Citizens as Earth Observation Sources: a Workflow for Volunteered Information Sensing.

Bertrand De Longueville
Février 2016

Thèse présentée en vue de l'obtention du grade de docteur en sciences agronomiques et ingénierie biologique

Faculté des Bioingénieurs

Université Catholique de Louvain

Jury Members

President	Bruno Delvaux (UCL)
Supervisor	Pierre Defourny (UCL)
Readers	Patrick Bogaert (UCL)
	Franck Ostermann (UTwente)
	Stéphane Roche (ULaval)

à Cécile, mon cinquième élément

Acknowledgements

I would like first to thank the members of my Steering Committee: Pierre Defourny, Patrick Bogaert and Massimo Craglia. They have given a frame and directions to my research.

I would also like to thank all the co-authors with whom I wrote conference and peer-reviewed journal articles, and most notably Robin Smith, Nicole Ostländer, Gianluca Luraschi, Sven Schade and Jean-François Chevalier. Team working with them was not only efficient: it was pleasant!

I would like to thank my family and friends for their support and understanding. Cécile, in particular, has provided essential advice and encouragements to me.

Finally, I would like to thank the members of my Jury: Bruno Delvaux, Pierre Defourny, Patrick Bogaert, Frank Ostermann and Stéphane Roche. It has been a pleasure to exchange views and take advise from them in the final stage of my PhD process.

Abstract

This research introduces the notion of *Volunteered Information Sensing (VGI Sensing)* as the set of standards, methods and techniques required to streamline georeferenced contents published online by citizens into a timely, reliable and cost-effective source of Geoinformation for Earth Observation purposes.

VGI Sensing is proposed as an emerging sub-field of research at the conjunction of Geographic Information Science, Data Mining and (Web) Knowledge Discovery. It is expected to have many practical applications requiring pervasive and/or real-time geospatial data such as health epidemics, crisis management, environmental monitoring, crime analysis, or socio-economic studies.

After presenting background works and formulating research objectives in the Introduction, this thesis explores the information potential of VGI (Chapter 1) in the context of natural hazards management, then proposes a generic workflow for VGI Sensing (Chapter 2) – which is exemplified to a real-life use case. Technical optimisations of key steps of the VGI Sensing workflow are then studied in details (Chapter 3), and finally, the concept of VGI Sensing is presented in the wider perspective of the Digital Earth Nervous System (Chapter 4).

By doing so, it gives significant contribution to the sub-field of Geomatics that aims at converting information shared on the Internet by citizens as a reliable source of Earth Observation data, and opens perspectives for further research - which are discussed in the final chapter. An additional commentary is then proposed, addressing the questions related to the limitations and ethics of VGI Sensing.

Table of Contents

Introduction 1
1. (Volunteered) Geographic Information Science 1
2. (Big) Data Mining 7
3. Knowledge discovery on the (Social) Web 10
4. Detection and characterisation of (Natural) Disaster events 13
5. Objectives 18
6. Outline of the thesis 21

Chapter 1 - Proof of Concept: the informational value of Volunteered Geographic Information 23
1. Introduction 23
2. Previous works 24
3. Case study: the Marseille Fire 28
4. Results and Discussions 30
5. Conclusions 36

Chapter 2 – A Volunteered Geographic Information Sensing Workflow 39
1. Introduction 39
2. Previous work 41
3. Case study: recent floods in UK 48
4. Description of the workflow 48
5. Comparison of output with independent data sources and discussion 57
6. Conclusions and future works 67

Chapter 3 – Filtering and Clustering Volunteered Geographic Information 70
1. Introduction 70
2. Material: VGI preparation 75
3. Methods: Clustering algorithms benchmark 81
4. Results 88
5. Discussions and Conclusions 98

Chapter 4 – Perspective: the Earth's Nervous System 102
1. Introduction 102
2. Background 104
3. A Nervous System for the Digital Earth 105
4. SWE for VGI Sensing: Implementing DENS 110
5. A Forest Fire scenario 115
6. Conclusions and Future Works 119

Chapter 5 : Conclusions and Perspectives 122
1. Research objectives and results 122
2. Discussion 126
3. Future Research 132
4. Closing note 136

Author's list of publications.....138

References 140

List of Figures

Figure 1 Chronology of the Marseille Fire, number of related tweets per hour and selected tweets' contents (Sources: information provided during and after the fire by AFP and La Provence, and information retrieved from twitter.com 31
Figure 2 : Hectares of burnt area reported in tweets over time 32
Figure 3 : Location, frequency and time of the first citation of place names cited in tweets, and estimated total burnt area 33
Figure 4 : Number of users that published tweets by type 34
Figure 5 : Number of users that published tweets by user type 34
Figure 6 : Number of unique cited URLs by domain type 36
Figure 7 : Overview schema of the integration workflow 49
Figure 8 : Location and time of retrieved Flickr pictures 51
Figure 9 : ranking value per cluster 56
Figure 10 : Ranking values of each VGI clusters and clusters that correspond to a large flood event compiled by the Dartmouth Floods Observatory 59
Figure 11 : VGI ranking value compared to the number of press articles about floods for each date class. 60
Figure 12 : Ranking values and number of press articles for each VGI cluster 62
Figure 13 : Study areas used for Remote Sensing – VGI comparison 63
Figure 14 : Correspondence between VGI Day Clusters (circles) and GFDS flood signal (line) in the Manchester-Blackburn area 64
Figure 15 : Area “in flood” (red pixels) near Manchester on 12 February 2009 65
Figure 16 : Correspondence between VGI Day Clusters (circles) and GFDS flood signal (line) near Exeter 65
Figure 17 : Correspondence between VGI Day Clusters (circles) and GFDS flood signal (line) near Exeter 66
Figure 18 : ROC curve of the classification model which aims at identifying irrelevant pictures in its learning set (815 pictures). 80
Figure 19 : Quantitative assessment matrix for VGI clusters. 87
Figure 20 : Visual definition of qualitative measure 88
Figure 21 : Cluster size distribution. 94
Figure 22 : Density function of time for pictures of cluster assigned to GROUSE fire according to DB no sem (continuous lined), DB sem (dashed line) and SatScan (dotted line). 95
Figure 23 : Interplay of SWE components. 111

Figure 24 : SWE for VGI sensing. 113
Figure 25 : a research agenda for VGI Sensing..... 133
Figure 26 : the DIKW pyramid 136

List of Tables

Table 1 : Date Classes resulting from Jenks Natural Break analysis of the VGI dataset 54
Table 2 : Large flood events reported by the Dartmouth observatory for the study period and area 58
Table 3 : Percentage of pictures by georeferencing precision level. .. 77
Table 4 : Basic filtering operations and their result. 78
Table 5 : Features selected by the model with their weight in the regression formula. 81
Table 6 : Parameter sets considered in the benchmark and main characteristics of clusters 89
Table 7 : Comparison of cluster characteristics by method. 92
Table 8 : Functional comparison of the human nervous system and the Digital Earth nervous system 109
Table 9 : Mapping between general concepts of the nervous system and SWE for VGI sensing. 115

Acronyms

AFP	Agence France Presse
AMSR-E	Advanced Microwave Scanning Radiometer-EOS
API	Application Programming Interfaces
CCM2	Catchment Characterisation and Modelling 2
CIA	Central Intelligence Agency (of the United States of America)
DART	Deep-ocean Assessment and Reporting of Tsunamis
DB Scan	Density-Based Spatial Clustering of Applications with Noise
DE	Digital Earth
DENS	Digital Earth Nervous System
EFFIS	European Forest Fire Information System
EMM	European Media Monitor
EOS	Earth Observing System
ESA	European Space Agency
ESRI	Environmental Systems Research Institute
EUTFR	European Forest Fire Tactical Reserve
GDACS	Global Disaster Alert and Cooperation System
GFDS	Global Flood Detection System
GI (Science)	Geographic Information (Science)
GIS	Geographic Information Systems
GML	Geographic Mark-up Language
GPS	Global Positioning System
I(C)T	Information (and Communication) Technologies
JRC	(European Commission's) Joint Research Centre
JSON	JavaScript Object Notation
KML	Keyhole Mark-up Language
LBSN	Location-Based Social Networks
LBSN	Location-based Social Networks
MIC	(European Commission's) Monitoring and Information Centre
MODIS	Moderate-resolution Imaging Spectroradiometer
NASA	National Aeronautics and Space Administration (of the United States of America)
NLP	Natural (human) Language Parsing
NSA	National Security Agency
O&M	Observations and Measurement
OASIS	Organization for the Advancement of Structured Information Standards

OGC	Open Geospatial Consortium
OLAP	Online Analytical Processing
OLAP	OnLine Analytical Processing
OPTICS	Ordering Points to Identify the Clustering Structure
PHP	(Programming) Hypertext Pre-processor
POI	Point Of Interest
PPV	Positive Predictive Value
REST	Representational State Transfer
ROCC	Receiver Operating Characteristic Curve
RSS	Really Simple Syndication
RT	Re-Tweet
SAS	Sensor Alert Service
SDI	Spatial Data Infrastructure
SES	Sensor Event Service
SMS	Short Message Service
SOS	Sensor Observation Service
SPS	Sensor Planning Service
SSTP	SatScan Space-Time Permutation
SWE	Sensor Web Enablement
UK	United Kingdoms of Great Britain and Northern Ireland
URL	Uniform Resource Locator
US(A)	United States of America
USFS	United States Forest Service
VGI	Volunteered Geographic Information
VoIP	Voice over Internet Protocol
WCS	Web Coverage Service
WFS	Web Feature Service
XML	eXtensible Language Mark-up

Symbols

α_{km}	weighting factor for the spatial distance
α_s	weighting factor for the semantic distance
α_t	weighting factor for the temporal distance
d_m	spatial distance between 2 VGI items
$ds_{A,B}$	semantic distance between VGI item A and VGI item B
d_t	temporal distance between 2 VGI items
Δkm	normalised spatial distance between 2 VGI items
Δs	normalised semantic distance between 2 VGI items
Δt	normalised temporal distance between 2 VGI items
epsilon	maximum neighbourhood distance set for DB Scan
MinPts	minimum number of points to define a cluster centre in DB Scan
$n_{j,A}$	number of occurrences of token j in A's textual metadata
$Q_{p,ds}$	quantile p% of the semantic distance among all pairs of pictures in the scope
RD	arbitrary spatial distance threshold considered as maximum
RT	arbitrary temporal distance threshold considered as maximum
w_j	weight of word j which is inversely proportional to its frequency

Introduction

This research is situated at the intersection of 3 scientific fields: Geographic Information Science, which finds its origins in the digitalisation of cartographic disciplines and the development of Earth Observation techniques; Data Mining, which is deeply rooted in statistical science and grew exponentially in conjunction with development of the data production and storage technologies; and the Web Knowledge Discovery, which can be seen as a specialisation of Natural Language Processing, adapted to the specific architecture of the Internet.

This research does not ambition to bring major breakthrough in any of these three scientific fields taken separately. Rather, it combines their key concepts and techniques to contribute building a specific sub-field –coined in this research as *Volunteered Geographic Sensing* -, which aims at converting heterogeneous information retrieved from the Web into a timely and reliable source of Earth Observation data. As such, the concept and techniques of VGI Sensing that are developed in this thesis can be seen as an advance of Geographic Information Science.

1. (Volunteered) Geographic Information Science

Technologies related to Earth Observation knew a fast evolution in the second half of the 20th Century, firstly with airborne photography - which was played a prominent role in Military Intelligence during Wold War II (Deuve 2013)-, then with satellite imagery - initiated with the launch of the CIA's Corona programme in 1959 (Corson & Palka 2004).

As set of methods – usually referred to as Remote Sensing - to store, analyse and interpret such data which were applied during the 1960's and 1970's in non-military fields such as forestry, geology or urban planning (Kramer 2002).

In support to research on Remote Sensing, computer systems for geographic information handling and processing were imagined (Tomlinson 1963) and implemented. This led to the development in the 1980's and the 1990's of the Geographic Information Science, a discipline where geographers, computer scientists, and environmental

or social researchers apply the latest digital technologies to geographic information (Coppock & Rhind 1991) .

At the dusk of the 20th Century, GI Science embraced the deeply transformative power of the Internet. In his speech titled "*The Digital Earth: Understanding our planet in the 21st Century*" the US Vice-President Al Gore (Gore 1998) depicted in a powerful metaphor how increasing computing power, abundant geographic data, broadband network connections and interoperability standards can be combined in the next generation of connected Geographic Information Systems, coined as the *Digital Earth*.

Academic and governmental actors ambitioned to materialise such vision by creating Spatial Data Infrastructures at national level (Grant 1999; Masser 2000), as potential building blocks of a future Digital Earth system (Rajabifard et al. 2000) . But efforts to build Spatial Data Infrastructures – notably the NSDI in the US (Rhind 1999) and the INSPIRE initiative in Europe (Annoni 2004) – as a constellation of large scale and state-owned computer systems, were challenged in the mid 2000's by a series of paradigm shifting inventions. As a sign of the times, these inventions were developed independently from government programmes and from the academic community.

Firstly, the Web technologies evolved – thanks to the vivid adoption of Open Standards (Maxwell 2006) – from a one-directional text broadcasting media to a two-ways information sharing platform. This phenomenon, coined as the *Web 2.0* (O'Reilly 2005), had a geospatial dimension since its initial phase. Flickr, for example, was one of the precursors of the Web 2.0 by proposing a picture sharing service and allowed geotagging of contents (Terdiman 2004); similarly Wikipedia, the crowdsourced encyclopaedia, allowed assigning geographic coordinates to articles referring to locations (Völkel et al. 2006).

Secondly, the 3D geo-browser *Google Earth*, launched in 2005, embodied the concept of Digital Earth better than any former effort of academia, governments or international organisations (Foresman 2008) and attracted user-generated contents based on KML, its Open Standard for geographic information.

Thirdly, the iPhone 3GS, launched in 2008, was the precursor of a generation location-aware mobile devices (Wilson & Fenlon 2009)

which made the creation of online geo-located contents a trivial action for the numerous consumers who adopted smartphones in their daily lives. They became *de facto* '*neogeographers'*, as (Turner 2006) named non-expert persons using modern technology to create geospatial information and design (online) maps.

Although participation of non-experts in the collection of geographic data is not a recent phenomenon (Stamp 1937; Lee 1994), the rise of neogeography posed existential questions to a number of researches in the GI Science field, as in every community of experts confronted with the lowering of technical barriers around them (Gould 2008). GI Scientists were facing the dilemma of embracing or dismissing this phenomenon of non-experts investing their field the same way 'normal people' used Wikipedia to publish the richest-ever encyclopaedia as a demonstration of the *wisdom of the crowds* (Sui 2008). Therefore, this phenomenon was seen by visionary researchers as an opportunity to further develop participatory GIS and empower citizens in geospatial decision-making (Ciobanu et al. 2007).

A consensus emerged among GI Scientists to welcome the neogeography actors for the Open Source tools they provide, and the vast and timely amount of data they can generate, but to disparage them for the lack of credibility, of trustworthiness, of quality of such data (Walsh 2008). The lack of expertise of neogeographers was stressed as a fundamental issue by academic geographers, but as an issue that can be overcome by appropriate techniques, in order to combine the best of both worlds (Goodchild 2009).

It is in this context that Goodchild (2007) coined the term *Volunteered Geographic Information (VGI)* as the vast amounts of geolocated data (e.g. Flickr pictures, tweets, Wikipedia pages, blog articles, *etc*.) posted on a voluntary basis by non-experts on the Internet. Unsurprisingly, the question of VGI credibility has been stressed as a central research issue (Flanagin & Metzger 2008). Various approaches have been suggested to overcome the alleged lack of credibility of VGI contents.

Firstly, volunteers can follow a pre-established protocol, and be trained until they develop the appropriate level of expertise to become trusted data providers, as featured since decades in the UK's national Christmas Bird Count involving a large network of amateur

ornithologists (Goodchild 2007). By design, this approach tends to decrease the potential number of contributors (since only the ones willing to be trained and to respect a set of rules are allowed to participate), and may antagonise with the spontaneous nature VGI can have (which is of primary importance in crisis situations, as discussed in section 4).

Secondly, the quality control itself can be set up as a volunteered process, where community of users can act as quality filters for VGI the same Wikipedia articles are curated by the community of users until they reach the appropriate level of quality (Bishr & Mantelas 2008). This *Wiki* model can also include the notion of reputation, where information provided by volunteers who performed well in the past is considered as more trustable (Maué 2007). Such approach involves that other volunteers can share expertise on the subject (and/or judged for their own), and therefore is not applicable when the data provider is the only person able to value the quality of his own data (which is often the case in emergency situation), when expertise is not an absolute intrinsic characteristic of the person (in the onset a crisis, the victims become suddenly the most qualified observers) and the time for volunteers to cross-check each other's data quality is a limiting factor (i.e. it excludes applications involving real-time data processing). Interesting recent research aims to avoid such caveat by analysing VGI contributor's behaviour *in lieu of* their (alleged) level of expertise in order to assess their trustworthiness (Bégin et al. 2013).

A third option could be to turn the challenge of data abundance into an opportunity, where reliable information is extracted from vast amounts of VGI with uncertain quality from numerous sources by applying cross-validation mechanisms. In other words, the data quality problem of VGI can be addressed by "aggregating input from many different people" (sic) (Mummidi & Krumm 2008) and by processing these VGI clusters to evaluate their relevance in a given context. Although this approach seems intuitively the best fitting the nature of the VGI phenomenon as described below, it is based on the assumption that higher quality information can be derived from vast amounts of low quality data (such assumption has been verified in the abundant corpus of scientific literature discussed in the next sections), and poses numerous methodological questions on how to perform properly such '*alchemical*' process.

Contributing to tackle such methodological questions is precisely the aim of this research, which coins the term *VGI Sensing* as the set of standards, methods and techniques required to streamline geo-referenced contents published online by citizens into a timely, reliable and cost-effective source of Geo-Information for Earth Observation purposes.

When this research started, with the publication of an article highlighting the potential value of VGI in a Forest Fire Use Case[1] (De Longueville et al. 2009) the main focus in the GI Scientific community was on the conceptualisation of the phenomenon (Budhathoki et al. 2008; Coleman et al. 2009), and on the analysis prominent initiatives such as OpenStreetMap (Chilton 2009) or Geo-Wiki (Fritz et al. 2009).

In the same time, another community of researchers was unveiling the information potential of properly processed contents from the Web 2.0, but without a specific geospatial perspective (see section 3). Interestingly, several research initiatives developed contemporaneously the intuition that geospatial information can be mined from the Web 2.0 (Pultar et al. 2008; Jones 2008; Crandall et al. 2009; De Longueville et al. 2009; Intagorn et al. 2010).

Interestingly, the field of Earth Observation science, which has an historical focus on satellite imagery, seems to increasingly exploit the potential of mobile devices technologies as a complementary source of information, e.g. for Land Cover mapping purposes (Fritz et al. 2009). This phenomenon of Earth Observation scientists embracing principles such as crowdsourcing, open data and citizen science (Ferster & Coops 2013) creates the conditions for the convergence of a new generation of Remote Sensing applications - coined as Earth Observation Science 2.0 (O'Neill 2015) – with VGI Sensing.

In the early 2010's research allowed deepening the understanding of VGI's key concepts, such as its value (Feick & Roche 2013) in the light of its fitness-for-purpose and Open Source nature, trust (Van

[1] See Chapter 1 - Proof of Concept: the informational value of Volunteered Geographic Information.

Exel & Dias 2011), visualisation (Roick et al. 2012), and dynamics of participation (Rehrl et al. 2013).

This research contributed to this development, by coining the term of *VGI Sensing* (De Longueville et al. 2010), conceptualising the data mining approach of VGI[2]. Such approach knew an increasing interest in the GI Science community, especially in the context of Sensor Web Enablement research (Resch 2013; Schade et al. 2013), and more generally in a range of applications requiring pervasive (Mooney et al. 2012; Spinsanti & Ostermann 2013) and/or real-time geospatial data (Thatcher 2013; Zhang et al. 2014). As this latter example shows, VGI Sensing seems to be adapted by design to crisis situations; this will be discussed in further details in section 4.

Latest research in the field seems to feature a fast growing number of Case Studies and field applications of VGI Sensing, related e.g. to floods (Kongthon et al. 2014; Schnebele et al. 2014), traffic monitoring (D'Andrea et al. 2015), or crime (Kounadi et al. 2015).

A limited number of publications from the GI Science field seem to be at the forefront of VGI Sensing research, by exploring in details algorithmic issues that are specific to the processing of spatiotemporal information retrieved from the social web. Cheng & Wicks (2014), for example, analysed the question of clustering VGI – which is also at the core of this research -, while Hahmann et al. (2014) tested machine learning techniques to correlate geo-located tweets with existing Points of Interest, and Bimonte et al. (2014) are exploring how techniques known as OLAP (OnLine Analytical Processing) can be applied to VGI.

In the conclusion of the book *Crowdsourcing Geographic Knowledge: Volunteered Geographic Information (VGI) in Theory and Practice* (Sui et al. 2012) gathering contributions from most prominent GI scientists, the editors called future VGI research for more interdisciplinary with other data-intensive disciplines (Elwood et al. 2013). This research is fully in line with this statement, and supported since its earliest stage integration of concepts from the Data Mining field – as suggested by, e.g. Fischer 2012-, and from the Natural

[2] See
Chapter 4 – Perspective: the Earth's Nervous System.

Language Parsing field – as studied by, e.g. Ballatore & Bertolotto 2011. The following sections further discuss such inter-disciplinary integration.

2. (Big) Data Mining

The statistics discipline appeared in the middle of the 18th century, in conjunction with the development of modern scientific approach, which enabled improved data collection methodologies. Initially focusing on the nation-state level – therefore the name *stat*istics – its primary fields of applications were economy and demography (Desrosières 2002).

The 19th Century saw an unprecedented development of mathematical concepts – e.g. in France under the auspices of Napoleon I – while the Industrial Revolution extended the production of statistical data to the private sector (Tresch 2012). It is not surprising, in this context to see Karl Marx constantly referring to national statistics and factory inspectors' reports to support his theoretical claims in his famous book *The Capital* (Cockshott et al. 1995).

Like Marx's observations, nevertheless, statistical studies in the 19th Century were mostly of an empirical nature – i.e. the careful examination of figures collected at a given moment for a given purpose (Stigler 1986). The systematic collection – and processing – of large amounts of data started in the first half of the 20th Century, often in support of new regulatory policies at national level (Piketty 2013).

The conjunction of technological advances discussed earlier – especially the exponential growth of computing power, data storage capabilities and broadband networks – drastically transformed statistical science in the second half of the 20th Century (Press 2013). These advances reverted completely paradigms: from a rigorous endeavour of collecting reliable data and extracting as much knowledge as possible from scarce sources, the issue became to extract meaningful knowledge from a overwhelming amounts of data – in essence, a shift from patient data gathering to real-time surfing the data tsunami (Shah et al. 2010). Retrospectively, we can only admire the first generation of Data Scientists who were not *digital*

natives (Prensky 2001) and reinvented a centuries-old science by *thinking computationally* (Levesque 2012).

Data Mining appeared at the eve of the 21st Century as a response to this paradigm shift. It is defined as 'the science of extracting useful knowledge from large data sets, by combining statistics, databases management, and artificial intelligence mostly' (Hand et al. 2001). This field of research is experiencing an exponential progress since then, with:

- the development of Machine Learning – one of most prominent field of Artificial Intelligence research in Data Science (Bishop 2006) ;
- the emergence of cloud computing – increasing to almost infinite proportions the available storage space and computing power (Armbrust et al. 2010) ;
- the rise of the Internet of things – converting any object into a source of digital data (Igoe 2011).

The term *Big Data* has been coined to designate such boom, and since then it is seen as an escalating revolution inducing important practical changes (Mayer-Schönberger & Cukier 2013), but also major societal challenges such as mass technological surveillance (Crampton 2015).

It is important to frame the reader's expectations, at this stage: this research does not aim at revolutionising Data Mining research, or at acting as a game changer in the Big Data field. Rather, it will try to mobilise key concepts of Data Mining, while re-using and adapting proven algorithms to the specific context of GI Science. We will try to apply the Data Scientists paradigm, which says that *more is more* when it comes to data, and endeavour to contradict the sadly famous *garbage in, garbage out* statement in the context of Volunteered Geographic Information. On this respect, we will follow the lead of Anselin (1989) by applying his challenging question "*What is special about spatial data ?*" to a scientific issue of present time. We are also in line with Roche et al. (2012) who envisioned how the conjunction of pervasive technology, real-time data processing and next-generation location-aware services will enable a novel relation to spatiality for end-users in the context of *Smart Cities*.

Although we will use simple Machine Learning methods to train a quality filter on input VGI[3] our main incursion in the Data Mining field will be related to Data Clustering.

Data clustering can be defined as the unsupervised classification of patterns into groups called *clusters* (Jain et al. 1999). In spatiotemporal clustering, the position of the features in space and time (e.g.: latitude, longitude, date, time) are used as the key dimensions (Gong et al. 2006). On the basis of the similarity measurement between spatiotemporal features, various clustering algorithms (hierarchical, partitional, density-based, *etc*.) can be applied, depending on the nature of the events that are investigated (Getis & Ord 1992). A wide variety of spatiotemporal clustering techniques and algorithms have been applied to detect events in fields like epidemics (Rogerson 2001), crime analysis (Johnson 2010), or meteorology (Hsu & Li 2010).

But whereas spatiotemporal clustering techniques are usually designed to deal with discrete, comparable objects such as sensor observations or tabular data records (Miller & Han 2001), VGI can be heterogeneous in terms of quality and accuracy (Metzger 2007). In particular, Quesnot & Roche (2015), as well as De Longueville & Hardy (2010) emphasized that VGI often have place names as spatial reference (e.g., town, region, country, etc.), resulting in different levels of spatial accuracy when looked-up in a gazetteer. Oppositely, the temporal reference of VGI is usually accurate because of the creation of a time stamp when VGI is posted online.

In consequence, current spatiotemporal clustering techniques might benefit to be better adapted to data with heterogeneous spatial reference such as VGI. In addition, the spatial and temporal dimension of VGI can benefit to be combined with its Semantic and Social dimension. The aim of this research is to contribute to the development of clustering methods that are suitable to extract event-related knowledge from VGI.

Multidimensional clustering of VGI is a relatively recent scientific endeavour (see, e.g. Kisilevich et al. 2013). Nevertheless Cheng &

[3] See Chapter 3 – Filtering and Clustering Volunteered Geographic Information

Wicks (2014), as well as Craglia et al. (2012) and Zhao et al. (2014) clearly established the value of Scan Statistics (Kulldorff 1997) algorithms on that purpose, more specifically Space-Time permutations, which presents the advantage to automatically adjust to temporal trends (Sikder & Woodside 2007). The Chapter 3 – Filtering and Clustering Volunteered Geographic Information – will propose a benchmark of the SatScan state-of-the-art algorithm with DB Scan, a challenger inspired by Kisilevich et al. (2010). Zhao et al. (2014) implemented the idea of combining semantic similarity and spatial proximity to optimise results, an idea that will be further explored in Chapter 3 as well.

At this stage, it is useful to position this research with regards to Web-based Knowledge Discovery (section 3) before discussing its relevance for the detection and characterisation of Natural Disaster events (section 4) – which is the type of Use Cases we addressed in this research.

3. Knowledge discovery on the (Social) Web

Scientists started parsing human language with computers during the Cold War, with the objective of enabling Machine Translation from Russian to English (Bhattacharyya 2015). But the condition for a genuine exponential growth of the discipline where only met, as for Data Mining, in the 1990's – 2000's period with the advent of modern computing technologies. Indexing Web Pages – a specific use case of Natural Language Parsing (NLP) – was at the origin of one of the biggest commercial successes of all times, namely: Google, *Inc*. (Fast Company 2003). The rise of the Web 2.0 amplified this boom by providing both abundant material to researchers, and a wide number of applications – a compelling example being the automatic inference of the emotional state of a person from his postings in social media (Pang & Lee 2008).

In this research, we have mobilised specific existing expertise in the field, e.g.:

- to detect automatically the language use in VGI items[4];

[4] Using the Google Translate API https://cloud.google.com/translate/docs

- to extract automatically Named Entities – such as Place Names[5] - in VGI items;
- to divide phrases into single words or n-grams – e.g. *Forest Fire* – with the Stanford NLP Toolkit (Manning et al. 2014) in preparation of Semantic distance calculation.

The application of Data Mining and NLP techniques for Knowledge Discovery from the poorly structured contents of the World Wide Web has been coined as *Web Mining* (Mobasher et al. 1996). The main focus was, at this time, to organise a "*dynamic and chaotic set of human-readable documents available online*" (sic) (Etzioni 1996).

During the early 2000's, however, the Web evolved towards more structured contents thanks to:

- the adoption of standards allowing flexible data formatting - e.g. XML (Bray et al. 1998);
- the retrieval of online resources facilitation by developer-friendly network addresses - e.g. REST architecture (Fielding 2000) ;
- the offering Application Programming Interfaces (API) to interact with online services (Hoong & Buyya 2003);
- the development of in-browser coding patterns allowing smooth interaction of remote data within the Graphical User Interface - e.g. AJAX (Paulson 2005).

Such technologies acted as enablers for the Semantic Web vision (Berners-Lee et al. 2001): the World Wide Web evolved from a collection of loosely inter-linked documents to a machine-friendly data base of semi-structured objects (Daconta et al. 2003).

Communities of practices and *de facto* standards enabled numerous practical applications of such vision, from e.g. the aggregation of daily news with RSS feeds to the automatic detection of the most desired item on e-bay (Feiler 2007). Most notably, the content aggregation of processing from online resources – coined as *Web 2.0 Mashups* – attracted the early attention of few GIS researchers, thanks to the ability offered by the Google Maps API to geocode real-life addressed

[5] Using the Yahoo Boss API https://developer.yahoo.com/boss/

and to easily create interactive online maps (Miller 2006), which can act as an enabler for collective geospatial intelligence (Roche & Kiene 2008). The term *Geospatial Web* was used by these researchers to designate the numerous applications combining network addressing (e.g. URLs), location and other online data (Sharl 2007).

Then Web became suddenly *Social*, with the emergence in the second half of the decade of online Social Networks (Ellison & others 2007). Social Web Mining became a primary field of research for online Knowledge Discovery; early researches (summarised by Kleinberg 2007) highlighted the potential applications in Marketing, Sociological Studies (around the concept of *communities*) and Communication theory (around concepts like *viral information spreading* and *collaborative problem solving*).

Most notably, Social Web Mining developed a rich literature around the notions of streams, trending and branding (Kaleel & Abhari 2015). Social Web Mining consists most often in (near) real-time analysis of streams of online postings (on Twitter almost exclusively); its early applications had in general a weak geospatial dimension (Steiger et al. 2015). In a typical Social Media stream analysis, incoming text is continuously processed in order to identify bursts in certain keywords usage frequency (Mathioudakis & Koudas 2010). Various clustering and Natural Language Parsing techniques are then applied in order characterise candidate topics, and assess their relevance (Atefeh & Khreich 2013).

Some researchers analysed Twitter streams to specifically retrieve information related to real-life events (as opposed to 'trending topics' that cannot be assigned with a spatiotemporal extent). Interestingly, number of these researches does not take the spatiotemporal dimension into account for clustering Tweets into events. For example Becker et al. (2011) used only the temporal, textual and social dimensions to extract event candidates from Twitter streams. Aggarwal & Subbian (2012) followed a similar approach – although modelling the Twitter stream as a graph to better capture its social dimension – and concluded that the method fails to distinguish contemporary events of the same nature. Oppositely, Walther & Kaisser (2013) created purely spatiotemporal clusters from the stream of tweets posted in the last 24 hours, and then used their semantic contents to assess the nature of the detected candidate events.

Similarly, Cheng & Wicks (2014) demonstrated that real-life events can be detected from Twitter by identifying abnormal spatiotemporal patterns on Twitter data. Furthermore, Zhao et al. (2014) computed in parallel semantic similarity and spatiotemporal proximity and combined these dimensions to build semantico-spatiotemporal clusters, which proved to be an effective way to detect domain-specific events in an unsupervised manner.

The research presented in this thesis contributed to early efforts in exploring the spatiotemporal dimension of Social Web Mining techniques, in conjunction with e.g. (Pultar et al. 2008; Mummidi & Krumm 2008; Crandall et al. 2009; Sankaranarayanan et al. 2009). In the 2010's, these techniques were developed in numerous applications (under the umbrella of *VGI Sensing*, or of *Social Web-based Events Discovery*) in fields like, e.g. public health (Mooney et al. 2012), environmental monitoring (Resch 2013) or traffic analysis (D'Andrea et al. 2015).

The interested reader would benefit the effort for a providing comprehensive – yet focusing only on Twitter - overview of similar research in a Systematic Literature Review completed by Steiger et al. (2015), from which we can cite, *in extenso*:

"*Reviewed papers and their application domains have shown that the study of geographical processes by using spatiotemporal information from location-based social networks represent a promising and still underexplored field for GI Science researchers*".

In the next section, the discussion will focus on a special type of events: Natural Hazards, which constitute a prominent type of Use Cases for VGI Sensing/Social Web-based Events Discovery.

4. Detection and characterisation of (Natural) Disaster events

Disaster management has been one of the most prominent non-military applications of Remote Sensing technology since its early years (Johnson 1980). Aerial photography has been used for earthquake damage prevention (Kadomura 1968), and satellite imagery for damage assessment after forest fires (Smith & Woodgate 1985), and in support to disaster response operations (Rush et al.

1976), to cite only few historical examples of this primary field of Remote Sensing science and operations.

Over time, public authorities in charge of disaster management have developed information systems, such as the Global Disaster Alert and Cooperation System (GDACS, De Groeve et al. 2006) or the European Forest Fire Information System (EFFIS, San-Miguel-Ayanz et al. 2002). The typical purpose of such information systems is to process in a timely manner data from satellite and/or in-situ sensors (e.g., MODIS for active fires, AMSR-E for floods water surface, DART buoys for tsunamis) into actionable information such as dynamic mapping of burnt scars for forest fires, of immerged area for floods, or of real-time alerts for tsunamis.

It is important to have in mind the various phases of such events in order to understand the information needs in disaster situations. Researchers usually distinguish the following successive stages (Dynes 1970):

- *Stage 0: Pre-Disaster*, which is the 'normal' state of the social system before the impact, where preventive actions can be taken;
- *Stage 1: Warning,* where some signs of the coming disaster can be perceived;
- *Stage 2: Threat*, where the imminence of a disaster can be inferred from substantial information;
- *Stage 3: Impact,* which is the onset of the disaster, when it (begins to) occur(s);
- *Stage 4: Inventory,* where the initial stocktaking of the extent of the damages can take place;
- *Stage 5: Rescue*, where help is provided through spontaneous, local actions (first aid, mitigation measures);
- *Stage 6: Response,* where professional and organised help is provided (medical aid, logistical operations);
- *Stage 7: Recovery,* where rehabilitation and readjustment actions are performed in order to restore the social system to (at least) its pre-disaster level.

The duration and intensity of each stage can widely vary depending on the type of disaster and the surrounding conditions. These definitions are nevertheless very useful to understand the type of information products disaster managers need at a given time. In the case of forest fires, for example, risk levels can be computed from remote-sensed soil moisture value during the *Warning* stage in order to inform the population. During the *Threat* stage, public authorities can use land cover maps and weather forecast data to evaluate how a fire will spread, and which infrastructures (houses, roads, etc.) will probably be impacted. In the *Inventory* stage, the burnt scars can be delineated using Remote Sensing data in the visible and thermal infrared spectrum in order to identify the surface and type of damaged forest and infrastructure. In the *Rescue* stage, it is very important to know where are the impacted citizens (did they evacuate the danger zone, or not ?), what they are doing, what is their condition and their emotional state. In the *Response* stage, the inventory data of several contemporary fires can be compared to prioritise fire-fighting resources at regional level (e.g. Canadair planes). In the *Recovery* stage, satellite imagery and authoritative databases can be used to assess the value of damages, in order to calculate compensations paid by insurance companies.

In a context where relevant Remote Sensing data is readily available, where public authorities have developed information systems to exploit them in a timely manner, and where security forces have efficient communications systems (including, notably a robust emergency phone call system for citizens) one may question what could be the added value of VGI for disasters management.

Palen & Liu (2007) gave a visionary and substantial answer to the question of VGI relevance in disaster management. Visionary, because their research was published before the wide diffusion of technologies such as mobile broadband Internet connection, smartphones, and related real-time social media applications. Substantial because, by analysing typical online behaviours by citizen affected by crisis in the early 21st century, the researchers could highlight typical use cases of online citizen participation at each stage of disaster situations.

