be exchanged ubiquitously, and the data
exchange should be in a very fast and secure environment. The benefits of distributed GIS-based
data sharing in emergency management can be summarized as below:

(1) Disaster management requires large quantities of current and accurate digital data, yet it is not
necessary for each organization to try to possess all the data in their own GIS. In distributed
GIS-based emergency management system, the organizations involved in an emergency
(government, private sectors, universities, and other research institutes) can save significant
time, money and effort when they share the burden of data collection and maintenance. The
time savings in are particularly important for emergency management during the disaster
response phase, and may make a significant difference in reducing the number of victims and
the extent of hazardous areas.

(2) Shared data allows more users to work with data and consequently estimate the quality of that
data. In the preparedness phase, an increased number of users are able to find and correct errors
in the data. For the response phase, it is also important that many people are familiar with the
spatial data sets. Familiarity with the data reduces the number of mistakes made because of a
lack of time or haste imposed by the disaster.

(3) Sharing data of common interest reduces duplication and enables emergency forces to defray
some of the costs of producing and maintaining the data. Mechanisms to facilitate the use and
exchange of data are a major justification for developing and expanding a distribute GIS
system (Williamson et al, 2003).
Sharing of spatial information in the preparedness phase helps the policy makers develop high quality contingency plans because a complete overview of the availability and status of spatial data provides better understanding.

### 2.2.4.2 Uses of distributed GIS system in disaster management

In order to address problems in data collection, access, and dissemination in disaster management, Mansourian et al. have developed an SDI conceptual model and a prototype web-based system to facilitate geospatial data sharing based on their on-going disaster management research project (2006). Their pilot project illustrated that a distributed GIS system is useful for effective and efficient disaster management, especially in the response phase, from the perspectives of reducing the response time so that first responders can be dispatched quickly and avoiding chaos in coordination among different agencies.

Cai (2005) explored the potential of extending distributed GIS to support geo-collaborative crisis management. He studied a proof-of-concept system architecture that can integrate distributed web service-based computer resource, common GIS web service standards, and activity centered system design. Barbosa et al. (2010) studied the uses of a GeoWeb GIS prioritization model-based forest fire alert system for Carajas National Forest in Amazonian Brazil. The purpose of the project is to monitor potential hotspots of forest fire, so that by verifying the hotspots in the field, the forest managers prioritize the treatment of hotspots. Thus, the allocation of fire-fighting resources can be improved. For easy, valid, and prompt information and tools sharing and utilization among different firefighting forces, Kalabokidis et al. (2013) proposed a Web GIS-based virtual system for fire management professionals to locate fire service vehicles and other resource online and in real-time. The proposed system, which integrates real-time images that are captured by distributed cameras and real-time weather data from remote automatic weather stations and the weather forecasting system, can be used for fire prevention and early warning. With this system, end-users can design an operational plan to encompass the forest fire and identify the best way to put the fire out efficiently.

Karnatak et al. (2012) studied the uses of a GeoWeb application in flood in the Assam State of India in 2010. The GeoWeb application, which can consume various data and information service over the internet (either third parties sources or the central server) and real-time ground observation data collected through mobile devices, is for end users to use and to participate during a flood
event. Rathore et al. (2010) studied a Web GIS application for dams and drought in India. In their study, the precipitation, water resources, and drought information included locational information. Through the Web GIS application, water resource planners and engineers, who need information on dams and diversions, retrieve the related information to make decisions on how to distribute the water resource correctly.

By using collaborative open source geospatial tools and maps, Polliono et al. (2012) studied how distributed GIS system supports the response planning in disaster earthquake events. The researchers investigated and developed a simple geospatial methodology and related collaborative geospatial tools to predict and map vulnerability to seismic hazard to support the earthquake response planning. The proposed geospatial approach improves the effectiveness of the disaster monitoring, management and awareness. It facilitates the analysis of potential hazards related to a given earthquake magnitude. A further advantage of the system is that the additional data can be integrated immediately once new information becomes available. The example demonstrates that VGI data and collaborative tools can contribute to a better emergency plan and providing fundamental information for real-time response when disaster occurs. Motivated by the terrorist attack on 9/11, Kwan and Lee (2005) examined the potential of using real-time 3D GIS to develop and implement a wireless GIS based intelligent system, which can facilitate quick emergency response to terrorist attack on multi-level buildings/structures.

Assilzadeh et al. (2010), Xie et al. (2012), and Yu et al. (2012) studied the uses of Web GIS system for landslide catastrophes and how to reduce the possible disaster risks. Their study proposed a Web GIS-based decision support system to help emergency management teams better prepare for, and respond to, landslide disaster, including early warning, risk and vulnerability analyses, and critical infrastructure damage assessment.

In the case of earthquakes, Wang et al. (2012) studied the importance of building an Internet GIS-based computational infrastructure to support the quick processing and distributing of huge amounts of geospatial data, which were not handled efficiently in the case of the Haiti earthquake (2009) and Japan 3/11 earthquake, tsunami, and nuclear crisis (2011). The Indian Ocean tsunami (2004) and Hurricane Katrina (2005) motivated Laituri and Kodrich (2008) to examine the advantages of integrating remote sensing, Internet GIS, and GPS for disaster manage management. They argued that traditional systems, including GIS system, were not good enough to handle
disasters of the magnitude as Hurricane Katrina. The authors emphasized the importance of using Internet GIS, which is core of the proposed system, the Online Disaster Response Community, in disaster management. The research realized the importance of mobile/wireless GIS, which could be used for real-time data capture. They argued that the distributed GIS system should include the concepts of citizen as sensor, public participation GIS, and online mapping for decision-making (Laituri and Kodrich 2008).

2.2.5 Distributed GIS: Summary

A distributed GIS system shows its power in disaster management situations. The benefits of using distributed GIS in disaster management can be seen from several perspectives. First, the general public can easily access the implemented system by using desktop computers, laptops, mobile device, and smart phones, through ubiquitous Internet connections. The latest updates from officials are accessed easily by the general public through the client side the distributed system so that the affected populations are able to make the right decisions on how best to evacuate from the disaster-affected zones. Second, emergency responders can access the distributed system in the field in real-time. As emergency professionals, they can post their findings back to the servers, so that emergency managers can view the updates in real-time or near real-time. Third, a distributed GIS system provides emergency managers with the overall picture of a disaster event. They can dispatch related resources and staffs promptly and efficiently. Emergency managers identify the potential hazards and risks more effectively enabling preparations to be arranged such that the affected populations are informed and relocated quickly.

With the availability of distributed GIS in disaster management and the development of Spatial Data Infrastructure (SDI), emergency management professionals have easy access to geospatial tools and data-driven disaster management. Correct decisions during a disaster event require professionals to have real-time data in simulation and modeling tools through distributed GIS systems. On the other hand, the rise of volunteered geographic information (VGI), from average citizens, provides another, possibly important, real-time geospatial data source from the community. VGI could be provided by volunteers in real-time manner, which is critical in disaster management. The next section discusses the rise of VGI and its possible uses in disaster management.
2.3 Volunteered Geographic Information

The 2005 Indian Ocean Tsunami, Hurricane Katrina (2005), the 5/12 Wenchuan Earthquake, the 2009 Haiti Earthquake (Zook et al. 2010), and the Japan 3/11 Earthquake, Tsunami, and Nuclear Crisis (Krausmann and Cruz 2013) showed the huge devastation of disasters on human population and infrastructure and the extremely high cost of recovery of affected communities. But, at the same time, we have seen the power of modern technologies helping the response to disasters. For example, in the immediate aftermath of Hurricane Katrina (2005), high-resolution images which showed the detailed impacts of the disaster were readily available on Google Earth, which was released by Google in early 2005 (National Research Council, 2007). By using Google Earth, people from the other locations could explore and have a better understanding of the impacts due to Hurricane Katrina, such as the levee breaks in New Orleans, the extent of flooding and the damage to the buildings.

The major technology behind these phenomena is Web 2.0. Web 2.0 was first coined in 1999 by DiNucci, who wrote

“The Web we know now, which loads into a browser window in essentially static screenfuls, is only an embryo of the Web to come. The first glimmerings of Web 2.0 are beginning to appear, and we are just starting to see how that embryo might develop. The Web will be understood not as screenfuls of text and graphics but as a transport mechanism, the ether through which interactivity happens. It will [...] appear on your computer screen, [...] on your TV set [...] your car dashboard [...] your cell phone [...] hand-held game machines [...] maybe even your microwave oven”.


It was closely associated Tim O’Reilly after the O’Reilly Media Web 2.0 conference held in late 2004 (see http://www.web2con.com/web2con for example). With Web 2.0, users are not only consumers of websites' information, but also information producers through a website. That means, users can post their observations of a phenomenon, opinions about a given situation, the findings of an event, and understanding about a problem. Those posts are also called user-generated-content (UGC). Most importantly, UGC on a website are accessible to other users of the same website simultaneously. This key feature is critical in disaster management, especially in the response phase. With the popularity of location based service (LBS) and GPS devices, the user-generated-contents could have geographic information embedded. This kind of UGC is also named as the

Academic interest in UGGC started at the Workshop on Volunteered Geographic Information, which was held at Santa Barbara, CA on December 13-14, 2007, in which a group of international experts discussed the state of knowledge of VGI and developed a research agenda for VGI. A special issue of GeoJournal appeared in 2008, containing VGI research papers from the Workshop. Elwood, Goodchild, and Sui et al. (2007) defined VGI as “geographic information acquired and made available to others through the voluntary activity of individuals or groups, with the intent of providing information about the geographic world” (p. 5).

Scholars are studying the uses of VGI in emergency (disaster) management. For example, in the case of a series of wildfires that happened in Santa Barbara, California before and during 2010, Goodchild and Glennon (2010) examined the key issues associated with VGI and the potential role in disaster management. This section will first study VGI and its origin, then study the uses of VGI in disaster management. The third part will introduce the issues of using VGI in disaster management. The final part will discuss two types of errors as related to disaster management.

2.3.1 Key concepts of VGI

Four key concepts behind VGI, or user-generated-geographic-content, include Web 2.0, social media, neogeography, and crowdsourcing. Web 2.0 indicates the second generation of the World Wide Web, which focuses on the ability for users to collaborate and share information online. On a Web 2.0 powered website, e.g., users of Wiki applications are not only information consumers but also information producers. For example, a user can check out the information available as do other normal website users. A user can edit the existing posts (this capability might require certain type of privileges) and add new content, which becomes available to other users through the website. In short, instead of one-way information consumption, Web 2.0 enables a user’s two-way information communication role: information consumption and information production.

WebAppRater.com lists key features of Web 2.0. First, the traditional Web uses a pre-defined classification of information like category & sub category. Without being restrained by the existing classification framework, Web 2.0 allows users to create a classification free of any prearrangement of information. This is also known as social tagging, e.g., the photo sharing site
Flickr. Second, unlike the traditional Web-based static pages, Web 2.0 uses Ajax (Asynchronous JavaScript + XML) to present a dynamic, rich experience to users. AJAX allows a users’ post to be seen right away on the host website. Third, a user of a traditional website is only a consumer of the information, which is often provided by the site owner, of the website. A Web 2.0 user can also contributes to the content by means of posting, evaluation, review, and commenting. The typical example is Wikipedia.com. Users can add new article to the website or edit an existing article. Fourthly, the contents are solely provided by the host of the traditional web-based site, but the Web 2.0-based website permits the users to participate in content sourcing, which is also known as crowd sourcing, e.g., Wikipedia, Twitter, YouTube. A stricter sense of crowdsourcing applies to the solicitation of information from a massive pool of users, to aggregate data and intelligence. The popularity of Web 2.0, along with the increasing use of blogs, wikis, and social media, has led many in academia and business to engage in applying Web 2.0 technologies in existing concepts and fields of study.

In short, Web 2.0 includes three parts: (1) a rich internet application (RIA), which can bring the desktop experience to the browser on the client side; (2) a web-oriented architecture (WOA), which defines how Web 2.0 applications expose their functionalities so that other websites or applications can use those functionalities; and (3) the social web, which defines how Web 2.0 interacts with end users and makes end-users an integral part of a website. According to Ahlqvist et al. (2008), Web 2.0 is one of the three key elements (i.e., content, communities, and Web 2.0) of social media. Due to the official emergence of Web 2.0 in 2004, social media has gained broad popularity in an exponential manner. User generated content (UGC) is the way in which users make uses of social media, including the various forms of media content created by the general public and are accessible by the general public.

Kaplan and Haenlein (2010, p. 61) defined social media as “a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of user-generated content”. Social media refers to the interaction among the users of an online social network system and the information creation, exchanging, and commenting over the social network system (Ahlqvist et al. 2008). Social media plays more and more important roles in everyday life, transforming the way we live and do business (Qualman 2013). Brenner and Smith (2013) estimated that 72 percent of online U.S. adults use social networking sites (although
the PEW Research Center reports only 65% in 2015) and 18 percent of them are Twitter users. (http://www.pewinternet.org/2015/10/08/social-networking-usage-2005-2015/). In any event social media has undergone a 10 fold increase in the past decade. The concepts of social media and research about social media have attracted more and more academic, business and industrial interest e.g., computer science, emergency management, political science, social science, etc. In the community of GIScience, numerous scholars (e.g., Goodchild 2007, 2009, and 2010; Elwood, 2008a, 2008b, and 2011, Sui, 2011, and Sui, Elwood and Goodchild, 2013) have recognized the importance of social media.

Neogeography, coined by Turner (2006) in his book “Introduction to Neogeography”, is another key concept associated with VGI. Put simply, Neogeography means “new geography”. Turner states that “Neogeography combines the complex techniques of cartography and GIS and places them within reach of users and developers” (Turner, 2006, p. 1). With the help of a set of techniques and tools that are not in the realm of traditional GIS, neogeographers can use mapping API (Application Programming Interface), such as Google Maps API, Yahoo Maps API, ESRI JavaScript/Flex/Silverlight API, to make maps, and can geotag their photos and post texts to map their journals or special events.

As with Web 2.0, neogeography blurs the distinction between data and map producers, communicators, and consumers of geographic information (Goodchild 2009). The core element of neogeography is how the general public use and create their own maps, not by using complicated and expensive GIS software packages, which are designed for geoprofessionals, e.g., ESRI’s ArcGIS, but by using public accessible toolsets, e.g., mapping API, and platforms, e.g., Google Earth and Mappedia. Neogeography can help citizens share location information with others, shape the context of an event, and conveying understanding of an event through knowledge of place. Within the realm of neogeography, VGI witnesses the increasing interest in using “the Web to create, assemble, and disseminate geographic information provided voluntarily by individuals” (Goodchild 2007, p. 211).

In the June 2006 Wired magazine article, “The Rise of Crowdsourcing”, Howe described the emergence of crowdsourcing, a term that was coined by him and Mark Robinson (Howe, 2006). After they studied 40 definitions of crowdsourcing, Estellés-Arolas and Ladrón-de-Guevara (2010) developed an integrating definition of crowdsourcing as “a type of participative online activity in
which an individual, an institution, a non-profit organization, or company proposes to a group of individuals of varying knowledge, heterogeneity, and number, via a flexible open call, the voluntary undertaking of a task” (p. 197). Crowd sourcing derives from the analogy outsourcing to the crowd (Howe 2006). Crowdsourcing indicates that data are solicited, acquired, and disseminated by large groups of citizens who might not be trained experts. Sometimes the terms crowdsourced geospatial data and volunteered geographic information are used interchangeably, but crowdsourcing has the additional implication of a solicitation or request for information, and is often associated with the business community. A good example of geospatial crowdsourced data application is WAZE, which solicits traffic information from subscribers.

In the GIScience community, crowdsourcing geospatial data means “[geospatially related] data acquisition by large and diverse groups of people, who in many cases are not trained … and who do not have special computer knowledge, using web technology” (Heipke 2010, p. 551). The subsequent tasks for using crowdsourcing in GIScience community include (and are not limited to): integrating crowdsourcing geospatial data into the current National Spatial Data Infrastructure (NSDI), subsequently after integration, processing the collected crowdsourced geospatial data to generate further information in the context of geospatial data simulation and modelling.

2.3.2 Two types of production of geographic information

There are two approaches to producing geographic information: one is top-down and the other is bottom-up. In the top-down approach data is produced and disseminated through a hierarchy of users by higher authorities. Accordingly, users at the lower levels receive data in the role of pure consumers. Based on the top-down format, geographic information is produced by national official mapping agencies or other authorized organizations and the generation of information follows elaborate protocols, standards and specifications. The related authorities employ well-trained agents and professionals to assure data quality and control. Geographic data and information produced in this manner is called authoritative or authorized geographic information (Goodchild 2007). Goodchild and Glennonn (2010) have argued that the new geospatial technologies, including GIS, have done very little to change the production forms of the geographic information in maps or atlases; the main change being that paper format maps have been transformed into to digital format for dissemination. Quintessential examples are the production of Digital Raster Graphic (DRG) copies of USGS topographic maps and, the more recent, background map services.
Two factors limit top-down geographic information production: the necessary expertise in map-making and data collection, and the associated high costs of map-making or digital data production. The first factor dictates that only experts can make maps, and the second that digital data are very costly, when hardware, software and labor are considered.

Web 2.0 promotes another approach to geographic information generation; the bottom-up approach. In the top-down approach, data producing and dissemination takes considerable time. There is limited real-time data availability, although some authorities such as the British Ordnance Survey (OS) have increased turnaround speeds of digital topographic mapping, for example in the OS Mastermap© product. Conversely, in the bottom-up approach data is acquired and disseminated by individuals at a lower, or grass-roots level, which can subsequently be aggregated as crowdsourced data. Users at different levels can access the crowdsourced data in real-time or with a very short-delay, depending on how the data are processed and what technologies are used. The geographical information generated through the bottom-up approach is termed user-generated-geographic-content (UGGC), or VGI. This kind of geographic information is considered to be asserted geographic information since it has no citations, metadata, references, or other accreditations (Goodchild 2007).