Such typical examples include :

- In the immediate aftermath of the 9/11 attacks in New York City (USA), public authorities used all available electronic communication means (email, VoIP, SMS, …) to collect '*welfare check*' information from citizens on the ground about missing persons (*Stage 4 – Inventory*) (Dawes et al. 2004);
- The hurricane Katrina disaster, which devastated the city of New Orleans (USA), on the 29/08/2005 saw an novel form of disaster communication, where citizens were spontaneously asking and offering assistance to each others via dedicated web sites and online *fora* (*Stage 5 – Rescue*) (Palen & Liu 2007);
- During the shootings on the Virginia Tech campus on 16/04/2007, Palen & Liu (2009) observed the emergence of '*distributed problem solving*' activity, where citizens advised each others in real time about the zones of the campus to avoid (*Stage 3 – Impact*) and by setting up a wiki page giving a collaborative space for assuming the function of '*welfare checks'* (*Stage 4 – Inventory*);
- Sakaki et al. (2010) demonstrated that the reporting via Twitter was so timely and accurate that it could serve as a credible basis for earthquake and typhoon detection system in Japan (*Stage 1 – Warning*);
- Experts from the EFFIS research group (De Longueville et al. 2010) identified VGI Sensing as a good complement to Remote Sensing for the early detection of forest fires (*Stage 2 – Threat*), for the characterisation of simultaneous events conditioning the allocation of fire-fighting aerial resources (*Stage 6 – Response*), and for the detailed damage assessment conditioning possible financial support from the public authorities to the restoration process (*Stage 7 – Recovery*);
- Opgenhaffen & Smets (2012) observed an advanced behaviour of *distributed problem solving* over social media when a major thunderstorm hit suddenly an open-air music festival ('Pukkelpop') near Hasselt (Belgium) involving several causalities and putting about 60.000 persons in a distress situation in the middle of a scarcely-populated rural area (*Stage 5 – Rescue*).
- Daume et al. (2014) argued that social media monitoring could contribute to a better awareness of forest ecosystems

conditions – including their social dimension – which in return can support actions for better resilience and disaster preparedness of such ecosystems (*Stage 0 – Pre-Disaster*).

Additional examples and further analysis of usage of VGI in crisis situations can be found in a recent survey by Imran et al. (2014). Such examples confirm in practice the conceptual issues discussed earlier: the necessity to apply a quality filter to VGI, the opportunity to adopt web mining techniques to perform such filtering in a timely manner, and the relevance of devising a typology of VGI Sensing use cases (depending on disaster stage, VGI sources, actors involved, disaster type, etc.) – each posing specific research challenges.
.

Atefeh & Khreich (2013) proposed to clearly distinguish use cases involving Events Discovery from those involving Retrospective Analysis (even if in nearly-real time). In this research, such distinction applies, since we did not invested in priority in the development of real-time processing capability, as opposed to abundant research on e.g., Twitter streams. As a consequence, the operational relevance of this research has to be situated between *Stage 4 (Inventory)* and *Stage 7 (Recovery)* of the disaster cycle, although further research could complement it in order to address the question of computing efficiency of the proposed methods. This focus on (early) retrospective analysis is in line with Roche et al. (2011) assessment that "*the potential of [VGI] for crisis management relates essentially to the response and recovery phases*" (sic).

Based on examples cited above of usage of VGI Sensing in the context of disaster management, the beneficiaries if VGI can be devised in two categories:

1. *Public Authorities*: Examples involving the Global Disaster Alert and Cooperation System (GDACS) and the European Forest Fire Information System (EFFIS) are developed in next chapters (see respectively chapter 2 and 4). They represent typical use cases of VGI Sensing where public authorities (at e.g. regional, national or international level) have developed thematic Information Systems which rely usually on Remote Sensing data and on authoritative data; VGI Sensing can be for such systems a novel source of information, improving

situation awareness and reactivity at every stage of disaster they are responsible of coping with.

2. *Citizens* : the outcome of VGI Sensing can be also returned directly to citizens, which is of primary importance in the Rescue phase (stage 5) were the disaster relief relies essentially on spontaneous actions from citizens. The 'Pukkelpop' case mentioned above (Opgenhaffen & Smets 2012) give a typical example, where the usage of filtered geolocated Twitter feeds allowed local citizens to give shelter and assistance to thousands of music festival participants which were hit by a major storm. Another example from the hurricane Katrina episode (Miller 2006) shows similar pattern: citizens spontaneously organise VGI, some (technology-aware) citizens deploy a specific web platform for the event, boosting distributed problem solving initiatives on the field.

After having situated precisely this research at this intersection of several research fields and application types, the next section will describe in the most concise and accurate possible manner the objectives of this thesis.

5. Objectives

The overall objective of this thesis is to develop robust processing methods that convert heterogeneous VGI into a timely, reliable and cost-effective source of Geo-Information for Earth Observation purposes.

This work specifically focuses on the contribution of VGI to situation awareness in the context of Natural Hazards, although the developed methods aim to be generalizable.

The usual sources of Earth Observation data include satellite imagery, in-situ sensors and field observations conducted by experts (e.g. fauna inventory). In an era of ubiquitous computing, an ever growing number of individuals can connect to the Internet almost anywhere and at any time; they post online vast amounts of openly accessible geolocated media (text, pictures, video), which may constitute a novel source of ground-based information that can contribute to the understanding of geospatial phenomenon.

This thesis therefore proposes to develop *VGI Sensing*, defined as a set of standards, methods and techniques required to streamline georeferenced contents published online by citizens. The development of VGI Sensing methods is a major challenge, as VGI is often regarded as insufficiently structured, documented, or validated and is available in often extremely important quantities.

To tackle this general objective of developing VGI Sensing, four research questions have been addressed in this thesis.

Which specific informational value does VGI present, that could complement usual sources of geoinformation ? What are the strengths and weaknesses of VGI and its typical Use Cases ?
There is a growing consensus about the potential of VGI as a source of geoinformation for various purposes like environmental monitoring, socio-economic studies or crisis management. The first step of this research aims thus at characterizing the informational nature of VGI through a Case Study, where the spatial, temporal, semantic and social dimensions of VGI produced in the context of a large Forest Fire incident are analysed and discussed (*Chapter 1*).

In a data overload context, what strategy could allow to tackle the credibility issue VGI is facing ?
The issue of VGI credibility is developed in Chapter 2 (more specifically in its section 2.2); in a nutshell, it is argued that researchers often highlight VGI lacks the usual characteristics of good quality scientific data such as trustworthiness of the source, measurable level of accuracy or compliance with a recognised observation methodology. The background research presented in the Introduction highlights three possible strategies to overcome VGI's credibility challenge. Firstly, it is possible to reinforce the control on the production chain by establishing a standardised data creation method and by working with a limited number of well-trained volunteers. Secondly, the quality control itself can be set up as a volunteered process and the community of users can act as quality filters for VGI as can be found for collaborative media like Wikipedia. A third option could be to turn the challenge of data abundance into an opportunity, where reliable information is extracted from vast amounts of VGI with uncertain quality from numerous sources by

applying cross-validation mechanisms. The concept of VGI Sensing that is developed in this thesis follows the third strategy.

What would be a typical chain of processing for converting VGI into reliable geoinformation and what are the research challenges to optimise such workflow ?
Following the *'Divide and Conquer'* engineering principle, this research proposes a succession of independent but complementary processing steps allowing to collect, format, enrich, filter, cluster, and validate VGI in order to convert individual heterogeneous information items into a consolidated geoinformation dataset. By doing so, specific research tracks have been individuated, each contributing independently to the general objective of developing VGI Sensing. Such VGI Sensing workflow is proposed in *Chapter 2*, and the various subsequent research challenges are discussed in the light of a specific Use Case in the context of Flood events in United Kingdom.

Specifically on the Clustering step of the VGI Sensing Workflow, what spatiotemporal clustering algorithm would provide the most satisfactory results with heterogeneous but semantically rich VGI ?
The fourth objective of this thesis aims to address one of the key challenges identified while defining the VGI Sensing Workflow, namely the clustering of VGI items. The approach is that Event Clusters can be created by aggregating VGI items that are close not only in terms of spatial and temporal proximity but also that have similar contents (semantically close) and/or that emanate from group of individuals with specific ties (socially close). If properly generated, such Event Clusters would then constitute collections of VGI items relating to a particular event on the Earth surface (e.g. a Forest Fire) that may contribute to enrich situational awareness (e.g. for supporting the decision to dispatch additional Fire Fighting resources or to perform a detailed post-crisis damage assessment). *Chapter 3* explores the question of VGI clustering in depth, by presenting and discussing a benchmark of a selected set of spatiotemporal clustering algorithms with the purpose of detecting and characterizing forest fires at a North American continent scale.

6. Outline of the thesis

The general methodology of this thesis follows the classical material-methods-results scheme: experiments involve the collection of VGI, their processing with a specific purpose in mind, and the analysis of the outcome allowing to draw conclusions that are applicable to further VGI sensing endeavours.

This is applied throughout the chapter with a gradual focus on 'sharp' methodological questions: chapter 1 is an exploration of the VGI material, while chapter 2 implements the entire VGI Sensing Workflow and compares the outcome with other sources of geoinformation, such as media and Remote Sensing. Chapter 3 focuses on optimising a specific step of the VGI Sensing Workflow (namely: the clustering step) by benchmarking algorithms and parameter sets. Chapter 4 then zooms out again, by looking at the wider perspective around VGI Sensing.

The outline of the thesis thus reads as following:
In **Chapter 1**, a proof-of-concept of the informational value of VGI in disaster situation is presented, using the 2009 Forest Fire in Marseille (France) as a Use Case and the micro-blogging platform Twitter as a data source.

In **Chapter 2**, the design of a generic VGI Sensing workflow is presented, and key research challenges are discussed through the analysis of its application on a floods detection and characterisation Use Case in the UK during the summer 2007, with the pictures-sharing site Flickr as a data source.

In **Chapter 3**, the specific question of VGI clustering stage of the VGI Sensing workflow is addressed in detail, with a benchmark of state-of-the-art algorithms applied to the detection and characterisation of Forest Fire events in Canada and the continental US during the summer 2009, with the pictures-sharing site Flickr as a Data Source.

In **Chapter 4**, this thesis puts in perspective the concept of VGI Sensing by further discussing its role as a complementary source of geoinformation next to Remote Sensing in the context of a conceptual Digital Earth Nervous System.

Key **conclusions** of this research, and perspectives for future works are then discussed, while a final commentary addresses the question of ethics for VGI Sensing.

Chapter 1 - Proof of Concept: the informational value of Volunteered Geographic Information[1]

Abstract

The emergence, in the first decade of the 21st Century, of interactive web applications, often labelled as Web 2.0, has permitted an unprecedented increase of content created by non-specialist users. In particular, Location-based Social Networks (LBSN) are designed as platforms allowing the creation, storage and retrieval of vast amounts of georeferenced and user-generated contents. Geographic Information specialists can thus see LBSNs as a timely and cost-effective source of spatiotemporal information for many fields of application, provided that they can set up workflows to retrieve, validate and organise such information. This chapter aims to improve the understanding on how LBSN can be used as a reliable source of spatiotemporal information, by analysing the temporal, spatial and social dynamics of Twitter activity during a major forest fire event in the South of France in July 2009. By doing so, the informational value of Volunteered Geographic Information in the context of crisis management is highlighted.

1. Introduction

Recent evolution of the Internet has permitted an unprecedented increase in content created by non-specialist users thanks to a reduction in technical barriers (O'Reilly 2005). When such user-generated contents have a geographical dimension, these are commonly referred to as Volunteered Geographic Information (VGI), having a huge potential to engage citizens in place-based issues and provide significant, timely and cost-effective source for Geographer's and other spatially-related fields of research and management

[1] This study has been published in the Proceedings of the International Workshop on Location Based Social Networks, 2009, Seattle, Washington, USA under the title *"OMG, from here, I can see the flames!": a use case of mining location based social networks to acquire spatiotemporal data on forest fires.* The text presented here is slightly modified from the original publication for layout and terminology harmonisation.

(Goodchild 2007). For this latter group, Location-based Social Networks (LBSN) are expected to become a rich source of VGI, as they combine the functionalities of Social Networking Services with a location-based technologies. In addition, such content may play an increasing role in Spatial Data Infrastructures (SDIs), and needs to be properly handled to ensure its appropriate use, particularly in time-critical issues such as crisis management and disaster response.

This chapter aims to contribute to this growing body of literature by studying how Twitter[1] can be used as a source of spatiotemporal information. By focusing on a recent real-life case of forest fire, we aimed to demonstrate its possible role to support emergency planning, risk assessment and damage assessment activities. Specifically, the analysis draws on publicly available Twitter messages published during a forest fire event that took place near the French city of Marseille in July 2009, with a particular focus on the identification of the content's temporal, spatial and social dynamics. Although the study only involves one use case, it is argued that the richness of the information provided in a real event by users from different backgrounds will provide generalizable outcomes to a range of scenarios and related LBSNs. Although the collection and analysis were performed almost 7 years before the publication of this thesis, the key observations and conclusions remain valid, as no fundamental change on the Twitter platform and its usage patterns occurred since then (although the number of users worldwide grew considerably).

This chapter is structured in four main sections covering previous works in the topic (section 2), a description of the use case (section 3), and the result of the analysis of the Twitter material (section 4). In order to provide some context, the following section begins by introducing the platform, Twitter.

2. Previous works

2.1. Twitter

Twitter can be defined as a 'micro-blogging' platform, a special type of Social Networking Service that puts emphasis on simplicity and openness (Marks 2009). Twitter allows users to post very short

[1] http://www.twitter.com

messages (maximum 140 characters), called *tweets*. By default, all the tweets are visible on a public timeline, where an asymmetric *following* system allows users to see their personal timeline for tweets they consider to be interesting. In addition, users have created a specific syntax for short messages. For example, inserting *@username* in a tweet means that this message is a response to a user called "username". Similarly, the *RT* code tells readers that a message has been 're-tweeted' (similar to "forward" in many e-mail clients). Finally, the 'hash tag' syntax has been introduced recently to ease topic-related searches. In this use case, for example, keywords like #incendie and #marseille were frequently present in the tweets, thus allowing every interested user to easily retrieve messages related to the event. It is such syntax that offers an important filter to extract useful content from Twitter and re-use it in geospatial application areas and resources such as SDIs. Although a Tweet is *sensu stricto* a (very) short message, it is literally 'packed with metadata' (Perez 2010) and carries numerous information items such as timestamp, time zone, unique ID, unique ID of the tweet it is a response to (if any), number of times the tweet has been re-tweeted, author's name, author's language, author's location (if enabled), … and more.

Twitter initially generated a lot of hyperbole in the media, as the most prominent example of the 'social' trend the Web 2.0 features (Anon 2009), that "will change the way we live" (Johnson 2009). More specifically, Twitter is presented as a primary source of '*citizen journalism*', as "every day in the US, people randomly witnessing an exceptional or dramatic event (crime, protest or accident) use their mobile phone to broadcast real-time information from the field on Twitter [translated by the authors]" (Eudes 2009). Twitter is also presented in the media as a highly dynamic means to communicate between citizens affected by mass convergence events, such as hurricane Gustav (Ulrich 2008) or the recent troubles surrounding the Iranian elections (Cardwell 2009).

Although, the body of scientific literature about Twitter is abundant (see Introduction for details), its potential for spatiotemporal information has still to be exploited. Most studies have focused on its social dimension by studying users motivations (Java et al. 2009), interactions (Huberman et al. 2009) or collaboration (Honeycutt & Herring 2009),with an article from the crisis informatics field examining Twitter adoption during mass convergence events (Hughes

& Palen 2009). When originally published, this paper aimed to provide a stimulus for further exploration of the role of LBSN sources such as Twitter for crisis management and to consider the valuable geospatial component they can contain, particularly for time-critical events. It seems it has been effective, since it contributed to a wide corpus of literature published since then (see Introduction for details).

Twitter is notable in its design in relation to both time and space. Tweets are organised in *timelines* (i.e., series of tweets sorted and displayed in reverse chronological order) and the time each tweet has been published is available with a level of accuracy of 1 second. The spatial dimension of Twitter is more complex, where georeferencing takes several basic forms. Firstly details can be provided in relation to tweets indirectly or directly. In an indirect form, a user's *location* is provided on their profile page but this *location* is expected to be the place were they live and not their location when a tweet is made. Notably, applications running on GPS-enabled smartphones allow users to automatically update this location *field* each time a tweet is posted, thus converting Twitter into a genuine LBSN. For example, a user living in San Francisco can tweet from his GPS-enabled smartphone and allow his Twitter app to disclose his precise location as metadata of his tweets ('direct location'), this would be referred to as 'geotweeting') (Stone 2009),. Oppositely, he might tweet from a desktop computer on which the browser is configured to not disclose any location information (although the Twitter server could guess it from the IP of the client computer, this information is not disclosed to third parties). In this case time zone,, no location will be available for this particular tweet and only the reference to San Francisco in the user profile can be used to guess ('indirect location') from where it has been published (although the user can be travelling anywhere in the world at that particular time)., It has to be expected that Twitter features are used in an heterogeneous manner by users, depending on their smartphone's settings, their privacy concerns and their technological literacy.

Third parties can access and import Tweets (including their metadata) via a specific Application Programming Interface (API). The Twitter API implements a RESTful architecture (Fielding 2000), so queries take the form of typical web client requests (http 'get' requests) to a specific URL. Such query can contain parameters (in the form of URL parameters in full text – e.g. "&user=john") that allow filtering the

relevant tweets (based on contents, time, user, location – if available -, etc.). Responses consist in lists of Tweets with their metadata[1], in a structured text format such as XML of JSON. The free version of the Twitter API (referred to as 'garden hose') does not guarantee the comprehensiveness of the results while the paid version ('firehouse') does. The Twitter API reference does not give specific information about the factors influencing the comprehensiveness of results, but advises 'that properly configured queries will fetch most of the relevant tweets'. It is nevertheless usually admitted in the research community that results from the 'garden hose' constitute a large and representative sample of all published tweets.

2.2. Harvesting spatiotemporal information from the web

The idea of harvesting spatiotemporal information from the web has seen some early endeavours that were contemporaneous with this research. For example, it has been demonstrated that general purpose *Points of Interest* (POI) can be automatically derived from users' map annotations (Mummidi & Krumm 2008) and vague geographic regions (e.g., Midlands, or Middle West) delineated (Jones 2008). As well as numeric and textual data, georeferenced pictures from the photo-sharing website Flickr have been processed in terms of their density to show where the most famous landmarks are for a given location (Crandall et al. 2009). In addition, a Geospatial Exploratory Data Mining Web Agent that retrieves geographic information from web pages (related to outdoor activities), has also been discussed (Pultar et al. 2008). As such, this chapter aims to explore the role of Twitter as another source of spatiotemporal information for such workflows, helping to advance existing capabilities for monitoring natural hazards.

In the case of obtaining data for the present study, the Twitter Application Programming Interface (API) has been used to retrieve tweets and related metadata in an xml format in response to a specific query. We wrote in PHP scripts for Data mining and web-crawling, we then applied them to the sample of tweets to create organised, meaningful content (including basic summary statistics) such as: a list of users' locations; a list of geocoded place-names cited in the tweets; lists of domains related to the full URLs contained in the tweets; etc.

[1] See https://dev.twitter.com/ for a comprehensive reference about Twitter API features.

Having adopted this methodology, the specific case of the Marseille forest fire can be introduced.

3. Case study: the Marseille Fire

3.1. The Marseille forest fire

The Marseille Fire took place on the 22nd and 23rd of July 2009 near the French city of Marseille, the second most populated city of France (1,6 million inhabitants) situated on the Mediterranean coast. According to information provided during and after the fire by the media agency *Agence France Presse* (*AFP*) and the local newspaper *La Provence*, the fire started at 13:34 the 22nd of July 2009 in an unpopulated and mountainous area, 20km from Marseille. The fire was started accidentally by soldiers during an exercise near the camp of Carriage and progressed rapidly towards Marseille. At around 16:00, its front crossed the pass of the Mont Latin and by 18:00 it was getting closer to densely populated areas in the East and Southeast of the city. Later in the evening (around 20:30), several isolated houses had to be evacuated and through the night, hundreds of citizens, frightened by the dense smoke, left their homes despite advice from the police to stay inside. The fire was reported as being completely under control by 7:00 on the 23rd of July; up to 10 houses had been destroyed, there were no fatalities but between 1100 to 1300 hectares of forest and Mediterranean scrubland had been destroyed.

The Marseille fire was chosen for three main reasons. Firstly, the Twitter usage is different from other studies and contexts, as most previous use cases have involved incidents in the United States. Instead the focus is on a non-English speaking European country with only of few Twitter users (0.9% in France) (Cheng et al. 2009), compared to the larger numbers found the United States (62.14%), United Kingdom (7.87%) and Canada (5.69%). Secondly, the Marseille Fire took place near a very densely populated area and thousands of citizens were, or at least appeared to feel, directly affected. Lastly, the event attracted a lot of attention from the media, allowing the research to explore something that should equally have attracted a lot of attention in Social Media such as Twitter.

3.2. Assumptions/hypothesis to verify

In order to provide focus to the study, the following hypotheses were set out. They are based on the hyperboles from the Media (see section 2.1) in an aim to verify the (degree of) veracity of such claims in the context of the Marseille fire.

H1: Twitter is an extremely fast information dissemination platform.

H2: As an LBSN, Twitter provides accurate and useful spatiotemporal information.

H3: Users use Twitter to communicate with each other in widely open conversation; as a result, it is a primary source of information from citizens.

H4: Twitter is used as information broadcasting and brokerage platform during crisis events.

The first three are addressing hyperbole of recent newspaper articles, whereas H4 is one of the conclusion points of (Hughes & Palen 2009).

3.3. Material: Tweets about the Marseille Fire

The Twitter API was used to collect material about the forest fire. The observation period started on the 22nd of July at 12:00 (local time = GMT+2) and ended on the 23rd of July at 12:00. This ensured that content was gathered more than 1 hour before the fire started and ended 5 hours after the fire had been declared 'under control' by official sources. In order to select appropriate tweets in the local language, the keyword '*incendie*' was used, as it specifically means, "fire that causes important damage" and is also widely used as a technical term to designate forest fires and wildfires (*incendie de forêt*). As such, the material harvested is a minimised but focussed set of Tweets on the key topic of interest.

From this search, 346 tweets were collected. A further filter was applied to the content to ensure only those *incendie*-tweets were directly connected to the Marseille Fire. From the 33 removed items, 24 notably had spatial references in their text to aid their exclusion (e.g. "Catalonia" and "Corsica"). Moreover, 20 removed tweets were sent between midday and 13:34, when the Marseille Fire actually started. Thus, such 'noise' in the data's signal was seen as readily identified and removed to eventually have a sample of 313 relevant tweets provided by 127 individual users. This number of tweets is relatively low compared to the 4000 of Tweets per day related to the Gustav hurricane in the United States in 2008 (Hughes & Palen 2009).

However, the richness of the content of the tweets over a short period of time arguably allowed for a decent representation of the event and a qualitative analysis of such content in terms of the temporal, spatial and social dynamics of the information reported and an assessment of the providers and a tweet's content for this emergency-related event.

4. Results and Discussions

4.1. Analysis #1: temporal dynamics

Figure 1 shows a time line with the major events related to the Marseille Fire and the number of relevant Tweets per hour, with examples of tweets (translated by the authors) taken at key moments in the event. The first Tweet mentioning the fire was published at 15:08, about one and a half hours after the fire started. It refers to headline published on the website of the local newspaper, *La Provence*, at 14:08. It would seem that well informed local journalists were still faster than 'the crowd' in reporting the fire, something that would challenge the idea that Twitter provides a rapid means to disseminate information (relating to H1). Part of this, however, may be explained by the role local media may play in actively contacting civil protection authorities to find new stories or even participate in emergency planning events. In addition, the fire started in an unpopulated area and this trend of low comment remained in place for some time until the fire began to threaten densely populated places.

A 'lag' of two and a half hours of comments from the public also seems to have been highlighted by a limited initial response from a 'citizen journalism' platform focussed on the event. After this point in time, the Twitter activity was then in line with the situation in the field: as the fire came closer to populated areas, more Tweets were published. The peak in posts around 01:00 on the 23rd of July corresponds with the most critical moment in the event when highly visible flames, smoke and flying ashes frightened hundreds of citizens out of their homes against recommendations from the police. The intensity of this period included a lot of direct messages between users (using the @ syntax) and exclusive information from citizens being forwarded to others (using the RT syntax), highlighting a lot of direct communication (related to H3). Between 3:00 and 7:30, very few tweets were published, until the morning (8:00 to 11:00) when

headline news about the fire was being widely commented upon by citizens.

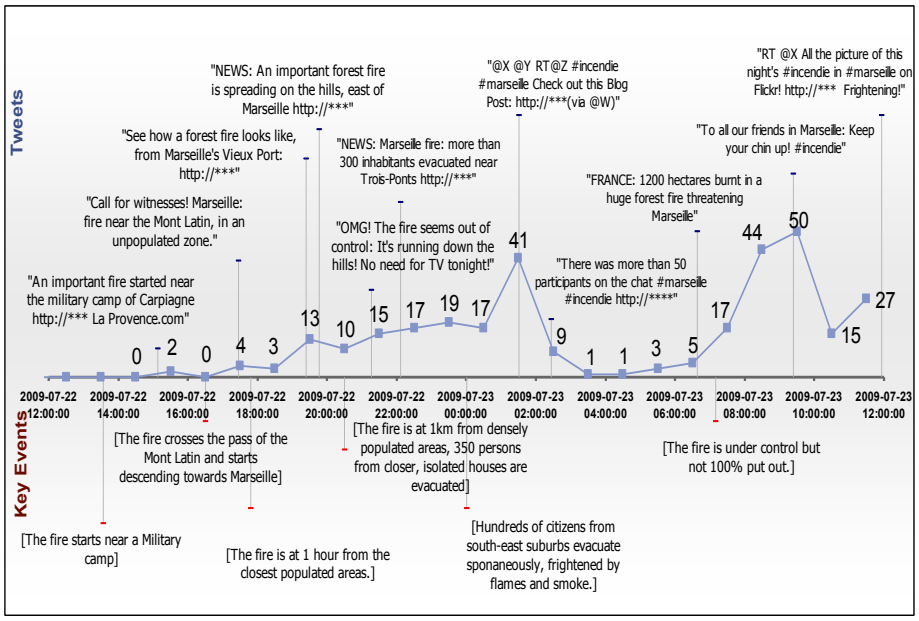


Figure 1 Chronology of the Marseille Fire, number of related tweets per hour and selected tweets' contents (Sources: information provided during and after the fire by AFP and La Provence, and information retrieved from twitter.com

4.2. Analysis #2: spatial dynamics

As mentioned above, there are two ways of acquiring spatial information from Twitter: the user location (if it is updated when a tweet is published) and the geographic coordinates and place names cited in the Tweets. The first source was not available for this use case, as API calls for retrieving users' location have to be made at the exact moment the Tweet was published, provided that a location can be updated at any time. Instead, API calls were made several weeks after the events, in order to calculate the proportion of users dynamically updating their location, and therefore using Twitter as a genuine LBSN. It is found that only 5 users out of 127 were providing accurate geographical coordinates as locations, which seems to contradict the idea that useful GI is readily provided (H2), although better georeferencing is expected to be added to each tweet (Stone 2009). It is also interesting to note that only 23 users have Marseille (or a nearby place) as location and 18 are from "Paris" (most likely

corresponding to media corporations' headquarters), whereas 26 users have not provided such details in their profile.

Information contained in the tweets also provided a chronology of burnt areas (in hectares) to be uncovered, offering some spatially related content (see Figure 2). Around 18:00, the figure of 60 hectares is cited once; it becomes 120 hectares around 19:30 (cited 3 times) and 400 hectares around 20:30 (cited 2 times). The figure of 1000 hectares was cited once around 23:00. It reaches 1200 hectares at 01:00 on the 23rd of July and remains stable during the whole night (cited 12 times). At 8:37, a tweet reports 1100 burnt hectares, and then 9 other Tweets provide the same figure during the morning. Finally, 1 Tweet mentions the figure of 1300 damaged hectares at 11:51. A difference between information providers is also present here, as figures are typically provided by official sources to the media and citizens seem less likely to be able to make burnt areas estimations in real time. Therefore, Twitter should be considered as a secondary source of information in this respect.

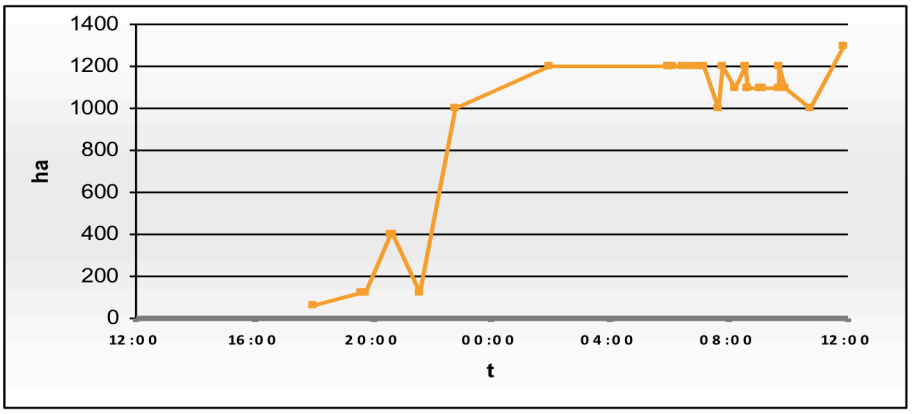


Figure 2 : Hectares of burnt area reported in tweets over time

Although no geographic coordinates were cited in any of the 313 tweets, place names were cited over time by users (see Figure 3). The yellow area surrounded by a red outline represents the estimated total burnt area (source: *La Provence*). The size of each symbol represents the number of citations that can be found in the 313 Tweets (given by the number), and their colour represents the time they have been cited for the first time (lighter means closer to the start of the event).

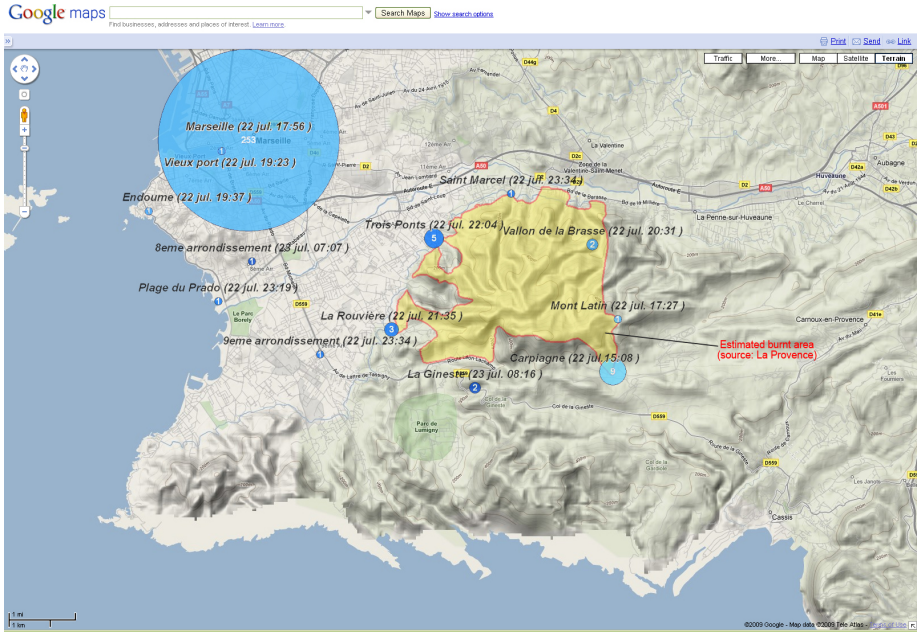


Figure 3 : Location, frequency and time of the first citation of place names cited in tweets, and estimated total burnt area

The most cited place name is, by far, "Marseille". Indeed, in the majority of the 313 tweets (80.8%) the keywords *incendie* and *Marseille* are used to describe this fire event, however these take different forms:

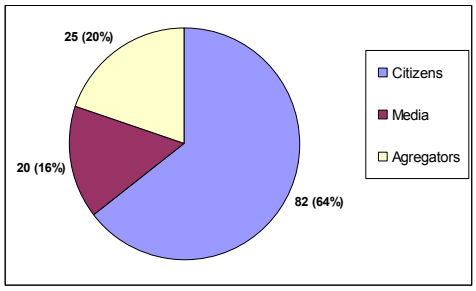
- short phrases: "*l'incendie de Marseille*" ("the Marseille fire")
- emotive phrases : "l'incendie aux portes de Marseille" ("the fire at the gates of Marseille")
- community code: #incendie #marseille.

This provides useful information to situate the fire on a wider scale and to discriminate from tweets related to other fires. To follow the progress spatially, other place names referring to local landmarks (neighbourhoods, valleys, mounts, *etc*.) provide interesting spatiotemporal information. The origin of the fire, for example, was cited 9 times in early tweets. Then, physical features (the *Mont Latin* and the *Vallon de la Brasse*) were cited later in the afternoon, showing that the fire spreads in the mountainous area and moves towards Marseille. The most exposed neighbourhoods are cited several times during the evening (Saint Marcel – 1 citation, Trois Ponts – 5 citations and La Rouvière – 3 citations). However, several nearby places outside the damaged area are also cited for various reasons: the Vieux

Port and the Plage du Prado (touristic landmarks, found in tweets like "I can see the fire from the Vieux Port"), Endoumes (where the fire-fighting Canadairs pumped water), La Gineste (referring to a local road closed on the 23rd of July as a consequence of the fire) and the 8th and the 9th *arrondissement* (administrative subdivisions of the city, which were close to the event).

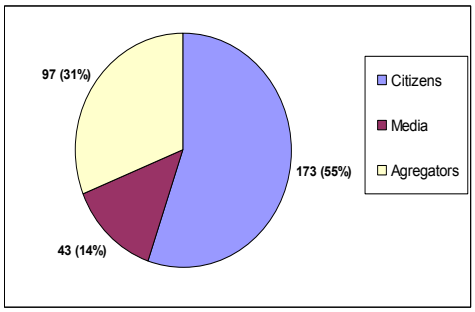
4.3. Analysis #3: social dynamics

To better understand the type of information that is available on Twitter, it is important to characterize who actually tweets. Indeed, a notable proportion of the 127 users had a name which referred to well known French speaking media corporations (e.g., TF1, Le Figaro, RMC). Based on the information provided in the publicly available user profile of each user, 3 categories are suggested: citizens, media and a role between these two as 'aggregators' (see Figure 4 and Figure 5).



Citizens are physical persons acting on their own behalf (64%; even if an unknown proportion of them may work as journalists without mentioning it in their profile), who contributed to 55% of the total amount of tweets.

Figure 4 : Number of users that published tweets by type



This tends to contradict the idea that Twitter is exclusively a primary source of information from citizens (H3). In contrast, the presence of well-known traditional media (newspapers, TV networks, radio) involved 16% of users.

Figure 5 : Number of users that published tweets by user type

Aggregators do not create new information but compile it into specific news-feeds that they broadcast to a targeted audience. Their user profiles often point to news portals that have a specific local focus

(e.g. Marseille's News), thematic focus (e.g., natural hazards) or to news-related and 'citizen journalism' blogs. Aggregator users can use tools like TwitterFeed[1] to automatically re-publish the contents a RSS feeds into tweets, and thus very easily reproduce information contents on Twitter without human intervention. In this research's sample, nearly 1 tweet out of 3 (31%) has been published by an aggregator. This finding is important to understand the apparent redundancy of every piece of information. It seems that aggregators act like a delay effect and propagate a redundant signal with limited added value. If for example a piece of information is published by a media agency, and then updated because it was erroneous, it is not guaranteed that the *errata* follow the same re-publication path via aggregators. The same applies to citizens that use the *RT* syntax when they 're-tweet' information; in the sample 18.8% contained the "*RT*" code). This can create problems to set up quantitative quality filters on top of Twitter: the fact that information is tweeted numerous times may be not be interpreted as a proof of veracity, or other sense of 'truth'. Such 'echo effect' poses a key methodological issue to VGI Sensing: when numerous VGI items are pure repetition of the same information, the cross-validation may become ineffective to assess its credibility (in a similar manner as someone giving more credit to a rumour if it is widely spread). Oppositely, VGI Sensing should be designed assess as credible information that relates the same facts independently from numerous sources.

4.4. Analysis#4: URL analysis

Further analysis revealed that 75% of tweets contained a URL. This is a very important proportion compared to previous findings - 13% (Java et al. 2009) and 25% (Hughes & Palen 2009) - and provides strong evidence for accepting H4. It is a common practice on Twitter to use abbreviated URLs[2]; where an *ad hoc* script was used to resolve full URLs before further analysis. Those 236 links pointed towards 148 unique pages (i.e., each link has been cited on average 1.6 times) and towards 62 unique domains (i.e., on average, each website has 2.4 cited pages). The cited domains were sorted according to the following classes:

[2] Using URL shortening services such as http://bit.ly/

- *Forum, Blogs, Chats*: this involved all domains corresponding to services that focus on user-generated text and discussions between users (e.g., blogspot.com, tinychat.com).
- *Social Media*: this involves services dedicated to share pictures or video between users (e.g., flickr.com, twitpic.com).
- *Media*: including websites from well-known media corporations, newspapers or broadcasting from television and radio (e.g.: france-info.com, lemonde.fr, tf1.lci.fr). It is interesting to note that no news agency's website – like reuters.com or afp.com - was present in the cited URLs
- *News Portals*: involves news aggregators, as noted above. Such news portals are not directly connected to 'traditional' media and, again, typically do not act as primary sources of information.

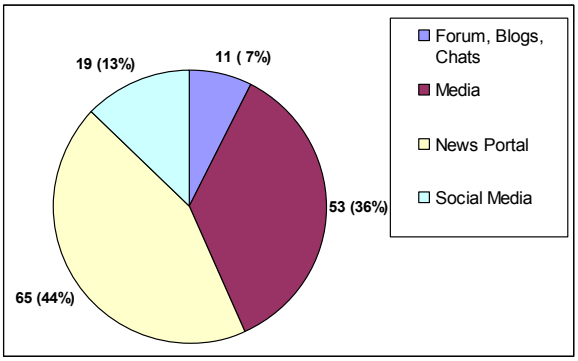


Figure 6 : Number of unique cited URLs by domain type

These results (see Figure 6) show that, even if citizens are indeed sharing personal reports on Twitter, 80% of the referenced material came from existing media and news portals, perhaps challenging H3.

However, this small proportion of links pointing towards *fora*, blogs, chats and social media led to additional useful material. Dozens of pictures of the fires taken and published on Flickr or Twitpic were accessible in nearly real time. Citizens used Twitter to call all interested participants to join a live chat on the events on TinyChat.com. A couple of blog posts from 'citizen journalists' relayed the situation in the field minute-by-minute, generating hundreds of comments from other citizens, thus contributing to a form of 'situational awareness'. Importantly, all such content is being delivered through one channel, a Twitter timeline.

5. Conclusions

This chapter has covered the application of Twitter as a source of spatiotemporal information for crisis events, following the example of

a recent fire in France. It presents an innovative use of LBSN content and explored the dynamics of its creation through four main axes.

Firstly, the analysis of the temporal dimension revealed that content was inherently accurate due to time-stamps but additionally well synchronized to actual events. However, this is not true for the initial phase of the event, which was first reported by the media, contradicting the hyperbole of Twitter as an extremely dynamic tool for citizens to report exceptional events. This can in part be explained by the sparsely populated location of the fire, which in itself raises issues about what is surveyed by 'citizens'.

Secondly, as Twitter users choose to provide a geographic dimension (either directly on indirectly) to events they record, it offers a valuable resource of GI following four main types: spatial terms (e.g. "burnt areas" coded by unit of measurement), direct place names ("Marseille"), coded place names (#marseille), location pairing ("the fire [over there] seen from [my location]"). Although the Twitter activity monitored involved only a few examples of accurate user-positioning, planned developments for the platform and wider penetration of smart phones on the mobile phone market could make this more accurate and abundant in the near future.

Thirdly, social analysis revealed 3 major roles of those who tweet: citizens, media and aggregators, where the latter do not produce primary content but compile existing sources into specific news-feeds that they broadcast to a targeted audience. This categorization is important to better understand the type of contents those re-using such content will be faced with. Specifically, the phenomenon of information replication in Social Media – which we nicknamed 'echo effect' – poses a methodological issue to the VGI Sensing approach consisting in considering co-occurrence of similar VGI as an estimate of its credibility. To overcome this issue, VGI Sensing methods should endeavour to distinguish independent observations from citizens as primary sources of information from pure repetition of the same information *en masse* (secondary sources).

Fourthly, further analysis of cited URLs revealed that the share of genuinely user-created content was even smaller than the proportion found in the social analysis, where only 20% of the contents that can be crawled comes from blogs, chats, *fora* and other citizen generated

information. Although scarce in proportion of other types, such contributions should be recognized as rich in content and, therefore, valuable. Another feature of URLs was elevated redundancy, where news items where repeatedly cited over time. Although this creates echo effects as discussed above, the massively aggregative role of Twitter from many information sources ensures that important primary content is presented in a single channel, thus easing information retrieval processes from a single time line.

It can be seen that such 'tweet channels' could offer promising seeds (starting-points) for crawlers to collect event-related data, where time and location matter. Future work should consider the categorization of such content in relation to other Web 2.0 platforms. This chapter aimed to support further development of automated content retrieval and processing workflows, helping to provide useful, contextualized and sought-after VGI to enrich the content of expert-driven Spatial Data Infrastructures. Just as we readily accept the processing of satellite data as an input to many geospatial analyses, we should also aim to better interpret the abundant and freely available signals provided by citizen-sensors. Such processing workflows are the main topic of the following chapter.

Chapter 2 – A Volunteered Geographic Information Sensing Workflow[1]

Abstract

In this chapter, the research question of how credibility in VGI can be increased is addressed, so that it becomes a valuable resource, within domains such as natural hazards. This is done by proposing a generic workflow that uses prior information about the phenomenon of interest and reasoning techniques to improve the reliability of the VGI; thus creating a useful source for scientific and technical investigation. This workflow has been developed for a particular case study: the use of pictures from the photo-sharing portal Flickr as a 'signal' that allows the pinpointing recent flood events in the United Kingdom. A comparison of the workflow's output with independent information provided by scientists and journalists about these floods is also presented in order to uncover some of the strengths and weaknesses of VGI in the present case study. We conclude that the proposed workflow allows VGI to be turned into a reliable data source.

1. Introduction

Initially designed as a one-directional text broadcasting media, the Internet has rapidly become an information-sharing platform commonly referred to as "Web 2.0". One of the key consequences of this evolution has been to increase participation, with lowered technical barriers permitting an unprecedented increase of content created by non-specialist users (Rinner 2008).When such user-generated content have a geographic dimension, it is commonly referred to as Volunteered Geographic Information (VGI), which has a huge potential to engage citizens and to be a significant, timely and cost-effective source for geographers' understanding of the earth (Goodchild 2007).

[1] This study has been published in Geomatica vol. 64, no. 1 (2010) under the title *Citizens as Sensors for Natural Hazards: A VGI integration Workflow*. The text presented here is slightly modified from the original publication for layout and terminology harmonisation.

To emphasize its potential as a novel source of information, O'Reilly (2005) described user-generated content created through blogs, wikis, and media-sharing platform as the *wisdom of the crowds*. However, such information is often regarded as insufficiently structured, documented and validated to be a reliable source of information to perform scientific and technical analysis (Flanagin & Metzger 2008). This is why Mummidi & Krumm (2008) emphasized that "one of the potential problems of VGI is ensuring its quality" (p. 215), while Craglia et al. (2008) identified the development of "collaborative frameworks allowing the emergence of hybrid infrastructures combining both voluntary and institutional data" (p. 162) as a major research challenge with emerging concept of a 'Next Generation Digital earth'.

In this chapter, the research question how credibility in VGI can be increased is addressed so that it becomes a valuable resource to study natural hazards. This is done by proposing a generic workflow that uses prior information about the phenomenon of interest and reasoning techniques to improve the reliability of the VGI, thus supporting scientific and technical investigation. This data integration workflow includes retrieval of information, its filtering, formatting, validation, ranking, clustering, and final conversion into a geographic information layer.

As a proof of concept, we applied the proposed workflow to a real-life case study. User-generated information from the popular photo exchange website Flickr[1] has been retrieved and processed through its Application Programming Interface[2] (API) to locate in space and time flood events that occurred in United Kingdom between the 1st of January 2007 and the 31st of March 2009. Ad hoc scripts have been written in PHP by the author to this end.

Floods are typified by a situation of crisis, where responsible authorities need up-to-date situational awareness in order to effectively coordinate response. Much of the information that is traditionally gathered for such situations comes from official, trusted sources (e.g. emergency services, local authorities, mapping agencies).

[1] http://www.flickr.com

[2] http://www.flickr.com/services/api/

Web 2.0 potentially empowers citizens to complement these trusted data sources with their own, dynamic, on-site observations (De Rubeis et al. 2009; Hughes & Palen 2009).

A qualitative and quantitative comparison of the workflow's output with data coming from entirely independent information sources is also presented in this chapter. By doing so, we aim to raise awareness about the strengths and weaknesses of VGI as a novel information source. The first validation dataset is the Global Active Archive of Large Flood Events from Dartmouth Floods Observatory[1], which catalogues floods that had a large impact. The second validation dataset is the archive of press articles classified as related to 'floods' and 'United Kingdom' by the European Media Monitor system[2] developed by the European Commission's Joint Research Centre (JRC). The third validation dataset is the JRC experimental Global Flood Detection System[3], which measures floodwaters extent globally and with a daily frequency using passive microwave sensing.

The remainder of chapter is structured as follows: the next section describes the concepts and technologies that are the basis of this research. A case study is then described (section 3). Before turning to, a detailed description of the VGI data integration workflow (section 4), and in section 5 the outputs to information from independent scientific and media sources are compared. This is followed by conclusion and future work items, in section 6.

2. Previous works

In this section previous works in the field of VGI initiatives are described. In particular, he issue of VGI's quality control is discussed, and its use in the context of natural hazards, as well as the Flickr platform as a source of VGI.

2.1. Volunteered Geographic Information

According to Goodchild (2007), the term Volunteered Geographic Information (VGI) is used to designate any user-generated content that has a relation to the surface of the Earth. There are various VGI

[1] http://www.dartmouth.edu/~floods/
[2] http://emm.jrc.it/overview.html
[3] http://www.gdacs.org/floods/

applications that allow users to upload and browse information in various media (text, pictures, videos, documents, *etc*.), where such information becomes 'spatial' through links to a spatial reference.

OpenStreetMap[1] is one of the most famous VGI initiatives. The OpenStreetMap product is a free, editable and general-purpose street map. It is created by collaborative methods, where users can upload new streets through GPS tracks or modify existing information (e.g. for the purpose of quality enhancement). This VGI initiative aims to extend the geographic coverage of the OpenStreetMap product across the globe. In 2009, it contained more than 22 million kilometres of roads, covering 114 countries over the 5 continents (OpenStreetMap 2009).

The WikiMapia initiative[2] was inspired by the success of the online multilingual encyclopaedia, Wikipedia. However, unlike Wikipedia, WikiMapia focuses on providing information strictly related to a particular geographic location (i.e. about towns, cities, lakes, regions, *etc*.). It offers a map interface to browse the content. Users can create bounding boxes, or more detailed polygons inside a bounding box. They can also insert a title, a short description, and a link to a Wikipedia page that allows more information about the described item to be documented.

Google Maps, the geographic interface to the Google search engine, allows users to create VGI in the form of all-purpose personal maps[3]. Such maps (called 'My Map') are collections of points, lines or polygons that are associated with media items (e.g. text, html documents, photos, and videos). The contents of such maps can be searched by other users who selected the option 'search user-created contents'.

These examples illustrate how VGI can involve vast amounts of data, and be applied across various domains. In addition, geotagging (i.e. associate geo-localization information to a piece of information) is

[1] http://www.openstreetmap.org/
[2] http://wikimapia.org/
[3] http://maps.google.com/support/bin/answer.py?hl=en&answer=68480

also popular for blog posts[1], short messages, photos and videos sent directly from GPS-enabled smart phone (Jones 2009).

2.2. The challenge of using VGI in expert-driven Information Systems

Although the term VGI has been recently coined, notions associated with citizen-based data collection, validation and generation have been long-standing. For example, annual bird counts in the 1900s in the United States (Lee 1994) and British land use survey in the 1930s (Stamp 1937) utilized indigenous knowledge to generate detailed content about the local environment. More recently, technological advances have led to something of a renaissance of such projects in environmental monitoring (e.g. (Oscarson & Calhoun 2007; Monk et al. 2008; Fritz et al. 2009), in line with earlier developments of Public Participation GIS (Tulloch 2008).

During this long history of public participation, data quality has always been recognized as a major concern which also applies to VGI research (Elwood 2008). VGI is often based on perceptions rather than measurements, and its quality cannot, therefore, be only measured with objective criteria like positional accuracy (Flanagin & Metzger 2008). This is why the notion of VGI credibility is used in this chapter, which is defined by these authors as a subjective notion that describes whether or not a piece of information can be believed in, considering any possible intentional or unintentional error, omission, or exaggeration. While credibility applies to each piece of VGI individually, the term reliability is used to designate if a platform for creating VGI provides a significant amount of credible information, and thus if it can be used as a valuable source of information.

Reasons for a perceived lack of credibility of VGI are various. Firstly, citizen-generated data can appear as insufficiently documented compared to scientific observations. Gouveia et al. (2004) highlighted that, in the context of citizen-created data, "data quality is often unknown; metadata on data sampling and collection are also scarce, making potential users sceptical about the data" (p.139). Secondly, Flanagin & Metzger (2008) note this credibility issue is mostly due to the apparent lack of control of the data creation process. Many

[1] http://bloggerindraft.blogspot.com/2008/12/new-feature-geotagging.html

parameters often remain uncertain while dealing with VGI, such the data creator's level of expertise, their motivations, or the data creation method and its expected maximum accuracy level. In addition, the same authors argue that in the data abundance context that characterizes VGI, traditional mechanisms that tend to increase trust in data, like credibility of the sources and certified information gatekeepers, are ineffective. But whereas a VGI item can be wrong (e.g. because of an error in positioning, or because the textual information contains inexact facts), VGI is often extremely rich in context, and the ratio benefits *versus* risks might be very high in a wide range of use cases where ground knowledge from local citizens is valuable (Goodchild & Glennon 2010).

Several strategies are possible to overcome the credibility challenge of VGI. Firstly, it could be possible to reinforce the control on the production chain, by setting up a standardized data creation method and by working with a limited number of well-trained volunteers (Lee 1994). Secondly, the quality control itself can be set up as a volunteered process, and the community of users can act as quality filters for VGI as for Wikipedia (Bishr & Mantelas 2008). A third option could be to turn the challenge of data abundance into an opportunity, where reliable information is extracted from vast amounts of VGI with uncertain quality from numerous sources via cross-validation. In other words, the data quality problem of VGI can be addressed by "aggregating input from many different people" Mummidi & Krumm (2008, p. 215), and by processing these VGI clusters to evaluate their relevance to a give goal.

The works presented in this thesis is developed following this third strategy.

2.3. VGI for natural hazards and crisis management

Natural hazards are typically monitored using instrumental observations, such as seismological networks for earthquakes. However, such observations produce information on the hazard characteristics, not on the impact. A natural disaster occurs when a local society is disrupted by a natural event and losses are so large that outside help is necessary. Information on the impact of a disaster is usually estimated using scientific models or reports from media and local governments (De Groeve et al. 2006). In this context, VGI can be a valuable alternative for information about subsequent impacts.

VGI can play a role in most phases of the disaster cycle: preparedness, mitigation, early warning, response, and reconstruction. In preparedness, VGI can be used to compile data on population and infrastructure at risk. For instance, geo-referenced picture databases, such as Flickr and Panoramio, can provide an accurate representation of an affected region. More than geographic databases, VGI can capture specific details, such the construction materials of buildings (e.g. adobe or reinforced concrete), which can be important to understand local vulnerability. In mitigation, VGI can be used to obtain early warning of a slow-onset disaster (Gendron & Hoffman 2009). One example is the Ushahidi platform for crowdsourcing crisis information, which has been deployed for applications as diverse as monitoring election problems and tracking the spreading of H1N1 influenza (Bahree 2008). In the reconstruction and rehabilitation phase, VGI can provide complementary information to satellite imagery to assess the extent of any damage (in the planning phase) and document implementation status of assistance (in the execution phase).

The most straightforward application of VGI is expected in disaster response, to obtain situational awareness in time critical conditions. In order to respond effectively to an event, crisis managers need up-to-date situational awareness, which is traditionally built through trusted information sources, e.g. by sending assessment teams to the disaster site, through existing networks (e.g. police) or using media reports. However, many cases can involve citizens present at the disaster sites where they provide VGI (Hughes & Palen 2009).

While this VGI can be timely, its value to the situational awareness is unproven. VGI can contain false information, interpretation, rumours and, in general, the information accuracy is unknown. Therefore, the crisis management community is somewhat wary of using it for decision-making. On the other hand, VGI can provide authorities with an understanding of how local citizens are reacting to a disaster (Palen et al. 2009), after which they have an opportunity to better focus public communication.

One example in situational awareness where issues of quality have been overcome is macro-seismic intensity data collection, where seismological institutions gather data on earthquake intensity through

"Did you feel it" reports. Even if filed by untrained citizens, research has demonstrated its sufficient quality for scientific use, including improving seismic wave propagation models (Wald 1999; De Rubeis et al. 2009). Further research and case studies are needed to extend this to other disaster types. For flood extent assessment, VGI could be a valid complement to aerial or satellite imagery.

2.4. Flickr

Flickr is an online application that allows uploading, store and organizing digital photographs[1]. Since its creation in 2004, Flickr has been recognized as one of the most innovative user-generated content sharing platform (Terdiman 2004) and a reference implementation of Web 2.0 principles (O'Reilly 2005).

Flickr offers numerous features that make it an interesting VGI platform. The first is the multiplicity of uploading options. Users can upload their photos online via the Flickr.com website, or via dedicated software distributed by Flickr and by third parties. Users can also send pictures by e-mail to a dedicated functional mailbox. It is even possible to upload pictures directly from camera-enabled mobile phones to a Flickr account. Such devices are becoming ever more accessible to the mass market and many of them also include built-in GPS sensors. It is therefore expected that Flickr will contain a growing number of geo-referenced content that will be available shortly after a photograph has been taken.

The possibility for users to geotag pictures is another important feature. Indeed, the wide majority of cameras currently in use do not include a GPS device that automatically inserts location in the image file metadata. Flickr users can thus manually add this information using an online map interface. Thanks to this feature, Flickr can be considered a major VGI repository. Recent research showed, for example, that the analysis of geo-tagged pictures taken in a given city clearly renders the location of its most famous landmarks (Crandall *et al.*, 2009). When this paper was originally published, the Flickr repository contained more than 3 billion pictures[2] of which 100 million were geotagged[3].

[1] http://www.flickr.com/about/

[2] http://blog.flickr.net/en/2008/11/03/3-billion/

[3] http://blog.flickr.net/en/2009/02/05/100000000-geotagged-photos-plus/

Flickr allows users to associate keywords – called 'tags' - to their pictures. This feature, which is supported by many Web 2.0 portals, is known as 'folksonomy' because the indexing of contents is the consequence of end-users' actions instead of a top-down classification process (Voss, 2007). Folksonomy can be criticized for the potential lack of coherence that it can generate. As users are totally free to associate keywords with a given resource, they can duplicate information (synonyms), or use the same word with different meanings (homonyms), produce alternative correct spellings (e.g. color and colour), use different languages, and make spelling errors. However, tags are valuable information to perform queries in the vast amount of images that can be found on Flickr.

It is also interesting to note that, in addition to the semantic structure provided by user-created tags, Flickr contents can be browsed like a Social Network. Indeed, Flickr offers users the possibility to create groups or join existing ones, and to create 'friend' relationships with other users. Flickr groups are pools of pictures users contribute to on a voluntary basis (i.e. there is no automatic retrieval). The focus of such groups can be very diverse: a photography technique (e.g. 'Night shots without flash'), a personality (e.g. 'David Bowie by You'), a centre of interest (e.g. 'Parks, arboretums and botanical gardens'), or an event (e.g. 'Obama inauguration 2009', 'Gloucestershire floods July 2007').

Thanks to its (Geo)RSS feeds and its Application Programming Interface[1](API), Flickr is not only an information silo: it can be accessed, viewed, updated, retrieved and analysed in many ways and for many purposes. This research takes advantage of the capacity offered through the Flickr API to submit complex queries, including spatial, temporal and semantic criteria, and to retrieve results as a structured dataset.

All those features make Flickr a genuine VGI platform, where users can upload and share geo-referenced content (for instance pictures with a title, description and tags) and where this content can be retrieved and analysed by specialists through the Flickr API.

[1] http://www.flickr.com/services/api/

The 'Volunteered' aspect should be highlighted: geotagging and tagging pictures requires active participation from the author. In addition, Flickr enables fine-grained privacy settings, allowing users to restrict access to a picture, or its location, and to prevent its picture retrieval through the API, if a user wants to make them visible on the Flickr website alone.

3. Case study: recent floods in UK

The case study presented in this chapter relates to the use of VGI from Flickr to map flood events that took place in the United Kingdom between the 1st of January 2007 and the 31st of March 2009. There were several reasons for this choice. Firstly, floods are events that have a precise location (compared to, for example, a seismic event) and that can be easily photographed (in contrast to fast-moving forest fire, for example). Secondly, the choice of the United Kingdom allowed a single language to query Flickr, while multilingualism would have increased complexity. The time period, finally, was chosen to cover a recent period that was known to include several flood events of varying extents.

In this case study, the actual pictures retrieved from Flickr are not analysed nor exploited to document flood events. The focus is only on the presence of pictures related to floods at a given time and place. This is used to derive the likelihood that a flood event took place in the past. It can be seen as a calibration exercise, as its purpose is to better understand how a VGI signal can be exploited to detect flood events using the proposed workflow. On this basis, further application can be developed in the future, including monitoring flood events in nearly real time, exploiting the actual picture contents to improve situational awareness during events, or improving damage assessment after them.

4. Description of the workflow

4.1. Principle and overview

The proposed workflow aims to convert raw VGI into a reliable information source through several configurable steps. The succession, definition and specific processing operations for each step constitute an original scientific contribution of this research, although existing algorithms can be adapted to the VGI Sensing context. The

author of this thesis coded specifically data processing scripts for each step.

In principle, each piece of VGI is considered as a part of a signal sent by citizen-sensors on a voluntary basis to a central repository, and this signal is processed to obtain reliable and useful information for a given purpose. In the present example, the purpose is to locate (in space and time) important flood events that took place in recent years in the United Kingdom. It is important to underline that the primary goal is thus not to make a collection of pictures related to flood, but to interpret their presence as an indication that a flood took place.

In practice, the workflow is based on the retrieval and processing of a subset of information contained in a wide scale user-generated data repository. Queries are sent to a web-based repository in order to extract relevant information. Results then go through several processing tasks that are designed to fit a particular need.

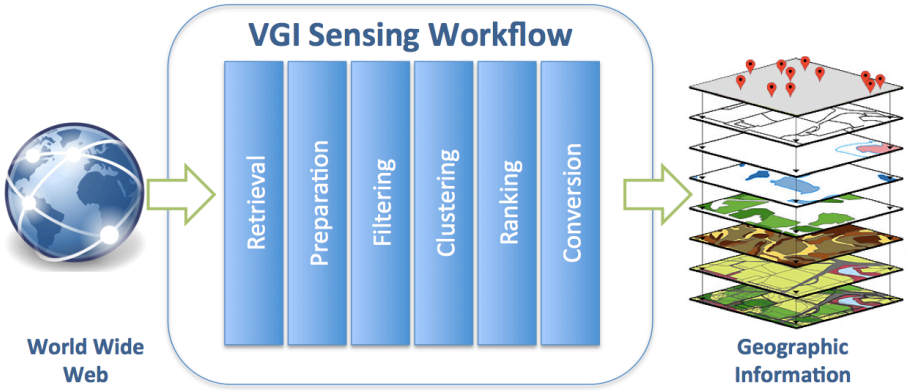


Figure 7 : Overview schema of the integration workflow

As shown in Figure 7, the proposed workflow includes 6 steps: retrieval, preparation, filtering, clustering, ranking, and conversion.

In the following sections, each step is described in detail.

4.2. Step #1: retrieval

The first phase aims at collecting information about pictures related to floods, *inter alia* to the time and place where they have been taken. This raw material will be used as input for the workflow that will derive reliable information about the presence of floods.

During the retrieval phase, queries are submitted through the Flickr API, and their results are saved locally for further processing. The Flickr API offers numerous options to submit queries using the *flickr.photos.search* method. Research parameters can include the date the picture has been taken, the date it has been uploaded, portions of text to be searched in its tile and description, the presence of one or several tags, the id of the group it belongs to, the id of the user that uploaded it, the place were it has been taken (bounding box or distance radius around a given location)..

The query used for this research used the following parameters:

- Tags including *floods OR flood OR flooding*.
- Date taken between the *1st of January 2007 and the 31st of March 2009*: a limited period of time has been chosen, that would cover several flood events.
- Geographic bounding box = *-11,50,2,60*(minimal longitude, minimal latitude, maximal longitude, maximal latitude): this bounding box also includes Ireland and a small portion of France; results had then to be manually refined in order to keep only pictures taken in the United Kingdom (thus excluding 33 images).

By using a geographic bounding box, all the pictures that have not been properly geo-tagged have been excluded *de facto*, even if they include geographic information in their metadata (e.g. a town name in the title or tags). This can look restrictive, as only 3 to 4% of Flickr pictures are geotagged (see section 2.4). However, it has been decided to focus on information provided on a voluntary basis, instead of generating it with more sophisticated procedures (e.g. by geocoding place names that appear in titles and tags). Following the same principle, the keywords search was restricted to user-created tags instead of using any text associated with the picture (title, description and tags), as it was assumed that the tags would contain less non-relevant use of the searched keywords.

This query returned a set of 1990 pictures located in the UK. A map showing their location can be seen in Figure 8.

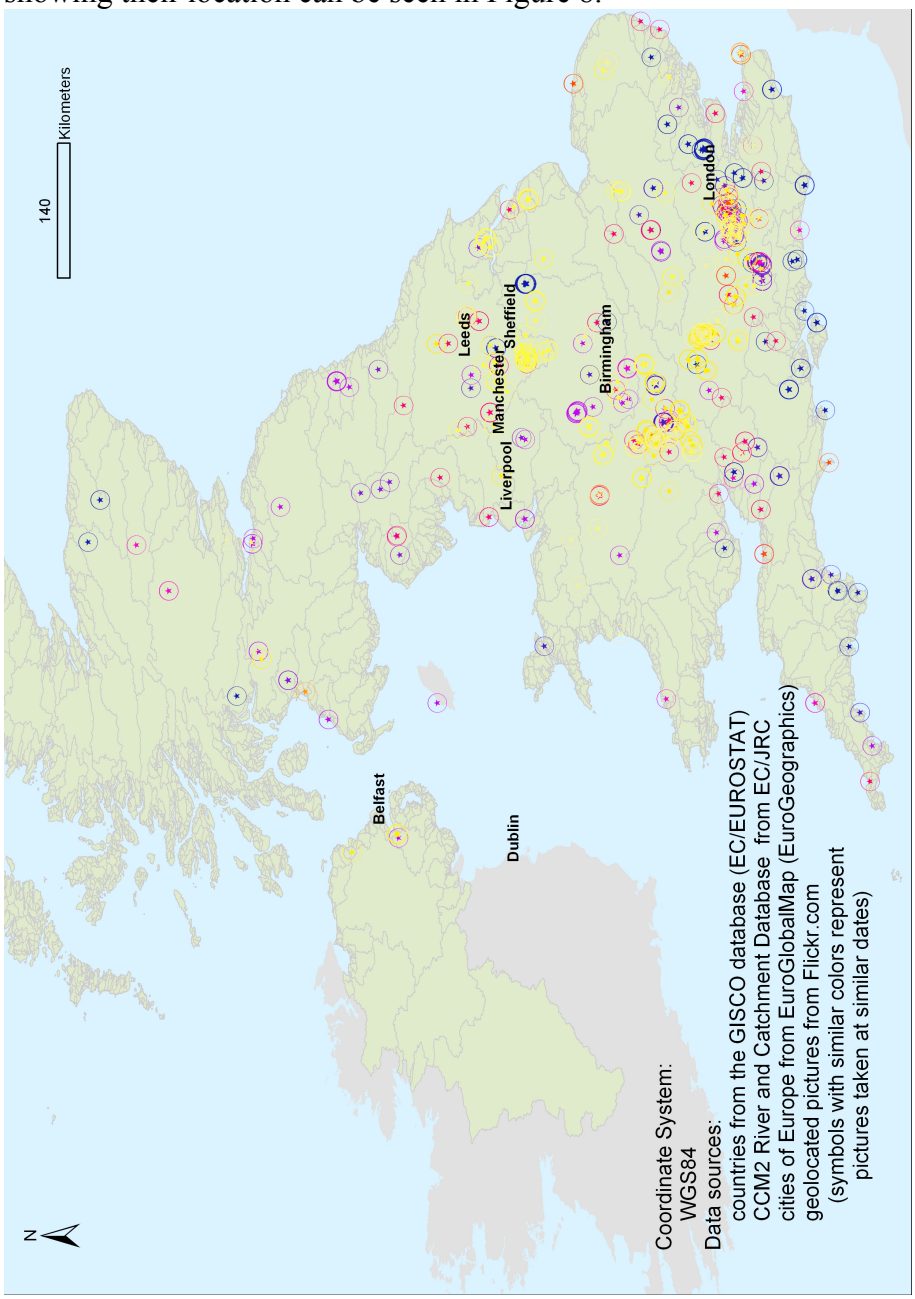


Figure 8 : Location and time of retrieved Flickr pictures

4.3. Step #2: Preparation

In this step, the results are converted from an XML format (.xml files) into a tabular format (.dbf file). Every XML response file from Flickr restricted to a maximum 250 results. It was, thus, necessary to retrieve every one of the 9 response pages (using an additional *page* parameter in the API query) and to concatenate them into a single table containing 2023 records (1990 in the United Kingdom and 33 in the Republic of Ireland).

The formatted table includes following fields:

- *ID*: the unique ID of the picture in the Flickr Database that can be used to visualize the picture itself or its complete set of Metadata;
- *Owner*: a unique identifier of the picture's author;
- *Title*: the user-created title for the picture;
- *Latitude*: the latitude at which the picture has been taken;
- *Longitude*: the longitude at which the picture has been taken;
- *WOEID*: the identifier of the named place where the picture has been taken. WOEID is an acronym for 'Where on Earth ? ID' and refers to the ID in the GeoPlanet Database[1], a Gazetteer system developed and maintained by Yahoo;
- *DateTaken*: the date and time at which the picture has been taken.

Optional processing can be applied in the preparation step in order to enrich the data, e.g. to look up certain attributes in gazetteers, or detect the language of the title through Natural Language Parsing (NLP) techniques.

4.4. Step #3: Filtering

The filtering is a formal step to check if the minimal information required to process the data is available in the proper format. It has a limited added value to the processed information, although it is a formal quality control step.

[1] http://developer.yahoo.com/geo/geoplanet/

Validation tests highlighted that 126 of the 1990 pictures had both latitude and longitude equal to 0. This was surprising, as all the results returned were expected to fit within a bounding box not including the location (0,0) and that a simple examination of these images confirmed that they were taken in the United Kingdom and not in the middle of the Gulf of Guinea.

No clear reason was found to explain why pictures that were supposedly correctly geo-tagged (as they were retrieved in response to a query that includes a geographic criteria) had no latitude and longitude information. Possible reasons are that a privacy setting prevented to retrieve the actual location even if the Flickr querying system was able to access the information, or that the querying system was able to 'guess' that the picture was taken inside the bounding box (using title, description and tags), even if no geographic coordinated were associated with it.

These 0 values have, thus, been treated as 'no data' values. As a consequence, those 126 images have been filtered out from the VGI dataset, which will finally contain1864 records.

4.5. Step #4: Clustering

In this next phase, heterogeneous information from various sources with uncertain quality is aggregated to obtain spatiotemporal clusters of VGI material. The assumption is that VGI elements created at the same place and time refer to the same *event* (in this case, the same flood). Starting from this step, the focus shifts from single photographs, to most-likely flood events represented by a pool of pictures. This has important consequences on the validation process: instead of wondering if a single picture is credible using only the few available metadata items about this piece of VGI, it is possible to apply quantitative methods to assess whether or not a VGI cluster corresponds to an event of interest.

Clustering data consists of generating sets of implicit classes that describe the data (Jain et al. 1999). In particular, Spatiotemporal clustering is widely studied in fields that rely on 'events' analysis, like epidemiology or crime analysis (Miller & Han 2001).

The clustering method used for this workflow included two steps. Firstly, purely temporal clusters are created, using a Natural Breaks

classification (Jenks & Coulson 1963). This classification scheme determines the optimal arrangement of values into classes by iteratively comparing sums of the squared difference between observed values within each class and class means. It identifies breaks in the ordered distribution of values that minimizes the within-class sum of squared differences. The advantage of this method is that it does not require *a priori* knowledge of the data distribution or, for instance, the expected duration of events. The disadvantage of the method is that the number of classes has to be decided in advance. We choose to create 12 temporal classes. The table 1 gives an overview of the reparation of Flickr Pictures in time classes.

Date Class	Begin Date	End Date	duration (# days)	VGI count (# pictures)
1	01/01/2007	11/02/2007	42	93
2	12/02/2007	02/05/2007	80	96
3	03/05/2007	07/07/2007	66	337
4	08/07/2007	18/08/2007	42	645
5	19/08/2007	11/10/2007	54	17
6	12/10/2007	12/12/2007	62	67
7	13/12/2007	29/02/2008	79	151
8	01/03/2008	10/06/2008	102	50
9	11/06/2008	13/09/2008	95	109
10	14/09/2008	14/11/2008	62	52
11	15/11/2008	11/01/2009	58	91
12	12/01/2009	31/03/2009	79	156
		TOTAL	821	1864

Table 1 : Date Classes resulting from Jenks Natural Break analysis of the VGI dataset

The second step of the clustering method consists in dividing the 12 temporal classes into sub-classes according to spatial criteria. The pan-European CCM2 River and Catchment Database (Vogt et al. 2007) has been used for this purpose, as river catchments have been considered as the most natural territorial subdivision to classify flood events.

A set of 156 clusters has been created, each one representing a pool of pictures taken in similar periods of time in a given river basin.

4.6. Step #5: Ranking

As explained in the previous section, aggregating VGI items into spatiotemporal clusters allowed the creation of most likely events, whose relevance can be quantified, instead of trying to assess the credibility of every single picture. Clustering and Ranking are the core steps of the workflow, which provide the most added value by converting isolated VGI elements with unknown credibility into most-likely events with measured reliability.

The Ranking step aims to quantify the relevance of each cluster, using automatic means. In other words, the ranking score reflects the likeliness that a flood took place in the time period and in the river basin each cluster refers to. The ranking value can be used to reduce noise (i.e. by eliminating clusters that are most likely to not correspond to flood event) by applying a threshold beyond which a cluster is ignored for further analysis.

The ranking score of each cluster has been calculated by retrieving all the tags associated with each picture it contains, and by summing the number of occurrences of words from a pre-defined list[1] related to floods.

This ranking method is based on two assumptions. Firstly, it has been considered that the relevance is related to the number of pictures taken (pictures in the dataset include at least the tag *floods* or *flood* or *flooding*, so each of them contributes to at least one ranking point to the cluster score). This choice has been motivated by the fact that the importance of a natural risk is usually measured by a combination of the extent of the risk (e.g. height of the flooding water) and the sensitivity of the exposed zone (e.g. population that lives nearby the flooding river) (ORCHESTRA 2008). In the present case, it was assumed that if there were more persons affected by the flood, there would be more pictures uploaded on Flickr. Our second assumption is that the use of several relevant keywords in the tags (e.g. *flood*, *river*, *deluge*, *torrential*) is a stronger signal than a single keyword (e.g. flood together with irrelevant tags like *sunset*, *light*, *bar*).

[1] The list of tags retained as contributing to the ranking of a cluster is: "bridge, brook, canal, damage, deluge, downpour, flood, floods, flooded, flooding, floodings, submerged, stream, rain, river, torrent, torrential, water"

No threshold has been applied to eliminate clusters for this study, as the intention was to keep the entire dataset for further analysis, and, therefore, have a better understanding of the criteria that could be used to refine the ranking process.