2.3.3 Characteristics of VGI

This section introduces some essential characteristics of VGI, including the methods used to capture VGI, the degrees of participant willingness in data contribution, and the degree of interaction and involvement in the process of data collection.

Traditionally, geographic information was produced by well-trained professionals with the help of sophisticated equipment. Both training and advanced devices are important factors in capturing geographic information. Without costly equipment and expertise, the geospatial professional cannot capture spatial information to the necessary specifications. As technologies advance, the cost of spatial information capturing devices, i.e. global positioning systems (GPS), drops greatly. Also, it is possible to integrate GPS capability into mobile devices, e.g., smart phones, tablets, and other PDAs. A second way of integrating location information is through Web 2.0-based mapping applications, such as Google Map/Google Earth, Bing Map, Yahoo Map, and MapQuest. By using those mapping applications, users can easily pin-point locations of interest by either panning or
using geocoding, a process that locates a place name or an address on map. The third way of recording location information is via cellular phone networks, which permit triangulation methods on cell tower locations. On a Web 2.0-enabled website, users can upload/post their own contents that have embedded location information. To reiterate: the general public has become not only geographic information consumers, but also geographic information producers. Most importantly, the geographic information can be consumed by other users of the same website in a real-time or near real-time manner.

Based on the definition of VGI, the word “volunteered” indicates assumed willingness in data transactions. In reality, in the domain of VGI, data contribution is not always voluntary and involves four degrees of willingness. The first degree of willingness includes information that is knowingly produced by volunteers on purpose and is freely available to everyone. At the second degree of willingness, people are asked to participate in data collection for a specific event or a special occasion, with a predefined set of criteria or within a limited duration or geographic extent (Seeger 2008). For example, in the event of Haiti Earthquake, in order to provide a complete Haiti street network for quick response and resource dispatch, a call for voluntary participation for updating Haiti street network was answered by Ushahidi (Ushahidi.com), a loose organization of civilians enabled by many professionals using OpenStreetMap. This type of data contribution is known as facilitated VGI (f-VGI, Seeger 2008). The third degree of data contribution willingness refers to the contributed data for only certain recipients in a designated circle of friends (Elwood 2008a). This degree is common in social network applications. Finally, the fourth level of willingness happens when contributors are mainly unaware that they are contributing (Elwood, 2008a). For instance, when using GPS-enabled cameras to take pictures and post them on Flickr, an online social picture sharing network, some users are unaware that the pictures’ location is embedded in the image files, as a geotag, and could be used by unknown individuals for malicious or other purposes. Much geotagging occurs without user awareness of the full nature of the transaction.

Traditionally, data collection is a one-way process, meaning that data producers and data consumers are distinct. In the VGI domain, data collection and dissemination is not one-way but two-way or multi-way. In two-way data collection, while data/information collected by other users is consumed, users can produce spatial-information enabled data and post those data to a website.
On the other hand, other users can access the posted data in a real-time or near real-time manner. As Goodchild (2008) points out, n-way data collection is now very common, since producers and users are often synonymous in the VGI context. An n-way interaction in data collection means that users can cooperate with each other by commenting or editing each other’s contribution. In the case of MapMedia, a user posts a piece of information, which other users can see simultaneously. If they think that editing is necessary, every user, including the information’s original producer, can edit this piece of information. Therefore, unlike one-way data collection, two-way and n-way are useful data collection interactions in VGI collection and dissemination.

2.3.4 VGI continuum and types of VGI

Based on Deparday’s (2010) study, the VGI continuum has two levels. At the base, the first level is built on two purposes for the production of VGI (Elwood 2008b). The first purpose of VGI production is to verify and update official data that is incomplete, out-of-date, or incorrect. Also, it is possible that VGI could be used to completely replace or duplicate existing official data. In this circumstance, VGI has the same structure as the official data, but is produced by volunteers, who are willing and able to follow guidance, rules and standards in data collection process, OpenStreetMap is a good example. The second major purpose of VGI production is to create a completely new form of knowledge, such as multimedia (Elwood 2008b). In the VGI continuum, Deparday (2010) defines VGI based on the first purpose as “Conventional GIS knowledge” and VGI based on the second purpose as “Unconventional spatial knowledge”.

The second level of VGI continuum is based on the fundamental difference of the nature of the contributed knowledge. Three types of VGI content are scientific knowledge, local knowledge, and personal knowledge. Scientific knowledge refers to knowledge that is produced by trained professional volunteers. VGI in this category has one or more of the following features: (1) the data is scientific in nature; (2) the data is the description of an event or earth surface phenomenon, such as soil classification; and (3) the data is produced in a reproducible, manner. Scientific knowledge falls into the conventional GIS knowledge at the first level of the VGI continuum. Local knowledge refers to knowledge that is created by volunteers from their local perspective and experience. Such knowledge could be conventional GIS knowledge, qualitative observations or opinions that are produced by volunteers from their personal perspective. In this research, the data to be examined are Tweets from Twitter. As social media-based VGI, spatial information enclosed
Tweets (or geo-tagged Tweets) is volunteers’ personal observation of a local event or phenomenon and is considered to be unconventional spatial knowledge.

### 2.3.5 Value of VGI in emergency management

Web 2.0-based geospatial technologies have been used as key tools in crisis management. The importance of Web 2.0-promoted VGI in crisis management has been recognized by scholars and other practitioners and experts. In the case of wildfire for example, Goodchild and Glennon (2010) have studied the issues and advantages of using VGI in disaster response; Goolsby (2010) studied the uses of social media to generate community crisis maps; Longueville et al. (2010) studied VGI in the context of a forest fire scenario; Roche et al. (2013) studied various issues and perspectives of using VGI in crisis management; Schade et al. (2011) studied VGI in wild fire management in France in 2010; and Zook et al. (2010) assessed the use of VGI in the Haiti Earthquake.

In the emergency management community, the use of VGI, especially social media-based VGI gradually gained practitioners’ attention at first, then with accelerating interest. On June 4th, 2013 and July 9th, 2013, the US house Committee on Legislative and Oversight Activities of the Committee on Homeland Security held hearings to examine the impact of social media and new technology on emergency response, recovery, and preparedness (US Gov. Printing Office, 2014, House Report 113-314). The witnesses were emergency management professionals from public sectors, such as FEMA, American Red Cross, state emergency management offices, local emergency management offices, and from the private sector, Google and Palantir, and others that contribute to emergency management. By using Hurricane Sandy (2012), Oklahoma tornado (2013), and other disaster events as examples, the witnesses’ testimonies explicitly expressed that: (1) social media provide real-time or near real-time information for emergency management; (2) VGI or more specially, social media-based VGI have been widely used in emergency management; and (3) emergency responders are some of the biggest beneficiaries of social media and social media-based data. Agencies are building social media monitoring centers which provide real-time information to emergency management professional. For example, on March 7, 2012, American Red Cross and Dell opened the Digital Operation Center, which is claimed to be the first social media monitoring facility in the United States (http://www.redcross.org/news/press-release/The-American-Red-Cross-and-Dell-Launch-First-Of-Its-Kind-Social-Media-Digital-Operations-Center-for-Humanitarian-Relief, last accessed 10/20/2015).
In disaster management, the importance of VGI can be seen from the following points of view. First, VGI is real or near-real time information. The traditional way of collecting and disseminating geospatial information takes a long time, including collecting data from the field, processing data in the office and disseminating data to the end users. On the contrary, VGI can be collected, processed and disseminated in a real or near-real time manner. In an event, there are nearly always citizens in the field with mobile devices. Those citizens have the capability to upload their findings or observations to their interested social media websites, e.g., Twitter or other specific websites, e.g., OpenStreetMap. These uploaded data can be accessed by other users of the same platforms simultaneously. This feature is particularly critical in disaster management, which requires live information about an event.

Second, VGI incorporates knowledge of a location from local perspectives. During an emergency event, it is unrealistic for official agencies to deploy a sufficient number of agents to the event location specifically to collect data. First, there is an inevitable delay for officials to grasp that unexpected events are happening. Second, the officials do not have enough field workers who can be sent to the field for data collection. Conversely, average citizens who happen to be at the event location can observe it at first hand and can become sensors. Compared with the official’s comprehension of the situation, there is no delay (or very small delay) for an at-the-scene citizen to observe and report about an event. The data collected by those volunteers is from the individual’s personal perspective and is based on their understanding. With the advances in technologies GPS-enabled mobile devices, including smart phones, tablets, and GPS data collectors, volunteers have the capability to post their observations to Web 2.0-based applications, such as Twitter, Facebook, and Flickr.

Third, VGI can be a significant data source. In the domain of VGI, average citizens have the potential to be sensors (Goodchild 2007) for data production. Potentially, the 5 billion human beings are 5 billion sensors. For a specific social media, especially the most popular ones such as Twitter, the users who create/edit the posts could be in huge numbers creating vast amounts of information. As of February 2012, Twitter reported more than 500 million registered users with 100 million monthly active users world-wide, and as of March 2011, more than 200 million Tweets per month were generated on Twitter. Twitter is just one of many popular platforms; other platforms, such as Facebook, Flickr, Google+, MapPedia, hold high volumes of VGI.
Fourthly, VGI is rich in context. Traditional geographic information is in attribute (text) or spatial data (geometry) format. In the VGI domain, geographic information could be in other formats, such as images, audio, video. For example, photographs captured using GPS-enabled cameras carry location information; or users can specify the location by using Web 2.0-based mapping application when they upload the pictures, such as Flickr, or Google Map/Earth. Another example, users can upload their geotagged voice videos to YouTube for other users to access them.

2.3.6 Issues with VGI

VGI are produced by volunteers. Under some situations, volunteers do not realize their participation in data production. The ways of data production and dissemination of VGI lack quality control and metadata. VGI does not typically include traditional measures of accuracy or inaccuracy. Modes of quality control in VGI are entirely different from those used in traditional production of geographic information, in large part because these data are produced through crowdsourcing. And, there is no valid way to handle/control VGI sources. The quality of VGI is more appropriately measured by the number of peers who have reviewed or edited its content, a principle called Linus’s Law (Elwood et al. 2011). The issues that associated with VGI include the data quality, the data credibility, the data heterogeneity, and the privacy issue.

The first issue discussed here is the quality issue. Quality issue is also called the accuracy issue or reliability issue, or the uncertainty issue. Geographic information is subject to measurement error, loss of detail, vagueness of definition, and many other sources of uncertainty and inaccuracy (Zhang and Goodchild 2002). The quality issue of VGI can be discussed from two perspectives: one is the quality of spatial information - the location information; the other is the data producer’s description of an event. Different from other user generated content, VGI has a location component. As the accuracy of acquiring methods improves, the location accuracy increases accordingly. For example, as GPS accuracy increases, the location accuracy embedded in VGI collected by users who use GPS or GPS-enabled mobile devices increases automatically. Another example is location information from geocoding service. As the street network data or address data that used in geocoding service become more and more accurate, the geocoding result will become more and more accurate. Thus, web mapping applications that use geocoding service provide users the more accurate location information.
As they are exposed to certain environmental conditions more frequently, data producers’ knowledge about that environment increases. The description of a phenomenon from data producers becomes more accurate with time. Also, as the computer technology, the software in this case, becomes more advanced, the software used to parse the users’ description becomes much smarter. It can extract the desired information from the description by using context or other reference information.

The second issue to be examined is the credibility issue of VGI. The credibility of data regards its believability (Flanagin and Metzger 2008). Traditional geospatial data are produced by government agencies, including cartographers, geographers, or other agents in an authoritative manner. In this way, credibility of these entities is agreed in the relevant scientific community and is tested and proven over a long time period. Also, in order to maintain their credibility, those entities need to continue to submit their reputations to a high data quality standard. On the other hand, the data credibility of VGI is under serious limitations. VGI is contributed by the public who are not experts, as are cartographers, surveyors or geographers. The average citizen has no intrinsic trustworthiness in producing data. When they produce data, there is no guideline to follow except common sense, and a visit to any social-media site will strain even this low level of credibility. Even though under certain circumstances there may be a recognized or de facto standard, volunteers will not follow it or they cannot follow it under the prevailing situation.

In the domain of VGI, data are contributed not just from one individual or one entity, but from any person around the world who are willing to contribute. This introduces another VGI related issue: heterogeneity. It is not realistic to expect data produces to follow a particular data producing standard, even if such a standard exists. While individuals have different perception, skills and vocabularies, heterogeneity is embedded in VGI. Individuals who are experiencing the same event will likely have different observations if they are at different physical location or from different angles. Also, the data contributors could be expert in the subject, or know nothing at all about the subject. Moreover, gender, age, race and economic status each contribute to the heterogeneity of perception and cognition. As for vocabularies, in the case of emergency management, for the word of shelter which is used by professionals, the average citizen could use some other words which have the same meaning, e.g., place to live, house, tent, hotel, motel, or others. Semantic ontology is a complex issue that will arise in this research in the form of key-word ambiguity.
Heterogeneity increases as the data volume increases. Heterogeneities contribute to considerable difficulties in using VGI data. First, the huge volume of VGI means that it is not realistic to manually interpret every piece of data. Data has to be interpreted by computer programs. It is a challenge to develop an algorithm that can parse the tremendous heterogeneities that exist in the dataset, one that is outside the scope of this dissertation (Carley 1993; Krippendorf, 2013).

The liability or privacy is also a concern in using any form of data, while the issue is more obvious in VGI. With crowdsourcing, data collection and dissemination become much easier. It has been argued that traditional geographical data could threaten citizens’ privacy as one piece of information – the location information – is embedded in the data. There are two kinds of privacy invasions. First, information producers endanger the privacy of other citizens. For example, Elwood (2008) cited a website - “RottenNeighbor.com”, which is not available now. RottenNeighbor.com allowed users to post and search for information about potential and current neighbors before moving. It “was met with controversy over concerns of names or false claims being posted to the website and fears of the website harming or impeding home sales” (Elwood 2008). In this case, producers may be aware of the intrusion into the others privacy, but, they might not be aware of how malicious or widespread it could be. Secondly, information producers’ privacy may be eroded without their awareness. That means, data producers do not know that their privacies they care about, are accessible by others. For example, a Twitter user is willing to post his/her comment to the Twitter platform, but he/she do not realize that his/her location information, which he/she might not want others to know under most situations, is seen by all other users.

For the first kind of privacy invasion, it is hard to protect the privacy of the others because the social media platforms have no clue whether a piece of content hurts another’s privacy or not. For the second of privacy erosion, many social media platforms recognize the issue and take some steps to protect their users’ privacy. For example, in order to protect users’ location information, Twitter gives its users a conscious choice of whether to share or not-share their location information. If the location information is hidden, Twitter will not display the location information of a message to others.

2.3.7 Two types of error in VGI

The previous section discusses VGI issues which can be addressed as methodologies evolve. Besides the discussed issues, VGI, a kind of geographic information produced by volunteers, have
some inherent errors, which result from the data not being verified by officials. For example, in the activities in disaster management, especially the response phase, it is clear that users of VGI must be conscious of the need to balance the rapid availability of VGI with the unverified nature of its content. A homeowner who evacuated based on information from an online map created by a volunteer might be responding to a false positive, and by waiting for official, verified information might have avoided the need to evacuate altogether. Figure 2.3 demonstrates the uncertainty inherent in VGI. In the first quadrangle, there are many VGIs and the event has high possibility of existence, the data indicate a correct interpretation. Again, in the third quadrangle, if there are few pieces of VGI and the event has low possibility of existence, the data represent correct information. VGI in those quadrangles I and III have no problems.

I: High occurrences and high possibility
II: Low occurrences and high possibility (False Negative Error)
III: Low occurrences and low possibility
IV: High occurrences and low possibility (False Positive Error)

Figure 2.3: Types of Error

On the other hand, there are different types of error in VGI. This research will focus on two types of errors as indicated by the figure above: false positive error and false negative error (Goodchild and Glennon 2010). False positive error, the fourth quadrangle of the figure above, denotes that
there are many VGIs about an event which has not in fact happened, or has low probability of occurring. It is possible that volunteers have post false or misleading information about an event, e.g., volunteers might post information that describes a bridge collapse (which is not true) based on hearsay, or even maliciously. In false negative error, the second quadrangle of the figure above, there is no or very little VGI reported about an event that just happened or is still on-going. This error occurs when nobody is posting about a situation that is in fact occurring, or some people are aware of an event, but show no interest in it, or think the event is too insignificant to report.

The two types of errors may result in different costs. In the case of false positive error, which indicates that there is an event going on, the emergency management team may dispatch resources, including a rescue team and logistic resources, to a scene where no event is happening. The cost of making this error will be an unnecessary waste of resources expressed as an opportunity cost.

In the case of false negative error, based on the lack of geographical data indicating a need, a management official will fail to dispatch resources to a place where there is in fact a demand for resources. The cost to this error could be more serious, even disastrous. In emergency management, it could leave victims in a dangerous situation. Comparing to the costs caused by false positive (locking up resources which will be unused), the cost caused by false negative errors (failing to provide needed resources) is potentially more serious. The topic of error is more subtle and complex than presented here and will be discussed further in Chapter 7.