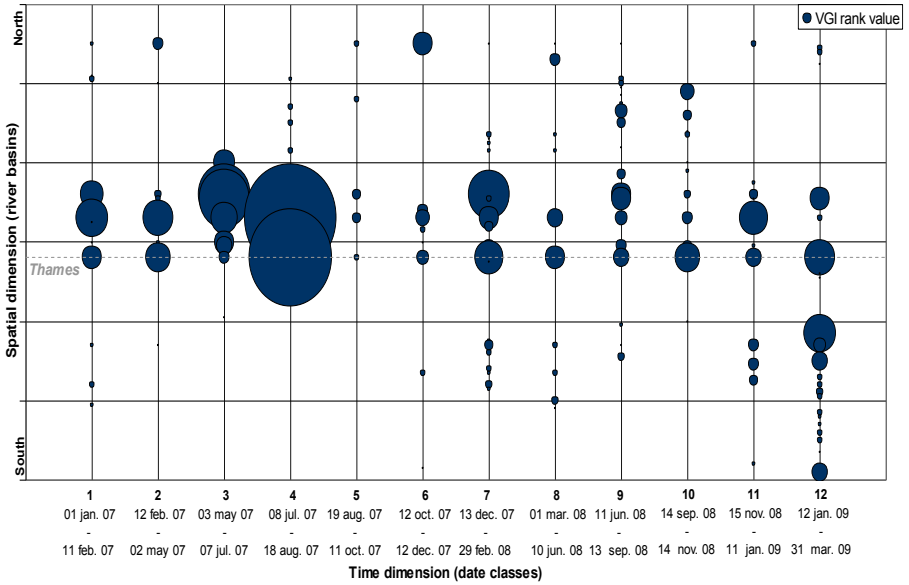


Figure 9 : ranking value per cluster

In Figure 9, the clusters are distributed according to their temporal (following to the date classification available in Table 1) and spatial distribution. River Basins are sorted along the Y-axis according to, but not proportionally, the latitude of their centroid. As a consequence, the clusters related to river basins located from the south to the centre of the United Kingdom (e.g., the Exe, the Thames, the Severn) are displayed lower on the graph, while river basins from the centre to the north (e.g. the Ouse, the Tweed, the Clyde) are displayed upper. The size of each cluster is proportional to its ranking score.

4.7. Step #6: Conversion

The conversion step is a formal process where the final dataset is converted into a geographic information layer. The final dataset contains 1864 point features classified into 156 ranked clusters. The attribute table contains the flickrID that allows each picture and associated metadata to be retrieved on Flickr, the date the picture has been taken, its geographic coordinates in WGS 84 latitude/longitude,

the ID of the river basin where it was taken, the ID of the cluster it is part of, and its contribution to the ranking value of the cluster.

5. Comparison of output with independent data sources and discussion

In this section, the output of the workflow described in previous sections is analysed in light of independent data sources. By doing so, elements of discussion are provided on the comparison of VGI with traditional sources of trusted information: media and experts. It has been chosen to perform several relatively simple analyses instead of developing a single in-depth analysis in order to provide an overview of strengths and weaknesses of VGI as a novel source of information based on multiple comparisons.

5.1. Analysis#1: the large flood events from the Dartmouth Floods Observatory

The Global Archive of Large Flood Events from the Dartmouth Floods Observatory has been used as a first comparison. This international reference laboratory for floods' monitoring compiles a list of major flood events every year, which is "*derived from a wide variety of news, governmental, instrumental, and remote sensing source*"[1]. This information can be retrieved online in XML format and can be considered as a typical example of trustable expert-driven data.

6 major events are reported for the period of interest in United Kingdom. Each event is reported with a start and end date, and additional information about its severity (e.g. number of victims, estimated cost of damage). The geographical information associated with each event is relatively imprecise, as it consists of a list of town names, counties and/or rivers affected by the flood event (see Table 2).

ID	Location	Began	Ended
3447	River Thames	07/02/2009	12/02/2009
3420	South-western U.K., Devon, Cornwall, Somerset, Dorset, Wiltshire, North Cotswolds	13/12/2008	14/12/2008

[1] http://www.dartmouth.edu/~floods/Archives/index.html

3260	England - Midlands and North England; *Rivers:* Severn, Avon, Leam, Frome, Stour, Ray, Aire, Dearne, Irwell, Rheidol, Calder, Alt, Dove, Ancholme, Ouse, Roch, Mersey. Thames and Severn canal	15/01/2008	26/01/2008
3137	*Counties*: Gloucestershire, Worcestershire, Oxfordshire, Berkshire, Bedfordshire, Herefordshire, Warwickshire, Lincolnshire. *Rivers:* Thames, Severn, Avon, Ock, Ouse, Evenlode, Windrush, Wye.	21/07/2007	30/07/2007
3110	Northern England; *Rivers:* Waring, Dearne, Don, Sheaf, Rother, Lud, Corve, Teme	25/06/2007	03/07/2007
3099	Britain - Yorkshire and Midlands areas; *Rivers:* Ouse, Tame, Dearne	15/06/2007	21/06/2007

Table 2 : Large flood events reported by the Dartmouth observatory for the study period and area

This information was used to associate every large flood event with one or several river basins from the CCM database, the reference dataset used for spatial analysis. The accurate time information has been used to associate every event with a temporal class that were created for the pictures' clustering. It is interesting to underline that none of those major events corresponded with more than one time class, which is evidence of the validity of the statistical approach (i.e. Jenks' Natural Breaks) used for the time-based classification of VGI data.

As a consequence, large flood events reported by the Dartmouth Observatory could be matched to one or several VGI clusters, and see how those important events were ranked. Figure 10 shows the ranking value of each VGI cluster, with an emphasis on clusters that correspond to a large flood event compiled by the Dartmouth Floods Observatory.

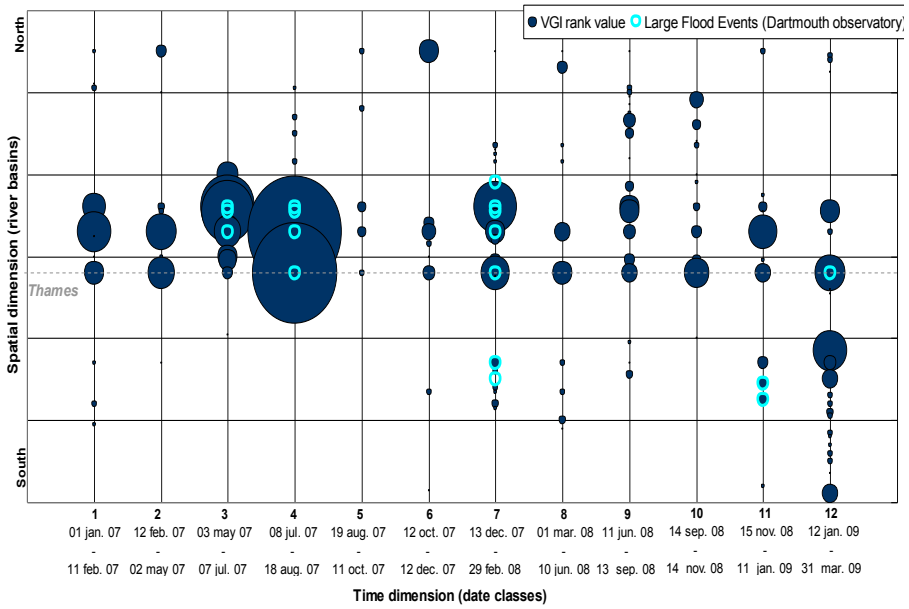


Figure 10 : Ranking values of each VGI clusters and clusters that correspond to a large flood event compiled by the Dartmouth Floods Observatory

Figure 10 appears to display good overall correspondence between VGI and the data provided by Dartmouth. In particular, the important flood events that occurred in Northern England (e.g. Sheffield) in the end of June – beginning of July 2007 (date class 3), and in the Thames (e.g. Oxfordshire) and the Severn (e.g. the Gloucestershire and the Worcestershire counties) river basins in the end of July 2007 (date class 4) are clearly visible in the VGI dataset, where the 5 best-ranked clusters could be found for these periods and locations. Similarly, the floods that took place in the second half of January 2008 (date class 7) in the Severn River Basin and in North England also correspond to clusters of relatively high importance.

However, there is a clear lack of VGI for large floods reported in the South West of the country in January 2008 (date cluster 7) and December 2008 (date cluster 11). This can be explained by the fact that it affected relatively sparsely populated river basins, so few citizens were present to report the floods.

On the other hand, numerous clusters that do not correspond to large flood events reported by the Dartmouth Floods Observatory have a good VGI ranking value. The following analysis demonstrates in

which proportion they correspond to less important flood events that are not reported by the Dartmouth Observatory.

5.2. Analysis #2: EMM temporal analysis

The European Media Monitor (EMM) developed by the European Commission's Joint Research Centre was used as source of media information to be compared with VGI. EMM harvests, on a daily basis, articles and news from thousands of online media sources (including news agencies, major national journals and television networks). Using semantic analysis of their contents, EMM then automatically associates all those news items with names of personalities, organizations, themes (e.g. floods, ecology, armed conflicts, immigration), and places. Depending on the place names present in the article, the geographic component of each article can be set as a town, county, region or at the country level. Information retrieved from EMM, thus, has heterogeneous spatial accuracy. 1637 press article related to floods where found in the EMM archives for the studied period in United Kingdom. Among those, only 360 articles were located precisely at the town or city level. The vast majority can be located a national (UK) or lower (England, Wales, Scotland, *etc*.) level. On the other hand, EMM data presents a fine temporal granularity, as every article clearly mentions its date of publication.

Therefore, as a first approach, it has been decided to compare VGI with media information on the temporal component only (see Figure 11).

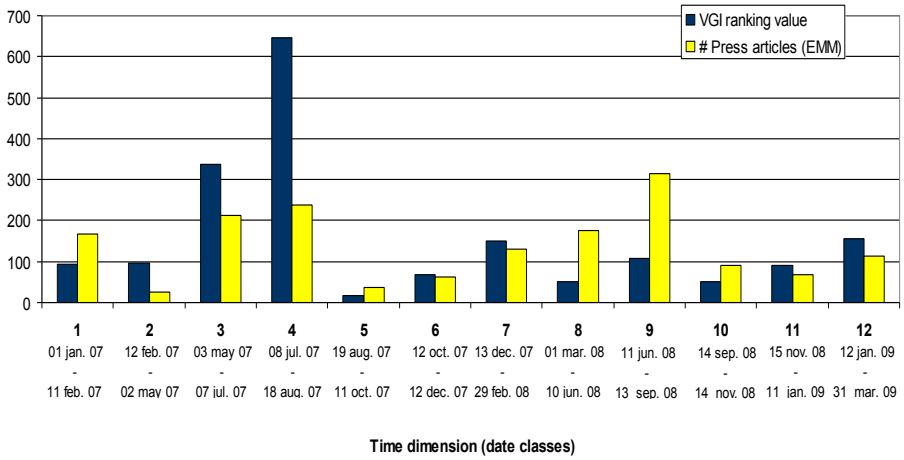


Figure 11 : VGI ranking value compared to the number of press articles about floods for each date class.

Figure 11 shows an overall correlation between VGI and news content. In other words, the periods when the media talks the most about floods in the UK correspond to the periods when Flickr users most intensively upload georeferenced pictures related to floods.

It can be noted that the events of June-July 2007 generated an increase of VGI, while press coverage about these floods was notably higher, but not in the same proportions.

Clusters 8 and 9 reveal an exceptional amount of EMM information compared to VGI A closer look at the news from this period (March, April and May 2008) suggested that there where no severe floods for these dates but important flood alerts from the authorities had been issued (a lot of articles had a title similar to "Flood warning in Britain", "'Stay-Inside warning in England and Wales" "Heavy storm threatens UK").

5.3. Analysis #3: EMM spatial-temporal analysis

This analysis is based on the 360 accurately geo-referenced press articles extracted from the EMM archive database for the study period in the UK. As this information was precisely located in space and time, it was possible to group them in clusters using the same delimiters as for the VGI dataset, and therefore perform a spatiotemporal analysis based on a cluster values comparison.

Figure 12 shows the ranking value of each VGI cluster and the number of press articles retrieved for the same time period and river basin.

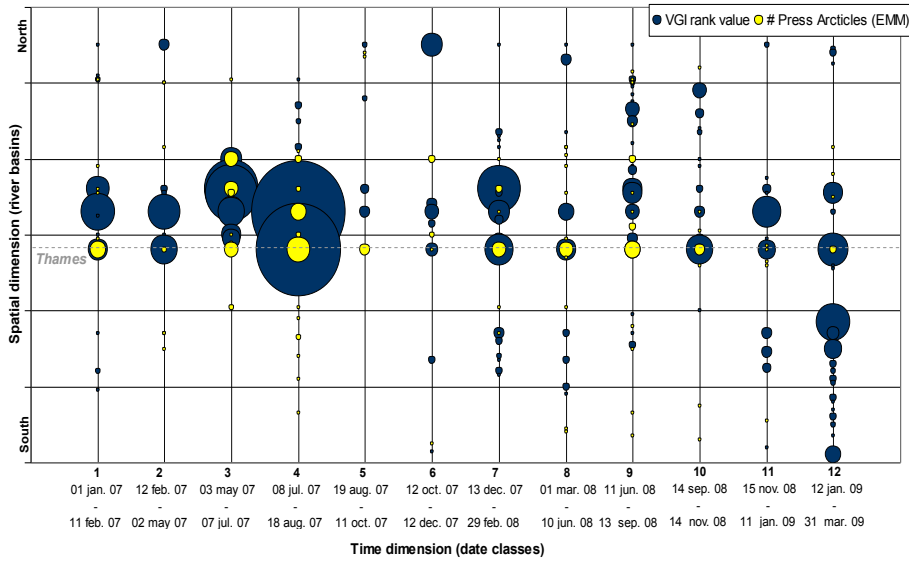


Figure 12 : Ranking values and number of press articles for each VGI cluster

The best-ranked VGI clusters appear to be those that correspond to the biggest number of press articles. However, EMM clearly tends to over-estimate the flood activity for the Thames river basin. The explanation can be found in the georeferencing methodology of EMM that is based on place names cited in the article, and by the fact that many events tend to be attached geographically to London, as numerous public and private decision centres are situated in the capital city (for example: "LONDON (Reuters) - Forecasters issued a severe weather warning on Monday after predicting more heavy downpours would hit Britain this week less than a month after the worst flooding for decades.", from Reuters news agency 13/08/07). It is interesting to underline that by their own nature, press articles are not always related to an accurate location on the earth surface, while the georeferenced pictures that were used as VGI do.

5.4. Analysis #4: GDACS on 2-3 selected sites

JRC's Global Flood Detection System (GFDS) monitors water surface changes using microwave remote sensing. The Advanced Microwave Scanning Radiometer (AMSR-E) on-board of the Aqua satellite platform scans the Earth's brightness temperature on a daily basis. The 36.5GHz band (vertical polarization) is particularly sensitive to surface water and less sensitive to atmospheric water. Brakenridge et al. (2007) developed a method to filter atmospheric and other signal noise, retaining indicative surface water data. De Groeve & Riva

(2009) modified the methodology to make it applicable globally. The GFDS observes in near real-time the Earth's surface water with a resolution of 10x10km^2. Time series were calculated back to 2002.

To test the VGI cluster data, three observation sites were established (see Figure 13) near Manchester (red), Sheffield (green) and Exeter (orange) respectively. Each area consists of 9 by 10 pixels, each 10x10km in size. For each pixel and each day, water surface anomalies are calculated by looking for extreme values in the time series. For each day, the number of pixels with a positive water surface anomaly were counted, resulting in a percentage of the area likely to experience floods. It is important to underline that due to the low spatial resolution, GFDS measurements are more sensitive to large-extent floods. In fact, the system was built to detect large floods needing humanitarian intervention. Although small floods are recorded, their computed magnitudes are much smaller. In the time series used in this chapter, pixels were counted as "in flood" when the value exceeded 4 standard deviations above the mean (i.e. a probability of 0.003%, under the hypothesis of a normal distribution).

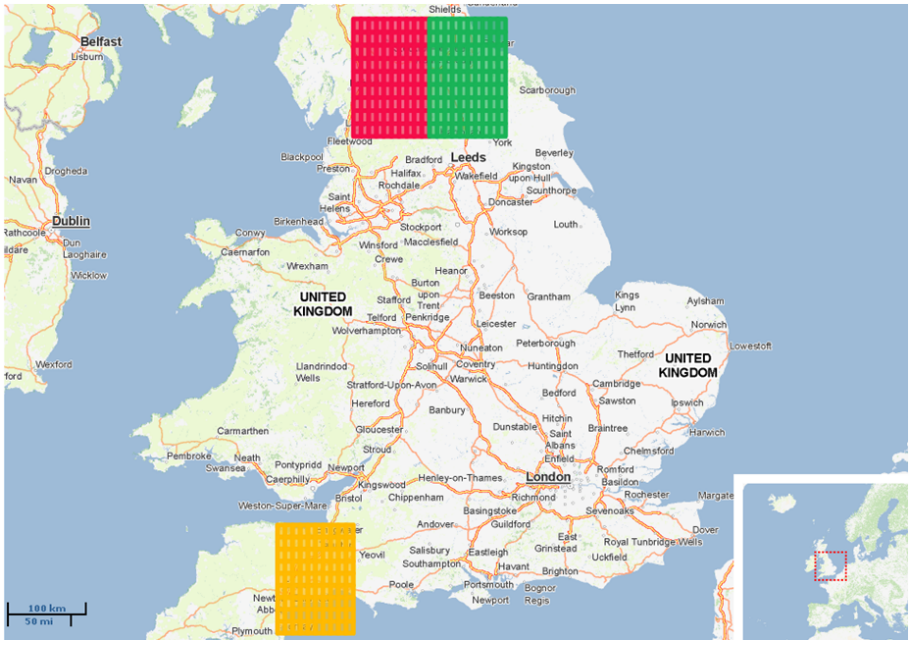


Figure 13 : Study areas used for Remote Sensing – VGI comparison

This analysis has been chosen to compare VGI with a purely sensor-based approach, which is based on objective measurements of

physical characteristics of the observed sites. It also allowed us to perform a finer analysis compared to those presented above, thanks to the spatial and temporal resolution of the remote sensing data. Therefore, VGI Day Clusters have been created for each study area, with values involving the sum of ranking scores of each individual picture for each day and study area.

When displaying the VGI Day Clusters on the same graphs as the GFDS surface water record (Figure 14), one can see reasonable correspondence. In the Manchester - Blackburn area, nearly all flood peaks are consistent with VGI Day Clusters. However, two large flood peaks (on 7 December 2008 and 12 February 2009) do not correspond with any VGI cluster. This can be explained when considering for example the flooded area for 12 February 2009, where one can see that it affects rural areas outside of Manchester (Figure 15). This seems confirm an assumption experts usually have concerning VGI: it requires a combination of events and the presence of people to witness them. A flood event that takes place in a sparsely populated area, or with limited impact, will most likely not be detected by VGI means.

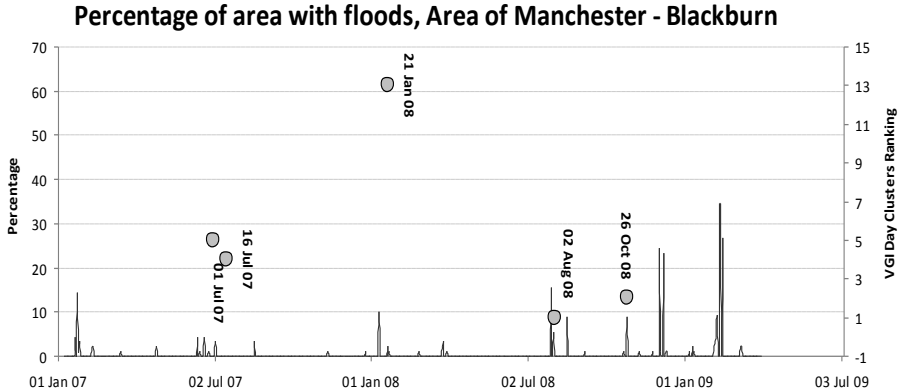


Figure 14 : Correspondence between VGI Day Clusters (circles) and GFDS flood signal (line) in the Manchester-Blackburn area

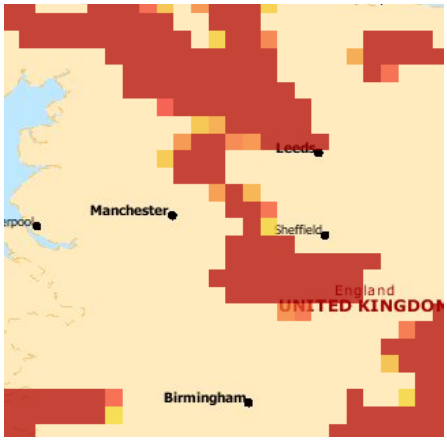


Figure 15 : Area "in flood" (red pixels) near Manchester on 12 February 2009

For the Sheffield - Leeds area, most VGI Day Clusters correspond to small or medium events, although some (11 October 2007, 6 September 2008) do not (Figure 16). For the two most significant flood peaks (January 2008 and February 2009) there are VGI clusters, although the latter has a low ranking. On the other side, VGI shows important flood activity in July 2007 while it can be seen as a relatively limited flood event while looking only at remote sensing data. This comparison shows that although both methods converge when study the presence/absence of a flood event, they can diverge when assessing the importance of such events.

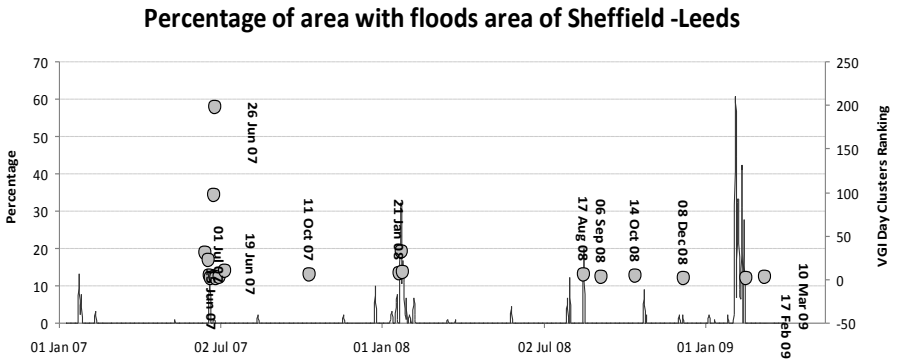


Figure 16 : Correspondence between VGI Day Clusters (circles) and GFDS flood signal (line) near Exeter

For the sparsely populated Exeter - Bridgewater area, the situation is less clear (Figure 17). As it could be expected, the VGI information is a lot less abundant and a maximum of 1 or 2 pictures taken in a single

day can be found. This does not allow robust statistical analysis to be performed. Care must be taken, though, with the GFDS data in this area, since the ocean / land boundary can introduce anomalies in the data.

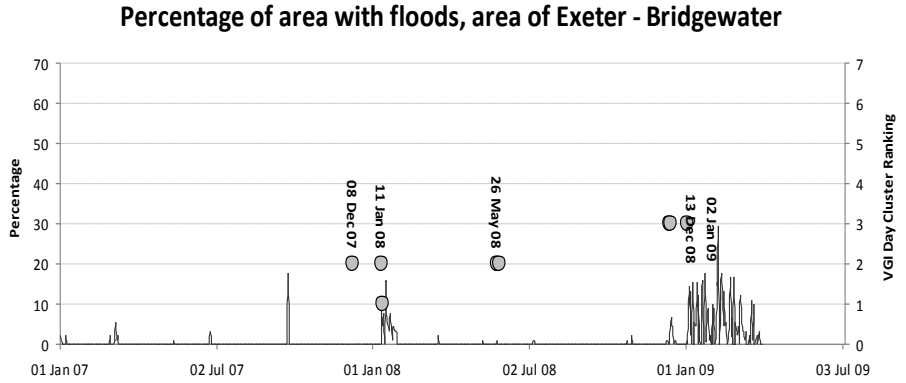


Figure 17 : Correspondence between VGI Day Clusters (circles) and GFDS flood signal (line) near Exeter

5.5. Analysis #5: Clusters' content analysis

This last analysis was performed to collect further information and no dataset was used for comparison. It consisted of visual interpretation of the pictures contained in 60 low-value ranked clusters, in order to better comprehend to what extent these clusters can be considered as information 'noise' or if in contrary they contain few but highly valuable information that merit more focus. Roughly half of these low-value ranked clusters contained pictures that were not relevant or not clearly relevant.

Among clusters that can be considered as relevant, two typical situations were frequently observed. In the first situation, the pictures represent small local floods in rural areas; citizens have photographed them most likely for esthetical reasons (e.g. partially submerged trees, reflection of sky in waters, old stone bridge in the middle of a 'lake') or because they faced were with an uncommon situation (e.g. a torrent that crosses the road, or the road becoming a torrent itself). In the second situation, pictures represented floods with limited extent but in populated area (e.g. a road in a village, the car park of a supermarket).

Among the not relevant clusters, the reasons for the tags 'flood', or 'flood' or 'flooding' even if the pictures do not represent a flood can have several explanations. Firstly, there can be a semantic confusion

in the use of terms like flood of floods. For example, pictures representing Toby Flood, the famous international rugby player were found in the VGI dataset (e.g. picture 2765965537). Two 'flood-lights' clusters were found; flood-lights are used in stadiums (e.g. picture 2765965537) or for stages (e.g. picture 2765965537). Secondly, the term flood can be used to describe domestic floods of very limited extent, caused by a broken pipe or a leaking roof. This has been observed in 2 clusters. In other cases, many of which were not relevant clusters corresponded to coastal areas, where the keyword 'flood' seems to describe natural pools on the beach after a tide, and that are photographed for esthetical reasons (e.g. reflection of the sky). In one situation, no clear explanation was found: 52 pictures taken by a user from February 2007 to November 2007 tagged with heterogeneous tags (e.g. flood, bar, sunset) with no clear connection to the picture. Finally, an interesting and frequent case reveals pictures that are related to floods events, but that do not show the flood itself. For example, pictures that show cleaning activities after a flood; or pictures of signs that show the maximal height of a major past flood; or old pictures of historical floods that have been scanned. 'Pictures of pictures' were even found, taken in an exhibition at the South London Art Gallery representing a flooded Mac Donald's in an unidentified place and time (e.g. picture 3309622848). This problem of indirect witnessing of flood events merits further work as roughly 10 clusters (one third of the not relevant clusters that were studied) depicted in one way or another past floods from 1912 to 2005.

6. Conclusions and future works

In this chapter, it has been demonstrated that VGI could be turned into a reliable source of information about natural hazards, if retrieved and processed with appropriate methods. Such methods have been described in detail, and a generic workflow that can increase the reliability related to VGI has been proposed. By applying retrieval, formatting, validation, clustering and ranking procedures, pictures uploaded by users on the photo-sharing website Flickr have been converted into a dataset locating floods in the United Kingdom between the 1st of January 2007 and the 31st of March 2009.

The resulting VGI-based layer was compared to datasets coming from independent sources, based on news headlines, experts' reports and

remote sensed data from satellites. These analyses allowed us to emphasize several strengths and weaknesses of VGI compared to more traditional information sources.

Firstly, it is important to highlight that VGI has been clearly successful at detecting major floods affecting large numbers of citizens. If this information can be uploaded and rapidly retrieved after or during an event, VGI can have an important role to play in innovative flood-related information systems, providing an interesting direction for future works, as the integration of VGI in real-time appears to have a wide field of application in several phases of disaster management activities.

Secondly, VGI has shown that its spatial and temporal accuracy makes it a valuable complement to traditional information sources. By their own nature, press articles do not always relate to a precise location, while VGI can, when properly georeferenced. As a consequence, trustable 'story-based' information from the media has a heterogeneous spatial resolution but can be complemented with GPS-accurate VGI. Similarly, satellite sensors offer homogenous and objective measurements of various phenomena, but they have a limited spatial (i.e. pixel size) and temporal (i.e. periodicity between two consecutive images) resolution. VGI can be a good complement to such information, as it can be taken at local level at any time. Complementarities between traditional information sources and VGI should be further studied, as it appears that the weaknesses of one can be compensated by the strengths of others. Furthermore, use cases can be envisaged where citizens are encouraged to collect specific information - like welfare checks or building damage assessment –in order to enhance complementarity with other sources, and/or use specific keywords or hash tags to ease VGI retrieval. Such concept of 'tasking VGI sensors' will be further discussed in chapter 4.

This research has, however, highlighted several weaknesses of VGI that partially confirms assumptions about the lack of trust VGI can generate. In particular, analysis has shown that the retrieval of irrelevant VGI can have several causes. Therefore, the design and implementation of generic noise-reduction filters for VGI could be a complex task. For that purpose, the conclusion is that future research that aims to improve the efficiency of the ranking system could be of particular value.

Another weakness of VGI that has to be emphasized is the difficulty to convert it into quantitative measurements. Indeed, we demonstrated that VGI was in most cases successful in detecting the presence of flood events. But how could the VGI 'score' of each event reflect their importance, in term of damages, impacted population or geographic extent, for example ? From this research, we cannot exclude the possibility that there could be important distortions between the perception of an event witnessed by VGI and its objective measurable attributes, and between such perceptions and the information actually posted online. One obvious example of such bias has been uncovered in this chapter: VGI signal is weaker where population is scarcer. Although the intensity of the signal could be normalised using parameters based on population density or socio-cultural factors (in order to take into account possible effects of 'digital divide'). Nevertheless, it should be considered that VGI Sensing cannot intrinsically provide a comprehensive image of all phenomenon of interest; cases will inevitably occur where relevant events are missed by lack of citizens willing – or being able – to report it online.

In any case, future work based on robust statistical analysis should further investigate how quantitative measurements of the phenomenon of interest could be derived from VGI. The research presented in the next chapter aims at contributing to the development of such methods.

Chapter 3 – Filtering and Clustering Volunteered Geographic Information[1]

Abstract

This research aims at retrieving useful spatiotemporal information from the vast amounts of online data posted on a voluntary basis by citizens. It uses the photo-sharing website Flickr as a data source and the characterisation of Forest Fire events in North America as a Use Case. Following a state-of-the-art VGI Sensing Workflow, Web Mining techniques for collection, formatting and enrichment of semantic, social and spatiotemporal data about Forest Fires are described and discussed. More specifically, a benchmark of existing clustering algorithms is presented, in order to identify optimised techniques for Web-based Events characterisation. Results suggest that density based spatial clustering is fitter-for-purpose than the state-of-the-art SatScan Space-Time Permutation algorithm for Volunteered Geographic Information (VGI). Moreover, it is highlighted that the semantic dimension of such data can significantly contribute to improved results, compared to strictly spatiotemporal clustering algorithms. Although this research is based on retrospective data analysis, the near-real time application potential is discussed in the final sections.

1. Introduction

This chapter focuses on the Filtering and Clustering steps of the VGI Sensing Workflow described in the previous chapter. The Filtering step is essential to reduce the noise in the VGI signal, while the Clustering step is key in the identification and characterisation of Events.

[1] This study is currently under review for the International Journal of Geographic Science under the (provisional) title *"What, When, Where » a clustering algorithm for event characterisation with Volunteered Geographic Information"* The text presented here is slightly modified from the original publication for layout and terminology harmonisation.

The filtering step will implement simple machine learning methods to exclude irrelevant VGI. VGI clusters will then be created, and considered as possible Forest Fire events; it will be then possible to assess the performance of each clustering algorithm by comparing its output (VGI 'Forest Fire' clusters) with the location of real-life Forest Fire Events (as documented by relevant Public Services). The overall objective being to detect and characterise events from heterogeneous data through careful content aggregation, the measure of performance of algorithms will be based on the idea that VGI clusters should ideally correspond to genuine Forest Fires on the field, and vice versa.

More specifically, this chapter will explore how the three main dimensions of VGI (namely: space, time, and semantics) can be best co-exploited to contribute to situational awareness in Crisis Management. In other words, how can the 'Where', the 'When', and the 'What' of information[1] shared on the Web can contribute, through a fine tuned algorithm, to automated Event characterisation.

To this end, a benchmark of clustering algorithms is proposed, based on a VGI dataset about Forest Fires extracted from Flickr and covering continental USA and Canada on the summer 2009. It compares the results obtained using the state-of-the-art SatScan space-time permutations algorithm with 23 different combinations of the spatial, temporal and semantic dimensions for the DB scan algorithm.

VGI Clustering

Data clustering can be defined as the unsupervised classification of patterns into groups - called clusters (Jain et al. 1999). In spatio-temporal clustering, the position of the features in space and time (i.e.: latitude, longitude, date, time) are used as the key dimensions (Gong et al. 2006). On the basis of the similarity measurement between spatio-temporal features, various clustering algorithms (hierarchical, partitional, density-based, *etc*.) can be applied, depending on the nature of the events that are investigated (Getis & Ord 1992). A wide

[1] It should be noted that the Social dimension has been intentionally left aside. Indeed, in the advent of Crisis Events, social analysis notions like power of influence and relationships graphs can be poorly relevant, since legitimate VGI sources are people being impacted, whatever their previously recognised expertise and their friendship habits.

variety of spatio-temporal clustering techniques and algorithms have been applied to detect events in fields like epidemics (Rogerson 2001), crime analysis (Johnson 2010), and meteorology (Hsu & Li 2010).

But whereas spatio-temporal clustering techniques are usually designed to deal with discrete, comparable objects such as sensor observations or tabular data records (Miller & Han 2001), VGI can be heterogeneous in terms of quality and accuracy (Metzger 2007). In particular, De Longueville & Hardy (2010) emphasized that VGI often have place names as spatial reference (e.g., town, region, country, etc.), resulting in different levels of spatial accuracy when looked-up in a gazetteer. Oppositely, the temporal reference of VGI is usually accurate because of the creation of a time stamp when VGI is posted online.

In consequence, current spatiotemporal clustering techniques might benefit to be better adapted to data with heterogeneous spatial reference such as VGI. In addition, the spatial and temporal dimension of VGI can benefit to be combined with its Semantic and Social dimension. The aim of this research is to contribute to the development of clustering methods that are suitable to extract event-related knowledge from VGI.

Multidimensional clustering of VGI is a relatively recent scientific endeavour (see, e.g. Kisilevich et al. 2013). Nevertheless Cheng & Wicks (2014), as well as Craglia, Ostermann & Spinsanti (2012) and Zhao et al. (2014) clearly established the value of Scan Statistics (Kulldorff 1997) algorithms on that purpose, more specifically SatScan Space-Time Permutations (SSTP), which presents the advantage to automatically adjust to temporal trends (Sikder & Woodside 2007). This chapter proposes a benchmark of this state-of-the-art algorithm with a challenger inspired by Kisilevich et al. (2010) which will be described later in this section.

Event detection* vs. *Event characterisation

Numerous examples of VGI usage in crisis situations have been provided in previous chapters; additional examples and further analysis can be found in a recent survey by Imran et al. (2014). Such examples confirms in practice the conceptual issues discussed earlier: the necessity to apply a quality filter to VGI, the opportunity to adopt

web mining techniques to perform such filtering in a timely manner, and the relevance of devising a typology of VGI Sensing use cases (depending on disaster stage, VGI sources, actors involved, disaster type, etc.) – each posing specific research challenges.

For example, Atefeh & Khreich (2013) proposed to clearly distinguish use cases involving Events Discovery from those involving Events Characterisation (i.e. retrospective analysis, even if in nearly-real time). In this research, such distinction applies, since we did not invest in priority in the development of real-time processing capability, as opposed to abundant research on e.g., Twitter Streams. As a consequence, the operational relevance of this research has to be situated between *Stage 4 (Inventory)* and *Stage 7 (Recovery)* of the disaster cycle (see p. 13 for a description of each stage), although section 4.4 will discuss the question of computing efficiency of the proposed methods, in a view of adapting them to real-time VGI stream analysis applications.

SatScan Space-Time Permutation (SSTP) and DB Scan algorithms
Section 3 will provide all necessary details and explanations on why and how SSTP and DB Scan (see below) have been used in the benchmark exercise that this chapter is based on. Following paragraphs describe the key features of both.

The SatScan algorithms are based on a cylindrical windows of varying radius (space) and height (time), which move across the study area. This process is repeated until all possible space-time locations have been visited (Block 2007). Each window is viewed as a potential cluster; with the number of incidences (data points) within each being compared to the number of expected incidences for that window, according to a pre-defined theoretical points distribution. On this basis, a P-value is calculated for each candidate cluster. This P-value describes the likelihood that the high number of observation within a window is an abnormal aggregate (e.g. a disease outbreak) and not the pure effect of chance.

SSTP has been specifically designed for detection of disease outbreak – for which it presents the advantage to not require any population-at-risk data to estimate the expected distribution of disease occurrences (Kulldorff et al. 2005). It has nevertheless be demonstrated that SSTP is suitable for other applications including VGI clustering (Craglia et

al. 2012). The method consists in iterating over a finite number of geographical grid points while gradually increasing the circle radius from zero to some maximum value. The height of the cylinder (representing the number of days) can also be set as variable with a maximum value., For each window position and size, the number of observed cases (points) is compared to what would have been expected if the spatial and temporal locations of all cases were independent of each other and randomly distributed. To this end, an arbitrary number of replications of the dataset with random spatio-temporal locations are created, called Monte-Carlo replications. The higher the number of replication, the more statistical significance will take the outcome, but the computing cost will increase in proportion; it is usually admitted that a number of 999 Monte-Carlo replications is suitable while running SSTP (Kulldorff 2006). As a result, a P-value (i.e. a probability, expressed as a decimal number between 0 and 1) of being a cluster is calculated for each Window position and size; an arbitrary threshold is then applied to P-Values in order to keep only higher probability clusters.

To summarise, the clusters identified with SSPS are thus groups points contained in spatio-temporal windows of various size and positions, for which the probability that they are more numerous than if it was by pure chance is higher than a given value.

DB Scan (Density-Based Spatial Clustering of Applications with Noise, Ester et al. 1996) has a slightly different approach, but it also performs iterations through a dataset in order to test if some conditions of points density are met in order to decide on the presence of a cluster. As opposed to the moving window system of SatScan, DB Scan simply iterates through each point of the dataset. For each point in the dataset, it counts the number of points that are within a distance of 'Epsilon' (an arbitrary combined distance threshold set as a key parameter of the algorithm). The value of Epsilon being be a combination of distances in the various dimensions, we can intuitively compare the cylindrical windows of SatScan with the spherical neighbourhood of points in DB Scan. The point and his neighbours will then be considered as part of a cluster only if their number exceeds 'MinPts' (another arbitrary threshold which the second and last parameter DB scan require to be set).

So, if a point has more than 'MinPts' points within a neighbourhood distance of 'Epsilon', it is considered as a central point for a cluster. Each point, which is within the 'Epsilon' distance of a neighbour of such central point, is then also considered as part of the cluster. The iteration continues until no point can be added to the cluster (i.e. when there are no more point that is not yet in the cluster at a distance smaller than 'Epsilon' of any point – central or not - of the cluster).

In addition to the careful selection of threshold values ('Epsilon' and 'MinPts'), which should be based on knowledge of the phenomenon of interest, it important when using DB Scan for multidimensional clustering, to carefully design the point distance function. The aim of such function is to compute multidimensional point-to-point distance value. In a purely spatial clustering application, this is intuitively easy, since the distance function will simply aggregate metres (or millimetres, or kilometres, …) and the Euclidean Geometry provides an appropriate distance function (which equals the square root of the sum of the squares of distances in x, y and z). Spatio-temporal clustering, however, requires a distance function combining space and time in a single distance metric, thus de facto aggregating days (or seconds, or years, …) with meters. Section 3.2 will describe in details how this issue has been addressed in this research by normalising values, and testing various weightings of each dimension in the benchmark exercise.

2. Material: VGI preparation

2.1. Collection and enrichment of a VGI dataset

The VGI dataset used for this research has been harvested from the web through the Flickr photo sharing website. In total, 12 911 pictures were retrieved, which were taken between the 1st of June 01 2009 and the 1st October 2009 and their title, description or tags contained the words 'forest' and 'fire' – or translations and their synonyms in French, Italian, Spanish, Portuguese, Greek and Catalan. Flickr has been chosen at that time because it was a prominent VGI platform with important amounts of geolocated contents compared to other photo-sharing platforms. Languages aimed to cover most parts of North America as well as the Mediterranean region in Europe. It is considered that, although the sample was collected several years before this research is published, its conclusions are still timely, since

factors influencing its spatio-temporal distribution and its semantic contents are more cultural than technological (and therefore evolve at a slower pace). The main changes that can affect VGI Sensing since then would be the increased volume of available VGI (because of the larger penetration rate of web 2.0 services) and its average spatial accuracy (because of much larger market share for GPS-enabled smartphones).

The next step was to determine the language in which each picture's metadata (title, description, tags) are expressed. The Google Language API (replaced meanwhile by the Google Translation API) was used on that purpose, although the Flickr API returns also a 'language' parameter. However, it is based on user settings, which are most often set to the default value (English). Therefore, determining the actual language on a picture-by-picture basis was judged more reliable. English was by far the most used language: 89,4% of the retrieved pictures had their metadata in English. The other significant languages used in the VGI dataset were unsurprisingly the ones used in the retrieval query: Greek (3,2%), Spanish (2,3%), French (1,7%), Italian (1%) and Portuguese (0,8%). Only Catalan is present in very small proportions (0,1%).

A set of 1742 pictures had a geographic reference expressed in latitude and longitude coordinates in Flickr (13,5% of the total), while 7 718 (59,8% of the total) had one or several place names in their metadata that could be looked up in a gazetteer (for instance, Yahoo Placemaker, replaced meanwhile by the Yahoo Boss Geo Services). For these pictures, an estimated location (latitude, longitude and granularity measure) could thus be inferred from textual metadata. Yahoo Placemaker was chosen because it was at that time the only natural language geocoding service claiming to use grammatical analysis to improve quality of place names extraction (i.e. "I am travelling to Bath" would return a place name – a city in UK- , while "I am taking a bath" would not.) Finally, 3451 pictures (26,7%) did not have any spatial reference and were then discarded from the VGI dataset. The remaining 9460 pictures had heterogeneous geolocation precision. Table 3 gives an overview of the granularity (precision level) of geolocation in the VGI dataset. It is important to note that this table gives a view on precision, not on correctness of the geolocation. Yahoo Placemaker computes this precision level by

matching each type of geographic feature (e.g. town, city, country, etc.) to a precision level on an arbitrary basis.

A set of 1559 pictures was manually analysed by an operator, who concluded in 87% of the cases that the places (there can be more than one) associated by Yahoo Placemaker to a picture were correct. Such assessment was nevertheless subjective and non-comprehensive by nature, i.e. it means a human having the same information as the algorithm reaches the same conclusion, but it did not mean the information was correct, or that they both interpreted it correctly.

Precision	**#**	**%**
Latitude and longitude coordinates from Flickr	1742	18,4
Very precise place name (Point of Interest, estate, suburb, …)	1421	15,0
Precise place name (town, district)	4613	48,8
Relatively imprecise place name (county, province, region)	628	6,6
Imprecise place name (state, island, country)	1050	11,1
Very imprecise place name (continent)	6	0,1
TOTAL	**9460**	**100**

Table 3 : Percentage of pictures by georeferencing precision level.

The last enrichment operation was performed to prepare the data for the calculation of semantic distances. To this end, the textual metadata of each picture (namely, their title, description, tags, name of the sets they belong to and name of the pools they belong to) were tokenised using the Toolkit of the Stanford Natural Language Processing Group (Klein & Manning 2002). Such freely available toolkit is widely used in the NLP research community and is well documented. It was considered as suitable to execute the simple and generic text analysis that we required to prepare the VGI dataset. Tokenisation consisted in dividing the textual metadata into single words – or pairs of words, e.g. Los Angeles – corresponding to identifiable lexical elements (nouns, proper nouns, verbs, adjectives, etc.). This allowed to create 278 247 picture-token pairs, to discard irrelevant token types (f.i. punctuation or determiner) and to identify Named Entities among these. For example, 13 145 of these picture-token pairs were referring to locations, and 9 283 to (parts of) dates according to the Stanford NLP Toolkit. Tokenisation and named entity recognition allowed to transform unstructured text from users into a comprehensive set of

semantic data associated to VGI items. This semantic data was used to build VGI item features for filtering as described in the next section. Furthermore, the measurement of semantic distance between 2 pictures was based on the comparison of tokens associated to them, as explained in section (3.3)

It has to be noted that the tokenisation was performed only on the pictures identified as having metadata in English (89,4% of the total). As explained in the next section, this was not an issue as the non-English pictures have been filtered out.

As a result, the VGI dataset consisted in a series of records with a geographic (i.e. point features resulting from geographic coordinates, or centroid of a polygon feature as explained above), temporal (i.e. timestamp) and semantic (i.e. collection of tokens from which a semantic distance between VGI items can be calculated, see section **Erreur ! Source du renvoi introuvable.**3.3).

2.2. Filtering the VGI dataset

As stated earlier, the purpose of the filtering step is to remove poor quality and irrelevant VGI items. In this research, filtering also allowed concentrating on a delimited geographic area for which homogenous and reliable ground truth information was available. Indeed, the Northern America (continental US and Canada) has been considered as a suitable study area because such conditions were met.

2.2.1. Basic filtering

The Table 4 provides details about the basic filtering operations, which aimed at 'scoping' the VGI dataset by keeping only exploitable items that are located in the area of interest. As it can be seen, about the half of VGI items initially collected were kept after basic filtering.

Filter	Number of items	% of total
(no filter)	12911	100
could be geolocated	9460	73,3
geolocation accuracy at Town or District level at least	7776	60,2
metadata in English	7400	57,3
located in Northern America (continental US or Canada)	6377	49,4

Table 4 : Basic filtering operations and their result.

2.2.2. Noise reduction filtering

Since VGI is by definition an uncontrolled data source, the presence of poor or irrelevant items into the dataset is an issue that needs specific attention.

Although the optimisation of filtering methods is not the main objective of this chapter (which is the optimisation of clustering methods), significant effort has been dedicated to the suppression of poor quality and irrelevant VGI items in the dataset.

The binary classifier Rank – which basically consists in a cross validated regression on a linearized space (Van de Merckt & Chevalier 2008) - was used to filter noise (i.e. to remove irrelevant VGI items from the scope). Rank is commercial software developed by a team which one of the authors of this research was part of, so the software and know-how were readily available. This presented in practice a major advantage on possible alternatives (such as e.g. using machine learning R packages).

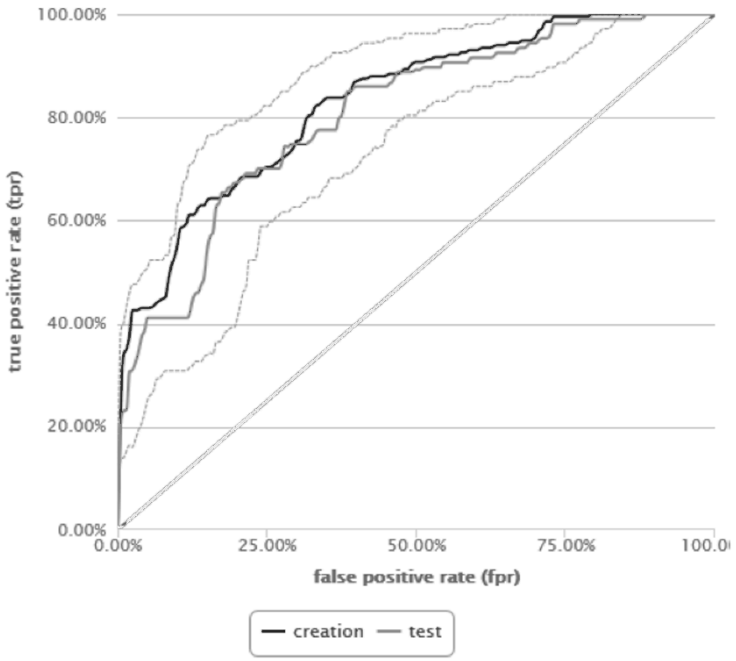
A model to learn a manual classification of 815 pictures (chosen randomly) among which 323 were identified as not related to Forest Fires (Target) was trained. 180 features were created in a semi-automated manner (i.e. by specifying rules for feature creation to the Rank software) by analysing the Title, Description and Tags associated to each picture in order to give quantified indices of potential relevance, such as:

- the number of occurrences of keywords of interest in the picture's metadata (e.g. 'firemen', 'smoke', 'ashes', 'burning', etc.);
- the number of occurrences of certain types of named entities (e.g. person's names, place names, etc.);
- the size of the provided metadata (e.g. number of tags, length of description, etc.).

The software then selected the best combination of these features to estimate a probability a given picture is related to an actual fire.

Figure 18 shows the model performances, expressed by a typical Receiver Operating Characteristic Curve – or ROC (Bradley 1997) where the trade-off between True Positives (Sensitivity) and False

Positives (1-Specificity) can be visualised. The model performance is evaluated on its training set and on a test set, which has been left apart for both features selection and weight optimization. A confidence interval of the test performance is constructed using percentile 1 and 99 of 2000 bootstrap samples.



AUC	
Creation	83.05%
Test	80.54% ±2.59%

Figure 18 : ROC curve of the classification model, which aims at identifying irrelevant pictures in its learning set (815 pictures).

The model required 8 of these 180 features. Some other features were informative but were excluded because of their risk of overfitting, i.e. they were too strongly related to the studied period (e.g. the presence of word 'California'), and therefore were considered as a potential obstacle to the application of the model in another context (e.g. not in the US). Table 5 shows the features that have been retained for the final model, and their respective normalised weight as calculated by the RANK model (Van de Merckt & Chevalier 2008). Note that, due to the feature recoding method (linearization of the space) the weight

is always positive no matter the relationship between the feature and the target. It worth also noting that this research did not include an analysis of the Flickr picture itself, but solely of its metadata. Optical characteristics of the pictures (which could be correlated with the presence of e.g. flames, smoke, tree, *etc*. on the image) have not been considered (by lack of specific skills in computer vision, and to keep the scope of the research), although it may offer interesting improvement in future works.

Variable	**Normalized weight**
Presence of word "camping"	1,00
Absence of word "wild fire"	0,84
Number of proper nouns (e.g. "Thomas Jefferson")	0,50
Absence of word "fire"	0,45
Presence of named entity expressing a duration (e.g. "5 years ago")	0,43
Number of places identified by Yahoo place maker for the picture	0,37
Length of textual metadata (Title, description, tags,…)	0,29
Number of adjectives (e.g. "nice", "easy")	0,03

Table 5 : Features selected by the model with their weight in the regression formula.

Applying the model to the scope, an estimation of the probability for each picture to be irrelevant can be inferred (i.e. for the picture to be part of the noise) and pictures having this probability higher than 70% can be filtered. This threshold has been chosen empirically since it allowed eliminating a significant proportion (about 10%) of poorly relevant pictures, while a lower threshold would have significantly increased the risk of excluding relevant pictures (e.g. 60% threshold excludes 25% of the pictures). Accordingly, 5,677 pictures were retained.

3. Methods: Clustering algorithms benchmark

A practical approach known as benchmarking has been used, which consists in comparing the result of various means (for instance algorithms and parameter sets) to reach a pre-defined objective (for instance detecting and characterising Forest Fires).

As explained earlier, each VGI Cluster is considered as a potential Forest Fire event characterisation. Performance of clustering algorithm can thus be measured by comparing the location of VGI clusters to the location of real-life Forest Fires. The next section describes in detail how such measurements were performed in this research.

Section 3.1 explains why certain algorithms have been selected, and section 3.2 provides then explanations on the choice of the parameters that were used. Section 3.3 then describes how the respective performance of algorithms will be measured.

3.1. Choice of benchmark algorithms

SatScan Space-Time Permutation (SSTP) has been identified by previous research as the most suitable spatio-temporal clustering algorithm for VGI (Craglia et al. 2012; Cheng & Wicks 2014). It is thus a natural choice as reference for our benchmark exercise.

DB Scan has been chosen as the challenger algorithm for various reasons:

(1) it is specifically designed to deal with noisy data - while VGI is by its own nature noisy (Yang et al. 2014);
(2) it does not require *a priori* knowledge for the estimation of the number of clusters, unlike e.g. kmeans-based algorithms (Tuia et al. 2009) – while the number of Forest Fires during a given period cannot be predicted;
(3) it does not require *a priori* knowledge of the events distribution unlike e.g. Scan Statistics (Kisilevich et al. 2010) – while distribution of Forest Fires in space and time depends on many concurrent factors and is difficult to model;
(4) it combines all considered dimensions into a single metric ('Epsilon') – which provides a practical way to test and compare the possible combinations of Spatial, Temporal, and Semantic dimensions in the clustering results.

Since it requires 1 to 1 distance calculation for all the features in the sample, DB Scan is computing intensive, which is its major disadvantage. This issue can be overcome, however, as discussed in section 4.4.

Further alternatives could have been considered, however, no algorithm benefiting a readily available implementation in R has been identified as matching the level of suitability to VGI Sensing that both SatScan and DB Scan present intuitively.

3.2. Choice parameter sets

As stated in the introduction, SatScan is used as the benchmark, since it is the state-of-the-art algorithm for spatiotemporal clustering VGI (Cheng & Wicks 2014). The parameters set used in this research is the one defined by Craglia et al. (2012) who demonstrated the value of SatScan Space-Time Permutation algorithm for VGI clustering, for instance:

- time precision and aggregation : per day;
- maximum spatial widow radius : 50 km;
- maximum temporal window size : 10% of sample duration (so, 9 days since the VGI was collected over a 90 days period);
- number of Monte-Carlo replications : 999;

DB Scan was chosen as challenger algorithm for the reasons exposed in section 3.1. It requires a 1 to 1 distance calculation, which in this case combines information about the 'Where' (spatial dimension), the 'When' (temporal dimension), and the 'What' (semantic dimension)[1].

Intuitively, a distance function consisting in a weighted sum of normalised distances in the various dimensions has been judged as suitable, since it allowed parametrisable aggregation of various values

[1] The opportunity of exploiting the social dimension ('Who') has been considered. However, it can be expected that two pictures taken by the same user would also close in term of 'Where' and 'When' (since once person can be only in one place at time). In addition, social analysis notions like power of influence and relationships graphs can be poorly relevant for crisis management use cases, since legitimate VGI sources are people being impacted, whatever their recognised expertise and their friendship habits. We therefore refrained to try modelling social behaviours in a 'who dimension which would be expected to be very limited added value from. Nevertheless, the 'pictures pool' social feature of Flick has been exploited within the semantic dimension. Pools are groups of pictures various users can create and/or contribute to (e.g. 'Calif. Station Fire') by adding their own pictures. The name of the pools a picture belongs to is part of its textual metadata, which have been taken into account in the semantic distance. As a consequence, pictures belonging to the same pool(s) will be semantically closer.

(for instance, time intervals, spatial distances and semantic similarity scores) with very limited computation complexity.

To this end a distance measure for these 3 dimensions has been defined and normalised, then various weightings could be tested and the ones providing the most satisfactory results could be assessed. The normalized distance of the 3 dimensions were defined as follows:

(1) The 'Where' is the spatial distance between the 2 pictures. As the interest goes on close pictures, the distance has been normalised as follows:

$$\Delta km = \min (d_{km}/RD, 1)$$

with d_{km} the distance between the two pictures. RD is an arbitrary parameter that expresses a maximal distance beyond which 2 pictures cannot be likely considered as related to the same event (i.e. the theoretical maximum 'radius' of an observation zone of the event). 150km has been considered as an appropriate RD for Forest Fire events, based on experience of past events observation distances.

(2) The 'When' is the time elapsed (in days) between the dates when the 2 pictures were taken:

$$\Delta t = \min (d_t/RT, 1)$$

with d_t the temporal distance (in days) between the two pictures. RT is an arbitrary parameter that expresses a maximal time lapse beyond which 2 pictures cannot be likely considered as related to the same event (i.e. the theoretical maximum duration of a fire event). 90 days has been considered as an appropriate RT for Forest Fire events, which is a very conservative value (since it covers the duration of the sample). The consequence of the choice of RT is however limited since Δt is used as an relative measure. Since d_t cannot be greater than RT, it only means Δt will never be equal to 1 in this sample.

(3) The 'What' is computed comparing the tokenised textual metadata of the different pictures (see section 2). The distance is computed, based measures the percentage of common

tokens between the 2 pictures. For pictures A and B the semantic distance is defined as:

$$ds_{A,B} = \Sigma_{i\varepsilon\{Common\ w\}} \max(n_{i,A}, n_{i,B}) w_i \ / \max(\Sigma_{j\varepsilon\{A\ words\}} n_{j,A} w_j, \Sigma_{k\varepsilon\{B\ words\}} n_{k,B} w_k)$$

where $n_{j,A}$ is the number of occurrences of token j in A's textual metadata, and w_j is the weight of word j which is inversely proportional to its frequency. The semantic distance is further normalized:

$$\Delta s = \min(\max(d_s - Q_{0.1,ds} \ / \ Q_{0.9,ds} - Q_{0.1,ds} \ , 0), 1)$$

with $Q_{p,ds}$ stands for the quantile p% of the semantic distance among all pairs of pictures in the scope.

The 1-to-1 DB Scan distance between two pictures is then calculated by the weighted average of the 3 retained dimensions:

$$d = (\alpha_{km} \Delta km + \alpha_t \Delta t + \alpha_s \Delta s) / (\alpha_{km} + \alpha_t + \alpha_s)$$

with $\alpha_{km} + \alpha_t + \alpha_s = 1$ except when no semantic distance can be computed (no info for one of the pictures). In that case $\alpha_s = 0$.

In order to test a wide range of proportions between the 'Where', the 'When' and the 'What', 23 DB Scan parameter sets were considered, from purely temporal ($\alpha_{km} = \alpha_s = 0$) to purely spatial ($\alpha_t = \alpha_s = 0$), from spatiotemporal ($\alpha_{km} = \alpha_t = 0.5$ and $\alpha_s = 0$) to semantico-temporal ($\alpha_s = \alpha_t = 0.5$ and $\alpha_{km} = 0$), from unbalanced combination (e.g. $\alpha_{km} = 0.6$, $\alpha_t = 0.3$ and $\alpha_s = 0.1$) to balanced combination ($\alpha_{km} = \alpha_t = \alpha_s = 0.33$), *etc.*

In addition to a distance, DB Scan also requires the definition of a MinPts (minimum number of points to define a cluster centre) parameter and an *epsilon* (maximum neighbourhood distance) parameters. The MinPts parameter was set to 1 in order to allow clusters of a single picture. The *epsilon* was set to 0.05, based on the characteristics of the events of interest (i.e. forest fires last several days, and span to dozens of km^2, as orders of magnitude) and the normalization methods presented above.

3.3. Assessment of algorithms performance

Assessing the spatiotemporal location of clusters in respect to the ground truth allows evaluating the performance of every algorithm and parameter set. Two types of assessments were performed by comparing VGI clusters locations and actual Forest Fires:

3.3.1. Quantitative assessment.

The Large Fires Incidents Data as reported by the US Forest Service was used on that respect, which cover the continental US territory as well as Canada on a daily basis, using Remote Sensing data (MODIS) and "value-added" information from US and Canadian fire management agencies in order to provide critical, timely and comprehensive fire data and information[1].
Using this dataset, Predictive Value and Sensitivity could be calculated by counting the number of Fires for which a cluster exists and vice-versa. A Cluster was considered as 'matching' a Forest Fire if it satisfied the following conditions:

(1) Its earliest picture was taken not earlier than 2 days before (to compensate possible inaccuracies or delays in the reference data) the Fire was reported as started, and its latest picture was taken not later than 10 days after the Fire was reported as put out; although only pictures taken shortly after fire starts are relevant for detection purposes, pictures taken during the whole event duration are relevant for monitoring and characterisation purposes.

(2) The taker picture from the cluster that was the closest to the fire could potentially view it and/or its smoke plume. It has been considered, on an arbitrary manner based on actual observation distance of past fires, that the maximum distance a fire can be viewed from depends on its size. This maximum distance is between 50km (for the smallest fire in the dataset) and 100km (for the largest fire in the dataset) and varies proportionally to the fire size (in hectares).

If a cluster can be matched with two fires, only the closest one was retained. Having the fire-cluster matches, the following matrix (see

[1] Source : http://activefiremaps.fs.fed.us/

Figure 19) could be created for each algorithm/parameters set, making possible quantitative assessment of their performance.

FIRE presence → **↓ VGI CLUSTER presence (Fire detection)**	**YES**	**NO**
YES	XX.x % successful detection (TP = True Positive)	YY.y % unfiltered noise (FP = False Positive)
NO	ZZ.z% unsuccessful detection (FN = False Negative)	n.a
Sensitivity = *TP/(TP+FN) = % of actual fires that have been detected* Positive predictive value = *TP / (TP + FP) = % of detected events that were actual fires*		

Figure 19 : Quantitative assessment matrix for VGI clusters.

In this context, the following Quantitative assessment measures can be defined:

The **sensitivity** is the percentage of actual fires that have been detected.

The **Positive Predictive Value** (PPV) is the percentage of VGI clusters that could be matched to actual fires.

To be noted, the main characteristics of clusters themselves were also studied, and main findings are discussed in section 4.3.

3.3.2. Qualitative assessment

In order to deepen the understanding of the strengths and weaknesses of the algorithms and parameters choices, 4 selected areas were analysed in detail: the first in North California, the second in Alaska (Fairbanks) and Yukon (White Horse), the third around Seattle, and the fourth in the Yosemite National Park. In total, 65 fires were reported during the whole study period for these zones, and 1133 Flickr pictures were located in these zones. For each picture, it has been verified manually if the picture was related to an active forest fire (if yes, which one from the USFS data), and if it was correctly georeferenced. The clustering behaviour could then be compared to a fully manual classification of pictures related to a given fire. For each fire and each clustering method, the most likely cluster can be defined

as the cluster regrouping the highest number of pictures showing the same fire. Then, the following measures have been computed (see Figure 6):

The **conviction** is the percentage of pictures of fire A that are present in the corresponding most likely cluster. This indicates the part of available information about fire A captured in his associated cluster. In Figure 20, Conviction = 5/10 = 0.5.

The confidence is the percentage of pictures from the most likely cluster that represent fire A. This indicates the part of the information from the cluster that is relevant (not noise). In Figure 20, confidence = 5/6 = 0.83.

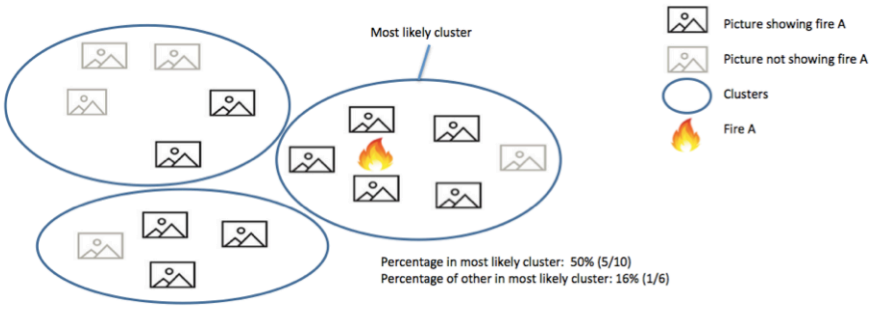


Figure 20 : Visual definition of qualitative measure

4. Results

4.1. Performance measures

One SatScan parameter set and 23 DB Scan parameter sets were considered for the benchmark. Table 6 summarises the 23 parameter sets considered for DB Scan, and the main characteristics of the results.

ID	α_{km}	α_t	α_s	# of clust.	avg. pictures./clust.	avg days/ clust.	sensitivity	PPV	conviction	confid.
DBS1	0,025	0,225	0,75	22	258,0	5,5	0,005	0,136	1,00	0,07
DBS2	0,05	0,45	0,5	477	11,9	0,5	0,099	0,128	0,95	0,08
DBS3	0,075	0,175	0,75	951	6,0	2,4	0,178	0,116	0,82	0,56
DBS4	0,075	0,675	0,25	1168	4,9	0,5	0,225	0,119	0,80	0,65
DBS5	0,125	0,125	0,75	1032	5,5	3,0	0,202	0,121	0,81	0,59
DBS6	0,15	0,35	0,5	1204	4,7	1,0	0,226	0,116	0,80	0,66
DBS7	0,175	0,075	0,75	1022	5,6	4,4	0,197	0,119	0,80	0,63
DBS8	0,1	0,9	0	1073	5,3	0,6	0,210	0,121	0,80	0,64
DBS9	0,225	0,025	0,75	984	5,8	6,6	0,186	0,117	0,80	0,66
DBS10	0,225	0,525	0,25	1303	4,4	0,5	0,236	0,112	0,79	0,80
DBS11	0,25	0,25	0,5	1231	4,6	1,2	0,226	0,114	0,78	0,77
DBS12	0,333	0,334	0,333	1307	4,3	0,8	0,237	0,112	0,75	0,73
DBS13	0,35	0,15	0,5	1221	4,6	1,8	0,233	0,118	0,75	0,67
DBS14	0,375	0,375	0,25	1328	4,3	0,8	0,239	0,111	0,75	0,74
DBS15	0,3	0,7	0	1260	4,5	0,6	0,236	0,116	0,75	0,80
DBS16	0,45	0,05	0,5	1141	5,0	4,6	0,218	0,118	0,75	0,64
DBS17	0,525	0,225	0,25	1306	4,3	1,1	0,239	0,113	0,75	0,73
DBS18	0,5	0,5	0	1284	4,4	0,8	0,241	0,116	0,75	0,79
DBS19	0,675	0,075	0,25	1226	4,6	3,0	0,223	0,113	0,72	0,64
DBS20	0,7	0,3	0	1261	4,5	1,3	0,239	0,117	0,72	0,71
DBS21	0,9	0,1	0	1182	4,8	3,4	0,225	0,118	0,69	0,71
DBS22	0	1	0	1	5677,0	122,0	0,000	0,000	1,00	0,07
DBS23	1	0	0	1081	5,3	7,4	0,210	0,120	0,71	0,70
SatScan	n.a.	n.a.	n.a.	1466	3,9	2,3	0,215	0,091	0,71	0,50

Table 6 : Parameter sets and main characteristics of clusters and results of the quantitative and qualitative assessments

Table 6 shows the result of the Quantitative and Qualitative assessments for the 24 algorithms. On this basis, a number of key observations can be made:

(1) Overall, Sensitivity scores may seem very low (about 20% of the fires are detected). This is due to the presence, in the validation dataset, of a large proportion of fires located in totally unpopulated area, and which did not cause any significant damage to persons, ecosystems or property (natural grassland and bush fires not damaging trees are also reported in the database). This highlights a bias of VGI Sensing, i.e. it relies on presence of citizens willing – and able – to report online. This has already been commented in Chapter 2, and it will be further discussed in the conclusion chapter. No consequences are expected on the outcome of the benchmark exercise since all methods are equally impacted by this bias.

(2) Positive Predictive Value features also very low overall scores (only about 10% of clusters match actual fires). This is due to the conjunction of three factors. Firstly, despite the simple noise filtering processing described in section 2.2.2, the VGI dataset still contained a significant proportion of irrelevant pictures (estimated at 10 to 20% by manual verifications) thus leading to the creation of false positive clusters. Secondly, even smallest clusters (1 picture) were kept in order to avoid supressing true positives. As a consequence, one relevant picture that would be wrongly geolocated (i.e. by using an erroneous place name in its description) would result in a false positive. Thirdly, the reference database only reports large fires[1], which means VGI clusters corresponding to real fires causing possible significant damages (e.g. 30 hectares of forest) would be counted as false positives. This has no consequences on the outcome of the benchmark exercise since all methods are equally impacted by these issues.

(3) Too small α_{km} leads to clustering with poor results. Indeed, in such case clusters tend to cover a very wide geographic area, and therefore to mix simultaneous events.

[1] Defined on the USFS website as wildfires of 100 acres (about 40 hectares) or more occurring in timber, or wildfires of 300 acres (about 120 hectares) or more occurring in grass/sage. (http://activefiremaps.fs.fed.us/mapterms.htm)

(4) The Impact of Semantic distance on performances seems to be poorly significant. Even more, large α_s (0.75) have a negative impact on performances. This can be explained by two factors. Firstly the Semantic dimension can be more subject to noise than space (at least when georeferencing is reliable) or time (usually defined by an accurate timestamp from the camera). Secondly, adding more dimensions in clustering may be counterproductive, the distance in a given less discriminant dimension attenuating the effect of more discriminant ones - a phenomenon known as *curse of dimensionality* issue.

(5) Besides the extreme choices for the parameters α_{km},α_t and α_s (e.g. purely temporal, semantico-spatial) DB Scan consistently provides better results than the benchmark algorithm, both in the Quantitative and the Qualitative measures.

(6) The fact that results seems better with $\alpha_{km} < \alpha_t$ can be explained by the higher precision of the Time dimension in comparison to the Spatial dimension. Indeed, most of the pictures are geolocated using a text-to-places parser while an exact timestamp is given by the camera's internal clock. In addition, pictures can show a fire from a high range of different distances: from far away smoke to close burning tree.

With α_s value and the ratio α_{km}/α_t not too close to 1 or 0, performance differences are not significant and probably reflect more the sample data than the configuration performance. Hence, we further concentrate only on the results on the 2 following clustering methods, which appear to be the most suitable DB Scan parameter sets:

- DB Scan no semantics: $\alpha_{km} = 0.3$, $\alpha_t = 0.7$, and $\alpha_s = 0$
- DB Scan semantics: $\alpha_{km} = 0.225$, $\alpha_t = 0.525$, and $\alpha_s = 0.25$

4.2. Cluster Statistics

Table 7 gives further details of the results and cluster characteristics for the 2 selected DB Scan parameter sets as well as for the benchmark algorithm.

Measure	DB Scan no sem.	DB Scan with sem.	SatScan
Proportion of spatial dimension (α_{km})	0.3	0.225	
Proportion of time dimension (α_t)	0.7	0.525	
Proportion of Semantic dimension (α_s)	0	0.25	0
Number of Clusters	1260	1303	1466
Sensitivity	0.236	0.236	0.215

Positive Predictive Value	0.116	0.112	0.091
Avg clusters by fire	1.952	2.062	2.977
Avg photos by cluster	4.506	4.357	3.872
Avg photos by fire	18.397	18.438	20.060
Avg time span	0.6	0.5	2.3
False fire	1114	1157	1333
Fire Detected	146	146	133
Conviction (avg % in most likely)	75%	79%	71%
Confidence (avg % relevant in most likely)	80%	80%	50%
Fire missed	473	473	486

Table 7 : Comparison of cluster characteristics by method.

From these results, following elements should be highlighted:

(1) This further analysis confirms the observation stated earlier: the added value of the semantic dimension to the clustering process is overall poorly significant. The only notable difference is that semantic increases the Conviction, expressed as the percentage of fire pictures included in the most likely cluster for a fire (79% with semantics, instead of 75% for purely spatio-temporal DB Scan). In other words, DB Scan with semantics allows a slightly better comprehensiveness by retaining more relevant pictures for a given fire. Since this is achieved without prejudice of other performance aspects (Sensitivity and Positive Predictive Value are very similar) this can be considered as a (minor) advantage of using semantic information for clustering.

(2) The average time span is considerably higher for SatScan (2.3 days, compared to 0.5 and 0.6 for DB Scan methods). Considering that the phenomenon of interest has a clear spatiotemporal extent, this seems to be a disadvantage of the use of SatScan in the context of this research. (It should be noted that due to the wide number of very small clusters that tend to have very short time span – see next section -, an average of 2.3 days means that clusters of bigger size tend to have a time span that is considerably longer that the usual forest fire duration: several weeks – or even month – instead of several days.) The spatiotemporal permutations mode of the SatScan algorithm allows detecting events without a priori knowledge of their expected scale; this is valuable for epidemics detection where such dimensions can vary

considerably (an entire village suffering the same disease the same week or 2% of European population catching the same virus over one winter are both considered as an epidemic event), but for forest fires that have a typical temporal (and spatial) span, a purely density-based approach seems more appropriate as this benchmark analysis consistently suggests.

(3) In addition to the significant differences in Sensitivity and Positive Predictive Value between DB Scan and SatScan, the considerable difference in Confidence, expressed as the percentage of significant pictures in most likely fire clusters should be highlighted. Indeed the quality of clusters is the key difference between DB Scan and SatScan. When a cluster can clearly be considered as a fire event, four fifth (80%) of the picture it contains are truly relevant for such event with DB Scan. With SatScan, it is only one half (50%) of the most likely fire cluster pictures that have been relevantly assigned to such fire.

4.3. Topology

In order to further refine the analysis, it is proposed propose in this section to look at the key characteristics of the clusters generated by the various benchmark algorithms, namely their size and their shape.