2.3.8 Conclusion

The emergence of web 2.0 has promoted the rise of VGI (volunteered geographic information), or UGGC (user-generated-geographic-content). VGI has demonstrated its power in emergency management, such as VGI’s timeliness is critical in disaster management. VGI is local knowledge from volunteers’ perspective, it is potentially a significant data source; and it is rich in context, including text, image, audio, video format. VGI also attracts researchers to study related issues, such as quality, credit, heterogeneity, and privacy. For example, in disaster management, though VGI can provide critical real-time information about an event from a local perspective, users should always refer to other reference information to make better decisions or identify and discard problematic data so that decisions are more reliable.
Chapter 3: Research Objectives

3. Research Objectives

Chapter 3 establishes objectives and goals of this research study on the basis of the literature on mass evacuation, distributed GIS, and VGI as reviewed in Chapter 2. Mass evacuation involves a large number of human beings relocating within a large geographic region, often within a short period of time, such as hours or days. Usually mass evacuations occur following large scale natural or man-made catastrophic events. During a disaster event, mass evacuation place enormous requirements on logistic resources across a wide region, including the evacuation ingress area, but the ingress areas, especially rural communities, do not necessarily have sufficient resources ready on short notice. Under such situations, an emergency operation center (EOC) and evacuees require real-time information so that resources can be dispatched to where they are needed. As discussed in Section 2.3.2, the traditional top-down data generation model, is unable to produce and disseminate the necessary logistic resource demand data in a timely manner. This process is slow due to limited resources (including human resources and technical resources), strict standards and specification, and the establishment of credibility. It is difficult for the top-down model to satisfy urgent real-time needs during an emergency event.

On the other hand, a bottom-up or grass-roots model has the potential to acquire and disseminate data more efficiently. In this model, the general public can generate geographic data, or volunteered geographic information, or user-generated-geographic-content, through Web 2.0-based websites and location-service enabled devices. During the response and recovery phases of a disaster event, the average citizen can produce data in real-time where events are occurring. Data uploaded by the local users are hosted on corresponding servers/websites. A more detailed discussion about the two data generation models was presented in in Section 2.3.2. This research study investigates the collection and dissemination of grass-roots data in near real time.

While VGI exist on their own designated websites or servers, the current study investigates the means of acquiring relevant VGI to support requests for logistic resource support during mass evacuation. The overall objective of this research is to investigate the management, analysis, mapping, and evaluation of VGI in the context of a crisis or mass evacuation. While it was made clear in Chapter 2 that distributed GIS is critical to disaster management, this research further
investigates the integration of real-time VGI into a distributed GIS system to support the needs of a mass evacuation. To this end Chapter 3 discusses the major research goals. Before proceeding to a detailed discussion of the goals, section 3.1 discusses the conceptual framework derived from the literature review.

3.1 A conceptual framework

As discussed in the literature review, distributed GIS can be used in every phase of the disaster management cycle; namely the response, recovery, mitigation, and preparedness phases. Of the four phases, the response phase has the most urgent requirements for timely data input. This research studies VGI for the support of mass evacuation, a response activity that requires live data feed. The data collection capability must embedded into a Web GIS platform so that relevant geo-tagged Tweets from Twitter can be stored into a distributed geodatabase, where they may be mapped along with other GIS data, used for further analyses and by other applications.

A conceptual framework has been created to illustrate the context of VGI, disaster management, and distributed geospatial technologies (Figure 3.1). A nested hierarchy is proposed that embeds the current research in the broader context of distributed technology and disaster management. As shown, this study focuses on the response phase of emergency management for which activities real time information is required for informed decision making.
Within the response phase, the study focus is establishing a series of tasks for a work flow to acquire and disseminate relevant VGI. The work flow can be described as follows: first, a VGI collection component is used to extract VGI based on specific parameters; secondly, the collected VGI is stored in a distributed geodatabase so that mapping and further analysis steps can be performed.
3.2 Research goals

After the discussion of the importance of VGI in disaster management, the first goal is to identify the relevant VGI data that will be utilized in this study.

While there are many different sources of VGI data, some such as OpenStreetMap, provide reliable geo-scientific knowledge contributed by professional or trained personnel. On the other hand, social media such as Facebook and Flickr, do not provide scientifically validated knowledge, but rather observations from the data provider’s current perspective at particular location. While such data cannot be as accurate or precise as scientific knowledge generated by trained professionals, social media-based data is produced in-situ and may be harvested much more quickly. As the quality of timeliness is most critical during the response phase of disaster management, this study focus on harvesting real-time data for rapid decision-making during a disaster, emergency or a crisis, which requires the dynamic updating of data about an on-going event. To meet these conditions, the most suitable VGI data source is social media-based VGI. Among most common social media, the data provided by Facebook, Flickr, and other social media are more complicated than those provided by Twitter, which are called Tweets. Compared to photo-format, audio-format, and video-format data, Tweets, the text-format data, are more straight-forward to be processed by computer programs, although perhaps at the cost of data richness. Additionally, this study focuses on advances of utilizing VGI in mass evacuation and while the data format is not essential to the success or failure of the research, the use of a more straight-forward data format will enable an effective proof-of-concept.

After realizing the first goal of identifying suitable data, the second goal of this study is to investigate the access of the selected VGI. VGI can be accessed online or offline, but online VGI is more valuable because of the need for timeliness in the context of the response phase of disaster management. Most of the social media websites provide an Application Programming Interface (API) for those who have an interest in accessing VGI online to develop interfaces. Twitter provides the REST API which is under the public domain and the Streaming API, which is a secured, proprietary domain. While the REST API retrieves Tweets that have been uploaded to the Twitter server, the Streaming API can collect immediately the request is initiated. While the Streaming API is a powerful tool, this study uses the REST API to retrieve Tweets from Twitter. The decision not to use the Streaming API included the complexity of accessing the API and the
required subscription fee for using the API and data. More detailed discussion of Twitter and Tweets can be found in Section 1 of Chapter 4. REST (Representational State Transfer) coined by Fielding, 2000), is an architectural style, which relies on a stateless, client-server, cacheable communication protocol. REST uses four HTTP operations - CRUD (Create/Read/Update/Delete) to read (GET), to create (POST), to update (POST), and to delete (DELETE) resources. Of many REST APIs, the Twitter Search API is used to retrieve Tweets in order to meet the third research goal.

The Search API can specify a search query and a search region (as a circular buffer). That means the Search API can be used to study a specific region for a specific purpose. But the Streaming API can only specify either a search region (in a rectangular shape) or a search query. Once a specific region is specified, the rest of the query is disabled automatically. The Streaming API is good for studies that have no specific location specifications. In this study, Twitter’s REST Search API is used because a location specification is desirable.

Having identified the method for data collection, the third goal is to identify and extract the relevant VGI. To this end, Twitter Search API considers parameters used in a search request to identify the appropriate VGI. The parameter “q” is used to specify keywords for content search and the parameter “geocode” is used to specify the geographic area of the search.

The fourth goal is to map the extracted geo-tagged Tweets. Unlike other non-geo-tagged user-generated-content, VGI has a spatial component which enables a mapping capability. This study will build a Web GIS platform to display and map the collected VGI. In contrast with other VGI mapping applications, which only map VGI on-the-fly for visualization purposes, this study saves the collected Tweets permanently as a feature class in a distributed geospatial database, and thus is able to handle the Tweets in the same manner as other GIS data. Furthermore, tweet data analysis can also be performed in the same manner as other traditional geospatial data.

Implementing the Tweets’ near real-time-saving goal requires two preprocessing steps; first to build a tweet feature class with predefined fields, and save it into a distributed geospatial database; second, a feature service is published after the feature class is saved in a map document with pre-defined symbologies designed to represent selected query keywords, or a keywords category. Those processes are completed through ESRI’s ArcGIS software suite.
The next goal is to categorize the VGI data for various purposes. As mentioned in section 3.2, the collected VGI is saved into a geospatial database as a “feature class” that can be a map layer in ESRI’s ArcMap map document. The ArcMap map document is published as a map service through ESRI’s ArcGIS Server context. At the time of collection design, the scheme of the VGI feature class, other key information such as “SearchKeyword” and “Keyword Category”, are included in the feature class scheme. At the map design phase, one of those two fields is used to indicate the purpose of the Tweets.

After the collection of real-time VGI data, the next goal is to study the possibility of exporting VGI into a mass evacuation model for analysis. At the cost of real-time display, this study stores collected VGI in a distributed geospatial database as a normal GIS feature layer. This VGI layer can be used as a typical GIS layer for near real-time mapping, spatial analyses and models. In the absence of an ongoing crisis event, similar VGI data could be made available for testing during the preparedness phase of crisis management. The data can be collected as hypothetical or training data, and subsequently can be used to develop and test spatial models or analyses that can then be used to aid decisions during the response phase of crisis management. Since it is built on real VGI data, the model or the spatial analyses could be expected to run smoothly in a real crisis event.

The final goal is to demonstrate the benefits of VGI in the context of mass evacuation. VGI provides critical live information that is not available from official sources in real time. During the course of a crisis event, given a lack of timely authoritative data, VGI, as live and dynamic, crowd-sourced information, provides crisis management professionals an augmented understanding of the real situation. This study uses a case study to illustrate the application of real-time collected VGI into previously defined spatial analyses models. It will investigate how critical information can be retrieved based on live data and how correct decisions can be drawn by emergency management professionals on the basis of timely dynamic information. It is possible that average citizens using the application could improve their ability to direct emergency management official’s response. With the successful cases (Elwood 2008, Goodchild and Glennon 2010, Longueville et al. 2010, MacEachren et al. 2011, Poser and Dransch 2010, Roche et al. 2011, Starbird 2011, Zook et al. 2010), it has been recognized that VGI has its own problems.

One objective of this research is to evaluate the potential roles of VGI in a disaster or crisis event. As an asserted data source, VGI raises several concerns, such as those of data quality, credibility
and privacy. Among those issues, data quality is the major concern, since VGI carries no assurances that are explicit in official data. In the case of a disaster event, especially a life-threatening catastrophic event, the cost of responding to problematic data could be smaller than making no response due to lack of data (Goodchild and Glennon 2010). Goodchild and Glennon pointed out there are two kinds of error possible in VGI: false positives, which indicates the posting of data about a non-existing phenomenon during an event, and false negatives, which means there are no postings about an existing and on-going phenomenon during an event (Goodchild and Glennon 2010). These conditions have to be distinguished from true positive and true negative conditions. A sample of Tweet data in the case study will be evaluated in regard to accuracy in regard to false positives. Unfortunately in the absence of observations, the question of false negative error is not subject to statistical testing.

3.3 Summary

Based on the literature review of Chapter 2, this chapter has proposed a conceptual framework, which generates a list of goals by which the overall research objective may be accomplished. The overall research objective is to provide the means to incorporate real time VGI to support the requirements of logistic resources during a mass evacuation. Consequently, the goals necessary to achieve the objective are to identify suitable data, to investigate access to the selected VGI, to extract the required information, to export the information to a model, and to evaluate the potential roles. To implement those objectives Chapter 4 will discuss the methods related to the identified goals.
Chapter 4: Research methods

4. Research methods

Chapter 4 proposes the methods for the implementation of the goals identified in the prior chapter. The chapter discusses the practical means necessary to identify, collect and store real-time VGI in a distributed geodatabase so that VGI can be operated as a normal GIS data layer. The last section in this chapter considers the discussion of the system architecture and implementation of live tweet collection and storing.

4.1 A possible VGI source: Twitter and Tweets

OpenStreetMap and social media are possible VGI data sources. OpenStreetMap is a type of VGI that is able to provide geospatial scientific knowledge under certain conditions. On the other hand, social media, such as Twitter, Facebook, and Flickr, can provide VGI that includes local knowledge, but which cannot be scientifically verified. In case of Twitter, Tweets are timely local knowledge. In the case of disaster most original Tweets are created by users who are involved or are at the scene of an event. Potentially, those users have better observation of an event happening in their surroundings. While the question of re-tweeting must be considered, this study will focus on initial Tweets in Twitter.

Twitter is a “real-time information network” that connects all its users over the Internet. On Twitter, users can post “the latest stories, ideas, opinions and news” about what users think are interesting or significant (Twitter 2014). As an online social networking service and microblogging service, Twitter allows its users to post and read text based message, known as “Tweets”, which can hold up to 140 characters. Tweets are the basic atomic unit of Twitter. Even though the size is small, a tweet may send not only text, but also photos and videos. Twitter has more than 700 million registered users world-wide as of 2015, of whom 77% are outside the US. More than 200 million Tweets were generated as of March 2012 and more than 1.6 billion search queries were handled every day as of May 2011 (Dugan 2012). Information on Twitter can be accessed by non-Twitter users. It is not necessary to contribute in order to be able receive / consume Tweets. Because anybody can access Tweets, which are users’ real-time observation of an interesting or a significant event, Twitter helps us to better understand unexpected events that are happening around the world.
The fields included in a tweet entity are not useful in all contexts, but several fields are useful in different situations:

- **created_at**: the time that this tweet was created. This information is important since time is critical in this research.
- **from_user**: the username of a Twitter user who created the tweet.
- **from_user_id**: the unique identifier of a Twitter user who created the tweet.
- **id**: the unique identifier of a tweet.
- **text**: this is the actual message. This information is the major subject that this research focuses on.
- **geo**: if present, this field indicates that this tweet is associated with a precise location. With this information, a tweet can be mapped and saved as a feature of a GIS file.
- **location**: if present, this field indicates that this tweet is associated a place, from which it did not necessarily originate. With this information, a tweet can be mapped to the location it indicates, and saved as a feature of a GIS file.

The **coordinates** field and the **place** field are used to store the location information. Twitter stores the location data that is passed by the client in two ways: (1) All location information that a user chooses to publicly display will be stored. It means if a user Tweets using their exact coordinates, Twitter will store the exact location in a tweet as long as the tweet exists (or until the user clears the location history in the setting page in his/her Twitter User Profile page). (2) If a user turns on the location information, but chooses not to include the exact location in a tweet, Twitter will not include the location information. A tweet without location information (geo or location) cannot be mapped. In this research, such a tweet is saved for the purpose of content data analysis from non-spatial perspectives.

The following code is a sample of a tweet. The username, user id, user’s profile image URL, tweet’s id are replaced by a dummy value in order to protect the user’s privacy.

```
"created_at":"Sat, 23 Feb 2013 23:59:33+0000",
"from_user":"somename",
"from_user_id":11111111111,
"from_user_id_str":"11111111111",
"from_user_name":"11111111111",  
"from_user_name":"11111111111\222222222",
```
The next section will discuss how to retrieve Tweets from Twitter’s server and other third party host companies.

4.2 Tweet collection

Twitter provides APIs (Application Programming Interfaces) for developers to access and use Tweets. An API is a library that may include specification for routines, data structures, object classes, and variables. As a protocol, an API is used as an interface by software components to communicate with each other. In the context of web development, an API is a set of HTTPs (Hypertext Transfer Protocols). HTTP is used to request messages and to receive the responses, which are usually in an Extensible Markup Language (XML) or JavaScript Object Notation (JSON) format.

There are several types of APIs in the Twitter’s development environment. Two major ones are the Stream API and REST API. The REST API enables developers to access some of the core primitives of Twitter including timelines, status updates, and user information, with a few RESTful API calls. It also enables developers a multitude of integration opportunities to interact with Twitter. The Streaming APIs is for developers with data intensive needs in a real-time manner, e.g. building a data mining product or performing an analytics research. It enables developers to establish and maintain a long-lived HTTP connection. The difference between the REST API and
The Streaming API is that the streaming API requires a persistent HTTP connection to be open (see Figure 4.1) while the REST API establishes the connection upon users’ request and such connection will be killed once the data is received by the client (Figure 4.2).

![Diagram](image)

**Figure 4.1:** REST API Tweets request and response diagram. Source: Twitter.com

The streaming API uses codes to maintain a streaming connection process. This connection process is typically separated from the process that handles HTTP requests. It establishes the connection between Twitter, receives streamed Tweets, performs the data processes, and stores the results in a data store (see Figure 4.2).
The Streaming API is more attractive for an emergency management application as: (1) it allows access to real-time data; (2) it has better accessing speed and higher or no rate limits; (3) it supports advanced querying needs, e.g., a larger number of keywords to be specified and tracked, retrieving geo-tagged Tweets from a certain region, or having the public status of a user set returned. While the data collected from both API are the same, the Streaming API requires more complicated configurations and has subscription fees. Although this research focuses on the accessibility of the Tweets in an emergency management context the REST API, which is under public domain, is the focus of study for reasons explained subsequently. The Twitter Search API, which returns a collection of relevant Tweets matching a specified query, is one sub-category of Twitter REST APIs. The study uses the Twitter Search API for the collection of Tweets.

4.2.1 Access Tweets through Twitter Search API

The Twitter Search API focuses on accessing Twitter’s content, i.e., retrieving Tweets with specific keywords, accessing Tweets from a specific user. It is dedicated to search against real-time Tweets. In the Search API, the URL which is used to send the query is the following format: https://api.twitter.com/1.1/search/Tweets.json. The word “json”, which stands for JavaScript Object Notation, refers to the format of the Twitter’s response. JSON is a lightweight text format.
for data-interchange based on its properties such as being independent from programming languages, easy to read and write, and easy to parse and generate (json.org).