4.3.1. Cluster size

By definition, single pictures that could not be associated with any other cannot be considered *sensu stricto* as clusters, but rather as outliers. Nevertheless, such '1 picture clusters' have been kept in the analysis for the sake of comparison. As it can be seen in Figure 21, the number of such cases is much higher for SatScan than for DB Scan (1275, *versus* about 835 and 905). Oppositely, DB Scan identifies about 300 small clusters (respectively 357 and 333 without and with semantic) while SatScan identifies 132 small clusters (i.e. where size >1 and <= 10). The three algorithms return about 30 very big clusters (where size > 20).

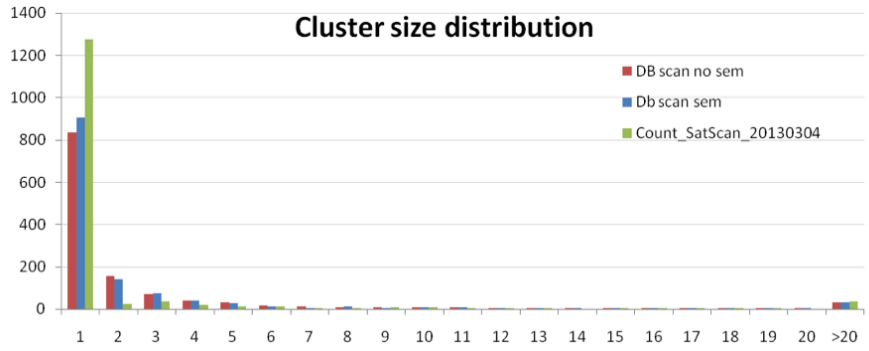


Figure 21 : Cluster size distribution.

These results highlights that DB Scan can specifically outperform SatScan for fires were only a few pictures are available. Such specificity is particularly relevant for detection of events from VGI, which consists often in finding 'diamonds in the rough'.

4.3.2. Cluster Shape

This section provides a brief comment on the shape of clusters resulting from the 3 benchmark algorithms for a typical fire event that took place early July 2009 in the Yosemite National Park (the Grouse Fire). More specifically, Figure 22 shows the temporal distribution of the pictures that are part of the cluster that could be associated to the Grouse Fire event. DB Scan shows a sharp peak around the fire date (early July), while the temporal distribution of pictures associated to the fire via SatScan is much more regularly distributed before and after the event (dotted line).

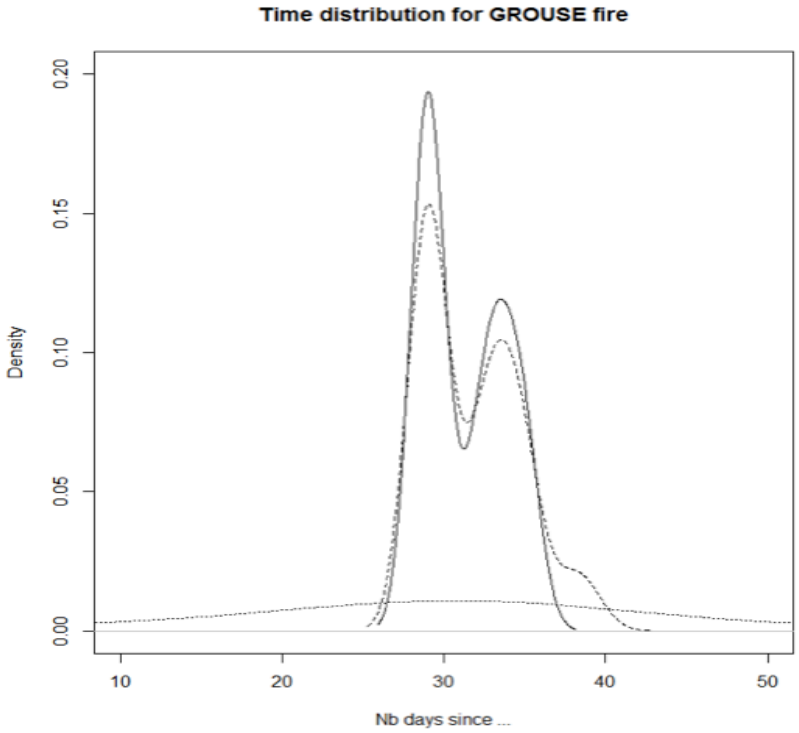


Figure 22 : Density function of time for pictures of cluster assigned to GROUSE fire according to DB no sem (continuous lined), DB sem (dashed line) and SatScan (dotted line).

As stated in the *cluster statistics* section, this can be explained by the nature of the SatScan Space-Time Permutation method, which is 'scale-independent' (i.e. since it uses windows of variable size, it can potentially detect clusters with very different spatial and/or temporal extent). This confirms the performance of the DB Scan method for event detection in noisy datasets.

Compared to DB Scan without Semantic, DB Scan with Semantic looks slightly less sharp, because it includes more pictures from several days after the event (dashed line, bottom right). This tends to confirm the observations made from cluster statistics: semantics helps associating more pictures that can be relevant. In this particular example could be pictures showing damages in the aftermath of the fire. The opportunity of enriching clustering with semantics depends thus greatly of the specific aim of the application, knowing that giving too much importance to semantics may lead to an overall degradation of the quality of the clustering results.

Similar behaviour is expected for the geographic distribution of pictures around the fire location. Nevertheless, the georeferencing method that has been used in this research may lead to particular data artefacts. For example many pictures of the Grouse fire were georeferenced as 'Yosemite National Park'; it can thus be expected expect a peak of density at the distance that separates the fire from the geographic centre of the park. For that reason, the spatial distribution of pictures around a fire location is often poorly relevant to characterise its shape accurately.

4.4. Computing costs and stream analysis

This section aims to complete the benchmark with specific considerations about the complexity – and therefore the computing cost – of the two algorithms that were tested. Indeed, since the findings of this research could be beneficial to applications involving real-time analysis of Social Media streams, it is important to have a basis of comparison of the cost of each algorithm, in addition to the comparison of results that has been provided in previous sections.

To be noted, no measures of computing times have been performed in this research. However, a conducting a purely theoretical discussion of the complexity of the algorithms is considered as a suitable way to have a good comparative basis of computing costs.

For SatScan algorithm, the complexity will at least be proportional to the size of the dataset, since each point (of a number of N points) must be indexed. In addition, the algorithm will require generating k random replications of the dataset (so-called Monte-Carlo replications) in order to calculate the P-Value of each candidate cluster. The algorithm also requires that many locations (noted as m) are visited, in order to assess if the local density of points is abnormally high. The complexity of SSTP algorithm can thus be estimated by O(k.m.N).

In this research, 999 Monte-Carlo replications have been generated, and the windows were set as having a maximum radius of 50 kilometres and to cover a maximum duration of 9 days (in practice, the average time span of clusters found with SatScan was of 5.9 days, as shown in Table 6). Provided the size of the study area (approximately a rectangle of 4500 km by 5000 km) and the duration

of the study period (90 days), an order of magnitude of 100.000 windows have been considered as potential clusters by SatScan. Considering the number of VGI items of a typical VGI Sensing use case (in hundreds of thousands, or even more), we can approximate that k.m has very roughly the same order of magnitude as N. As a consequence, the algorithmic complexity of SSTP as applied to VGI sensing can be estimated to $O(N^2)$.

In comparison, DB Scan as applied in this research would be more expensive by a factor N (so, $O(N^3)$). Indeed, it would require computing the distance between each pair of points ($O(N^2)$), and then loop through each point individually to take clustering decision (O(N)). In addition, the computational cost of the calculation of the semantic distance can significantly increase complexity.

It must be noted that both algorithms can be easily optimised to decrease drastically their respective computing cost. In SatScan, for example, very big windows can be analysed first in order to identify empty areas in which no data is present (Kulldorff 1999). This would drastically decrease the size of the factor m, especially if the points tend to be distributed very unhomogenously, as it is the case for VGI.

Similarly, sorting the points spatially before computing the one-by-one distances may result in a reduction of DB Scan complexity to O(Nlog(N)), as studied in detail by Ankerst et al. (1999).

In a view of performing real-time stream analysis, it must be noted that methods exist to ensure that each occurrence of a new point in the stream do not result in the re-calculation of values for all the points in the dataset. In other words, it is possible to ensure that each new point require only some point-specific calculations without the need of re-indexing the whole dataset. Zhao et al. (2014), for example have successfully applied such principle and calculated on the fly a so-called Local Modularity Scan Statistic inspired by Kulldorff's statistics in order to create real-time clusters of event-related tweets.

The cost of semantic distance computation requires special attention too. Unlike space and time, the semantic dimension of a VGI Item cannot be easily measured against a fixed reference (such as the datum for geographic coordinates or an universal epoch for time). As a consequence one-to-one calculation would still be required each time

a new point arrives in the stream. A simple way to overcome such issue would be to identify the possible clusters a new point could be added to on a purely spatio-temporal basis, and then to use one-by-one semantic distance calculation (calculated ONLY for the points from these candidate cluster) to refine choices taking topicality into account. Such two-stage approach has been followed e.g. by Zhao et al. (2014) as well as Cheng & Wicks (2014). Adapting such methods with DB Scan-inspired cluster allocation routine for new points (i.e. for each new point, consider all clusters which have points in distance smaller than 'Epsilon' from it) may result in a simple and straightforward VGI stream analysis process.

5. Discussions and Conclusions

This chapter aimed at contributing to the detailed definition of optimized pre-processing and clustering methods in order to extract Event-related information from VGI. By executing a benchmark of algorithms based on a real-life use case (namely, the detection and characterization of Forest Fires in North America on the summer 2009) a method that can significantly improve results of similar methods previously presented in the literature is suggested.

The benchmarking exercise allowed confirming the two main intuitions behind this research.

Firstly, a purely density-based clustering such as DB Scan is more appropriate for specific events detection from VGI, compared to the state-of-the-art SatScan Space-Time Permutation. Indeed, one of the main advantages of such SatScan method is to be scale-independent, which is of primary importance in epidemiology research (where an entire village suffering the same disease the same week, or 2% of European population catching the same virus over one winter are both considered as an epidemic event). Oppositely, when the nature and the orders of magnitude of a phenomenon are known (typically, a Forest Fire event last 1 to 10 days and spans over dozens or hundreds of hectares) DB Scan can, by design, outperform SatScan Space-Time Permutations (which tends to cluster contemporary fires in one event, or fires from different periods in the same area). This can be explained by the fact that the contribution of the various dimensions to the aggregated clustering distance (*epsilon*) can be calibrated to fit the

characteristics of the phenomenon of interest. This advantage in performance also leads to an intrinsic limitation: DB Scan parameters must be calibrated according to the typical extent of the Event-of-interest, and the method is less appropriate for fully generic event detection (i.e. for the detection of any event, even of unknown nature). To be noted also: this research did not aimed at optimising the parameters used with SatScan, but rather to compare its state-of-the-art implementation to VGI with a novel approach based on DB Scan. Further research might investigate alternative SatScan optimisations, which would give a more comprehensive definition of both algorithms respective performance in the context of VGI Sensing.

Secondly, the semantic dimension of VGI can play a significant role in the VGI Sensing workflow. In other words, VGI items are not simple points on a map; their rich meaning and the context in which they published can contribute to the wide picture, even if they are often poorly structured and noisy. More specifically, this research has given examples of usage of the semantic dimension of VGI for the pre-processing (f.i. enrichment and filtering), for which added value can be formally demonstrated in further works. Its contribution to the clustering *sensu stricto* is limited to a minor improvement of comprehensiveness of True Positive clusters (Conviction 4% higher with semantics).

This research suggests significant improvements of the VGI Sensing Workflow as presented by De Longueville, Luraschi, et al. (2010), but it also opens perspectives for further optimisations that should be explored in future works.

Firstly, the Noise Reduction Filtering method could benefit of further research: the machine learning techniques that were applied provided interesting results, which can be extended and consolidated. Improving the capability to discriminate relevant material from noise without discarding the ‘diamonds in the rough’ from the very beginning of the processing chain could greatly improve the final results.

Similarly, the segmentation of the VGI input (i.e. the automatic recognition of ‘types’ of VGI items) could worth being explored (as a basis for a differentiated treatment of each segment). While performing the qualitative assessment, clear and distinct cases of

Flickr pictures related to Forest Fires, e.g. smoke plume of a remote fire, Fire Fighters in action, burning trees, old fire scars, haze from a distant fire at sunset. Each type of such pictures should ideally be treated according to their specificities (i.e. actual picture of firemen in action is a direct observation of the phenomenon, while haze due to a distant fire is a secondary evidence). The main difficulty is the resulting typology would be only applicable in a given socio-cultural context for a given phenomenon of interest (for instance, Forest Fires). On that respect, the automated analysis of the image itself (e.g. for detecting presence of flames) would probably have a major impact.

The calculation method for Semantic Distance would also benefit further works. The method used in this research was simply based on co-occurrence of words (weighted by their overall frequency), but finer-grained Natural Language Similarity methods could certainly lead to improved final results (since it would increase the meaningfulness of the Semantic distance). Similarly, the social dimension could be better exploited, specifically when dealing with social-focused VGI sources, and with use cases other than natural hazards detection.

Last but not least, this research findings should be applied in the context of Social Media streams analysis. While the experimental setup of this research was purely retrospective, with no particular focus on optimisation of computing costs, its outcome is applicable to near-real-time applications. In Social Media streams analysis, algorithms are usually one-pass, and clusters are re-evaluated on-the-go, as new VGI items appear in the stream (Atefeh & Khreich 2013). So, similarly to Zhao et al. (2014) who calculated Kulldorff's Scan Statistic for each incoming item in the Twitter stream in order to re-compute on-the-fly spatiotemporal clusters, it can be easily calculated if a new VGI item is within *epsilon* distance (combining spatial, temporal and semantic dimension) of an existing point, and therefore if it may contribute to an existing candidate cluster, or in contrary if it should be considered – until possible 'close' VGI items appear – as an outlier.

Furthermore, and since this chapter aimed at (early) characterisation of events and not at their detection, candidate clusters can be initiated using other sources of information about new events (e.g. authoritative data from relevant Government bodies, or hot spots identified via

Remote Sensing). Such 'exogenous cluster seeding' approach has already been tested successfully on Twitter streams by Benson et al. (2011) in the context of cultural events in New York City. This could help reducing drastically computing costs, by prioritising calculation of clusters on spatiotemporal regions of interest. Interestingly, the next Chapter adopts such perspective of combining VGI Sensing with other existing sources of information, by suggesting the vision of a Digital Earth Nervous System.

As a final note, we can highlight that a perception is growing that big data is about 'few people' (e.g. marketing companies, governments) surveying many citizens. However, in crisis management facts have shown that it is often many people helping many (e.g. see Terpstra et al. 2012). This can be supported by big data operations that streamline the volunteer contribution of many into timely, useful and clear information. We hope this research contributed to provide such kind of methods. By doing so, it would contribute to the achievement of the Digital Earth vision as Al Gore drew it almost 2 decades ago, and affirm the difference of the Digital Earth Nervous System with the Big Brother vision that is *en vogue* nowadays.

Chapter 4 – Perspective: the Earth's Nervous System[1]

Abstract
Digital Earth is a powerful metaphor for the organisation and access to digital information through a multi-scale 3D representation of the globe. Recent progress gave a concrete body to this vision. However, this body is not yet self-aware: further integration of the temporal and voluntary dimension is needed to better portray the event-based nature of the world. We thus aim to extend Digital Earth vision with a Nervous System in order to provide decision makers with improved alerting mechanisms. Practical applications are foreseen for crisis management, where up-to-date situational awareness is needed. While it is traditionally built through trusted sources, citizens can play a complementary role by providing geo-referenced information, known as Volunteered Geographic Information (VGI). Although workflows have been implemented to create, validate and distribute VGI datasets for various thematic domains, its exploitation in real time and its integration into existing concepts of Digital Earth, such as Spatial Data Infrastructures, still needs to be further addressed. In this chapter we suggest to bridge this gap through sensor web enablement for VGI, where VGI sensing becomes a sense of the Digital Earth's Nervous System. This approach and its applicability in the context of a Forest Fire scenario are then discussed.

1. Introduction

By articulating the vision of Digital Earth (DE) as a "multi-resolution, three-dimensional representation of the planet, into which we can embed vast quantities of geo-referenced data" (Gore 1998), the former US Vice President Al Gore provided a powerful metaphor for

[1] This study has been published in the International Journal of Digital Earth vol. 3, no. 3 (2010): 242 - 259 under the title *Digital Earth's Nervous System for crisis events: real-time Sensor Web Enablement of Volunteered Geographic Information.* The text presented here is slightly modified from the original publication for layout and terminology harmonisation.

innovative earth observation systems. Ten years later, Craglia and colleagues (Craglia et al. 2008) published a position paper to argue that this vision has not yet been achieved. The main argument was that, in parallel to the increased availability and access to information, the need of better understanding interdependencies of environmental and social phenomena had also increased. For the authors, DE thus required more dynamic systems, new sources of information and stronger capacities for integration. Therefore the next generation of DE was not envisioned as a single system but multiple connected infrastructures based on open access and participation across multiple technological platforms and addressing the needs of different audiences.

In this chapter we want to contribute to this reformulated vision by suggesting a more dynamic view on Digital Earth characterised as a *digital nervous system of the globe*, which actively informs us about events happening on the earth's surface by connecting to sensors networks and situation aware systems. In addition, this implies link up with initiatives underway like Spatial Data Infrastructures (SDI), Sensor Web Enablement (SWE), and Volunteered Geographic Information (VGI). While SDI focuses on the distributed management of relatively static geospatial data (GSDI 2004), SWE concentrates on the observation of highly dynamic phenomena, such as weather and air pollution (Botts et al. 2008). VGI is complementary to both, by addressing user contributed geospatial content (Goodchild, 2007).

We present the *Digital Earth's Nervous* S*ystem (DENS)* focussing to the support to emergency response and disaster management field, which are in particular need of timely information. We focus on citizens as invaluable source of such information: they are (almost) everywhere, they are mobile, they perceive events, and thanks to recent technological developments, they can report them in real-time through the Internet. The implementation of DENS with VGI as one of its senses is discussed. We envision such VGI sensing in analogy to remote sensing. Just as the processing of satellite data as an input to many geospatial analyses is readily accepted, VGI sensing should aim to better interpret the abundant and freely available user-generated content (De Longueville et al. 2009). SDI provides the necessary structures on which DENS operates, especially the extensions of SWE as SDI-compliant standards provide means for integrating senses.

In this chapter, DENS is described in more detail and explain how a sensor web enablement of VGI contributes to its implementation. Crisis management serves an ideal setting for explaining the benefits. In the next section, required background on Digital Earth is presented, VGI, especially in respect to data quality, the added value that VGI contributes to crisis management. The vision behind (DENS) is developed in section 3, while a potential SWE-based implementation of DENS is discussed within the context of crisis management in section 4. Its applicability in a Forest Fire scenario at the European level is introduced, before the added value in the European context together with given constraints are specified (section 5). This chapter then ends with concluding remarks and directions for future work on DENS (section 6).

2. Background

This section provides background knowledge and related work on two large fields the metaphor of nervous system for the planet is built on: Digital Earth, and Volunteered Geographic Information. The potential of VGI is discussed, in general and in the context of crisis management in particular, touching upon the most frequently discussed issue in relation to VGI: the credibility of the information.

2.1. (Next Generation) Digital Earth

The vision of Digital Earth has been first formulated in a speech the former US Vice President Al Gore gave at the California Science Center in Los Angeles in January 1998 (Gore 1998). The "multi-resolution, three-dimensional representation of the planet, into which we can embed vast quantities of geo-referenced data" he described was at this time a powerful metaphor for innovative earth observation systems. Geobrowser technologies and virtual globes developed by private companies (e.g. Google Earth, Microsoft Virtual Earth or ESRI's ArcGIS Explorer), or by open source projects (notably NASA's World Wind) gave a concrete body to this vision, that can be described as a kind of 'Web Wide World' (Butler 2006).

Ten years after this speech, (Craglia et al. 2008) published a position paper to argue that this vision has not yet been achieved mainly because, in parallel to the increased availability and access to information, the need of better understanding interdependencies of environmental and social phenomena has also increased and this

requires more dynamic systems, new sources of information and stronger capacities for integration. In order to further develop such integration, these authors advocated for a Next Generation Digital Earth that can act as "collaborative framework allowing the emergence of hybrid infrastructures combining both voluntary and institutional data" (p 162.).
In this context, the Digital Earth can be seen as a powerful framework for developing novel flows of information, with a reconceptualised role for end-users (Budhathoki et al. 2008), aiming at promoting collaboration between expert users, as well as with non-expert users (Grossner et al. 2008).

2.2. Volunteered Geographic Information for Crisis Management

There is an increasing consensus to recognize the role of VGI in support to crisis management activities, as described in previous chapters. It has also been stressed that the credibility issue of VGI can be overcome by "aggregating input from many different people" (Mummidi & Krumm 2008, p. 215). The VGI Sensing concept further developed in this chapter in the context of Digital Earth follows such cross-validation strategy.

3. A Nervous System for the Digital Earth

Before devising a system for improved crisis management capabilities for the Digital Earth, some questions have to be addressed: How to create more dynamic, event-based, and quality controlled information flows ? In particular, how to integrate the wealth of heterogeneous geo-information generated by citizens to such information flows ? How to combine these with existing spatiotemporal information available, for example through SDIs ? How can such combined information become a suitable source to timely feed efficient decision-support systems ? How can a more dynamic DE be designed to become an integrated solution to these issues ?

The current vision of DE suggests a rich, but passive system where any information retrieval is triggered by the user, who has to interact with the system to answer a question. We extend this vision towards a system that actively notifies its users and informs them about situations that might require a reaction. Such even-based, situation-

aware system can be enabled by establishing a *Digital Earth Nervous System* (DENS) on top of the currently suggested elements of DE.

In order to develop the vision underlying DENS, we describe an existing system, which is recognised for its efficiency to build sophisticated decisions based on the combination of in-memory knowledge with a complex flow of real-time information, the human nervous system. A simplified and generalised overview of sensory processes is provided, based on few classical references from the field of cognition sciences, by describing the chain of activities in the nervous system that cause a reaction by a human on a stimulus. Please note that it is neither intended to reflect up-to-date research activities and debates from related fields, nor tried to describe rigorously such a complex system. The analogy shall help us to identify the requirements towards DENS and its implementation. The following brief narrative describes how a person perceives the stimulus of an insect moving on his arm:

> *John, who is allergic to bee-stings, is outside in the garden wearing a t-shirt. He feels something is moving on his arm: The sensor neurons in his skin are stimulated and create a kind of a 'mental image'. This contains the location of the stimulus on John's body, the estimated size of the stimulus and the fact that the stimulus moves. This sensation is compared to memory and implicit knowledge: the moving object is a small animal, an insect or spider. It might even be a bee. This context is important enough to cause a reaction: the situation is analysed and a lack of information is identified. John cannot be sure that it is not a bee. He turns his head to set his eyes on the object to find out.*

This little story tells us that a nervous system allows us (i) to receive stimuli from our close environment, (ii) to compose complex impressions of our environment, (iii) to set these into context, and (iv) to react on the situation. We want to apply exactly these capabilities for sharpening DENS. This biomimetic strategy is inspired by (Ross 2009) who coined the term *Social Nervous System* to describe the impact of recent evolutions of Internet and mobile technology on communication processes in modern societies. As a result of such benchmarking exercise, we propose a set of technologies and practices emphasising the dynamic dimension Digital Earth must develop in order to better capture the changing, event-based nature of the world.

Scientific research related to the understanding and modelling of human sensory activities have a long history - see, e.g. (Attneave 1959) - and have been stimulated by countless possible applications in artificial intelligence (Newell & Simon 1972). It is thus not surprising that cognition sciences and computer sciences widely share similar terminology, such as process, sensor, measurement, state, pattern recognition, system architecture, messaging, *etc*. - see, e.g. (Port & Van Gelder 1995). In their famous essay approaching human reasoning and emotions from a neuropsychological perspective, (Damasio & Sutherland 1996) describe the sensory process in several steps, where a mental image of sensory signals is first created, and then associations are created with in-memory knowledge to identify and characterise elements. In another example of particular interest, (Harnad 1987) describes human perceptions as a categorisation process, where the brain is able to assign a set of received stimuli to well-known entities (objects or events), and to use implicit knowledge to infer current characteristics of such entities.

On the basis of these previously cited works, we focus on the following concepts that can be used to describe both the human sensory system and DENS:

- *stimuli* are defined as changes in the environment that can be detected by sensors;
- *sensors* are specific components of the system designed to encode stimuli in a pre-formatted message;
- *sensations* are centralised, organised sets of sensor messages that result in complex measurements
- *perceptions* describe the set of features and their characteristics that have been obtained by recognising patterns in sensations and comparing them to implicit, in-memory knowledge;
- *attention* describes the prioritisation of perceptions according to context;
- *reaction* describes the planned set of actions resulting of the analysis of perceptions with highest attention level, involving additional nervous system functionalities.

The following stories illustrate how such concepts can work in practice, both for the human nervous system, and for DENS, They help us identifying received stimuli, compositions as impressions, context, and reactions.

> John, receives touch stimuli: sensor neurons specialised in pressure and itchiness measurement are stimulated. Based on the information from the different sensor neurons, a sensation is created; it is a set of organized measurements that forms a kind of a 'mental image': something small is moving somewhere on his skin surface. This sensation is used to create a perception: it is compared to memory and implicit knowledge derived from memory using reasoning. The result is that he has most likely an insect on his arm. This perception is compared to the context: John is outside in the garden wearing a t-shirt, plus he is allergic to bee-stings. This context is enough to create attention towards the perception, as is important enough to cause a reaction. In this case the reaction is that the given information is analysed and a lack of information is identified. John cannot be sure what kind of animal this is and whether it is a threat. The reaction is thus a mobilization of additional senses: he turns his head to set his eyes on the object and find out if it is indeed a bee. As a result, John receives a vision stimulus of the moving object…

Accordingly, DENS receives VGI stimuli: micro-blogging messages and pictures are posted through the World Wide Web. Based on the information from the different connected devices from the citizens, a sensation is created; it is a set of organized measurements that forms a kind of a 'digital image': a big cluster of messages and pictures with similar contents comes from somewhere on the earth surface. This sensation is used to create a perception: it is compared to memory and implicit knowledge derived from memory using reasoning. The result is that there is most likely a Forest Fire in Northern Spain. This perception is compared to the context: DENS did not know about this fire before, so it is a new event. Provided the season, the weather conditions and the amount of VGI usually coming from there, it must be an important fire. This context is enough to create attention towards the perception, as is important enough to cause a reaction. In this case the reaction is that the given information is analysed and a lack of information is identified. The reaction is thus a mobilization of additional senses: regional SDIs are queried for refining potential impact assessment, and a satellite is tasked to provide remote sensing images from this zone, thus providing additional stimuli to DENS …

A generalised picture of both, the central concepts of the human nervous system and the nervous system of the DE is provided in Table 8.

	The Human Nervous System		The Digital Earth Nervous System	
Concept	**Touch sense** (John's story)	**Vision sense** (succeeding John's story)	**VGI Sensing** (DENS-VGI story)	**Remote Sensing** (succeeding DENS-VGI story)
Stimulus	Something changes on the body surface, expressed in heat, pressure, itchiness and/or pain	Something changes in the environment, expressed in movement, shape and/or colour	Something changes where there are people that can and want to report it	Something changes that can be observed by a (constellation of) satellite-embarked sensors
Sensor	Specific receptors in the skin, tongue, throat and mucosa convert stimuli in electrochemical waves	Cones and rods cells situated in the eyes convert stimuli in electrochemical waves	Information is digitized (using smartphones, computers, GPS, *etc*.)	Waves reflected or emitted by remote objects are digitized according to sensor specifications
Sensation	An organised set of measurements is created in the brain	A mental image is created in the brain	Heterogeneous information is centralised and organised is a complex set of measurement results	n-dimensional remote sensing images are created
Perception	Patterns are recognised in touch sensations and compared to memory for identification and characterisation of features	Feature in the image are discriminated and compared to memory for identification and characterisation	Complex measurements are analysed for identifying events and situations	Images are processed to provide derived information, features are identified using prior knowledge
Attention	Relevance is assigned according to context	Relevance of such features is assigned according to context	Alerting mechanisms are triggered according to context	Alerting mechanisms are triggered according to context
Reaction	Further nervous systems functionalities are mobilised, and a prioritized list of activities is created, and including e.g.: - instinct (fast, simple) reactions; - evaluation of information lacks; - request to other senses for more information; - movements.		Sensor network information is integrated in crisis information systems, where appropriate tasks are prioritised, related to e.g.: - early response; - situation awareness, requ. for additional information; - mitigation actions; - damage assessment.	

Table 8 : Functional comparison of the human nervous system and the Digital Earth nervous system

4. SWE for VGI Sensing: Implementing DENS

In the early stage of crisis events responsible authorities need up-to-date situational awareness in order to effectively coordinate response. In such time-critical context, specific ways of collecting, organizing, accessing and communicating information have to be set up (Annoni et al. 2010). The main challenge is to make sense of a very dynamic stream of information (De Groeve et al. 2010). We argue that providing Sensor Web Enablement (SWE) for VGI is the next important step towards the implementation of DENS because it provides an SDI-compliant method to fulfill these requirements.

SWE offers straightforward ways to support the VGI sensing capabilities identified above. Sensor Observation Services provide continuous information, thus addressing the timing concern when such services are in already in place before a crisis onset; appropriate calibration of sensors and modeling of functional constraints widely contributes to the data quality of sensor data; while Sensor Alert Services offer a possible way of organizing and prioritizing flows of information crisis managers have to deal with. SWE standards can thus be seen as a possible technological solution for implementing DENS. Moreover, as depicted above, a wide consensus is emerging on the role VGI can play in the next generation of crisis information systems (section 2.2). In the following, the sensor web enablement activities of Open Geospatial Consortium (OGC) are introduced, which we consider most relevant for this work, and then discuss the implementation SWE-VGI to support crisis information systems in their need for timely, quality-controlled, accessible and easy-to-use information from citizens.

4.1. OGC Sensor Web Enablement

Previously installed and ad-hoc sensor networks can be a primary source for feeding crisis information systems with near-real time geospatial data (Jirka et al. 2009). In order to improve interoperability between risk management systems and sensor networks, the Open Geospatial Consortium (OGC)[1] provides standards for Web-based Sensor Web Enablement (SWE) (Botts et al. 2008). In this section, the SWE components that are relevant for this work (Figure 23) are introduced. We decided to use SWE, because it provides a well-

[1] Official Web page available from http://www.opengeospatial.org/

structured framework fitting our interests, it is based on open standards, and it has a growing user community. In the following, we use examples from current practices. These technologies will be projected to VGI later in this chapter.

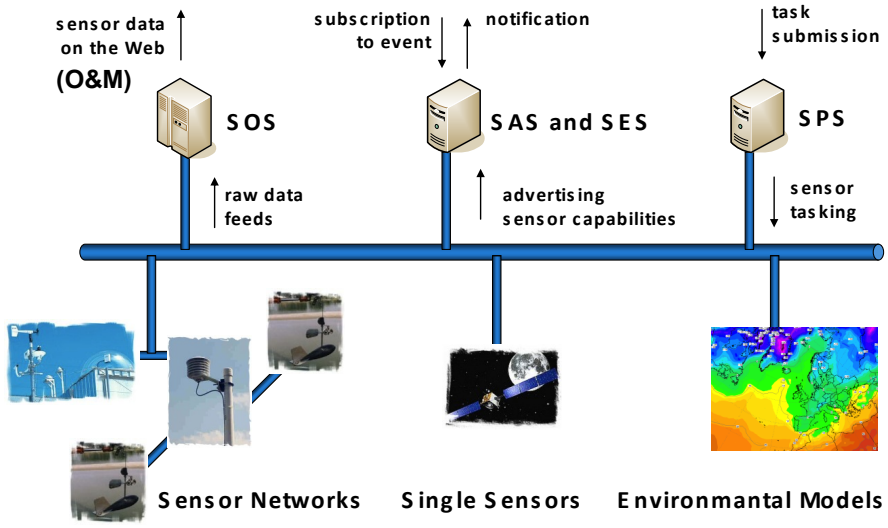


Figure 23 : Interplay of SWE components.

SWE provides a set of standards dedicated to the integration of time series data into classical Spatial Data Infrastructures (SDI). It complements established OGC standards, such as Geography Mark-up Language (GML) (OGC 2007a), and the data access components Web Feature Service (WFS) (OGC 2005) and Web Coverage Service (WCS) (OGC 2008b). The Observations and Measurement (O&M) standard serves the underlying data format, which can be used to encode a series of in-situ measurements, like air pollution, and remotely sensed information, like satellite imagery (OGC 2007b). It is the standard for encoding sensor data on the Web. The Sensor Observation Service (SOS) is the data access service exposing O&M (OGC 2007c). It is used to encapsulate raw data feeds from single sensors, sensor networks, sets of both, and simulations. As an OGC service, it follows the common interface specifications (OGC 2006)and extends them with descriptions of available sensors and access to dynamic geospatial data. Initially the SOS was used for providing raw data, such as the results of a series of temperature measurements. In the last years, the SOS is increasingly used for accessing value added products, i.e. processed observation results, like

daily averages of temperature, maxima of daily concentration of air pollutants, or results of dispersion models (Havlik et al. 2009).

Apart from pure data provision, which implies pull-based communication, information can be pushed to potentially interested users via the Sensor Alert Service (SAS) (OGC 2007d). The SAS allows clients to subscribe to events. Clients specify conditions under which they would like to be notified using a simple constraint model. The SAS monitors inputs from advertised sensors. Notifications are triggered each time a certain constraint is met, for example, is a specific pattern occurs. Users may be interested in high spatiotemporal frequency of pollutant concentrations above a specific threshold or of a significant raise of temperature. As soon as the SAS identifies an according pattern, is sends notification messages to all interested users. The SAS optionally supports the Common Alerting Protocol (CAP) of Organization for the Advancement of Structured Information Standards (OASIS) (OASIS 2005).Currently, the SAS is being generalized towards a Sensor Event Service (SES), which basically provides a richer constraint model, i.e. it allows for more complex patterns (OGC 2008a).

If in any case sensors or other processes should be tasked, for example if a new measurement series should be initiated, the Sensor Planning Service (SPS) can be used for calibration (OGC 2007e). In addition to the common OGC Web Service capabilities, the SPS provides sensor descriptions including information how sensors, sensor networks, and simulations may be tasked. For example, a sensor network can be set up to measure air pollution in intervals of five minutes or a satellite can be tasked to sense a specific region on the surface of the globe.

4.2. Suggestion a Mapping of OGC Sensor Web Activities

Any kind of spatiotemporally referenced resource, for example geo-tagged and time-stamped photographs posted on picture-sharing platforms, can be encoded in O&M. Such encoding can easily include additional attributes, such as a list of (thematic) keywords. In this way, every piece of VGI could be encoded as an observation or measurement using the existing OGC standard. We argued earlier that this most obvious use of the O&M standard means providing stimuli as observation results (section 4.1). However, due to the low level of information provide and to the sheer amount of data, we favour to use

O&M encoding at the *sensation* and *perception* level of DENS. In particular, O&M offers *ComplexObservation* as a construct to represent n-dimensional values of heterogeneous types and *ObservationCollection* as a means to represent sets of observations (OGC 2007). With these capabilities, O&M can be directly encoding collections of VGI items collected under pre-defined constraints. For example, such collection can be a set of geo-tagged photographs and related information (such as time stamps) retrieved from a picture-sharing platform. Spatiotemporal clusters of such information can be encoded in a similar manner.

The SOS is classically used to encapsulate single sensors, sensor networks, sets of both, and simulations. In principle, any other provider of O&M encoded data could be encapsulated in a similar way. According to the above, this does include pieces of VGI (each user is considered a sensor), complete platforms (considered as a sensor network) and value added information, such as organised streams or spatiotemporal clusters of VGI. The first two closely corresponds to a SOS offering raw data, while the latter provides another instance of a SOS providing access to value added products. We thus argue that the expressing the VGI activity in O&M and providing it as a SOS implements part of the *perception* step for VGI sensing in DENS. Accordingly, Figure 23 can be extended in order to account for VGI as SOS (Figure 24).

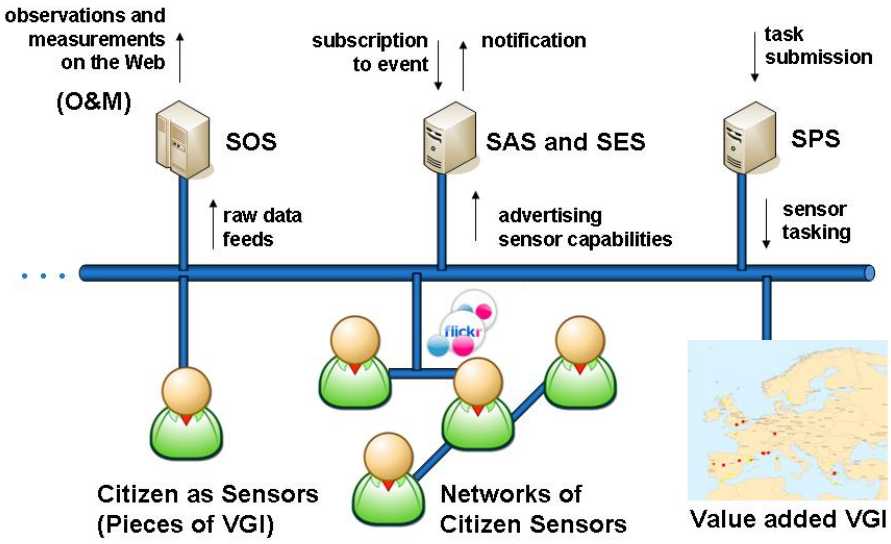


Figure 24 : SWE for VGI sensing.

The SAS allows clients to subscribe to alerts. Notifications are triggered if a certain constraint is met. This principle can directly be applied to perceptions and attention in the VGI sensing context. For example, if spatiotemporal clusters of VGI items related to a specific hazard are identified and exposed by a SOS (*perception*), and if the SAS identifies that the size of a cluster (expressed in number of VGI items, or other metrics assessing its relevance) reaches a certain threshold (still *perception*), then a 'hot spot' is detected and the SAS sends an alert to responsible authorities (*attention*). The Sensor Event Service (SES), which basically provides a richer constraint model, than the SAS, provides the frame for including complex patters of VGI.

Finally, the SPS can be used for calibration of value added VGI products. In the considered VGI setting, such SPS calibration may be applied to clustering algorithms that identify hot spots. For example, the threshold for triggering alerts tuned socio-economic indicators. For instance, densely populated zones with great internet connectivity will most likely generate more VGI than others, where relevant events may be reported by less abundant VGI. The importance of this calibration issue needs to be underlined. Unlike observations based on purely mechanical sensors (e.g. in satellite remote sensing), VGI sensing relies on human factors. Therefore, a particular care must be given to the calibration process, where cultural and technical constraints that lead to the creation of a VGI item can be modelled in order to measure the statistical significance of the derived information. If DENS identifies lack of information, an SPS may also be used to task additional senses, for example a satellite could be tasked (as already described above).

Table 9 provides a mapping of the notions of SWE and the more general concepts underlying the nervous system. The VGI examples on the right part of the table are an extract of previous examples.

Concept	SWE Notion	Example for VGI sensing
Stimulus	Observed property	e.g., environmental changes caused by a crisis situation
Sensor	Raw data **SOS** (encapsulated device that digitize stimulus and provide property values)	e.g., **SOS** encapsulated smartphone data
Sensation	Raw data collection **SOS** (encapsulated network of devices or time-series of a single device)	e.g., **SOS** encapsulated collection of VGI that is organised and has a given focus (time, geography and theme)
Perception	Value added information **SOS** (encapsulates clustering algorithms etc.) **SAS** (as situation-aware component)	e.g., **SOS** encapsulated identification of VGI clusters (thanks to additional data and contents analysis) e.g. **SAS** part setting hot spots into context, i.e. filter the important hot spots according to pre-calibrated constraints
Attention	**SAS** (as alerting component)	e.g., **SAS** part. using CAP to inform crisis manager
Reaction	Decision support system (can include **SPS**)	e.g., **SPS** for satellite tasking

Table 9 : Mapping between general concepts of the nervous system and SWE for VGI sensing.

5. A Forest Fire scenario

In this section, an example of Information System related to a particular type of crisis - Forest Fires – is presented. First, the current system based mostly on Remote Sensing is described. Secondly, a proposal of how VGI sensing can be integrated in such system is devised. Finally, it is explained how such example illustrates opportunities and constraints of remote sensing, VGI sensing and SDI integration.

5.1. Forest Fire Hot Spot Detection by Remote Sensing

The European Forest Fire Information System (EFFIS) of the European Commission (San-Miguel-Ayanz et al. 2002) provides support to forest fire fighting and prevention in Europe throughout the fire season. Every year an average of over 500.000 hectares of forests are burned by wildfires throughout Europe. The impact of these fires is significant both financially and in terms of human lives. The EFFIS system was developed to provide continuously updated forest fire

related geo-spatial information, during crisis situations to enhance international cooperation (e.g., by aerial fire fighting) and support decision-making.

The EFFIS Web portal[1] provides this fire information through a comprehensive but simple interface. Critical information is displayed over a Web map and rapid access to datasets and map services are provided using SDI standards. These include two major added-value information products of EFFIS: the hotspots (active fires) and burned area (monitored forest fires) layers.

Active fires are located on the basis of the hotspots product of MODIS sensors on board of Terra and Acqua satellites (Kaufman et al. 1998), which identifies areas on the ground that are distinctly hotter than their surroundings. The difference in temperature between the areas that are actively burning with respect to neighbour areas allows the identification and mapping of active fires. In addition, detected hot spots are overlaid with land cover information to retain only those taking place in forest areas. The mapping of active fires is performed to provide a synoptic view of current fires in Europe and as a mean to help the subsequent mapping of fire perimeters. Information on active fires is normally updated daily and made available in EFFIS within 3-4 hours from MODIS acquisition.

Forest Fires events are monitored by mapping dynamically burnt areas and by characterising fire events using several information sources. Firstly, the active fires map described above is used to locate areas requiring further analysis. Secondly, satellite images are used to delimit the burnt area perimeter. Thirdly, burnt areas are overlaid with a series of reference data to characterise the risks related to each fire event. Reference data includes protected areas, population density, sensitive infrastructures and local meteorological conditions. In addition, EFFIS includes a systematic collection of fire news from European media sources (using multilingual RSS feeds filtering and aggregation techniques), which completes with the overview of the fire's importance within short time delays. As a consequence of this processing chain, a list dynamic of forest fires with major importance is created.

[1] http://effis.jrc.ec.europa.eu/

Active fire detection and burnt areas mapping activities are feeding a dedicated application created by the EFFIS team with the purpose of providing geospatial information support to the EC Monitoring and Information Centre (MIC) on current major fire events in Europe and the Mediterranean area. Since 2009, the MIC has been in charge of tasking the European Forest Fire Tactical Reserve (EUFFTR). This fleet is composed of two Canadair CL 215 fire-fighting aircrafts stationed in Corsica. In the event of major forest fires and under request of a given country, the EUFFTR is deployed to support fire-fighting operations. If multiple concurrent major fires occur in different parts of Europe and two or more countries request EUFFTR assistance, the MIC has to evaluate risk and potential population affected to determine tasking priority. To aid this decision-making EFFIS provided and maintains an easy-to-use feature rich web application, which combines information from a number of sources.

5.2. Forest Fire Hot Spot Detection by VGI Sensing

This section describes how the current system can benefit in the future of VGI integration, thus becoming a VGI-enabled EFFIS. As shown in the description above, EFFIS operates on remote sensing information and reference SDI databases to provide situation awareness about major forest fires in Europe and support decisions for tasking aerial fire fighting capabilities. Current version of EFFIS does already of information available through the web, by completing fire-related assessment through consultation of online news.

In this context, VGI can be seen a rich and complementary source of information for the purpose of identifying active fires. As described in section 4.2, we propose to use Sensor-Web VGI to provide EFFIS with a dynamic and organised stream of VGI describing forest fire events. In other words, we suggest the creation of VGI-sensed hot spots to complement the hot spots obtained through remote sensing. In practice, this can be done, by triggering an automatic alert when threshold has been reached in the size of VGI clusters related to forest fires.

In addition to its potential contribution to active fires detection, VGI can be used to further document and characterise detected fire events, thus supporting the second major EFFIS dataset: burnt areas mapping. In this perspective, VGI can be seen as a powerful complement to traditional geographic information layers, and to the News

aggregation system already in place. Twitter feeds, citizens' contributed photos, personal blog articles, etc. would allow for an even richer picture of the fire situation which includes in one single view modelled fire risk, space observed data, and ground reported information. This could be of particular interest when fires are threatening populated zones, where VGI can be used to depict the situation the local population is faced with.

This example shows how VGI sensing can be used to support activities related to forest fires detection and characterisation, thus acting as an additional sense of DENS to enhance perception of a particular type of crisis event. In the following section, opportunities and constraints for a VGI-enabled EFFIS are identified, thus providing directions for other possible DENS implementation scenarios.

5.3. Opportunities and Constraints for a VGI-enabled EFFIS

The description of a VGI-enabled EFFIS illustrated how remote sensing and VGI sensing and SDI components can act as complementary senses supporting a crisis-related scenario[1]. By proposing an integrated view of the situation based on these DENS senses to support decision making, the proposed solution addresses an essential requirement of crisis-related information systems, i.e., to"make sense of a very dynamic stream of information" (De Groeve et al. 2010).

In EFFIS, added value is created on top of sensory information by using reference data bases, e.g. to check if a detected hot spot is situated in a forest area, or to assess the potential impact of a detected forest fire. SDIs typically provide access to stable, trustable data source that can be used as in-memory knowledge in the processes of perception. In addition, the distributed, decentralized nature of SDIs allows gathering information that can have a certain level of dynamicity even if not collected trough sensors (e.g. state of impacted infrastructures provided trough a Regional SDI), thus providing useful mechanisms to unlock information under responsibility of public authorities (Annoni et al. 2010).

[1] To be noted, this example did not include example of use of in-situ sensors (e.g. air quality or weather monitoring stations) that embody another important possible sense for DENS.

As noted above, EFFIS provides data visualization as web services following well-known SDI standards. We suggested further use of SDI standards, such as SOS for exposed added-value products as Observations and Measurements, and SAS, to trigger alerts to subscribed clients. Such standards adoption will create wide opportunities for other crisis information systems to inter-operate with EFFIS, thus further implementing the Next Generation Digital Earth vision as a "multiple connected infrastructures based on open access and participation across multiple technological platforms" (Craglia et al. 2008).

However, while standards compliance and provision of web services guarantees openness, information in a sensitive field such as crisis management should be distributed carefully. A trade-off must be found between openness and reactivity from one side, and security and confidentiality from the other side.

Another constraint for standards adoption relies on their evolving nature, as co-existence of various versions of similar standards can entail lesser interoperability of systems. This could be the case, for example, for SWE standards, as a new version (SWE 2.0) is still underway.

6. Conclusions and Future Works

In this chapter we proposed the Digital Earth's Nervous System as a new metaphor to extend the Digital Earth vision and to provide more dynamic capabilities that fulfill specific requirements of crisis management information systems. In complement to *in-situ* sensor devices and remote sensing platforms, we identified Sensor Web Enablement of VGI as an essential step towards the implementation of a Digital Earth Nervous System. Indeed, SWE provides a framework for integration of VGI in expert-driven systems in a timely way, providing a good synthesis of the situation to decision makers, while complying with reasonable level of quality control that is expected in this context.

We coined the term VGI Sensing to designate the set of standards, methods and techniques required to streamline geo-referenced contents published online by citizens as a new sense for the Digital Earth's Nervous System. Future works are required in this novel

research field. Firstly, specific techniques should be developed to crawl, retrieve, filter and coherently organize heterogeneous VGI data from the web, in order to capture a digital image of stimuli provided by citizens. Secondly, advanced pattern recognition and knowledge discovery methods are required to automatically interpret and characterize events from VGI, taking into account the specific semantics of such user-generated information. Thirdly, VGI sensing should include calibration methods to improve quality assessment and ranking of aggregated items, while considering socio-economic, sociological and cultural factors that shape event-related VGI creation and publication.

Interestingly, it should be noted that such factors triggering VGI postings could also be influenced, in a form of two-ways VGI communication. For the sake of clarity and focus, the VGI Sensing applications envisaged in this research were exclusively one-way. In other words, the focus was on the collection and processing of readily available VGI in a 'waterfall' workflow, without any interaction with VGI producers. However, the same way a satellite can be tasked to take high-resolution images of a specific phenomenon (e.g. a major oil spill), citizens can be requested to produce specific type of VGI to address a specific crisis management need.

Such pattern has been applied by See et al. (2015) which have provided online and mobile applications called - 'Geowiki' - to citizens in order to obtain field images at specific geographic locations in order to validate land cover and land use data. Salk et al. (2015) even conducted a two-ways VGI experiment based on gamification, by inviting citizens to participate to an image classification endeavors aiming at identifying croplands locations in the form of a game.

In parallel to such organized VGI collection campaigns, recent events show several examples of spontaneous two-ways VGI, where citizens with no formal authority encourage each other to post VGI following specific conventions. For example, the case of the Pukklepop Music Festival deadly storm, which was discussed earlier, featured the massive adoption of specific hashtags and geolocation in Tweets in order to match people in distress with people offering help (Opgenhaffen & Smets 2012). Similarly, the recent terrorist attacks in Paris saw the emergence of the hashtag #porteouverte (French for #opendoor) combined with addresses in order to indicate to people

where to find shelter as attackers were still active in Paris' streets (Radanne 2015).

Standards like Sensor Planning Services are specifically designed for tasking sets of technical sensors; further research should study if and how they can be adapted to citizen-sensors interactions, in order to enhance the DENS metaphor presented in this chapter with two-ways VGI Sensing workflows.

By introducing DENS we suggest to move from Digital Earth as a "multi-resolution, three-dimensional representation of the planet" to an active and dynamic multi-dimensional framework able to monitor changes, react to crisis and improve citizens' ability to contribute to situation awareness and decision-making.

Chapter 5 – Conclusions and Perspectives

1. Research objectives and results

This thesis proposed an original contribution to the Earth Observation field coined as *Volunteered Geographic Information (VGI) Sensing*. We defined VGI Sensing as the set of standards, methods and techniques required to streamline geo-referenced contents published online by citizens into a timely, reliable and cost-effective source of Geo-Information for Earth Observation purposes.

In its initial phase, this research aimed at better understanding the nature of VGI, its raw material, thus addressing the first research question identified in the introduction of this thesis:

Which specific informational value does VGI present, that could complement usual sources of geoinformation ? What are the strengths and weaknesses of VGI and its typical Use Cases?

A proof-of-concept developed after a major forest fire near Marseille (France) allowed stressing the informational value of VGI (in its spatial, temporal and semantic dimensions), but also to confirm the challenge its interpretation presents. VGI is by nature subjective, and its lack of control and validation leads to a credibly issue. This Marseille fire Use Case allowed highlighting the echo effect Social Networks can present (this will be discussed in further details in the next section); specifically, the role automated news aggregator's presence on Twitter has been uncovered. The limitations of VGI for early event detection have been also highlighted: in a context were emergency services can be contacted 24/7 through a wide and reliable mobile telephone network, the hyperbole describing VGI as a faster way to detect crisis situations should be relativized, at least for natural hazards events. It also demonstrated that VGI contains valuable information about how the crisis develops in real-time (e.g. progress of the fire towards populated areas) and how impacted citizens cope with the crisis situation.

The second Use Case involving flood events in the UK also confirmed VGI Sensing could provide useful information for situation awareness in crisis response phase, for e.g. the location and evaluation of flood

damages. This positive outcome should not conceal identified shortcomings, however. VGI is sometimes sparse - especially in poorly populated, or socio-economically less developed regions-, and always noisy - e.g. because a specific keyword can have several different meanings. VGI is also subject to several biases due to technology - e.g. smartphone application settings - or culture - e.g. citizens will tend to post more about events with wide media coverage (this will be discussed in further details in the next section).

After real-life applications allowed to characterise the value and limitations of VGI, the next question this research aimed to address was:

In a data overload context, what strategy could allow to tackle the credibility issue VGI is facing ?

This research brought an original answer to this question, and has been among the precursors applying Data Mining and Web Knowledge Discovery principles and techniques to user-contributed information from the Internet with a specific focus on its spatiotemporal dimension. While mainstream VGI efforts highlighted endeavours of highly trained and motivated volunteers (e.g. OpenStreetMap) or extended the role of volunteers from contributors to validators (e.g. WikiMapia), this research proposed a third way. By applying cross-validation mechanisms to vast amounts of VGI, it aimed at turning the challenge of data abundance into an opportunity.

Algorithmic foundations of such cross-validation mechanisms were pre-existing; e.g. Machine Learning models proved to be effective for filtering poorly relevant items from large datasets, and spatiotemporal clustering techniques were used successfully for aggregating co-occurring items in fields like crime analysis or epidemics. This research, however, endeavoured to tackle a question unaddressed so far – which is this thesis' third research question :

What would be a typical chain of processing for converting VGI into reliable geoinformation and what are the research challenges to optimise such workflow ?

Based on the understanding of VGI acquired through real-life applications described above, and having in mind the requirements of

decision-support information systems about natural hazards, this research has filled the gap between the two by designing an original VGI Sensing workflow. This consists in a succession of independent but complementary processing steps allowing to collect, format, enrich, filter, cluster and validate VGI in order to convert individual heterogeneous information items into a consolidated geoinformation dataset. A quote attributed to Henry Ford says: "*nothing is particularly hard if you divide it into small jobs*". Our approach followed this piece of advice, and the modular design of this workflow allowed individuating specific research questions, each one contributing to solve the wider issue of VGI credibility and over-abundance.

Such research questions were numerous, and covered each step of the VGI Sensing workflow (see section 3 for an overview of possible future research questions for each step). For example, how the collection process could be generic enough to cover a wide number of Social Media platforms (Twitter, Flickr, Youtube, Foursquare, Instagram, …) into a single VGI data model while taking into account the meaningful specificities of each (e.g. picture pools on Flickr, hashtags on Twitter)? And how to setup filters that can in the same time address the echo (VGI items duplication) issue, understand all the *nuances* of human language (e.g. Toby Flood is a rugby player, not a natural hazards event), and avoid being too specific (over-fitting) or too restrictive (i.e. avoiding muting relevant parts of the VGI signal)? The question of aggregating individual VGI items into candidate events, of which relevance can be assessed as a whole has been considered as central in this research. Indeed, the possibility to correlate (spatially, temporally, but also semantically) VGI items as a mean to address their individual credibility issue by cross-validation is a core feature of VGI Sensing; therefore, the fourth research question:

Specifically on the Clustering step of the VGI Sensing Workflow, what spatiotemporal clustering algorithm would provide the most satisfactory results with heterogeneous but semantically rich VGI ?

This research has then endeavoured to provide algorithmic advances on the question of clustering VGI; in particular, this research has explored how the spatial, temporal and semantic dimensions can be jointly exploited in specific, tailor-made VGI Sensing algorithms.

Compared to datasets on which spatiotemporal clustering algorithms are usually applied (e.g. crime statistics, disease occurrences, wildlife observations), VGI is heterogeneous (e.g. the precision of their spatial reference can vary) but semantically rich (i.e. they usually contain human language qualifying the observation). A benchmark of existing algorithms with various carefully designed parameter sets was performed, based on the characterisation of Forest Fires in North America with VGI from Flickr. This research highlighted that properly parameterised DB Scan algorithm can outperform the state-of-the-art SatScan Space-Time permutation. It also highlighted that the added value of the semantics dimension is mostly on pre- and post-processing of VGI, while its contribution to clustering *sensu stricto* seemed limited (due to a phenomenon known as *curse of dimensionality*).

Beyond the four research questions, this research has finally depicted a wider perspective for VGI Sensing in the context of the Digital Earth, and envisioned its combination with the other Earth Observation 'senses' at disposal. This thesis introduced the *Digital Earth's Nervous System* (DENS) as a conceptual and technical framework to integrate VGI Sensing, Remote Sensing, *in-situ* Sensing and expert surveys into a coherent situation-awareness system. Inspired by the human's perceptual and nervous system, we designed the DENS as an advance to the Digital Earth metaphor described by Al Gore in 1998.

Interestingly, the interest of the Earth Observation community – which focused mainly since last decades on satellite Remote Sensing – is currently shifting towards an integration of data sources analogous to the DENS vision. Most notably, the European Space Agency (ESA) recently invited Earth Observation scientists to consider VGI, crowdsourcing and citizen science in future applications[1], while partnering in the same time with Big Data and Cloud computing industrial actors to adopt new paradigms in Earth Observation applications[2].

---

[1] See http://eoscience20.org/ for details about ESA's Earth Observation Science 2.0 initiative

[2] See http://www.eo21.org/ict-for-eo-alliance/ for details about the ICT4EO alliance created in the frame of ESA's *Earth Observation for the 21st Century initiative*.

2. Discussion

2.1. Limitations of VGI

This research reached its objectives by demonstrating how volunteered online geoinformation from citizens can be converted into a valuable data source for Earth Observation purposes.

To harness the *wisdom of the crowds*, however, one should be well aware of the caveats of citizen-sensing. The following paragraphs highlight the most prominent of such limitations, in a view of benefiting future VGI Sensing endeavours with key issues uncovered in this research.

Firstly, it must be acknowledged that performing VGI Sensing is like aiming at a moving target. Since the earliest works of this research, numerous VGI platforms (also known as Location-Based Social Networks or LBSN) have been created and several have succeeded in challenging the prominent position of older ones (for example, who remembers MySpace?). This does not only poses technical challenges for VGI Sensing – f.i., adapting retrieval code to novel APIs. It also contributes to the evolution of the *meaning* of VGI, since each platform tends to create its specific cultural bias. We observed, for example, that Flickr being designed for photography enthusiasts, significant number of forest fire-related pictures was posted because of their aesthetical value (e.g. "the haze of a distant forest fire at sunset"). The same event would probably have a very different coverage in an LBSN promoting instant VGI with a lighter tone (e.g. Instagram or Snapchat). Furthermore, new platforms means new rules and habits for its users; for example, hashtags seem nowadays an obvious grammar for self-organisation of volunteered content, but what novelties will the next generation of platforms (or the next version of Twitter) bring? With permanent innovation in underlying online services, VGI Sensing research has to adapt constantly to new technologies and usage patterns.

Chapters 2 and 3 showed occurrences of the *echo effect* on social networks, where people tend to post information from each other (e.g. the 'retweet' mechanism is a core feature of Twitter). Since one of the key VGI Sensing intuitions is cross-validation of information from numerous sources, it is tempting to measure information credibility by

simple metrics like number of co-related (in space, time and semantics) VGI items. However this may lead to take the extent of a rumour (or 'buzz') as a feature of its trustworthiness. Such mistake is a typical problem of our time, were even seasoned politicians and journalists tend to give more credit to many people's gossiping instead of chasing the scarce valuable trusted sources. VGI Sensing should contribute to identify, in an application-specific manner, primary sources of relevant VGI, and filter out the VGI contributing to blur the picture captured by citizen-sensors.

Chapter 4 proposed an ambitious metaphor, which describes VGI Sensing as one of the 'senses' of the Digital Earth, analogue to the output of Remote Sensing or of networks of in-situ sensors. However, each metaphor has its limits: citizen-sensors cannot be compared to sensitive electronic components, especially when the question of calibration is considered. Indeed, a given phenomenon (e.g. a surface on the Earth reflecting electro-magnetic waves from the visible spectrum) can be translated to data in a stable and predictable manner by components having the same technical specifications. Oppositely, the subjective perceptions of citizen-sensors and their motivation to relay (parts of) them publicly online answer to a combination of socio-cultural factors that would be highly complex to model. For example, a major forest fire taking place in a scarcely populated area causing high environmental damage and some infrastructure or property losses may attract less VGI coverage than a much smaller one impacting the emotional state of some Hollywood celebrity.

Whereas subjectivity is an inevitable factor while dealing with human perception, deliberate manipulation can also present an important challenge to VGI Sensing. A recent article from Der Spiegle (Bidder 2015) alleged that the Russian government invested important resources to create an army of paid internet users (called 'trolls') posing as westerner citizens spreading pro-kremlin views on social media as a 9-to-5 job. In sensitive applications like those involving civil unrest, how a VGI Sensing process would discriminate trustable local information from fraudulent contents intended as a manipulation from hostile agents? Such social media manipulation risk should not be underestimated. By analogy, what would be the impact of a hacking of the MODIS satellite sensor resulting in presenting to European population and authorities the erroneous picture of massive forest fires all over the continent?

Such extreme scenarios raise the question of liability; since VGI Sensing aims to contribute to decision-making via situation awareness, what if it leads to take the wrong decision by wrongly depicting a situation? As in any technological innovation, a risk analysis needs to be performed before usage in operational context, having clearly in mind the limitations such technology inevitably has. Wise usage of VGI Sensing is thus essential: the blind cannot blame the ditch.

2.2. The Ethics of VGI

It is appropriate, in this concluding chapter, to initiate a discussion about the societal acceptability of VGI Sensing.

This research is rooted in the very end of the 2000's, when the *Zeitgeist* was overall positive about the transformative value of Information Technologies (IT). Mainstream media relayed consistently hyperboles depicting how IT will improve our societies and ultimately, us. The web 2.0 was described as an unprecedented enabler of freedom of speech, potentially turning every citizen into an honest journalist ready to report and fight blows to democracy, justice and fairness across the globe (Eudes 2009). Pervasive computing would allow building *smart cities*, where frustrations of the urban life (traffic congestion, pollution, insecurity) would be drastically diminished thanks to an optimised management of urban utilities and services based on real-time data flows (IBM 2010). At individual level, the combination of smartphones and specific sensors would allow the emergence of a *Quantified Self*, where our health, efficiency and emotional balance would benefit from fine-tuned algorithms transforming our personal data into to wisdom (Wolf 2010).

As a sign of the times, it did not take long before this grace period (or *hype* in modern terminology) decreases and criticism arises about the digitalisation of our lives. Interestingly, it is not a stoical tendency to contradict bold optimistic statements, nor a failure to deliver its promises that generated such disenchantment: it is the misuse of the transformative power of IT for mass surveillance purposes (Crampton 2015). Although the US Government is not the only actor allegedly intruding in our digital lives - see e.g. (Miller 2014; Venkataramanan 2014; Schneier 2015; Leloup 2015)-, the Snowden revelations about

mass surveillance by its National Security Agency (NSA) (Greenwald 2013) crystallised in the common imagery as the archetype of such misuse.

Concerns about privacy in a technological world are not recent, however. Joinet (1979) warned about the "freedom-destroying trap of informatics"[1], 30 years before privacy advocates use similar arguments against Big Data and Cloud Computing - see, e.g. Foucart (2008). In his visionary novel *Nineteen Height Four*, first published in 1948, Georges Orwell depicted the fictitious combination of totalitarianism and pervasive technology in disturbingly modern terms (Orwell 1948):

> *"[…] in the past no government had the power to keep its citizens under constant surveillance. The invention of print, however, made it easier to manipulate public opinion, and the film and the radio carried the process further. With the development of [telecommunications], and the technical advance which made it possible to receive and transmit simultaneously on the same instrument, private life came to an end."*

Recent revelations coined as the 'Hacking Team Leaks' (Hern 2015) substantiates this disastrous anticipation: it is now established that a private company had developed a technical solution able to transform targeted smartphones into surveillance instruments able to covertly *receive and transmit* information such as sound, image, data and location, and that this solution has been sold to repressive regimes, which meant their citizen's *private life came to an end*.

In this context, it seems healthy to question the societal acceptability of VGI Sensing. Indeed a critical mind may wonder if it is acceptable to massively compile information from citizens into a new dataset without any specific form of consent from them and without providing them any feedback about the information they contributed to. We would like the reader to forge its own opinion on the matter. Therefore, the following paragraphs do not aim to provide comprehensive answers, but rather 'food for thought' on the legal and ethical aspects of VGI Sensing.

[1] The original title, in French, is "*Les 'pièges liberticides' de l'informatique*"

From a legal perspective, first: the most relevant field of legislation seems to be the one on Personal Data Protection. The European Directive 95/46/EC (European Commission 1995 [1]) is probably one of the most stringent piece of legislation on the matter worldwide (Bellanova & De Hert 2009). It foresees – among others – that the Data Subject (i.e. the natural person about which the data in question can be used to identify, even indirectly) has the right to be informed and to object about any processing operation applied to his/her Personal Data. In particular interest in this context, the Data Subject should be notified if his/her Personal Data is used in a different purpose that the one for which they were initially collected. If Social Media postings were considered as Personal Data, this would be an issue for VGI Sensing – as well as for all third party applications interacting with Social Media platforms *via* their API. But legally speaking, it seems that our Tweets, Flickr pictures and other online postings are not considered as Personal Data *sensu stricto* whatever the degree of intimacy they can contain (Claypoole 2014).

Without exploring the question in detail, few elements could support the idea, at least for the two platforms used in this research (namely: Twitter and Flickr), that Social Media postings can indeed legally be considered as public information rather than Personal Data:

- Twitter and Flickr can be used anonymously; there is no obligation to post on its real name, and indeed the vast majority of users choose a pseudo-name and do not disclose their real identity (the exception being often people willing to promote their professional activity – e.g. as politicians, as journalists or photographers). It must be stressed, however, that the ability to combine anonymised data in order to link them undoubtedly to a unique natural person is a direct threat to privacy in a Big Data context (Venkataramanan 2014). As a consequence, the notion of anonymous use of Social Media can be subject to discussion.
- Both services state very clearly in their Terms of Use that the user's postings are "public by default and will be able to be viewed by other users and through third party services and websites" (sic) (Twitter 2015), while Yahoo, who operates the Flickr service, restricts the scope of their Privacy Policy to

[1] The legislative process aiming at replace this directive by a renewed legal text is on-going.

> "Registration Data and certain other information", i.e. not the Social Media Postings, for which there might be "transmissions over various networks; and changes to conform and adapt to technical requirements" (Yahoo 2015).

The fact that an operation appears to have no obvious legal objections does not mean it is fair, however; VGI Sensing's social acceptability should be considered from ethical angle too. The ethical perspective can be approached in two complementary ways:

- from 'downstream', the question would be whether the purpose of the VGI Sensing end product would be acceptable for Social Media users, since they contributed – even if unknowingly – to it;
- from an 'upstream' point of view, one may question the *Volunteer* dimension of VGI Sensing since VGI is collected without their explicit consent.

On the 'downstream' question, it must be stressed that the purpose of the VGI Sensing use cases envisaged in this research aim supporting Disaster Management processes (e.g. impact assessment, response, recovery) and therefore may sound more acceptable to the citizen than Social Media monitoring endeavours for e.g. marketing purposes (Venkataramanan 2014) or surveillance in a social unrest context (Wilkin & Balali 2015). Nevertheless, the legitimacy of a purpose is always a matter of interpretation - for example, totalitarian regimes tend to call their opponents as 'terrorists' (Vidal 2015). As a consequence, every VGI Sensing endeavour should adopt a code of conduct of not 'betraying' the intentions of citizens publicly sharing their views of the world. This can only be done on a case-by-case basis and would always contain a certain degree of subjectivity.

The 'upstream' question relates to the *Volunteered* dimension of VGI. Craglia, Ostermann & Spinsanti (2012) introduced the distinction between *explicitly volunteer* (or 'active') and *implicitly volunteer* (or 'passive') VGI. A typical example of explicitly volunteer VGI would be OpenStreetMap (Chilton 2009) where GPS enthusiasts build collaboratively reference cartographic datasets. Oppositely, when a geographic dataset is inferred from publicly available information from citizens, it will be referred to as *implicitly volunteered*, since citizens shared willingly information, but were not aware of its reuse for a specific purpose. VGI Sensing falls clearly in the second

category, for which it is important there is no ambiguity about the public nature of VGI contributions. Indeed, the fact that a picture is publicly available does not mean that it reflects the intention of its originator. Leaving aside acts of criminal nature (Dewey 2014), a typical reason for publishing unwillingly VGI contributions is by neglect to configure properly a smartphone's Social Media application (e.g., Instagram), which may lead to unwanted publication of geolocated contents with potentially unpleasant consequences such as burglary (Connolly 2015). Nerveless, although tracking Social Media contributions from a single user may clearly lead to privacy concerns (Shubber 2013), VGI Sensing has an opposite focus, i.e. on a large number of user's contribution without seeking to unveil anyone's identity and whereabouts. In any case, the anonymisation of user's contributions should be applied as a minimum safeguard within any VGI Sensing process.

While aiming to harness the transformative power of Information and Communication Technologies for socially acceptable purposes (i.e. Disaster Management), VGI Sensing is exposed to the risk of potentially illegal or unethical misuse – as any innovation. In this context, the quote of the British scientist and novelist C.P. Snow seems to apply to modern ICT (Snow 1971):

> *"Technology [...] is a queer thing. It brings you great gifts with one hand, and it stabs you in the back with the other."*

With our lives being increasingly digital, the consequences of technology misuse can be devastating for individuals, even lead to suicide (Chew 2015). Every computer scientist nowadays should take particular care of the ethics of the endeavours they contribute to, and ensure the result of their work does not *stab* anyone *in the back*.

3. Future Research

As any research aiming at pioneering to some extent in its specific domain, this research has probably raised at least as much new questions as it has solved existing ones. Suggestions for further works have been provided at the end of each chapter, however we summarise and discuss key ideas in the next paragraphs, while Figure 25 gives an overview of what could be a draft research agenda for VGI Sensing.

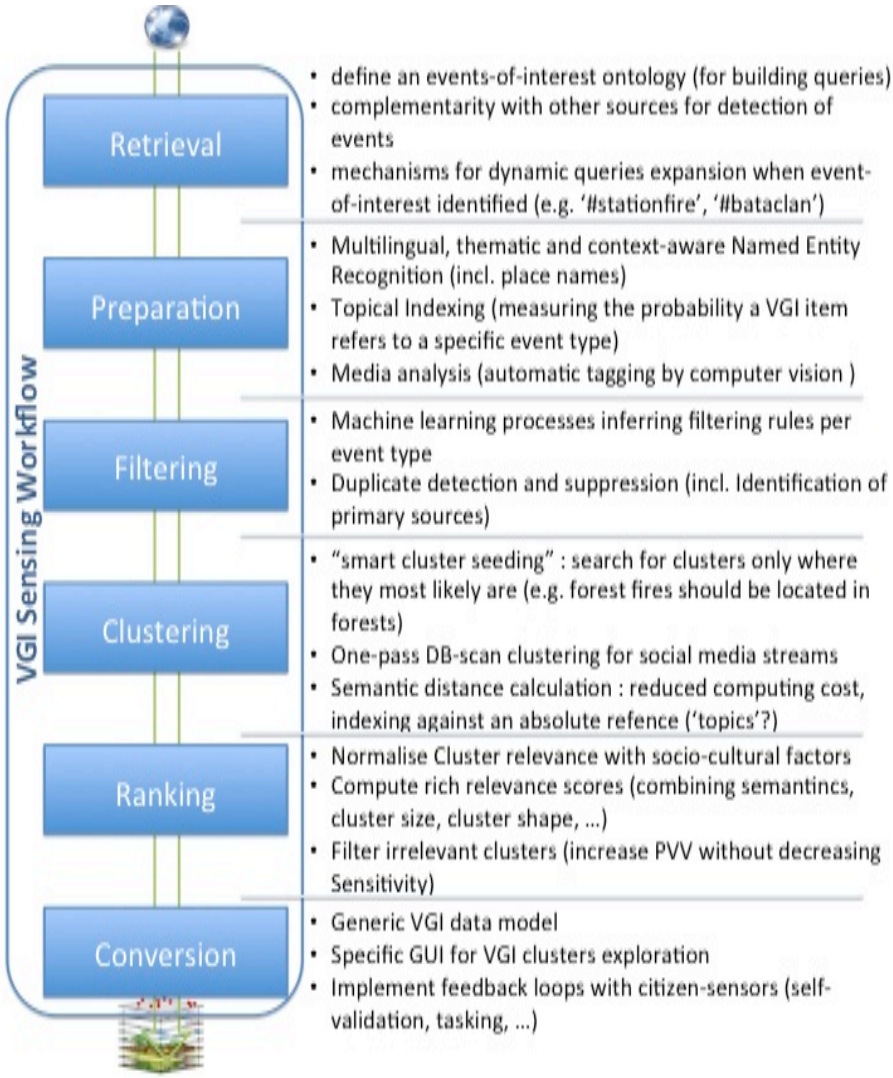


Figure 25: a research agenda for VGI Sensing

On the retrieval step, firstly, the question of VGI sources should also benefit further works. This research focused on Twitter and Flickr, which were the two main Social Media platforms broadcasting publicly available user contributions when our main VGI datasets have been collected. But VGI is by its own nature a moving target, and VGI collection efforts should constantly endeavour to include new leaders (e.g. Instagram[1], which surpassed Flickr for picture

[1] https://instagram.com/

sharing thanks to its smartphone-first approach) and emerging challengers (e.g. Yiha[1], which interestingly responds to privacy concerns by offering to post anonymous geolocated messages). To this end, further research should aim for generic VGI data model and for adaptable collection modules that would allow including new sources almost effortlessly, by design – although this would not address the 'cultural shift' limitation discussed in previous section.