The parameters included in the URL include \( q \) (query), \( \text{geocode} \), \( \text{lang} \), \( \text{result_type} \), \( rpp \) (records per page), \( \text{page} \), \( \text{until} \), \( \text{since} \), and others. Among all the parameters, only the parameter “\( q \)” (query) is required. The other parameters are optional. The parameter “\( q \)” is used to indicate the search keyword in this research, e.g., shelter, gas. The format for using this parameter is: \( q=\text{shelter} \). The maximum length of this parameter value is 1,000 characters. The operators used to form more precise search results are a blank space, quote, or, and, “-”, and many others. The blank space operator (the default operator), “ ”, e.g. \( \text{hurricane Sandy} \), selects Tweets that contain the text “hurricane” and “Sandy”. The quote operator (paired single quote or paired double quote), e.g., “\( \text{hurricane Sandy} \)”, selects Tweets which contain the exact phrase in the quote, \( \text{hurricane Sandy} \). The “or” operator, e.g., shelter or gas, selects Tweets that contain either “shelter” or “gas” (or both). The “-“ operator, e.g., shelter –gas, selects Tweets that contain the text “shelter”, but not the text “gas”.

The parameter “\( \text{geocode} \)” selects Tweets that are created by users that are located within given radius of a given location, which is given in latitude/longitude format. The format of this parameter is “latitude, longitude, radius”, e.g. 37.117121, -84.332299, 50mi. The search radius must be specified as mi (miles) or km (kilometers). Also, note that there is no space between the radius term and the radius units.

The term “result_type” is another important parameter. It is used to specify what type of search results that the developer would like to receive. There are three options for this parameter: “mixed”, indicates Tweets that are popular and real-time results in the response; “recent”, indicates only the Tweets that are the most recent results; “popular”, selects only the most popular Tweets.

The parameter “\( rpp \)” indicates how many Tweets per page. This parameter is only valid in Twitter API 1.0 and was discarded in Twitter API 1.1. While the Kent State case study is based on Twitter API 1.0, this parameter will be introduced as follows: The maximum is 100 Tweets per page. The parameter “\( \text{page} \)” indicates the number of pages the developer wants. One page is one response from Twitter according to the request. The maximum number of Tweets for one response is defined by the \( rpp \) parameter. The Search API can only return about 1,500 Tweets and this limitation indicates that the product of “\( rpp \)” and “\( \text{page} \)” cannot exceed 1,500. In Twitter API 1.1, the counter
part of the parameter “rpp” is “count”, which needs to be combined with another important parameter “max_id” to retrieve all the Tweets that satisfy the search requirements.

The parameters “until” and “since” are dates and are formatted as YYYY-MM-DD, which is four-digital year, two-digital month, and two-digital day, e.g., 2010-11-06. While the parameter “until” means to return Tweets that were created before the given date, the parameter “since” means to return the Tweets that were created after a given date. The two parameters can be combined together as a timeframe to get Tweets in a window from a given date to a given date.

According to the parameters introduced above, an example of the full format of a URL of a Twitter Search API 1.0 is:

```
http://search.twitter.com/search.atom?q=vote%20obama&since=2013-02-
17&until=2013-02-18&page=1&rpp=100&geocode=39.20443,-
76.94313,50km&result_type=recent
```

Under Twitter API 1.0, by inputting this URL in the address bar of a Web browser, a list of Tweets that satisfy the parameters will be returned in JSON format and displayed in the Web browser. Similarly, an example URL of a Twitter Search API 1.1 is:

```
https://api.twitter.com/1.1/search/Tweets.json?q=vote%20obama&since=2013-02-
17&until=2013-02-18&max_id=54321&count=100&geocode=39.20443,-
76.94313,50km&result_type=recent
```

Under Twitter API 1.1, there is a security layer between users’ tweet request and the Twitter’s servers. The Tweets request URL needs to be wrapped into a package that included the Twitter authorized token, keys, etc. By sending the package to the Twitter server, a Tweets request connection can be built, and the Tweets that satisfy the query parameters will be returned to users. While this authorization is under public domain, which means anybody who has a Twitter account can get it for free, the ways of getting Tweets is more complicated than that under Twitter API 1.0.

**4.2.2 The design and the workflow of tweet collection**

A prototype tweet collection program was designed to harvest Tweets in response to a hypothetical mass evacuation-related event in the Baltimore-Washington D.C. Metropolitan area target area. The study area also includes the West Virginia Eastern Panhandle which has four counties:
Berkeley, Hampshire, Jefferson, and Morgan as the surrounding rural communities of an urban area have high potential to be an evacuation destination. Evacuation management requires a variety of resources and a dynamic assessment of their availability. Based on the simulated situation in this study area, this design of the search center and radius specification is discussed as below.

Due to the rate limit, a term from Twitter.com and which indicates the number of Tweets that can be extracted in a defined time-window, the search center and radius specification need to follow some rules. First, the search attempts to search all the Tweets for the targeted area. Secondly, as the possible number of Tweets related to an event for the urban area is conceivably very large, the search radius for the urban area will be made small, e.g., 20 km. In response to the smaller search radius, more search centers will be designated in order to cover the target area completely. On the other hand, for the low-density, rural population areas, the possible number of Tweets will be fewer, so the search radius can be larger; thereby fewer search centers are required. An example of the distribution of the search centers and the radius resulting from the rules is shown as the Figure 4.3.

Figure 4.3: the distribution of search centers and radii in the Baltimore-Washington D.C. Metropolitan study area
Having established geographical coverage with the object of maximizing the number of messages, the Tweets are searched using keywords. The search subjects / keywords are emergency response-related resources, such as gas, shelter, water, food, or medical. The keywords are classified by categories. Each category includes the keywords related to the topic, for example, the keywords for the shelter category can be “shelter”, “place to live”, “tent”, “hotel”, “motel”; for the fuel category: “gas”, “gasoline”, “fuel”; for the medical category: “medic”, “medical”, “doctor”, “hospital”; the keywords for the water category: “water”, “drink”, “drinking”; and the keywords for the food category: “food”, “hungry”, “restaurant”, and so forth.

During data collection, the Twitter Search API-based work flow includes four major steps. The first step is to specify a date for data collection. In this step, if several days, i.e., $d$ days, are specified, the data collection program will programatically reset the timeframe to one day and performs the search $d$ times. The second step is to specify a location and the corresponding search radius. After the location information is geocoded, the latitude and longitude information is recorded. The number of times that the data collection program sends the request for Tweets will depend on the number of locations, $l$. The third step is to specify search keywords. The number of keyword categories, $c$, and the number of groups, $g$, in one category determines the number of times that the data collection program sends the request to the server. The fourth step is to save the results. All the searched Tweets will be saved in CSV files for further investigations. If a tweet has location information attached or geo-tagged, it will be saved to geospatial database and mapped.

In general, the total number, say $n$, of requests for Tweets can be expressed as: $n = d \times l \times c \times g$. A CSV file will be generated for each request. There will be $d \times l \times c \times g$ CSV files saved. All of the geo-tagged Tweets will be saved in one GIS file.

Under many situations, the request will get more than 100 Tweets, which is the maximum number of Tweets per page, and up to 1500 Tweets, which is the rate limitation imposed by the Twitter Search API. Under those situations, the same request will be sent to the several times with different page numbers until the number of Tweets on a page is fewer than 100. The possible number, $p_i$, of requests with different page number for one group of keyword is different. Overall, the number of times, $N$, that the collection will send request to the Twitter server will be:
\[ N = d \times l \times c \times g \times \sum_{i=1}^{d\times l\times c \times g} p_i \quad \text{where } p_i \in [1,15] \]

Figure 4.4 illustrates the work flow how the collection program works.

Figure 4.4: the flow diagram of Tweets search by using the Twitter Search API
Following the design of Tweets collection, its implementation is based on a Web-GIS platform, which provides the Tweets collection program basemaps, mapping interfaces, and analysis platforms.

### 4.3 Web GIS system

This section focuses on the method to build a platform for collecting, mapping and analyzing tweet-based VGI. This research has developed a prototype Web GIS platform which provides capabilities in three categories: (1) an online mapping capability, which provides users the capability to view the data on different basemaps and to be able to navigate the map; (2) tweet collection capabilities, which allow users to collect Tweets based on some prior specifications, e.g. search locations (including search radius), search timeframes, search keywords; (3) geo-tagged GIS-formatted tweet management capabilities, which include saving geo-tagged Tweets, putting time-stamp on the collected data, and performing geospatial functions.

The development of the prototype application is based on different Application Programming Interfaces (API). The API used for mapping is ESRI’s ArcGIS JavaScript API. The API used for Tweets collection is the Twitter Search API. For the Web GIS development, ArcGIS Server is not needed because the Web platform accesses basemaps that are published under the public domain or a free license agreement. What developers do need however, is a uniform resource locator (URL) by which developers can access the resources published from an ArcGIS Server machine, which is an advantage of RESTful web services. The next section will discuss enabling technologies, including RESTful web services, map cache, and server GIS and geodatabase.

### 4.2.1 Enabling technologies for building a Web-GIS application

Traditional or first-generation Web-GIS applications run very slowly when retrieving information and rendering online maps. Additionally, they do not have powerful functions due to the limitation of the technology. However, as computer technologies, network technologies, and information technologies have advanced, the next generation of Web-GIS becomes possible. The major features of next generation Web-GIS include faster map rendering and more powerful online GIS functions. The technology advancements that make the next generation of Web-GIS possible are RESTful web services, map cache, and server/database (geo-database) technologies, which are most important for in this research for implementation in the Web-GIS application.
In the context of ESRI’s ArcGIS Server, RESTful web services are accessed by the clients. Several different types of services are published using ArcGIS Server (esri.com), such as geoprocessing services, geocoding services, and map services. Among them, map services can be served in two formats: dynamic format, which is called ArcGISDynamicMapServiceLayer by ESRI in ArcGIS APIs, and cached format, which is called ArcGISTiledMapServiceLayer by ESRI in ArcGIS APIs. Map services in dynamic format are used for a map that has data that is updated frequently, e.g. event related data or the GIS-formatted geo-tagged Tweets in this study. Cached services are fast map services and are used for better map rendering which is useful in the data presentation component and it will be discussed in more detail in following section.

In ESRI’s ArcGIS Server, a map cache is a collection of pre-rendered map tiles that can be used to display a map service. A map cache allows a map service to quickly display maps because the map image does not have to be rendered on the fly; the time cost of rendering the image is paid only once when the cache is created. The server generates the entire map tile set for the whole map document at predefined scales and stores the map tiles in an Internet accessible location. Since the maps at different scales are physically stored in the server, the server can then distribute these pre-existing image tiles to the user, for a map at a given scale, without dynamically extracting the map data from the data server to generate maps. Since map tiles are created and saved at predefined scales, the resolution of the image does not noticeably affect how quickly the server can distribute data.

Distributed geodatabase technology is central to the research method. The collected geo-tagged Tweets are saved in GIS format and used for further analysis and management. The geo-tagged Tweets are pure text-formatted messages with location information enclosed in the message body. Mapping them is not a problem given the location information; however the challenge arises when trying to save these geo-tagged Tweets in a GIS format.

Usually, Web-GIS do not support data saving and online editing, which is a major accomplishment of this research. To achieve this capability, a geospatial-based relational database plays a key role. In the context of ArcGIS Server, the relational database is managed by ArcSDE, a spatial database engine. In this research, ArcSDE is powered by Microsoft SQL Server Express, which is a smaller and simpler version of Microsoft SQL Server, an enterprise relational database.
4.2.2 Components of the integrated application

The Web GIS application includes three conceptual components: the mapping component, the Tweets collection component, and the Tweets (the GIS-formatted geo-tagged Tweets) management and analyses component. Physically, there are two parts in the application interface: one is the mapping component; the other is the operational part, which is for the other three components. As the operation part is fixed at some width (e.g. 400 pixels), the mapping part takes the rest of the available web browser space; so that the greatest extent or more detailed information can be displayed on the map.

The mapping component displays maps, including basemaps, which are composed of framework data in cached map service format; and an operational map, which is composed of GIS formatted of geo-tagged Tweets and in dynamic map service format. The mapping component also provides the navigation functions, including pan (in the direction specified by the user), pan to 8 fixed directions (i.e. east, west, north, south, and northeast, northwest, southeast, southwest), zoom in/zoom out with mouse scrolls or by using navigation bar. The other function of the mapping component is the overview map. It indicates the location of the current map in a bigger map context.

The Tweets collection component is for specifying the Twitter Search API parameters, including search places, search keywords, and search timeframes. The design of this component is introduced in details in the design and the workflow of tweet collection section. The fourth component can save geo-tagged Tweets dynamically in GIS format and provide other functionalities for handling GIS formatted geo-tagged Tweets.

The geo-tagged Tweets management and analyses component holds the functions that handle the geo-tagged tweet GIS layer. The functions include: turning the tweet layer on and off, summarizing the data in different data collection category, playing the time-frame, and creating hot/cool spots. More detailed discussion can be found in the next section.

4.3.3 Integration of tweet collection component into the distributed GIS system

Tweet collection is one component of the Web GIS system. This component helps users to input search keywords, which is a required parameter for the Twitter Search API. Another important parameter, “geocode” which specifies the search area in the format of “latitude, longitude, radius”, is retrieved from the map component. Users can pan around the map to focus on their area of
interest. The map center will be returned to the system automatically, which provides the value for the “geocode” parameter. The Web GIS system provides a capability for users to switch to different basemap layers that provide reference data for the Tweets data. Without a basemap, standalone event-related data have no geographic context. A Web-GIS application provides a third capability to map the geo-tagged Tweets. In this research, the proposed Web GIS system will not map the on-the-fly geo-tagged Tweets, which will be lost after the application exits, but it will map the geo-tagged Tweets that are saved into the tweet geodatabase.

4.4 Saving geo-tagged Tweets in GIS format

In contrast to other research, which focuses primarily on mapping and display, this study has enabled the capacity for geospatial analysis based on real-time data. Analyses are performed from a geospatial perspective and the main focus is on the geo-tagged Tweets. In order to achieve this goal, the geo-tagged Tweets are saved in a relational database (Microsoft SQL Server Express, in this case) using geodatabase as a GIS layer. Saving Tweets in GIS format and operating on them as regular GIS features are major differences from previous studies.

In order to implement real-time data saving in a GIS format, the target layer needs to be managed by ArcSDE as a layer in an ArcMap document, which is then published as a map service, consequently, the Tweets GIS layer is stored in geodatabase and served as a layer in an ArcMap document. The ArcMap document is then published as a map service with the feature access capability enabled. Since the GIS layer is an ArcSDE feature class (a term for a GIS layer in geodatabase context), the GIS layer allows internet users to operate on it through a web service that is in RESTful style. The possible operations include SELECT from, INSERT into, UPDATE, and DELETE. In this study, the function used most often is INSERT into, which adds new features in a layer and SELECT from, which is used to read the records in the GIS layer.

The data structure of the Tweet GIS layer is predefined, including the name of the author, the date and time that a tweet is generated, the location of a tweet, the message body, the keyword(s) used to retrieve the tweet, the name of the search place, the search radius, and the tweet ID, which is unique to every tweet. The general step of saving and mapping geo-tagged Tweets can be summarized as follows: first, the geo-tagged Tweets are differentiated from non-geotagged Tweets; second, the geo-tagged Tweets are saved in the geodatabase; and third, the saved geo-tagged Tweets are mapped on the interface.
Since the geo-tagged Tweets are saved and displayed as other GIS layers, operations appropriate for regular GIS layers can be applied. The general Web-GIS functions include selecting features by geometry, by attribute and querying against features. The analyses functions included in this research are based on the real-time data and are performed on-line, which is a second major difference from previous research. The next section is the detailed discussion of the Tweets management component.

### 4.5 System architecture and implementation of live data saving

The prior section has discussed how to save live geo-tagged Tweets into a distributed geodatabase. Based on the discussion, an architecture is proposed in Figure 4.5. The system includes two parts: one is the Tweets collection component, and the other is the tweet storage component.

![System architecture for live tweet storage](image)

**Figure 4.5: System architecture for live tweet storage**

With the tweet collection component in the web interface, a user can send the Tweets collection request, which has the corresponding parameters specified, to the Twitter server, which will return the requested Tweets back to the client. The majority of existing VGI mapping projects finish at this point. They map the collected VGI on-the-fly, and the mapped VGI will be lost after the mapping projects close. This study stores data in database format and converts the geo-tagged Tweets into GIS file format in the tweet storing component. By doing this, we can avoid data loss,
save the data for future study, and feed the live data into another model. In this study, the tweet storing component includes four sub-components: 1) the Web server component, 2) server GIS Application server component, 3) Map server component, and 4) the Geodatabase server component (Figure 4.6). As an online system, the Web Server component, not only hosts the integrated system of Web-GIS component and tweet collection and management system, but also supports the Server GIS system, which includes the application server of Server GIS, the Map Server, and the Geodatabase Server.

After the client side receives data that are returned from the Twitter server, the geo-tagged Tweets will not be mapped right away. Instead, the collected geo-tagged Tweets are forwarded to Server GIS application server through Web server. The Server GIS application server sends the data to geodatabase server, and geodatabase server saves the received data into designated server. At the same time, the Server GIS application server communicates to the Map server which can map the data that is stored in the Geodatabase Server.

Following is the development of the system architecture for live Tweets saving; the implementation is based ESRI’s ArcGIS software suite including ArcMap, ArcGIS Server, ArcSDE, and ArcGIS JavaScript API (Figure 4.6).

Figure 4.6: Implementation of live tweet saving

To implement live data saving, the server side configuration needs to be performed first. The first configuration is to create a feature class in the distributed geodatabase. Besides the fields in a tweet message, the tweet feature class also includes other fields, such as “keywords” which is used to indicate the search keyword for retrieving Tweets, “keyword category” which is used to indicate
the purpose of the collected Tweets. The pre-created Tweets feature class is then added into ESRI’s ArcMap to create a map document. In the map document, the feature class has cartographic design based on the keyword category. The next step is to publish the map document as map service through ESRI’s ArcGIS Server. In the configuration of map service publishing, the feature access capability, which is a FeatureServer service in the ArcGIS Server context, needs to be enabled for online data editing. The feature access capability requires GIS data to be in a distributed geodatabase format, not a standalone shapefile format, personal geodatabase format, or a file-based geodatabase format.

On the client side, which mainly handles tweet collection, the implementation is based on the Twitter Search API and ArcGIS JavaScript API. The Twitter Search API returns Tweets based on the search parameters. For the returned Tweets, a geo-filter is used to separate the geo-tagged Tweets from the total Tweets population. The ArcGIS JavaScript API is employed to write the geo-tagged Tweets to geodatabase server through ArcGIS Server’s FeatureServer service that is created at the server side.

**4.6 Summary**

Identifying Tweets as the VGI for mass evacuation is the first step of this research. Therefore Chapter 4 has specified the methods for Tweets used to implement the series of goals that were proposed in general in Chapter 3. The methods selected are API-based and database-driven for real-time (or near real-time) tweet collection, live data saving, and live data mapping. The next chapter will carry out the proposed approaches and implement those methods in the case studies. Chapter 5 also evaluates the use of Tweets and the corresponding methods in the scenario of mass evacuation situation.
Chapter 5: System implementation and Case studies

5. System implementation and Case studies

Chapter 4 identified Twitter, a “real-time information network”, as a VGI data source for improving response decision-making in mass evacuation. It also discussed Twitter’s content and messages ‘Tweets’, including their structure, how to access them, how to extract them, and how to store them. Of most significance, Chapter 4 proposed methods to integrate and manage collected Tweets in a Web-GIS system to support mass evacuation. The chapter also discussed the technologies that support this research, and how to apply those technologies, including Web GIS and Server GIS, RESTful web service, map cache, geodatabase, and API. Additionally, the last chapter discussed the components for the proposed VGI collection and dissemination: mapping, basemap, Tweets collection, and the GIS formatted geo-tagged Tweets analysis components.

Chapter 5 will employ the methods in two case studies or prototypes, which at first glance do not appear to be related; however, an argument will be developed that the experience gained from the two studies can be integrated to provide functions valuable in emergency management. The first case study is an introductory study of VGI in a man-made emergency event, the 2014 Kent State University campus shooting. This study illustrates the utility and potential of using Twitter in emergency events. The Twitter API that was used was version 1.1. The second case study is of the application of Tweets in a disaster event, the 2012 Hurricane Sandy in Washington-Baltimore Metropolitan area. It illustrates how live data provide near-real time benefits for resource needs identification and dispatch during a potential mass evacuation in a catastrophic event. The available Twitter API in 2012 was version 1.0. While there are differences between versions of the Twitter Search APIs in the limitation of tweet acquisition, the versions have no difference regarding other query parameters, and the tweet text, which is most important in study of the use of VGI, remains the same.

Before continuing the discussion, two conditions that are necessary to enable citizens’ potential as data producers during a crisis event need to be identified. The first condition is that the supporting infrastructure in the affected area must be available during a crisis/disaster event. This condition ensures that real-time information created by volunteers in the affected area is able to be uploaded via Internet to their platform of choice, such as Twitter, Facebook, and Google+. Those platforms
are not normally located in the affected area; thus users in the affected area can access those platforms. The out-of-affected-area platforms also ensures that people can access information generated from the affected area. As well as the social network, the means of accessing the network must be in place in the form of wireless, Ethernet or cell-tower connections. The second condition, then, is that average citizens have access to the appropriate equipment for information dissemination. Without this capability, average citizens as data producers, cannot post information about their surroundings during the event, even if the network is functioning. In general, these two assumptions make sure that volunteers can generate real-time information and such information can be accessed by users who are not in the affected area.

5.1 Key search parameters in the Twitter Search API

Twitter Search API 1.0 limited the number of Tweets that may be returned per request. The modification of the acquisition limitation is in the design of the Twitter Search API’s parameters. Key search parameters include search location/region, search keywords, search timeframe, and search type. The search parameters will be discussed in the two different studies presented in this chapter. Of all the parameters, the limitation of the number of Tweets (1400) that can be returned per request is the most critical to the proposed use. The parameter design of the case studies is intended to overcome this limitation and retrieve as many Tweets as possible.

The general guidelines for the design to acquire the maximum number of Tweets are as follows: First, a large region must be divided into smaller regions, based on population density, to reduce the likelihood that the threshold of the maximum tweet number for one search will be exceeded. Multiple tweet requests, based on the smaller regions, will be issued.

Second, for the tweet search query parameter, one important caveat needs to be considered: Tweets are not described semantically (Abel et al 2011). A semantic search, which retrieves information by meaning, rather than literal strings, is complicated, requiring extensive research into this specialized domain (Fernández 2011, Fazzing et al 2011). Due to the depth of study required for semantic search, this study initially uses keyword-based search for information retrieval, while recognizing its limitation. The search term can be a single keyword or several keywords. For example, to retrieve electricity-related Tweets, the search terms “power”, “out of power”, “electricity”, “outage”, “blackout”, or “no power” may be applied, with the understanding that
relevant and irrelevant messages could be returned. In this study, the power category includes all Tweets that can be collected by the search keywords.

Thirdly, upon initiation, the collection process will collect the Tweets posted before the initialization time. The Twitter Search API can retrieve Tweets that are posted within the past seven days. Although they are not real-time information, the Tweets that are posted in near past could provide background information about the event of interest. As mentioned before, the Twitter Search API has a return limit. In order not to exceed the limited number of returned Tweets, the proposed method will set time-frame of a search to be as small as possible. In Twitter Search API, the smallest time-frame that can be specified is one day. The collection period and one day temporal resolution restrictions indicate that the collection process must request Tweets in a time-frame up to seven times to avoid exceeding the total single retrieval threshold imposed by Twitter.

Another critical parameter, “result_type”, specifies the type of search results to be returned. Three values correspond to this parameter: “recent” - only the most recent Tweets; “popular” - only the most popular Tweets, and “mixed” - Tweets that are both popular and recent. This study seeks real-time data for crisis or disaster events therefore it is necessary to set the parameter “result_type” to “recent” to ensure that the latest Tweets will be returned first if the total number of returned Tweets exceeds the rate limit.

5.2 An introductory study: The Kent State shooting of 2014

Compared to the real world, college campuses are generally much safer, but colleges face challenges when murders transform the campus from “a nursery for hope and promising young leaders into a graveyard of despair and lost youth” (Ferraro and McHugh 2010, 2). Campus shooting seems particularly tragic since it shatters the myth of the college campus as a safe haven and a place apart (Dungy and Reberts 2010). The 2007 Virginia Tech campus shooting brought campus safety under intense attention and scrutiny (Harwood 2007). Although the Kent State incident is not an extensive disaster, no injuries were reported and there was no mass evacuation incurred, this case study served to illustrate how to design the search parameters to collect the interested Tweets in a generic emergency event.

The 2014 Kent State University campus was put into lockdown in the evening of April 2nd, 2014 after a gunman fired a shot near the Bowman Hall classroom building (Timmons 2014). The
gunfire, which occurred at around 8:10 pm, was triggered by a dispute between a man (the gunman) and a woman. While the whole Kent State campus went on lockdown at around for about one hour (9:15 pm – 10:20 pm April 2nd), Bowman Hall and the Business Administration Building were evacuated and cleared at about 11:45 pm of April 2nd. The suspect was apprehended later.

In this introductory case, the Tweets harvested were stored in CSV (comma-separated value) format. CSV is plain text using a character set such as ASCII file. It is a common, relatively simple file format that is widely accepted by researchers, businesses and educators. Every line in the file is treated as a record, and the record is divided into different fields by delimiters, such as commas, semicolons, or tabs, and every record has the same sequence of fields. CSV file can be opened in Microsoft’s Excel and be imported into Microsoft Access or other database software.

**5.2.1 Twitter Search API’s parameter design**

In this case, the gunfire location was close to Kent State University’s Bowman Hall. The design of the geocode parameter is based on the location of Bowman Hall with a search radius of 200 km. The radius was based purely on the researcher’s choice and a large radius was selected to collect as many relevant Tweets as possible. If the radius is set too large, it is easy to reach the maximum threshold of Tweets. The value of the geocode parameter is expressed as “41.148693,-81.346292,200km”.
The keywords used in the 2014 Kent State University campus shooting include “kent state” – the name of the university; “campus” – the type of shooting occurring environment, “shooting” – the type of crime, “bowman” – the building at which the gun fire occurred, “gunman”, “police”, “safe”, and “evacuate”.

### 5.2.2 Results

Table 5.1 is a summary of Tweets collected for the specified keywords for the days from April 1\(^{st}\), 2014 to April 4\(^{th}\), 2014. From the summary table and the associated graph (figure 5.2) we can see that there is a dramatic increase in Tweet messages from April 2\(^{nd}\) to April 3\(^{rd}\), which indicates that Kent State shooting becomes a very hot topic in the region, especially during the period of the debate of gun control. There is also a dramatic decrease from April 3\(^{rd}\) to April 4\(^{th}\), which indicates...
that the Kent State campus shooting is not especially significant since there were no injuries and the assailant was arrested quickly. Such a short “attention span” is a characteristic of social media.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>bowman</td>
<td>52</td>
<td>58</td>
<td>4799</td>
<td>96</td>
<td>5005</td>
</tr>
<tr>
<td>campus</td>
<td>977</td>
<td>979</td>
<td>8324</td>
<td>700</td>
<td>10980</td>
</tr>
<tr>
<td>evacuate</td>
<td>15</td>
<td>14</td>
<td>107</td>
<td>4</td>
<td>140</td>
</tr>
<tr>
<td>gunman</td>
<td>16</td>
<td>133</td>
<td>197</td>
<td>43</td>
<td>389</td>
</tr>
<tr>
<td>kent state</td>
<td>74</td>
<td>75</td>
<td>5609</td>
<td>128</td>
<td>5886</td>
</tr>
<tr>
<td>police</td>
<td>1456</td>
<td>1397</td>
<td>5747</td>
<td>1470</td>
<td>10070</td>
</tr>
<tr>
<td>safe</td>
<td>1087</td>
<td>1424</td>
<td>3595</td>
<td>1032</td>
<td>7138</td>
</tr>
<tr>
<td>shooting</td>
<td>543</td>
<td>1240</td>
<td>2864</td>
<td>713</td>
<td>5360</td>
</tr>
<tr>
<td>summary</td>
<td>4220</td>
<td>5320</td>
<td>31242</td>
<td>4186</td>
<td>44968</td>
</tr>
</tbody>
</table>

Table 5-1: Summary of collected Tweets for event of 2014 Kent State Shooting
As the Kent State shooting was not a particularly significant event, the impacts that it posed to the community were only evident for a short time period, e.g. two or three days. For this event, studying the Tweets on an hourly-basis is more helpful than on a daily basis. Figure 5.3 depicts the number of Tweets collected hourly for 12 hours after the event initiation at about 8:10 pm on April 2\textsuperscript{nd}. From Figure 5.3, we can see that, the number of Tweets in each category posted by Twitter users is in rising; the number reaches at the peak at 1:00am on April 3\textsuperscript{rd} (which is from 1:00am to 2:00am); the number decreases from 2:00 am on April 3\textsuperscript{rd}; and at 6:00am, the number reaches the base level before the event. The finer temporal scale tells us more detailed and accurate information about the number of Tweets posted at different smaller time frame.
5.3 Case study 2: 2012 Hurricane Sandy

The Kent State University campus shooting was a short duration event with only local impact. The shooting happened at a specific time without duration. From Figure 5.3, we can see that the discussion of the event on the Twitter platform died down after less than 13 hours. The Kent State Shooting illustrates the design of the search parameters in the Twitter Search API during a crisis event. The second prototype study is for a large scale disaster event, Super-Storm Sandy, October 2012 as experienced in the Washington-Baltimore Metropolitan area (see Figure 5.4). This event lasted about 5 days from its first hit of Brigantine, New Jersey on October 29 to its dissipation on November 2nd. Although the eye of the storm came ashore in New York City and the New Jersey shore area, the storm was sufficiently large to trigger warnings over an extensive region. Hurricane Sandy swept through the eastern coast of the United States in late October 2012 and became a complex “super storm” as it merged with another weather system. Though the selected ingress region is not the most heavily impacted area, the storm left many dead (1 in Maryland, 12 in
Pennsylvania, 6 in West Virginia, and 4 in Connecticut) (Center for Disease Control and Prevention, 2013), and very significant property damage. As long-lasting catastrophic event, it put great demands on disaster management resources in the affected area. In-line with the original intent of this dissertation, the targeted possible evacuation ingress area is in West Virginia’s Eastern Panhandle (figure 5.3), although other areas could be designated. Even though no mass evacuation occurred during the disaster in the chosen study area, a hypothetical mass evacuation for learning the benefits of VGI for evacuation-related resource dispatch is a reasonable test of the ideas.

Figure 5.4: Study Area of Case Study Two– Washington-Baltimore Metropolitan Area and the possible ingress area of West Virginia’s Eastern Panhandle

This case study was based on a web platform which includes two components: a Web-GIS component and a VGI collection component (Figure 5.4). The Web-GIS component includes the following tasks: host the VGI collection component and map GIS data layers, including traditional GIS data and the crowd-sourced VGI data. The tweet collection component provides functions such as specifying search keywords, specifying search locations and search radii, specifying the search timeframe, and providing capabilities for VGI management. The design discussed in the following segments is based on these components or sub-components, including the Twitter Search API parameter design, the geodatabase design, and the map design (Figure 5.5). In the figure, the arrow indicates that the from-object sends requests to the to-object. For example, at the tweet collection component, the Client Application to the Twitter Server arrow indicates that the client
application sends requests to the Twitter server, and the Twitter Server to the Client Application. Arrow indicates that the Twitter Server responds to the Client Application. Another example, at the Tweet storing component, the Web Server sends requests to the Server-GIS Application Server, and the Server-GIS Application Server responds to the Web Server.

![Diagram of Twitter and GIS interaction](image)

Figure 5.5: Two major components of case study two: Tweets collection component and Tweets saving and VGI mapping component

### 5.3.1 Twitter Search API’s parameter design

This study was implemented in Twitter Search API 1.0. The Washington-Baltimore Metropolitan area is a heavily populated urban area with a population of more than 2 million. The high population density indicates a potential high tweet volume. In Twitter Search API 1.0, the maximum number of Tweets that can be returned per query is 1400. To ensure the retrieval of the total relevant tweet population, the targeted study area, the Washington-Baltimore urban area and its surrounding area, will be divided into small sub-regions which reduces the potential total number of Tweets in any sub-region. Thus the design of the search region is as follows: the sub-region will be smaller in the most heavily populated areas and in the less populated areas the sub-region will be relatively bigger. In this case study, there are twenty sub-regions and the design of the search region is illustrated in Figure 5.6. Every sub-region is represented as its center, e.g. latitude, longitude, and its radius. This collection of latitude, longitude, and its radius are stored in an array. The data collection component will loop through the array of sub-regions to retrieve Tweets for all the Tweets generated in the whole study area.
Search ‘keyword’ in the search query is another parameter that needs to be considered. This case study attempted to retrieve Tweets that could be related to a putative mass evacuation and associated logistic resources. While there could be other keyword categories, e.g., a flood group, which could include search keywords like “flood”, “flooding”, “flooded”, this study only collected Tweets in the following groups: fire, water, gas, medical, shelter, and food. This classification is based on the possible purpose of a tweet, e.g. a tweet is sent to indicate a need for water, for food, or for warning of fire. This classification does not mean there are only 6 keywords used in the data collection process; each keyword stands for a group of words that have the same or similar meaning. For example, to search Tweets that are related to fire, the actual keywords used in the Twitter Search API were set to “fire”, “burning”, “flames”, “smoke”; for Tweets related to potable water, the actual keywords used were “water”, “drink”, “drinking”, and “thirsty”; for Tweets related to gas, the actual keywords were “gas”, “gasoline”, “fuel”; and similarly for medical, shelter and food.

Figure 5.6: the Tweets collection sub-region distribution for 2012 Hurricane Sandy at the Washington-Baltimore Metropolitan Area
5.3.2 Geo-tagged-tweet feature class design

The system is designed to harvest VGI data in near real-time and convert them to a traditional GIS data format. This case study applies database concepts to implement live data harvesting and storing functions. There are several steps required to implement this goal. First, a GIS feature class is created in a distributed geodatabase, which supports online data editing; in this case, to add new geo-tagged Tweets into the existing feature class. The geodatabase concept is based on a relational database management system (RDBMS) that has a spatial component embedded into it. In this case, the selected database is Microsoft SQL Server Express and the spatial database engine (SDE) is ESRI’s ArcSDE.

The concept of feature class in a database context is a set of tables, which include a major table that stores the attribute information, and other tables that support the spatial information and are invisible to end-users. The data structure of the feature class is designed based on the tweet’s field information. Some key API fields include tweet_id (the unique id of tweet), keyword (which is used to retrieve this tweet), category (the category of keyword, e.g., hotel, motel are in category of shelter), location, message, and so on.

5.3.3 Map design

The tweet GIS feature class is added into ESRI’s ArcMap to create a map. As with other GIS layers in ArcMap, the symbology of the tweet feature class layer is customizable, and is based on different information fields. In this research, the cartography design of the tweet layer is based on the field – category, which could include different Twitter search keywords. Tweets in different category use different symbols (Table 5.2).

<table>
<thead>
<tr>
<th>Tweet Keywords</th>
<th>Tweet Category</th>
<th>Visualization Symbol</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fuel, gas, gasoline</td>
<td>Fuel</td>
<td>🏞️</td>
</tr>
<tr>
<td>Food, hungry, restaurant</td>
<td>Food</td>
<td>🍽️</td>
</tr>
<tr>
<td>Water, drink, thirsty</td>
<td>Water</td>
<td>🍷️</td>
</tr>
<tr>
<td>Motel, hotel, tent,</td>
<td>Shelter</td>
<td>🚁️</td>
</tr>
<tr>
<td>Medic, medical, hospital</td>
<td>Medical</td>
<td>🔥️</td>
</tr>
<tr>
<td>Fire</td>
<td>Fire</td>
<td>🔥️</td>
</tr>
</tbody>
</table>

Table 5-2: Tweet symbology design
After the map is created, it is published as a web map service for display purposes. At the same time, the Feature Access capability is enabled, which is only available when the map layer is enabled within a distributed geodatabase feature class. This capability is key for online real time data editing and saving.

**5.3.4 Web interface design**

The system as implemented includes major components: the data display component, the tweet collection component, and the VGI management component. The data display component integrates the display of VGI with other traditional GIS data (see the left part of Figure 5.7). It includes a map navigation component; an inset map component (which provides the overview of the current main map) and a map layer component, which provides different basemap formats, such as Bing Street map, Bing Image map, Bing Hybrid map, ESRI Street map, ESRI Image map, ESRI topographic map, and ESRI National Geographic map, and other reference maps.

The tweet collection component primarily handles parameter settings for tweet collection (see the right part of Figure 5.6). The dropdown list of “predefined place” stores the predefined sub-regions for the Twitter Search API’s “geocode” parameter. There are 20 sub-regions stored in this study (Figure 5.6). The keyword list is segmented into six categories, as discussed in the parameter design section. The timeframe dropdown list is a list of searchable days which is the seven previous days. The days are organized as “from 7 days ago to 6 days ago”, “from 6 days ago to 5 days ago”, and so on.

The VGI management component includes display choices of VGI in different categories; and whether or not to display the heat map based on the collected tweet data and the specified content. The displayed VGI may be based on different timeframes, which can be set manually or automatically. In the the Kent State University campus shooting case, the daily temporal scale and hourly temporal scale were done manually. This online mapping platform provide capability on automatic temporal scale setting.
5.3.5 Manual interpretation of Tweets content

The prior section discussed VGI mapping in the context of disaster management. While the mapping focuses on the location of Tweets, it is meaningful to examine the contents of the Tweets. Since it is not possible to check every individual tweet, this section examined a sample. The sample Tweets were collected with the Springfield, WV town center as the search area center and with a 65km search radius on October 3rd, 2012 during Super-Storm Sandy. Springfield, WV is a community that is a candidate ingress area for a potential mass evacuation originating from the Washington DC and Baltimore metropolitan areas. There were 4,780 Tweets collected and of these, only 361 Tweets were geo-tagged. The rate of Tweets being geo-tagged was 0.076. In other samples, the geo-tagged tweet rate falls between 0.05 and 0.10. By default, opt-out from geo-tag is the default setting as per Twitter privacy policy. Second, many users may not wish to share their location information, even if they know how to enable the geo-tagged information. The low return of geotagged Tweets is a major disadvantage for their use in emergency management models.
Among the 361 VGI records, 123 VGI records were keyword sandy-based, 42 records were keyword storm-based, 51 records were keyword are keyword storm-based, 48 records were keyword power-based or electricity-based. As no mass evacuation occurred in the Super-storm Sandy event, logistic resources related VGI, that were based on keywords, such as food, restaurant, hungry, starving, water, drink, drinking, hotel, and hotels showed in 40 records, which is 11.1 percent of the total VGI, returned. It is unsurprising that no records returned for keywords such as medic and evacuate.

While the semantic query of information is difficult in tweet collection, some records satisfied the keyword, but were not related to the Super-storm Sandy event. To determine the extent of the problem, individual Tweets were examined for content. For the 38 VGI records collected in the food and the water category, 28 records were not related to Super-storm Sandy. For example, for the geo-tagged Tweets collected in the Food category, none of the 20 records was related to Hurricane Sandy. For the 11 records collected based on keyword “water”, not one record was related to drinking water; 6 records were not about the event; and 5 records were related to the situation of flood, such as, “My backyards under water. #nbd **”, “Water is coming into the basement!”, “Foot of water in the basement.... Ain't nobody hot (sic) time of that! Throw me another beer”.

Another 20 VGI records from all other categories, which were not related the Super-storm Sandy event. For example, a keyword “snow” based record, “Watching snow white and the huntsman more people are welcome to join!” is not related to the disaster event, Hurricane Sandy. Another example, a keyword “sandy” based record, “@Jeinilys_x3 , You want me to! Imma rap , ft ; My girl Sandy ! Lls”, might be ambiguous whether Sandy means the Hurricane Sandy or a person named Sandy.

There was another troubling situation about the collected VGI, one VGI record could be collected several times by using different keywords. There were 39 records that were collected twice due to the fact that they were satisfied two different keywords. For example, a tweet, “aaaaaand there goes our power. ** WHY SANDY WHY”, would be returned twice by using keyword “power” and by using keyword “sandy”; a tweet, “@carlieenicole lol my house is getting raped right now by hurricane sandy”, was returned twice by using keyword “hurricane” and “sandy”. There are also two Tweets shown up 3 times respectively. For example, a tweet, “@MommaNagle40 68 isn’t
closed yet. Snow is still west of here but the rain is insanity. Cell service is sketchy and power been flickering”, was returned three times by using keyword “closed”, “snow”, and “power” respectively. If the duplicated Tweets were removed from the total VGI records, there were 306 (361-39-2×2) unique VGI records in this sample dataset.

Among the event-related VGI (313 records), more than half of them (172 records) were not related to an emergency situation, but simply described the event. For example, Tweets like, “hurricane sandy oooo”, “@JennyFletcher1 I get a hurricane day”, “If the power goes out, will we still have internet on our phones ? #AskTwitter”, “I have my sandy survival kit. I have my bottle of sleeping pills and a warm comfy bed. Goodnight”, were related to the Hurricane Sandy event but had nothing about emergency. There were even some Tweets that did not identify hurricane as an emergency, but as something fun. For example, “Cheers to sandy ************”, “Who wants to go play in Hurricane Sandy? A little football or capture the flag anybody?”, “Perfect time to catch up with some school work. #Sandy can have a good side.”, and “Wish this storm would just hurry and get here.....I want some damn snow!”.

According the study, there are 361 geo-tagged Tweets collected. Among the 361 Tweets, there are 313 Tweets related to the event. Among the 313 Tweets, only 141 Tweets related to an emergency situation. A summary of the sample Tweets can be found in Table 5.2.

<table>
<thead>
<tr>
<th>Keywords</th>
<th># of Tweets</th>
<th># of event-related Tweets</th>
<th># of event-related and emergency-related Tweets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sandy</td>
<td>123</td>
<td>122</td>
<td>47</td>
</tr>
<tr>
<td>Hurricane</td>
<td>52</td>
<td>52</td>
<td>15</td>
</tr>
<tr>
<td>Power/electricity</td>
<td>48</td>
<td>47</td>
<td>23</td>
</tr>
<tr>
<td>Storm</td>
<td>42</td>
<td>40</td>
<td>19</td>
</tr>
<tr>
<td>Snow</td>
<td>26</td>
<td>20</td>
<td>14</td>
</tr>
<tr>
<td>Food/restaurant/hungry/starving</td>
<td>20</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Closed/collapse/collapsed</td>
<td>18</td>
<td>16</td>
<td>12</td>
</tr>
<tr>
<td>Water/drink/drinking</td>
<td>18</td>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>Flood</td>
<td>8</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>---------------</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Tree/trees</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hotel/hotels</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>Summary</strong></td>
<td>361</td>
<td>313</td>
<td>141</td>
</tr>
</tbody>
</table>

Table 5-3: Summary for the sample VGI dataset

5.4 Evaluation

In this case study, the collected geo-tagged Tweets were plotted on a web map. The benefits of this application for resource management in mass evacuation can be discussed in the following sections.

5.4.1 Geo-tagged Tweets for resource dispatch.

First of all, the Tweets are collected in real-time or near real-time. In disaster management, real-time information is critical and is not available from official sources in a timely manner. Under the situation of no data, the real-time data from volunteers could benefit disaster management professionals. This kind of real-time data might have problems, but it provides valuable information especially under the situation of no data. Secondly, the data is stored in a distributed geospatial database; that means it can be accessed by users from different locations. Any interested agencies can access the (near) real-time data for their own analysis. Thirdly, as a web mapping application, it provides location information of the places that have requirements for different resources. The disaster management professionals can make corresponding decisions based on the collected data. If they think the data is reliable, the disaster management professionals can dispatch resources to the locations. If they think the data is suspicious, the disaster management professionals can double-check the data.

Fourthly, the web application groups the collected Tweets in different search categories. With this capability, users can focus on the Tweets of their interest. What’s more, the application can provide more information about the collect data. As discussed, a tweet has a field, “created_at”, which records the UTC time when this tweet was created. With the time information, this application can display the Tweets that fall in the desired time frame. The time frame display of the collected Tweets assists users to understand the real-time demand trends of the resources. This function displays not only the total Tweets, but also the Tweets in individual categories.
Besides providing the trend of Tweets in different categories, this web application also provides a heatmap capability for users to study the collected geo-tagged Tweets. The heatmap can help users to understand hotspots that have the most requests for resources. The heatmap is generated based on the visible geometrical features, that means the heatmap can be generated not only for all Tweets but also for the Tweets in different categories and for the Tweets at different specific time intervals. Figure 5.8 is a sample heatmap that is generated based on the geo-tagged Tweets in the medical category.

![Heat map based on VGI in the medical category](image)

**Figure 5.8: Heat map based on VGI in the medical category**

### 5.4.2 Integrating VGI with an online mapping application

To integrate VGI with other traditional GIS datasets, the system first needs to collect the Tweets based on predefined locations and subjects. Both case studies demonstrated that the proposed system has the capability to retrieve Tweets through the Twitter Search API. Tweets with location information are converted into a GIS feature class and are stored in distributed geodatabase. Tweets without geo-tags are saved as CSV files.

Secondly, to integrate VGI with other traditional geospatial data, social media should include location information. Tweets have two types of location information. One type is the actual longitude, latitude coordinates, and the other is the regional location information. Location information is generated principally by mobile devices that are GPS enabled. When a user Tweets with the location information option turned on, the location information will be automatically
embedded into the tweet entity. In this case, users’ mobile applications first identify a location, and then post the message in the popup window. This type of locational data is the actual location from where the tweet is sent. In the tweet entity, one of the tweet’s fields, “coordinate”, is used to store this information.

The second type of location is principally from non-GPS enabled devices, such as desktop computers, which Twitter users can specify the location information only to the level of state, city or town. In this case, the location information stands for the whole area. One of the tweet’s fields, “place”, is used to store this piece of information. For example, the location information of City of Morgantown, WV is stored as a rectangular bounding box.

5.4.3 Real-time data accessibility

Real time data is critical in social, crisis, or disaster events. It provides live information from which decision-makers and emergency professionals are able to make well-informed decisions. Both case studies employ the Twitter Search API for information retrieving. The Twitter Search API can collect Tweets that are less than 7 days old and that satisfy the search requirements. Those Tweets might not be exactly real time information, but they do provide important background and context information in a time frame more recent than authoritative data, which may be years old. During the initial phase of tweet collection, it is useful to collect Tweets that are 7 days old and newer for context.

After the initial data collection time, the collection function will retrieve all posted Tweets. In every subsequent time period, the collected Tweets are real time information. In this case, another Twitter Search API’s search parameter, “search_type” which has been discussed in Chapter 4, becomes important. In this study, the value “recent”, which stands for the most results, of the “search_type” parameter was used. In disaster events or other time sensitive events, the Tweets collection should retrieve the latest Tweets. That means the value of “search_type” parameter should be set as “recent”. The Tweets collected with “search_type” of “recent” are real time data. Geo-tagged Tweets collected in real-time are inserted into the Tweets feature class that are part of distributed geodatabase. This online-editing capability is provided through ArcGIS Flex API. As discussed in the previous section, the same Tweets feature class is in a web map service which is
in RESTful style and is accessible by other applications, including web mapping applications and other fat-client applications, such as ESRI’s ArcGIS Desktop applications (Figure 5.9).

![Diagram of data collection and live data access]

**Figure 5.9: Real time data collection and live data access**

### 5.4.4 Knowledge discovery

As discussed in the previous section, in the context of Twitter, selection is based on the five parameters: “q” (query for keywords), “geocode” (the search location), “since” and “until” (the time frame), and “search_type”. This keyword-based knowledge discovery is used to extract location-related and event-related VGI. In the Twitter Search API, the parameter “q” is mandatory and is used to specify the subject of knowledge. The parameter “geocode” is optional and used to specify the search location. The parameters of “until” and “since” are used to specify the timeframe, which is only useful at the initial phase, Tweets that have been posted within seven days are collected. After the initial phase, the parameters “until” and “since” are not important. The importance of “geocode” will be discussed in the next section, and the parameter “search_type” has been discussed in the previous section.

Through the Twitter Search API, the knowledge discovery of a specific subject is based on the search parameter of “q” for useful information. In this research, the possible keywords have been discussed in each cases. The usage of “q” parameter can be illustrated as following (more detailed uses of the query parameter can be found at Twitter’s website):

1. q=water, which is used to discover Tweets that have the word “water”.
2. q=water thirsty (without quote), which is used to discover Tweets that have both the words
“water” and “thirsty”, and the order of the words does not matter, and words can be separated.

(3) q="water thirsty" (with quote), which is used to discover Tweets that have the exact match of “water thirsty”.

(4) q=water or thirsty, which is used to identify Tweets that have either the word “water” or the word “thirsty”.

(5) q=water –thirsty, which is used to identify Tweets that have the word “water” but have no instances of the word “thirsty”.

In our case studies, each keyword is selected using option 4, to assure not to miss any related tweet words.

5.4.5 Symbolize Tweets layer

As a Web GIS system, mapping is the primary capability. In this research, mapping is not directly done on the client side, but also on the server side. The mapping capability is through the map that is created in applications such as ArcMap. The cartography design is done in ArcMap, and can be accessed and displayed by online GIS application. As a layer in ArcMap, the Tweets GIS layer can use different symbols to indicate different subjects. As a layer in the map document, the symbol of the features can be designated (see Table 5.2).

After the map document is created successfully, it can be published as a web map service in REST style. When this web map service is displayed on a Web GIS platform, the symbology can be displayed to the other users.

5.4.6 Storage and distribution of live data

The web application is designed to collect (near) real time geo-tagged Tweets to support the resource management in mass evacuation. As discussed before, the Twitter Search API can retrieve (near) real time Tweets based on the search parameters. At the same time, the live data are stored in a distributed geodatabase. First, the targeted geo-tagged Tweets layer, which includes the predefined data schema, such as “TweetID”, “CreatedAt”, “SearchKeyword”, “KeywordCategory”, “twtText”, and so on, is housed in a distributed geodatabase. Second, the geo-tagged tweet layer is added in ESRI’s ArcMap and the cartography is defined in ArcMap. Through ESRI’s ArcGIS Server, the created map document is published as a web map service with the feature service
enabled. The feature service allows online GIS feature editing, which is key for live data storing. With this capability, the web application can insert the collected geo-tagged Tweets into a database.

Live data distribution includes two perspectives. First, the raw data is accessed over the Internet. The database that houses the geo-tagged Tweets is a distributed geodatabase. That means users can access the saved data directly. The live new data can be accessed by users at the same time. Second, the web map service is in distributed access. Right after they are collected, the collected data can be published a web map service immediately. Users can display the live map in their own platform by using the URL of the web map service.

5.4.7 Handling possible duplicates

In the Tweets search component, a tweet could be collected several times if it satisfies different search terms. Under this situation, those Tweets will not be counted as duplicates. The duplicates discussed in this section are duplicated on both the tweet id and search keywords. As discussed in section 5.4.6, the geo-tagged tweet GIS layer in the distributed database includes TweetID and SearchKeyword fields, which are used to identify possible duplicates.

Figure 5.6 illustrates overlaps between the search sub-regions. It is possible that the Tweets generated in the overlapping areas will be collected several times. Even though some Tweets are collected several times, the Tweets storing component has an algorithm to identify and delete duplicates. The algorithm is simple: for the next collected tweet, the program will check whether this tweet exists in the database or not, based on the TweetID and SearchKeyword field. If the tweet does not exist in the database, this tweet will be inserted into the database; otherwise, the tweet will not be inserted into the database (see Figure 5.10).

![Figure 5.10: An algorithm to remove possible duplicate Tweets](#)
5.4.8 Addressing two types of VGI error

As discussed in Chapter 4, two types of error are possible: false positive errors and false negative errors. In a false positive error, a high volume of (spurious) VGI data misrepresents the existence of a phenomenon that does not in fact exist. For example, for a situation of a bridge collapse, the VGI data may indicate that a bridge at a certain location has been damaged. But in reality, this bridge is not damaged, the report has been made in error, or perhaps maliciously. The false negative error means there are no or very few VGI posts to indicate an existing on-going phenomenon. For example, for the situation of bridge collapse that does exist, but for some reasons, e.g. the collapse happens too soon and leaves no warning time, citizens that are at this location do not post this information online. As a result, there is no VGI data to indicate that bridge is, in fact, damaged. During a mass evacuation, the evacuees along the corresponding route will potentially be misdirected or stranded.

5.4.8.1 How VGI affects decision making.

First, we have to admit that, VGI cannot be the only data source for making correct decisions. Other information, such as daily operational data, framework data, data from official agencies, data from in-field workers, naturally will play important roles in decision-making. While the authoritative data might not be updated quickly and frequently, VGI can provide real-time information which is critical in mass evacuation and other disaster-related events and other social crisis events to reinforce and supplement such data. The two types of error may result in different costs. In the context of disaster management, the costs in responding to false positive information are a waste of time, personnel or other resources, which are limited during an event; and the costs in not responding in the case of false negative information are related to putting the victims in circumstances where no resources are available and this situation is not known by the appropriate agencies. While both errors mislead decisions, the cost of no response due to false negative errors could be higher and more risky.

5.4.8.2 Identification of potential errors

The overall cost of two types of errors could be costly or even disastrous. Identifying the potential errors existing in VGI are important. It could save time and resources due to false positive error,
or it could help identify the locations that have demand on resources, but no indication base on VGI data.

In order to identify the potential error, the corresponding disaster management professionals must have clear understanding of the areas that are under their management; in the preparation stage, a disaster management professional should compile a list of locations that may need extra attention or are known or likely to be problematic during catastrophic events. After compiling the list from different professionals, an overall location map could be created to indicate possible problem locations that need to be checked during an event, in other words, a vulnerability map. This ‘potential problem location’ map should be loaded into the online VGI mapping system. Users can compare the potential problem locations with the current VGI data layers for four possibilities. First, if a location in the potential problem location map matches the location of the VGI data, it indicates that such locations should be confirmed, meaning that the location is in the need of corresponding resources. Second, if a location in the potential problem location map has no VGI data, it could indicate the VGI data includes a false negative error at such location. Third, if a location has many VGI data, but such locations are not represented on the potential problem location map, it could indicate that the VGI data includes false positive errors. Fourth, a location could not be included in the potential problem location map and has no VGI data. Such locations could be locations without any problem and without needs for resources.

The first situation confirms the need for resources and the fourth situation indicates no need for resources, respectively. Those two situations do not require further investigation. The second and the third situations indicate that the data from the potential problem map and VGI sources do not confirm each other, thus indicating that further investigation is required. The disaster management professional can send an in-field worker to that location to resolve the uncertainty. A flow chart below can be used to indicate the error identification process (Figure 5.1).
5.4.9 Finer temporal scale

The Kent State shooting utilized two temporal scales to study VGI: a daily scale and hourly scale. While the daily study remains useful, a finer temporal scale can provide more important and accurate information for an event. The finer temporal scale study can help disaster managers to better understand the on-going event. The study of hourly VGI posted by Twitter users in Case Study 1 is done manually. For the second case study, the VGI management component of the online Web GIS platform includes a function of temporal VGI management. Users can view VGIs in different time frames and view the corresponding heatmaps, and can reset temporal scales coarser to several days or finer to several hours.

5.4.10 The design of search sub-regions

In this case study, due to maximum limits on Tweets return, the search area were divided into different sub-regions with different radii based on population density so that more Tweets can be collected. The region division is based on one assumption: more population indicates more Tweets. This study divided study area into different sub-regions manually. There is a possibility that the division can be done programmatically. A proposed way is to:

- load census blockgroup data of the study area
- find the population density for each blockgroup
- load community demographic and economic data of the study area
- intersect population density map and community index map
• identify the search center and radii based on the resulting polygons of population density and community demographics.

5.5 Summary

Beginning with the introductory case study of the 2014 Kent State campus shooting, this chapter discussed the implementations of two prototype case studies to assist in evaluating the pros and cons of the proposed system. The system is designed by using the Twitter Search API for tweet collection, by using ESRI’s JavaScript API for live Tweets saving, and by using distributed geodatabase for live tweet storing. From the evaluation, we can see that the system has the capability to collect and post data in near real-time and to store them in distributed geodatabase so that the collected data can be distributed for users over the internet to access them.

As indicated in the evaluation, the implemented system can help emergency management professionals on the distribution of resources in cases of mass evacuation. Since the geo-tagged Tweets can be mapped in a timely way on a distributed web map, the disaster management professionals can access the collected data from any places with Internet connection. Besides this advantage, this system has other unique benefits in resource management in disaster management. These advantages can be seen from the following perspectives: the Tweets are (near) real time information that is critical in mass evacuation especially at times when official information is not available; the geo-tagged Tweets have been plotted on a web map so that the disaster management professionals can easily distribute resources to the right locations; the system can help users understand the trend of the users requirements either for the total resources and resources in different categories; and the system can generate heatmaps (and other forms of geovisualization) for the geo-tagged Tweets for users to understand the areas that have most urgent requirements for various resources.

At the same time, we have to acknowledge that the collected Tweets have problems. This study will not address the credibility issue, liability issue and the privacy issue, but does address the accuracy issues: spatial accuracy and content accuracy. For the spatial accuracy, this research focuses on the geo-tagged Tweets, which use accurate longitude and latitude. If the location is from mobile devices, the spatial accuracy of location information depends on the location-based service providers and is beyond the control of this research. If the location is generated by using
embedded maps, the spatial accuracy depends on the online map service providers and is also beyond the discussion this research.

This research discusses the accuracy of the tweet content. To study the content accuracy, Section 5.3.5, “Interpret Tweets manually”, discusses the manual interpretation of collected Tweets. According to the study, which is based on one community (Springfield, WV on October 3rd, 2012) of the study area, there are 361 geo-tagged Tweets collected. Among the 361 Tweets, there are 313 Tweets related to the event, the Hurricane Sandy. Among the 313 Tweets, there are only 141 Tweets related to the emergency situation. As discussed above, the tweet collection is based on keywords due to the fact that the lack of semantic search support in Twitter’s API. In Twitter’s Search API, the search parameter “q” can be complex. But, no matter how complex the search term is, it is search against the message which is composed by a user that does not normally follow the structure of regular language.

Other than the accuracy issue, there are two types of errors in Tweets and other social media-based VGI as discussed in the Section 5.4.8, “Addressing two types of VGI errors”. We admit that the false positive error could be easier to be addressed. The false negative error is a real challenge to handle since there is no VGI available. For the resource management, the false positive error leads to a mis-dispatch of resources to un-needed locations. It’s a waste of resource. On the contrary, the false negative error could be more deadly. It will put the evacuees in the un-mapped area in a stranded situation. Compared to the false positive error, the false negative error could be far more costly.

This chapter has discussed some other aspects of the proposed system, such as the finer temporal scale of the Kent State campus shooting and the manual interpretation of a sample VGI from the Hurricane Sandy case, in addition to some key concepts for using VGI in disaster management including how to integrate VGI with other traditional/official dataset, how to implement real-time (or near real-time) data accessibility, how to discovery useful VGI, how to symbolize VGI, and how to address the potential two types of VGI errors. The next chapter will be a user guide to the on-line mapping system.
Chapter 6: The online geo-Tweets collection system

6. The online geo-Tweets collection system

Chapter five discussed pilot geo-Tweets collection systems for a 2014 shooting incident at Kent State University and Superstorm Sandy in the fall of 2012. Each prototype was developed to test various functions with the relevant Application Programming Interfaces (API). At the time of development the Twitter API was in version 1.0 and the most popular ESRI mapping API was ESRI Flex API. As both APIs evolved, the pilot geo-Tweets collection system no longer worked; consequently the research design had to be modified to accommodate changes in the APIs. Nevertheless, the two prototypes provided valuable information, and the lessons learned and their capabilities have guided and been incorporated into the current on-line version.

First, from the Tweets collection perspective, API 1.1 was released on May 7, 2013 and since then Twitter has no longer allowed direct connection to their server for data collection. To collect Tweets, developers need first to obtain keys and tokens, such as CONSUMER_KEY, CONSUMER_SECRET, OAUTH_TOKEN, and OAUTH_SECRET, to authorize Tweets collection from Twitter. Prior to sending the requests for Tweets collection, the required authorization tokens and keys need to be packed in a server-side programming technology (PHP, in this case) and be sent to Twitter’s server for authorization. The Twitter server will not process the data collect request until it detects and confirms the authorization. After the Tweets are returned to the client side, the collection system will keep and map the geo-tagged Tweets.

From the mapping perspective, due to limitations of Flex, many necessary applications are not supported by mobile devices. Hence, Flex-based development is not a suitable choice for web mapping when users wish to access the application on mobile devices. To accommodate both desktop web browsers and mobile web browsers, mapping applications now are being developed using JavaScript. Additionally, in the case of ESRI’s mapping APIs, ESRI no longer maintains Flex API for online mapping because Adobe no longer officially maintains Flex. Thus the current online geo-tweet collection system has been developed using ESRI’s ArcGIS JavaScript API.

The overview interface of the geo-Tweets collection system (Figure 6.1) includes three components: the mapping component (including a basemap sub-component), the Tweets collection component, and the Tweets management component (including a time management sub-
component). Figure 6.2 displays a set of Tweets, which were collected using specific keywords, and the corresponding heat map.

The rest of this chapter takes the form of a user’s manual. It introduces how to use this system, including how to specify the tweet search parameters, how to use the management tools to study the collected geo-Tweets, and how to use the mapping part of the system.

Figure 6.1: Overview of the online geo-Tweets collection system

Figure 6.2: Classified geo-Tweets with heat map
6.1 How to specify the search parameters

The application is accessed on the internet using URL http://www. The first time the system is used, the initial step is to start the tweet collection process (see Figure 6.3). However, if geo-Tweets are already collected and mapped, users can start another tweet collection process, append to the existing set, or browse and study the existing data. The next section introduces the specification of corresponding parameters for tweet collection. Before specifying the search parameters click on the “Search” tab to make it visible (Figure 6.3).

![Figure 6.3: specify Tweets search parameters](image)

First, the target search area (center and radius) needs to be identified. As indicated by Step 1 of Figure 6.3, this area can be set by using the zoom in / zoom out and pan tools on the map component. While the search center is set to the current map center and it is not changeable, users need to specify the search radius which is measured in miles. At present searches are restricted to circles.

Second, users click on the button at Step 2 of Figure 6.3 to generate a set of two text input boxes for search keywords. Due to semantic ambiguity, people can and do use different word(s) to indicate the same meaning; for example, for the meaning of ‘shelter’, Tweeters could use words,
such as, shelter, house, hotel, motel, place to live, tent, etc. The design of this section provides two text input boxes. The text input box on the left is used to indicate the generic term of a search keyword; and the text input box on the right holds a set of possible terms or synonyms that Twitter posters might use. Figure 6.4 and 6.5 illustrate the before and after status when clicking on the button. Users can also click on button to delete a keyword group that is not needed.

The third step is used to specify the search time frame as indicated in Step 3 of Figure 6.3. The default date for the To-box is the current date and the default date for the From-box is one day before the current date. Due to the fact that the Twitter Search API cannot retrieve Tweets that are more than 7-days old, boxes cannot have dates 7 days older than the current date. Also, the date in the From-box must always be older than the date in the To-box.

6.2 How to use the provided management tools

The major management functionalities (see Figure 6.6) include providing a heat map, and providing a classification for collected geo-Tweets, which allows the user to turn on/off individual tweet groups, providing the number of geo-Tweets in each keywords category. The application also provides sophisticated time management functionalities (see Figure 6.7).

The management tool assigns different symbols to Tweets collected with different keyword categories (or, alternatively, the generic term of keywords). It also automatically calculates the number of Tweets and displays the tweet count by the keyword category. The heat map is associated with the currently (selected) visible geo-Tweets. That means, if a geo-tweet category, e.g. flood, is turned off, the heat map will be generated based only on the remaining visible geo-Tweets. Furthermore, the heat map function is also integrated with the time management. In short,
if the Tweets are not within timeframe specified in the time slider, they will not be visible nor be used to generate heat map.

![Image of Geo-Tweets management interface]

Figure 6.6: Geo-Tweets management interface

### 6.2.1 How to use the time management tool

![Image of Time management tool]

Figure 6.7: Time management tool

Once the geo-Tweets are collected, the system will generate or re-generate a time slider on the time management component with the time extent of the collected geo-Tweets. There are two settings for configuration of the time slider. First, it is important to verify or reset the time interval. As shown in Option 1 on Figure 6.7, there are two time interval options, hour or day. Second, the following steps are necessary to reset time frame. As shown in Option 2 on Figure 6.7, there is a From-list and a To-list. These two lists hold all the dates between the oldest geo-Tweets and the latest geo-Tweets in the database. Users can specify the time interval to be a day or an hour, which determines the number of ticks on the time slider. If the time interval is measured in hours, the time slider may become too dense to be distinguished; users can use the reset time frame function to change time frame to a narrower time frame so that the ticks on the hour-interval-based time slider will be discerned clearly. This step is illustrated in Figure 6.8, at which the time interval is

97
an hour and time frame is 8 days, and Figure 6.9, at which the time interval is an hour and time frame is 2 days. The best practice is to set the time frame to be less than three days when time interval is set to hours.

On the time slider, the time difference between two ticks can be set manually while the default is one time interval (which is based on the choice of Option 1 on Figure 6.7). The collected time-sensitive geo-Tweets will show up on the map based on the two time ticks on the time slider. The other geo-Tweets that are not within the two time ticks will show up on the map. When the button is clicked, the time slider will slide automatically and geo-Tweets within the corresponding two times will show on the map; the heat map will be re-generated based on geo-Tweets currently visible on the map. Or, users may click on the button or the button to determine a range.

![Time Slider](image)

Figure 6.8: time slider with time interval in hour and time frame in 8 days

![Time Slider](image)

Figure 6.9: time slider with time interval in hour and time frame in 2 days

### 6.3 Map operation

The system provides some standard map manipulation functions, such as zoom in (click on the plus button on the navigation bar or scroll the scroll button on the mouse or hold SHIFT key and draw a box) zoom out (click on the minus button on the navigation bar), and pan functions. It also provides a selection of basemaps (Figure 6.10) so that users can have a better understanding of the collected geo-Tweets based on different framework data.
6.4 Summary

This on-line mapping system provides functions for geo-tweet collection and data management. As the Twitter Search API-based Tweets collection function is complete, more geo-Tweets management related functions can be foreseen for future development. For example, a time-sensitive histogram may be developed to display the Tweets counted during every time interval by keyword. For example, it will be helpful for users to view the tweet count of different keywords at different times to follow the development of an event. Another example is that the system can provide a function for manager to upload resource locations in CSV file format, shapefile format, or even, a web-map service format. With the resource location map, users, especially managers, have a picture of the current situation of resource requirements and resource distribution. Also, the system can be modified to provide an editing capability for managers to add the locations of new resources on the map so that users can identify added resource locations in real-time.
Chapter 7: Summary and Conclusion

7. Summary and Conclusion

Using a combination of a hypothetical mass evacuation scenario and real-world events, this study has investigated how volunteered geographic information, specifically Tweets, could benefit the efficient identification of unfolding events and effective dispatch of essential resources following a major disaster. Rural areas are candidate ingress and pass-through areas for evacuees seeking shelter, food and water, fuel and medical assistance, but may be unprepared or under-prepared to provide necessary resources, therefore real-time notification of needs is a valuable tool in emergency management. The first chapter of this dissertation introduced the research context, including an overview of current perspectives of disaster management and mass evacuation, the role and use of geospatial technologies in mass evacuation, and established the significance of timely information. Chapter one also introduced the overall research goals and objectives, and set out the tasks necessary for implementing these objectives. In the second chapter, the literature of mass evacuation and the importance of resources in ingress areas were reviewed, especially as related to rural areas. Additionally, the chapter enumerated the benefits and challenges faced in order to incorporate distributed GIS in disaster management. Most importantly, the concept of Volunteered Geographical Information (VGI) was introduced, including key ideas such as the characteristics of VGI, its potential value, and questions confronting the use of VGI in disaster management.

Having developed a conceptual framework, Chapter 3 proceeded to discuss the research questions necessary order to realize the overall objective: i.e. to provide real-time VGI to support the requirements of logistic resources during mass evacuation. Chapter 4 then discussed the methods employed to implement the proposed system, i.e. an online mapping application with near real-time tweet collection, live data storage, and live data dissemination capabilities. The methods discussed included the identification of the test data source – Tweets from Twitter.com, tweet collection technologies – the Twitter Search API, the Web GIS system, and the implementation of distributed geodatabases.

Based on the methods proposed in Chapter 4, the fifth chapter reviewed the implementation of two case studies as prototype systems. The 2014 Kent State shooting case was used to explore Tweets
collection using the Twitter Search API 1.1. The 2012 Super Storm Sandy, as experienced in the Washington-Baltimore metropolitan area, provided a case study using Twitter Search API 1.0 to illustrate the potential of using VGI (geo-tagged Tweets, in this case) to support near real time mass evacuation-related resource management. Finally, Chapter 6 presented the resulting on-line Tweets collection, management and display system in an application using real-world data associated with Hurricane Joachim in October 2015.

The following section contains an in-depth discussion of the contribution of the present research, the problems identified during its completion, together with their possible solutions, and the future research opportunities identified by this study.

### 7.1 The contributions of this research

In the response phase of the disaster management cycle, real-time information is critical, but, due to the limited personnel resources of disaster management agencies, real-time information may be sporadic and incomplete after the onset of an event. As real-time information from average citizens in the event area accrues in the form of VGI, managers have the potential to provide improved disaster management. The research investigates methods required to harvest VGI that is related to a crisis or disaster event to support resource dispatch in mass evacuation. Beyond mapping (near) real-time VGI in the form of Tweets, this research also has investigated how to implement the live storage of Tweets into a distributed geodatabase and how to redistribute the collected VGI in real-time. The following section discusses the unique contributions of this research.

Firstly, this research demonstrated how geo-tagged Tweets can help resource dispatch in mass evacuation by using a hypothetical mass evacuation triggered by Super Storm Sandy (2012) in the Washington-Baltimore Metropolitan area. The mapped geo-tagged Tweets contain location information of resource requirements and the resource types that are required. Based on this information in the context of an Emergency Management System, the emergency management office can dispatch resources to the victims for immediate assistance, or direct personnel to verify the situation. The unique value of VGI in resource dispatch in disaster management is its real time feature. The benefits of VGI in mass evacuation include several aspects: Real-time VGI plays important role at the response phase of disaster management when data is not available in a timely manner from official sources. Mapped Tweets can identify locations that have requirements for specific types of resources. Other than the direct mapping capability, the application provides a
temporal scale for the collected VGI, so that managers can understand the trend of victims’ requirements of different resources categories. The current application also provides a heat map for managers to visualize areas that have the most pressing requirements.

Based in part on the mapped Tweets, decision-makers determine when to dispatch resources and to where the resources will be dispatched. The decision “when to dispatch resources” could be based on a predefined threshold, which could be based on a given number of Tweets that require resources and previous experience. For example, decision-makers can sketch a specific area and the system will return statistical information about the specific area: the number of total Tweets and the number of Tweets in a resource category. The responses can be compared with the pre-existing deployment of resources for an informed decision. A further question is “to where should resources be dispatched?” It is neither rational nor realistic to dispatch resources to the location of every individual Tweet. It will be more reasonable to dispatch resources to the center of a specific area of multiple Tweets. The center could be either the geometric centroid (center of gravity) or the minimum travel time center of a tweet cluster. Once the actual location is determined, the decision-maker is able to add the location of new resources to the map. For the decision-makers, the next question is how to inform users who have posted requests for resources that a response has been made. There could be several solutions to communicating with requestors. One solution is to retweet every individual Tweet once the required resources are dispatched to a specific location. Another solution is to post a Tweet through the organization’s Twitter account. Through this kind of broadcasting, not only the followers can see this message, but also the other users can check out this message by browsing the emergency management organization’s Twitter Tweets.

Secondly, this research has illustrated how to collect real-time VGI for a specific event, including the location and the topic of interest. In this study, Twitter’s Tweets (especially, geo-tagged Tweets) were chosen. This research provides an interface for managers to specify the parameters corresponding to different requirements, such as keywords (q), location and region (geocode), time frame (until and since), and others. After the collect data command is issued, the collected geo-tagged Tweets are saved in the distributed geodatabase and mapped through a web map service. Collecting Tweets in real time and mapping them through a web map service are the unique features in which this study differs from other VGI mapping research.
Thirdly, the VGI collected were stored in a distributed geodatabase so that they can be used for other studies that have an on-going interest in the collection of Tweets. This research illustrates how to save and redistribute the collected Tweets. In contrast to other VGI mapping applications that map VGI on-the-fly (that delete the mapping data once the mapping application is shut down), this approach manages data by saving the Tweets in a distributed geodatabase so that the mapped Tweets can be accessed not only by this application, but also other applications. Of particular value is the opportunity for post-event analysis during the mitigation phase of the emergency management cycle. The live redistribution has been implemented from two approaches: the redistribution of data by direct access of the distributed geodatabase and the redistribution of map services that are composed of the collected data. This research has employed technologies, such as client programming technologies (ESRI’s Flex API or ESRI’s JavaScript API), ESRI’s server technologies, including RESTful feature service and RESTful map service. ESRI’s JavaScript API (or ESRI’s Flex API) and REST services play important roles in the implementation of online VGI.

7.2 Discussion

This dissertation has illustrated the advantages of using geo-tagged Tweets in resource dispatch in disaster management. This section discusses problems that the research faced and lessons learned during implementation and operation. Questions have been raised, such as how to ensure the availability of Twitter data, how to ensure the completeness of the dataset, how will possible future API evolution affect application development, tweet collection, how do possible VGI errors affect the resource dispatch, and how will semantic ambiguity affect the decision making. This section discusses these issues in depth.

Extending from the discussion in section 6.1, possible semantic ambiguities exist in the Tweets’ content. For example, search term “water” is used to retrieve Tweets related to the requirement of water in mass evacuation. But some Tweets were also retrieved, which are not related to drinking water, such as “My backyards under water. #nbd **”, “Water is coming into the basement!”, and “Foot of water in the basement.... Ain’t nobody hot (sic) time of that! Throw me another beer”. To a human reader those Tweets immediately indicate a flooding situation. The uncertainty of meaning results from the fact that semantic search is not supported in the Twitter API. The keyword search is simply a match for the actual words, but the Tweet content does not always indicate the meaning that the search term is seeking. This situation requires that the disaster
management professionals check whether the Tweet’s creator was requesting the particular resource, or was merely posting content that happened to include the search term. One possible solution is to make the search term more precise. For example, instead of using the single word “water” to search Tweets, a search phase like “need drinking water” could be used to get more precise results. But a more complicated search term will lead to filtering out of large amounts of data because people who use Twitter will likely use colloquial and abbreviated language. The language informalities of social media communication places the use of social media-based data in a difficult situation. As the search is keyword-based, the post would not be selected if the keyword is misspelled or abbreviated.

A second important issue is related to the existence of false positive and false negative errors. A false positive error is said to occur when a high volume of Tweets signifies and misrepresents the existence of a phenomenon that does not in fact exist. Likewise, a false negative error occurs when there are no or very few Twitter posts to indicate an actual on-going phenomenon. These two types of error are due to the nature of VGI, which are not official data but created by volunteers. A recent news report on using Tweets to track tornadoes stated that “not everything tweeted will turn out to be true” (National Public Radio April 20th 2015).

The two types of error may result in different costs. In the context of disaster management, the costs in responding to false positive information are a waste of time, personnel, or other resources, which are limited during an event. The costs of not responding to a real need because of false negative errors result in leaving the victims in circumstances that have no resources available and those needs are not known to the appropriate agencies. While both errors mislead decisions, the cost of no response due to false negative errors could both be higher and more risky. As is further discussed in section 7.3, the question of error is more complex than a dichotomy between true and false.

While this study focuses on geo-tagged Tweets, non-geo-tagged Tweets should be studied for a more complete view the data. As indicated in the study, geo-tagged Tweets comprise only 5 to 10 percent of the total search results. That means a majority of the retrieved Tweets were not available for location-based study because of a lack of location information (i.e. geo-tag) or a lack of the ability to map the content (textual information). Tweets without actual location information are collected initially as their creators’ do have location information in their profile. Under many
situations, the location information of a user’s profile does not mean the tweet is created at or for that location. Even though Tweets without actual location information can’t be mapped, they provide useful information for disaster management professional to understand the overall situation of the search area in terms of their number and content. The geo-tagged Tweets could be treated as a sample set of the total tweet population that satisfies the search criteria. While we focus on geo-tagged Tweets, it’s still meaningful to have a general understanding about the total population of Tweets. For example, we may learn the total requirements for resources based on the number of Tweets of each resource category or learn the trend of total resource requirements or for different types of resources. Another future study direction could be to investigate the non-geo-tagged Tweets in the context of resource dispatch in mass evacuation.

If no Internet connections or devices are available for users to post data, VGI of any type will be unavailable in the affected area. Without web access, data producers cannot create and post VGI, a situation that could be one possible reason for false negative errors. Without available data, there will be no resources dispatched to that location on the basis of tweet information alone. VGI availability makes two assumptions; the availability of an Internet connection; including wired connections, wireless connections (Wi-Fi), or G3/G4 connection from phone service carriers. The first assumption about VGI availability is that the infrastructure functions properly during an emergency event. In some extreme situations, communications infrastructure could be damaged; as a result the availability of VGI will be questionable. For example, right after the 2008 Wenchuan earthquake, the communication infrastructure, such as cell towers, fiber cables, and even the normal transportation network in the epicenter and other sub-centers were destroyed completely, whereas after the Haiti earthquake of 2010, cellular networks remained operational. The second assumption is that appropriate devices are on hand for citizens who are willing to contribute data. In practice these devices include mobile devices carried by citizens for normal communication, such as smart phones, PDAs, laptop computers and tablets. Under normal circumstances, the availability of those devices is reasonably certain and their adoption is widespread. But during extreme situations, personal devices might be damaged or not work properly, for example by running out of power. Dead batteries - a frequent and annoying condition in normal use – would be exacerbated in an emergency. The affected citizens would no longer have access to tools for regular communication, or any other device-based activities, including posting personal observations to a social network.
In this study, VGI is collected from a social media website to support resource dispatch and data completeness is a concern. In the case of Tweets, as indicated by Twitter.com, the Tweets collected by using Twitter API is not guaranteed to be a complete set based on the search parameters, but a sample set determined by Twitter. Even if the number of returned Tweets is within the threshold, Twitter will not guarantee the completeness of the data collected using Twitter’s APIs. Consequently, because Twitter’s RESTful API does not guarantee total retrieval, it remains uncertain that the collection process retrieves all the data that satisfy the search parameters. The data completeness issue then arises from the absence of relevant VGI, which is invisible to disaster management professionals. The effects of this issue can be seen from two perspectives. First, for a specific search, especially in a high population density area, it is possible that the number of Tweets exceeds the API limit. This limitation may be addressed by imposing temporal and spatial extent restrictions, which may require payment for service. Another important issue about the completeness lies in the nature of the Twitter API. To address the completeness issue from this perspective, Twitter recommends communication with its two other partner companies, Gnip and Datasift, which host the complete Tweet dataset. Retrieving Tweets from Twitter.com are free, but with some limitations, such as a search window and the maximum number of Tweets that may be retrieved. From the perspective of this research project, Tweets from Twitter API were enough for demonstration purposes. In the case of utilizing VGI on a production platform, the use of Tweets hosted by Twitter’s partners should be considered. Twitter’s partners provide their own API for accessing Tweets on their servers and can provide users the complete dataset based on your search parameters, but as commercial entities, Twitter’s partners charge for data.

Another serious issue has emerged recently. Beginning from Nov. 14th, 2014, the Twitter Search API will not return non-geo-tagged Tweets when a search area is specified. Only Tweets with explicit location information will be returned, again raising the completeness issue. It is possible that the total population of relevant Tweets is not returned. As of Feb. 6th, 2015, this issue was still under investigation by the Twitter API team. It is estimated that Twitter Search API will work properly from April 2015. While it has no large effect on research that focuses on geo-tagged Tweets, it poses issues to applications that are interested in the total number of keyword-related Tweets in a study area. For example, some researchers are interested in all the Tweets that satisfy the search parameters. API malfunctions could arise in different formats and cause difference
consequence. The current Twitter Search API returns only 5 percent to 10 percent of the total Tweets. It might have no large impact on mapping applications only handling geo-tagged Tweets, but it has a large impact on applications that are interested in all Tweets that could be returned.

An additional problem in using VGI is also API related. The methods used to harvest VGI depend on the APIs provided by the social network website. In the case of Tweets, as Twitter updates the APIs, the corresponding Tweet-collection methods will need to be updated accordingly. For example, the limit for Twitter Search API 1.0 was 1400 Tweets for a combined query of a keyword, a location and a time frame. After Twitter API 1.1, published in June 2014, Twitter API 1.0 was retired completely. Another example is the authorization for data collection. In the case of the Twitter Search API 1.0, an authorization was not required, except for the Twitter’s Stream API. Since Twitter API 1.1, all operations on API-based data acquisition require authorization. Some updates of Twitter API could make the Tweet-collection more complex and hard to control by the developers; for example, updates on the authorization. API updates do not necessarily make the task more difficult, but may require extra effort to comply with the new requirements.

On the other hand, development of all types of information technology is so rapid, that development carried out here may be quickly overtaken. One example is ESRI’s Social Media Template, see http://blogs.esri.com/esri/arcgis/2011/08/26/social-media-hurricane-irene-map/ for example, which illustrates an application similar to the one developed here, which specifically includes a Twitter template (ESRI, 2011).

7.3 Possible future research

This research uncovered problems in using geo-tagged Tweets in disaster management. One problem is the investigation of false negative and false positive errors. For simplicity, the distinction between errors has been presented as a crisp distinction between true or false. As discussed in the research, false positive error might be the easier to be tackled since Tweets at least exist. The more difficult research task is to identify and resolve false negative situations since this error is the result of no data being posted. It was suggested that a vulnerability layer be consulted to identify possible locations of need. Beyond a dichotomous True / False exists a continuum of
possibilities, reflecting a more fuzzy reality. For example, authoritative data used for comparison with real-time asserted data may itself contain errors of omission and commission.

In theory, a keyword search retrieves all Tweets that include the keyword in the tweet entity: not only in the message body, but also anywhere in the entity, such as the user name field or the place name field. For example, Tweets collected by using keyword “sandy” could include Tweets that do not have word “sandy” in the message body, but do have the word “sandy” as the user’s name. In this research, Tweets that do not have keyword in their message body were excluded before they were stored into the database. Another possibility for future research is to further investigate the collected Tweets. For example, the keyword “fire” is designed to retrieve Tweets that are related to disaster event such as forest fire, building fire, where fire means “the light and heat and especially the flame produced by burning” (Webster’s International Dictionary). But the collected Tweets might include some Tweets that could be about employment. In this case, fire means to be dismissed from a position. Or, the collected Tweets might include could be about guns. In this case, fire means “shoot”. Fire could also mean “remove from employment”.

The current study focused on geo-tagged Tweets. As discussed in previous chapters, geo-tagged Tweets formed a small portion of the total tweet population, thus Tweets without location information cannot be mapped, and the required resources cannot be effectively dispatched due to the lack of location information. While it is not supported by Twitter, a semantic search could be performed against the text content of harvested Tweets. For example, Okazaki and Matsuo (2010) used a support vector machine (SVM) to obtain a model to classify Tweets into a positive class (an event that occurred in reality) or a negative class (an event that does not occur). By using a faceted search, Abel et al. (2011) investigated ways to enhance search and content analysis, proposed strategies to infer facets and facet values on Tweets by enriching the semantics of individual Tweets, and introduced various faceted search for content exploration on Tweets. In an offline data analysis, Hambrick et al. (2010) employed content analysis to place collected Tweets relating to athletes (1962 Tweets) into different categories. Non-geo-tagged Tweets potentially provide information about the overall resource requirements of interested area so further research on content analysis is a priority task.
7.4 Conclusion

As discussed in section 2.3 of this dissertation, by now it is widely recognized that the use of VGI has some general problems or limitations, such as credibility, quality / accuracy / trustworthiness, privacy/ confidentiality, legality and security (Goodchild, 2007). At a Congressional subcommittee Hearing on “Emergency MGMT 2.0: How SocialMedia & New Tech are Transforming Preparedness, Response, & Recovery Disasters” (113rd U.S. Cong. House. Comm. on Homeland Security, 2013) witnesses from emergency management community agreed on the great power of using social media and social media – based data in emergency management. A major benefit of VGI – timeliness, shows VGI’s great potential in disaster management, which requires real-time information, especially in the response phase. However, we must strongly argue that VGI, such as Tweets, cannot be the only data source to support resource dispatch in disaster management and VGI should be incorporated with other emergency management–related operational data for acceptable decision making.

This study has demonstrated benefits of VGI for resource management, but has raised important topics that potentially restrict its value. The final on-line web application benefitted from the case studies of Kent State and Super-storm Sandy and is available for use in future crises and for further development. It has been demonstrated that geo-tagged Tweets collected and mapped in real-time can help emergency professionals to locate the emerging demand for resources so that needed resources can be quickly dispatched. The heat-map of resource requirement may help professionals to understand the current hotspots that have greater requirements for resources. After a disaster, the temporal display of collected Tweets can help professionals to understand the trend of requirements as the event evolved, and characterization of temporal trends in the demand for different types of resource may contribute to the preparedness for future disaster events.
References


Chamales, George (2013) "Towards Trustworthy Social Media and Crowdsourcing". Policy Memo Series 2, 12 DOI:


Feick, Rob and Stéphane Roche (2010). Valuing volunteered geographic information (VGI): Opportunities and challenges arising from a new mode of GI use and production. GeoValue,


Huang, Bo (2003) "Web-based dynamic and interactive environmental visualization". Computers, Environment and Urban Systems 27, 623-636 DOI:


How to effectively control evacuation traffic has emerged as one of the critical research


Lu, Qingsong (2006). "Capacity constrained routing algorithms for evacuation route planning" University of Minnesota (PhD).


Mooney, Peter, Huabo Sun, et al. (2011). VGI as a dynamically updating data source in location-based services in urban environments. UbiComp 2011, Beijing, China, ACM New York, NY, USA.


data sources, trends, accuracy and biases” Natural Hazards 67(2): 387-410.


Timmons, Lonnie (2014). Kent State University campus goes on lockdown after shot fired; suspect in custody. The Plain Dealer. Cleveland.


Tsai, Ming-Kuan, Yung-Ching Lee, et al. (2012). "Integrating geographical information and augmented reality techniques for mobile escape guidelines on nuclear accident sites“ Journal of Environmental Radioactivity 109 (July): 36-44.


Online resources


Online Resource. Brenner, Joanna and Aaron Smith. "72% of Online Adults are Social Networking Site Users". http://www.pewinternet.org/2013/08/05/72-of-online-adults-are-social-networking-site-users/. Last accessed on Feb. 1st, 1015


2014