On the preparation step, secondly, this research has stressed how the various dimensions of VGI can jointly contribute to the value of the final outcome, but has ignored two dimensions: the social dimension and the 'pixels' dimension. For the social dimension, the omission was intentional. As explained in the previous chapters, we decided not to follow the example of many Social Media Analysts who studied in the greatest details the social dynamics of online postings - mostly with marketing applications in mind. We argued that their conceptual framework of dividing a crowd of users into sub-networks with specific 'roles' (e.g. *trendsetter*, *influencer*, *groupie*, etc.) was not applicable to crisis situations where networks and personal relationships are re-shuffled extremely fast as the events occur. The 'pixels' dimension (i.e. the visual contents of VGI: pictures and videos) was only ignored by the lack of specific means. As said in the previous chapters, we think that analysis of the media contents (image or video) of VGI items with appropriate methods (e.g. to verify the presence of smoke or flames in a forest fire context) would be extremely beneficial to the interpretation of the meaning of VGI items – which is currently solely based on semantics contained in the textual (meta)data. In this respect, it would be very interesting to see an interdisciplinary research endeavour applying Computer Vision and Natural Language Parsing to advance VGI Sensing methods.

Speaking about Natural Language Parsing – and this is our third highlight for further research – we must acknowledge that this research took some shortcuts when analysing text contained in VGI items. The filtering step, for example, consisted is a rather simple scoring method where a Machine Learning algorithm deducted from a training sample, within the text of VGI items, the positive features (e.g. frequency of a wanted keyword like 'smoke') and the negative features (e.g. presence of an unwanted keyword like 'campfire', or the

[1] http://yiha.me/

presence of a past date) to take into account. Similarly, the semantic distance between two VGI items was calculated in an almost rudimentary way – by calculating co-occurrence of keywords, weighted by their frequency in the entire dataset. This thesis is clearly the contribution of a geomatician willing to enrich its research with concepts and methods from other fields. It appears that a next step should involve experts from the Knowledge Discovery field in order to further develop the analysis of the semantic dimension, as a complement to the spatial and temporal ones.

Fourthly, the real-timeliness of VGI Sensing should be further explored. Although the VGI Sensing workflow has been designed with real-time applications in mind, the use cases proposed in this thesis all involved retrospective analysis. This is due mostly to practical reasons, and it did not weaken the overall theoretical – since based on the intrinsic characteristics of single VGI items which do not vary over time -, or practical value – since retrospective analysis is useful in many crisis management situations, as detailed in previous chapters. Nevertheless, specific use cases may require (near) real time processing of VGI – we think for example of use cases involving mutual assistance between citizens in the aftermath of a disaster, and where online self-organisation efforts should be supported by reliable streams of information. In addition, real-time VGI Sensing methods could appropriately contribute to the very active field of Social Media Streams Analysis, which often tends to ignore the spatial dimension of the information. We have discussed in Chapter 3 how the proposed VGI clustering algorithm could be adapted to the one-pass requirement of real-time processing, and we are confident further research could run, with such design, successful real-time VGI Sensing applications.

The fifth highlight concerns the combination of VGI Sensing with other sources of information – most notably with satellite and in-situ sensors. Chapter 4 widely describes – both in conceptual in technical terms – why and how such combination could be beneficial; further research could support this vision with concrete use cases. For example, disaster alerts based on real-time satellite image analysis (e.g. thermal hot spots for forest fires or abnormal reflectance changes for floods) or in-situ sensors (e.g. specific buoys for tsunamis) could serve as a trigger for specific VGI Sensing processes to be executed. This could be done in the context of existing monitoring systems run

by public authorities, like the GDACS or EFFIS that were cited in previous chapters. Such triggered processes could involve richer interactions with citizens, aiming for, e.g. self-validation or on-request data collection ('tasking').

4. Closing note

It is frequent in information science to refer to the DIKW pyramid in order to represent the relations between raw *Data*, the *Information* that can be inferred from their careful analysis, the *Knowledge* of the real world such information enables, and ultimately the *Wisdom* we can implement in our knowledge-based decisions (Wallace 2007).

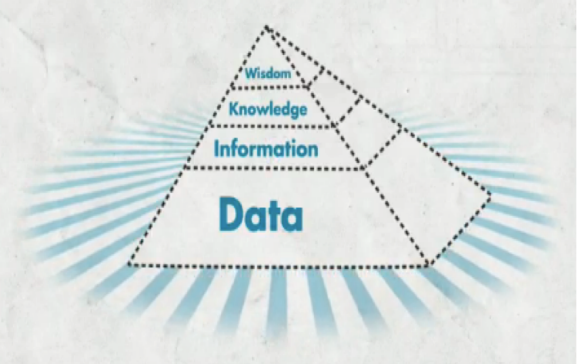


Figure 26: the DIKW pyramid

To many extents, this research gave us the feeling to literally climb the DIKW pyramid, certainly not until the top, but undoubtedly following an interesting path related to volunteered online contributions with a spatial dimension, a path we named *VGI Sensing*. We hope that our peers will find in this thesis at least a map of such path, which can facilitate their own climbing endeavour. We are pleased to imagine that some of them, while struggling to pass a steep section, will find adequately one of the algorithmic hooks we left here and there, and will rely on them in order to get to the next level, safe and upward looking.

Author's list of publications

Peer-reviewed Publications

De Longueville, B., Ostländer, N. & Keskitalo, C., 2010. Addressing vagueness in Volunteered Geographic Information (VGI)–A case study. International Journal of Spatial Data Infrastructures, Volume 5.

De Longueville, B., 2010. Community-based geoportals: The next generation? Concepts and methods for the geospatial Web 2.0. Computers, Environment and Urban Systems, 34(4), pp.299–308.

De Longueville, B., Luraschi, G., Smits, P., Peedell, S., De Groeve, T., 2010. Citizens as sensors for natural hazards: a VGI integration workflow. Geomatica, Special issue on VGI(64), pp.41–59.

De Longueville, B., Annoni, A., Schade, S., Ostlaender, N., Whitmore, C., 2010. Digital Earth's Nervous System for crisis events: real-time Sensor Web Enablement of Volunteered Geographic Information. International Journal of Digital Earth, 3(3), pp.242–259.

Schade, S., Diaz, L., Ostermann, F. Sprinsanti, L.. Luraschi, G., Cox, S., Nuñez, M. , 2013. Citizen-based sensing of crisis events: sensor web enablement for volunteered geographic information. Applied Geomatics, 5(1), pp.3–18.

De Longueville, B., Chevalier, J.-F., 2015, What, When, Where » a clustering algorithm for event characterisation with Volunteered Geographic Information. International Journal of Geographical Information Science (submitted)

Conferences and workshops

De Longueville, B., Smith, R.S. & Luraschi, G., 2009. 'OMG, from here, I can see the flames!': a use case of mining location based social networks to acquire spatio-temporal data on forest fires. In

Proceedings of the 2009 International Workshop on Location Based Social Networks. Seattle, Washington: ACM, pp. 73–80.

Ostlaender, N., Smith, R.S., De Longueville, B., Smits, P., 2010. What Volunteered Geographic Information is (good for) - designing a methodology for comparative analysis of existing applications to classify VGI and its uses. In Geoscience and Remote Sensing Symposium (IGARSS), 2010 IEEE International. Geoscience and Remote Sensing Symposium (IGARSS), 2010 IEEE International. pp. 1422–1425.

De Longueville, B., Annoni, A., Schade, S. & Ostlaender, N., 2010. Digital Earth's nervous system and Volunteered Geographic Information sensing: towards a self-aware planet. In Proceedings of the 3rd ISDE Digital Earth Summit. 3rd ISDE Digital Earth Summit. Nessebar, Bulgaria.

De Longueville, B. & Hardy, M., 2010. Clustering data with heterogeneous spatiotemporal reference: towards web-mining of event-related knowledge. In Proceedings of the 6th GI Science Conference, September 2010, Zürich, Switzeland.

Schade, S., Luraschi, G., De Longueville, B., Cox, S., Diaz, L., 2010. Citizen as Sensors for Forest Fires: Sensor Web Enablement for Volunteered Geographic Information. In Proceedings of the 2010 ISPRS WebMGS Workshop. WebMGS 2010: 1st International Workshop on Pervasive Web Mapping, Geoprocessing and Services. Como, Italy.

References

Aggarwal, C.C. & Subbian, K., 2012. Event Detection in Social Streams. In SDM. SIAM, pp. 624–635.

Ankerst, M. et al., 1999. OPTICS: Ordering Points to Identify the Clustering Structure. In Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data. SIGMOD '99. New York, NY, USA: ACM, pp. 49–60.

Annoni, A. et al., 2010. Earth Observations and Dynamic Mapping: Key assets for risk management. In M. Konecny, S. Zlatanova, & T. L. Bandrova, eds. Geographic Information and Cartography fore Risk and Crisis Mangement. Lecture Notes in Geoinformation and Cartography. Springer-Verlag Berlin Heidelberg, pp. 3 – 22

Annoni, A., 2004. Towards a European spatial data infrastructure: the INSPIRE initiative. Proceeedings of the 7th international global spatial data infrastructure conference, Bangalore, India, p.11.

Anon, 2009. Twenty years of the world wide web: What's the score? The Economist. Available at: http://www.economist.com/sciencetechnology/displayStory.cfm?story_id=13277389 [Accessed July 20, 2009].

Anselin, L., 1989. What is Special about Spatial Data?: Alternative Perspectives on Spatial Data Analysis, National Center for Geographic Information and Analysis.

Armbrust, M. et al., 2010. A view of cloud computing. Communications of the ACM, 53(4), pp.50–58.

Atefeh, F. & Khreich, W., 2013. A Survey of Techniques for Event Detection in Twitter. Computational Intelligence, 31(1).

Attneave, F., 1959. Applications of information theory to psychology: a summary of basic concepts, methods, and results., Holt, Rinehart, and Winston (New York).

Bahree, M., 2008. Citizen Voices. Forbes Magazine. Available at: http://www.forbes.com/free_forbes/2008/1208/083.html [Accessed December 10, 2009].

Ballatore, A. & Bertolotto, M., 2011. Semantically Enriching VGI in Support of Implicit Feedback Analysis. In K. Tanaka, P. Fröhlich, & K.-S. Kim, eds. Web and Wireless Geographical Information Systems. Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 78–93.

Becker, H., Naaman, M. & Gravano, L., 2011. Beyond Trending Topics: Real-World Event Identification on Twitter. ICWSM, 11, pp.438–441.

Bégin, D., Devillers, R. & Roche, S., 2013. Assessing Volunteered Geographic Information (VGI) Quality Based on Contributors'Mapping Behaviours. ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 1(1), pp.149–154.

Bellanova, R. & De Hert, P., 2009. Protection des données personnelles et mesures de sécurité : vers une perspective transatlantique. Cultures & Conflits, (74), pp.63–80.

Benson, E., Haghighi, A. & Barzilay, R., 2011. Event discovery in social media feeds. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics, pp. 389–398.

Berners-Lee, T. et al., 2001. The semantic web. Scientific american, 284(5), pp.28–37.

Bhattacharyya, P., 2015. Machine Translation, CRC Press.

Bidder, B., 2015. Paid as a Pro-Kremlin Troll: 'The Hatred Spills over into the Real World'. *De Spiegle,* June 2015 edition. Available at: http://www.spiegel.de/international/world/interview-with-ex-russian-internet-troll-lyudmila-savchuk-a-1036539.html [Accessed January 21, 2016].

Bimonte, S. et al., 2014. From Volunteered Geographic Information to Volunteered Geographic OLAP: A VGI Data Quality-Based Approach. In B. Murgante et al., eds. Computational Science and Its Applications – ICCSA 2014. Lecture Notes in Computer Science. Springer International Publishing, pp. 69–80.

Bishop, C.M., 2006. Pattern Recognition and Machine Learning, Springer.

Bishr, M. & Mantelas, L., 2008. A trust and reputation model for filtering and classifying knowledge about urban growth. GeoJournal, 72(3-4), pp.229–237.

Block, R., 2007. Scanning for Clusters in Space and Time: : A Tutorial Review of SaTScan. Social Science Computer Review.

Botts, M. et al., 2008. OGC Sensor Web Enablement: Overview and High Level Architecture. Lecture Notes in Computer Science, 4540(December), pp.175–190.

Bradley, A.P., 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition, 30(7), pp.1145–1159.

Brakenridge, G.R. et al., 2007. Orbital microwave measurement of river discharge and ice status. Water Resources Research, 43(4).

Bray, T. et al., 1998. Extensible markup language (XML). World Wide Web Consortium Recommendation REC-xml-19980210. http://www. w3. org/TR/1998/REC-xml-19980210, p.16.

Budhathoki, N.R., Bruce, B. & Nedovic-Budic, Z., 2008. Reconceptualizing the role of the user of spatial data infrastructure. Geojournal, 72, pp.149–160.

Butler, D., 2006. Virtual globes: The web-wide world. Nature, 439, pp.776–778.

Cardwell, S., 2009. A Twitter Timeline of the Iran Election. Newsweek Web Edition.

Cheng, A., Evans, M. & Singh, H., 2009. Inside Twitter. An In-Depth Look Inside the Twitter World.

Cheng, T. & Wicks, T., 2014. Event Detection using Twitter: A Spatio-Temporal Approach. PloS one, 9(6), p.e97807.

Chew, J., 2015. Ashley Madison Leak: Police Investigating Possible Suicides. Fortune. Available at: http://fortune.com/2015/08/24/ashley-madison-suicide/ [Accessed August 27, 2015].

Chilton, S., 2009. Crowdsourcing is radically changing the geodata landscape: case study of OpenStreetMap. In VV. AA., Proceedings of the 24th International Cartographic Conference (Santiago de Chile, 15-21 November 2009).

Ciobanu, D.-L. et al., 2007. Du Wiki au WikiSIG. Geomatica, 61(4), pp.455–469.

Claypoole, T.F., 2014. Privacy and Social Media. Business Law Today, January 2014.

Cockshott, P., Cottrell, A. & Michaelson, G., 1995. Testing Marx: some new results from UK data. Capital & Class, 19(1), pp.103–130.

Coleman, D.J., Georgiadou, Y. & Labonte, J., 2009. Volunteered Geographic Information: The Nature and Motivation of Produsers. International Journal of Spatial Data Infrastructures Research, 4, pp.332–358.

Connolly, A., 2015. F1 star and wife robbed: Instagram photo tagging could have been to blame. The Next Web. Available at: http://thenextweb.com/opinion/2015/08/07/batten-down-the-hatches/ [Accessed August 24, 2015].

Coppock, J.T. & Rhind, D.W., 1991. The history of GIS. Geographical information systems: Principles and applications, 1(1), pp.21–43.

Corson, M.W. & Palka, E.J., 2004. Geotechnology, the U.S. Military, and War. In S. D. Brunn, S. L. Cutter, & J. W. H. Jr, eds. Geography and Technology. Springer Netherlands, pp. 401–427.

Craglia, M. et al., 2008. Next-generation Digital Earth: A position paper from the Vespucci Initiative for the Advancement of Geographic Information Science. International Journal of Spatial Data Infrastructures Research, 3, pp.146–167.

Craglia, M., Ostermann, F. & Spinsanti, L., 2012. Digital Earth from vision to practice: making sense of citizen-generated content. International Journal of Digital Earth, 5(5), pp.398–416.

Crampton, J.W., 2015. Collect it all: National Security, Big Data and Governance. GeoJournal.

Crandall, D. et al., 2009. Mapping the World's Photos. In Proceedings of the 18th International World Wide Web Conference,. Madrid, Spain, pp. 761–761

Daconta, M.C., Obrst, L.J. & Smith, K.T., 2003. The Semantic Web: A Guide to the Future of XML, Web Services, and Knowledge Management, John Wiley & Sons.

Damasio, A.R. & Sutherland, S., 1996. Descartes' error: Emotion, reason, and the human brain, Papermac London.

D'Andrea, E. et al., 2015. Real-Time Detection of Traffic From Twitter Stream Analysis. Intelligent Transportation Systems, IEEE Transactions on, PP(99), pp.1–15.

Daume, S., Albert, M. & von Gadow, K., 2014. Forest monitoring and social media – Complementary data sources for ecosystem surveillance? Forest Ecology and Management, 316, pp.9–20.

Dawes, S.S. et al., 2004. Information, Technology and Coordination: Lessons from the World Trade Center Response, Center for Technology in Government, University at Albany, SUNY New York, NY.

De Groeve, T. et al., 2010. Mash-up or SDI: appropriate mapping tools for emergency situation rooms. In GI4DM - Geomatcis for Disaster Management. Torino, Italy. Available at: http://www.gi4dm-2010.org/program.php.

De Groeve, T., Annunziato, A. & Vernaccini, L., 2006. Modelling Disaster Impact for the Global Disaster Alert and Coordination System. In B. Van de Walle & M. Turoff, eds. Proceedings of the

3rd International ISCRAM Conference. Newark, NJ, USA, pp. 409–417.

De Groeve, T. & Riva, P., 2009. Global real-time detection of major floods using passive microwave remote sensing. In Proceedings of the 33rd International Symposium on Remote Sensing of Environment Stresa, Italy.

De Longueville, B., Luraschi, G., et al., 2010. Citizens as sensors for natural hazards: a VGI integration workflow. Geomatica, Special issue on VGI(64), pp.41–59.

De Longueville, B., Annoni, A., et al., 2010. Digital Earth's Nervous System for crisis events: real-time Sensor Web Enablement of Volunteered Geographic Information. International Journal of Digital Earth, 3(3), pp.242–259.

De Longueville, B. & Hardy, M., 2010. Clustering data with heterogeneous spatiotemporal reference: towards web-mining of event-related knowledge. In Proceedings of the 6th GI Science Conference, September 2010, Zuerich, Switzeland.

De Longueville, B., Smith, R.S. & Luraschi, G., 2009. 'OMG, from here, I can see the flames!': a use case of mining location based social networks to acquire spatio-temporal data on forest fires. In Proceedings of the 2009 International Workshop on Location Based Social Networks. Seattle, Washington: ACM, pp. 73–80.

De Rubeis, V., Sbarra, P. & Tosi, P., 2009. Web based macroseismic survey: fast information exchange and elaboration of seismic intensity effects in Italy. In J. Landgren & S. Jul, eds. Proceeding of the 6th International ISCRAM Conference, Gothenburg, Sweden, May 2009. ISBN - 978-91-633-4715-3.

Desrosières, A., 2002. The politics of large numbers: A history of statistical reasoning, Harvard University Press.

Deuve, J., 2013. Histoire secrète des stratagèmes de la Seconde Guerre mondiale, Paris: Nouveau Monde Editions.

Dewey, C., 2014. Don't worry about getting hacked. Worry about getting socially engineered. The Washington Post. Available at: https://www.washingtonpost.com/news/the-intersect/wp/2014/10/15/dont-worry-about-getting-hacked-worry-about-getting-socially-engineered/ [Accessed August 24, 2015].

Dynes, R.R., 1970. Organized Behavior in Disaster, Heath LexingtonBooks.

Ellison, N.B. & others, 2007. Social network sites: Definition, history, and scholarship. Journal of Computer-Mediated Communication, 13(1), pp.210–230.

Elwood, S., 2008. Volunteered geographic information: Key questions, concepts and methods to guide emerging research and practice. GeoJournal, 72(3-4), pp.133–135.

Elwood, S., Goodchild, M.F. & Sui, D., 2013. Prospects for VGI Research and the Emerging Fourth Paradigm. In D. Sui, S. Elwood, & M. Goodchild, eds. Crowdsourcing Geographic Knowledge. Springer Netherlands, pp. 361–375.

Ester, M. et al., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In AAAI Press, pp. 226–231.

Etzioni, O., 1996. The World-Wide Web: Quagmire or Gold Mine? Commun. ACM, 39(11), pp.65–68.

Eudes, Y., 2009. Twitter, les pirates et les diplomates. Le Monde.

European Commission, 1995. Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data, Available at: http://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:31995L0046&from=en [Accessed August 24, 2015].

Fast Company, 2003. How Google Grows...and Grows...and Grows. Fast Company. Available at: http://www.fastcompany.com/46495/how-google-growsand-growsand-grows [Accessed May 13, 2015].

Feick, R. & Roche, S., 2013. Understanding the Value of VGI. In D. Sui, S. Elwood, & M. Goodchild, eds. Crowdsourcing Geographic Knowledge. Springer Netherlands, pp. 15–29.

Feiler, J., 2007. How to do everything with Web 2.0 Mashups, McGraw-Hill, Inc. Available at: http://dl.acm.org/citation.cfm?id=1593760 [Accessed May 18, 2015].

Ferster, C.J. & Coops, N.C., 2013. A review of earth observation using mobile personal communication devices. Computers & Geosciences, 51, pp.339–349.

Fielding, R.T., 2000. Architectural Styles and the Design of Network-based Software Architectures. University of California, Irvine. Available at: http://www.ics.uci.edu/ fielding/pubs/dissertation/fielding_dissertation.pdf.

Fischer, 2012. VGI as Big Data. A New but Delicate Geographic Data-Source. Geoinformatics, issue 3, vol. 15, pp.46–47.

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

Foresman, T.W., 2008. Evolution and implementation of the Digital Earth vision, technology and society. International Journal of Digital Earth, 1/1, pp.4–16.

Foucart, S., 2008. Peut-on tout confier a Google? Le Monde 2, 248(Special High Tech), pp.30–40.

Fritz, S. et al., 2009. Geo-Wiki.Org: The Use of Crowdsourcing to Improve Global Land Cover. Remote Sensing, 1(3), pp.345–354.

Gendron, R. & Hoffman, E., 2009. Resource Scarcity and the Prevention of Violent Conflicts. The Peace and Conflict Review, 4(1), pp.37–51.

Getis, A. & Ord, J.K., 1992. The Analysis of Spatial Association by Use of Distance Statistics. Geographical Analysis, 24(3), pp.189 – 206.

Gong, B. et al., 2006. Event Discovery in Multimedia Reconnaissance Data Using Spatio-Temporal Clustering. In Proc. of the AAAI Workshop on Event Extraction and Synthesis (EES'06).

Goodchild, M., 2007. Citizens as sensors: The world of volunteered geography. GeoJournal, 69(4), pp.211–221.

Goodchild, M., 2009. NeoGeography and the nature of geographic expertise. Journal of Location Based Services, 3(2), p.82.

Goodchild, M, 2007. Citizens as Voluntary Sensors: Spatial Data Infrastructure in the World of Web 2.0. International Journal of Spatial Data Infrastructures Research, 2, pp.24–32.

Goodchild, M.F. & Glennon, J.A., 2010. Crowdsourcing geographic information for disaster response: a research frontier. *International Journal of Digital Earth*, 3(3), pp.231–241.

Gore, A., 1998. The Digital Earth: Understanding our planet in the 21st Century. In Speech given at the California Science Center, Los Angeles, California, on January 31, 1998.

Gould, M., 2008. GI field looking the wrong way - again? GEOconnexion International Magazine, dec07/jan08, pp.19–20.

Gouveia, C. et al., 2004. Promoting the use of environmental data collected by concerned citizens through information and communication technologies. Journal of Environmental Management, 71(2), pp.135–154.

Grant, D., 1999. Spatial Data Infrastructures: The Vision for the Future and the Role of Government in Underpinning Future Land Administration Systems. In Proceedings of UN-FIG International

Conference on Land Tenure and Cadastral Infrastructures for Sustainable Development. pp. 94–109.

Greenwald, G., 2013. NSA collecting phone records of millions of Verizon customers daily. The Guardian. Available at: http://www.theguardian.com/world/2013/jun/06/nsa-phone-records-verizon-court-order [Accessed August 24, 2015].

Grossner, K.E., Goodchild, M.F. & Clarke, K.C., 2008. Defining a digital earth system. Transactions in GIS, 12(1), pp.145–160.

GSDI, 2004. Developing Spatial Data Infrastructures: The SDI Cookbook D. D. Nebert, ed. Available at: http://www.gsdi.org/docs2004/Cookbook/cookbookV2.0.pdf [Accessed February 25, 2010].

Hahmann, S., Purves, R. & Burghardt, D., 2014. Twitter location (sometimes) matters: Exploring the relationship between georeferenced tweet content and nearby feature classes. Journal of Spatial Information Science, 2014(9), pp.1–36.

Hand, D.J., Mannila, H. & Smyth, P., 2001. Principles of Data Mining, MIT Press.

Harnad, S., 1987. Psychophysical and cognitive aspects of categorical perception: A critical overview. Categorical perception: The groundwork of cognition, pp.1–25.

Havlik, D., Bleier, T. & Schimak, G., 2009. Sharing Sensor Data with SensorSA and Cascading Sensor Observation Service. Sensors, 9(7), pp.5493–5502.

Hern, A., 2015. Hacking Team hacked: firm sold spying tools to repressive regimes, documents claim. The Guardian. Available at: http://www.theguardian.com/technology/2015/jul/06/hacking-team-hacked-firm-sold-spying-tools-to-repressive-regimes-documents-claim [Accessed August 24, 2015].

Honeycutt, C. & Herring, S.C., 2009. Beyond Microblogging: Conversation and Collaboration via Twitter. In System Sciences, 2009. HICSS '09. 42nd Hawaii International Conference on. System Sciences, 2009. HICSS '09. 42nd Hawaii International Conference on. pp. 1–10.

Hoong, D.C. & Buyya, R., 2003. Guided Google: A Meta Search Engine and its Implementation using the Google Distributed Web Services. arXiv preprint cs/0302018. Available at: http://arxiv.org/abs/cs/0302018 [Accessed May 18, 2015].

Hsu, K.-C. & Li, S.-T., 2010. Clustering spatial–temporal precipitation data using wavelet transform and self-organizing map neural network. Advances in Water Resources, 33(2), pp.190–200.

Huberman, B.A., Romero, D.M. & Wu, F., 2009. Social networks that matter : Twitter under the microscope. First Monday, 14(1).

Hughes, A.L. & Palen, L., 2009. Twitter Adoption and Use in Mass Convergence and Emergency Events. In J. Landgren & B. Van de Walle, eds. Proceeding of the 6th International ISCRAM Conference. Gothenburg, Sweden.

IBM, 2010. Smarter cities for smarter growth: How cities can optimize their systems for the talent-based economy. Available at: http://www-01.ibm.com/common/ssi/cgi-bin/ssialias?infotype=PM&subtype=XB&appname=GBSE_GB_TI_USEN&htmlfid=GBE03348USEN&attachment=GBE03348USEN.PDF [Accessed August 24, 2015].

Igoe, T., 2011. Making Things Talk: Using Sensors, Networks, and Arduino to See, Hear, and Feel Your World, O'Reilly Media, Inc.

Imran, M. et al., 2014. Processing Social Media Messages in Mass Emergency: A Survey. arXiv:1407.7071 [cs]. Available at: http://arxiv.org/abs/1407.7071 [Accessed May 11, 2015].

Intagorn, S., Plangprasopchok, A. & Lerman, H., 2010. Harvesting Geospatial Knowledge from Social Metadata. In S. French, ed. Proceedings of the 7th International ISCRAM Conference. International Conference on Information Systems for Crisis Response and Management. Seattle.

Jain, A.K., Murty, M.N. & Flynn, P.J., 1999. Data clustering: A review. ACM Computing Surveys, 31(3), pp.316–323.

Java, A. et al., 2009. Why We Twitter: An Analysis of a Microblogging Community. In Advances in Web Mining and Web Usage Analysis. pp. 118–138.

Jenks, G.F. & Coulson, M.R.., 1963. Class intervals for statistical maps. International Yearbook of Cartography4, 3, pp.119–134.

Jirka, S., Broering, A. & Stasch, C., 2009. Applying OGC Sensor Web Enablement to risk monitoring and disaster management. In GSDI 11 World Conference. Rotterdam, The Netherlands.

Johnson, D.S., 1980. Planning for a Civil Operational Land Remote Sensing Satellite System: A Discussion of Issues and Options. In Exploring Unknown: selected documents in the History of US Civil Space Program. National Aeronautics and Space Administration, p. 296.

Johnson, S., 2009. How Twitter Will Change the Way We Live. Time. Available at: http://www.time.com/time/business/article/0,8599,1902604,00.html [Accessed July 20, 2009].

Johnson, S.D., 2010. A brief history of the analysis of crime concentration. European Journal of Applied Mathematics, 21, pp.349 – 370.

Joinet, L., 1979. Les « pièges liberticides » de l'informatique. Le Monde Diplomatique. Available at: http://www.monde-diplomatique.fr/1979/03/JOINET/35052 [Accessed August 24, 2015].

Jones, C.B.. P., 2008. Modeling vague places with knowledge from the Web. International Journal of Geographical Information Science, 22/10, pp.1045–1065.

Jones, M.T., 2009. Unexpected Change - Transforming the GeoWeb for a New World. In Keynote speech. Geoweb 2009. Vancouver, BC, Canada.

Kadomura, H., 1968. Predicting the areas in danger from natural disasters due to soft ground conditions: an application of aerial photo analysis. Available at: http://www.repository.lib.tmu.ac.jp/dspace/handle/10748/3387 [Accessed May 21, 2015].

Kaleel, S.B. & Abhari, A., 2015. Cluster-discovery of Twitter messages for event detection and trending. Journal of Computational Science, 6, pp.47–57.

Kaufman, Y.J. et al., 1998. Potential global fire monitoring from EOS-MODIS. Journal of Geophysical Research D: Atmospheres, 103(D24), pp.32215–32238.

Kisilevich, S. et al., 2013. Towards Acquisition of Semantics of Places and Events by Multi-perspective Analysis of Geotagged Photo Collections. In A. Moore & I. Drecki, eds. Geospatial Visualisation. Lecture Notes in Geoinformation and Cartography. Springer Berlin Heidelberg, pp. 211–233.

Kisilevich, S., Mansmann, F. & Keim, D., 2010. P-DBSCAN: A density based clustering algorithm for exploration and analysis of attractive areas using collections of geo-tagged photos. In Proceedings of the 1st International Conference and Exhibition on Computing for Geospatial Research & Application - COM.Geo '10. the 1st International Conference and Exhibition. Washington, D.C., p. 1.

Kleinberg, J.M., 2007. Challenges in Mining Social Network Data: Processes, Privacy, and Paradoxes. In Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '07. New York, NY, USA: ACM, pp. 4–5.

Klein, D. & Manning, C.D., 2002. Fast Exact Inference with a Factored Model for Natural Language Parsing. In S. Becker, S. Thrun, & K. Obermayer, eds. Advances in Neural Information Processing Systems 15. pp. 3–10.

Kongthon, A. et al., 2014. The role of social media during a natural disaster: a case study of the 2011 thai flood. International Journal of Innovation and Technology Management, 11(03).

Kounadi, O. et al., 2015. Exploring Twitter to Analyze the Public's Reaction Patterns to Recently Reported Homicides in London. PloS one, 10(3), p.e0121848.

Kramer, H.J., 2002. Observation of the Earth and Its Environment, Berlin, Heidelberg: Springer Berlin Heidelberg.

Kulldorff, M. et al., 2005. A Space–Time Permutation Scan Statistic for Disease Outbreak Detection. *PLoS Med*, 2(3), p.e59.

Kulldorff, M., 2006. *SaTScan User Guide*, Available at: http://www.satscan.org/cgi-bin/satscan/register.pl/Current%20Version:%20SaTScan%20v9.3.1%20released%20October%208%202014.?todo=process_userguide_download [Accessed January 16, 2016].

Kulldorff, M., 1999. Spatial Scan Statistics: Models, Calculations, and Applications. In J. Glaz & N. Balakrishnan, eds. *Scan Statistics and Applications*. Statistics for Industry and Technology. Birkhäuser Boston, pp. 303–322.

Kulldorff, M., 1997. A spatial scan statistic. Communications in Statistics - Theory and Methods, 26(6), pp.1481–1496.

Lee, V., 1994. Volunteer monitoring: a brief history. The Volunteer Monitor, 6(1).

Leloup, D., 2015. Deux ans après Snowden, ce qui a changé pour la surveillance de masse. Le Monde. Available at: http://www.lemonde.fr/pixels/article/2015/06/05/deux-ans-apres-snowden-ce-qui-a-change-pour-la-surveillance-de-masse_4648014_4408996.html [Accessed August 24, 2015].

Levesque, H.J., 2012. Thinking as Computation: A First Course, Cambridge, Mass: The MIT Press.

Manning, C.D. et al., 2014. The Stanford CoreNLP natural language processing toolkit. In Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. pp. 55–60.

Marc Prensky, 2001. Digital Natives, Digital Immigrants Part 1. On the Horizon, 9(5), pp.1–6.

Marks, K., 2009. How Twitter works in theory. Epeus' epigone. Available at: http://epeus.blogspot.com/2009/03/how-twitter-works-in-theory.html.

Masser, I., 2000. Spatial data infrastructures in Europe. In Proceedings of the International Conference—Quo Vadis Surveying of the 21st Century, FIG Working Week.

Mathioudakis, M. & Koudas, N., 2010. Twittermonitor: trend detection over the twitter stream. In Proceedings of the 2010 ACM SIGMOD International Conference on Management of data. ACM, pp. 1155–1158.

Maué, P., 2007. Reputation as tool to ensure validity of VGI. In Workshop on volunteered geographic information. Available at: http://www.ncgia.ucsb.edu/projects/vgi/docs/position/Maue_paper.pdf [Accessed May 11, 2015].

Maxwell, E., 2006. Open Standards, Open Source, and Open Innovation: Harnessing the Benefits of Openness. Innovations: Technology, Governance, Globalization, 1(3), pp.119–176.

Mayer-Schönberger, V. & Cukier, K., 2013. Big data: a revolution that will transform how we live, work, and think, Boston: Houghton Mifflin Harcourt.

Metzger, M.J., 2007. Making sense of credibility on the Web: Models for evaluating online information and recommendations for future research. Journal of the American Society for Information Science and Technology, 58(13), pp.2078–2091.

Miller, C.C., 2006. A Beast in the Field: The Google Maps Mashup as GIS/2. Cartographica: The International Journal for Geographic Information and Geovisualization, 41(3), pp.187–199.

Miller, G., 2014. Edward Snowden asks Vladimir Putin about Russian spying on its citizens. Washington Post. Available at: http://www.washingtonpost.com/posttv/politics/putin-misleads-about-russian-surveillance--truth-teller/2014/04/17/dcf00572-c655-11e3-b708-471bae3cb10c_video.html [Accessed August 24, 2015].

Miller, H.J. & Han, J., 2001. Geographic data mining and knowledge discovery : an overview. In Geographic data mining and knowledge discovery. CRC Press, pp. 3–33.

Mobasher, B. et al., 1996. Web mining: Pattern discovery from world wide web transactions, Technical Report TR96-050, Department of Computer Science, University of Minnesota. Available at: http://eolo.cps.unizar.es/docencia/doctorado/Articulos/DataWebMining/webminer-tr96.pdf [Accessed May 18, 2015].

Monk, J. et al., 2008. Using community-based monitoring with GIS to create habitat maps for a marine protected area in Australia. Journal of the Marine Biological Association of the United Kingdom, 88(5), pp.865–871.

Mooney, P., Corcoran, P. & Ciepluch, B., 2012. The potential for using volunteered geographic information in pervasive health computing applications. Journal of Ambient Intelligence and Humanized Computing, 4(6), pp.731–745.

Mummidi, L.N. & Krumm, J., 2008. Discovering points of interest from users' map annotations. GeoJournal, 72(3-4), pp.215–227.

Newell, A. & Simon, H.A., 1972. Human problem solving, Prentice-Hall Englewood Cliffs, NJ.

OASIS, 2005. OASIS Common Alerting Protocol (CAP) - version 1.1.

OGC, 2007a. OGC GML Implementation Standard 3.2.1.

OGC, 2007b. OGC Observations and Measurements - version 1.0.

OGC, 2007c. OGC OpenGIS Sensor Observation Service - version 1.0.0.

OGC, 2006. OGC OpenGIS Web Service Common Implementation Specification - version 1.1.0.

OGC, 2007d. OGC Sensor Alert Service – version 0.9.

OGC, 2008a. OGC Sensor Event Service Interface Specification - version 0.3.0.

OGC, 2007e. OGC Sensor Planning Service (SPS) – version 1.0.0.

OGC, 2008b. OGC Web Coverage Service (WCS) Implementation Standard - version 1.1.2.

OGC, 2005. OGC WFS Implementation Standard 1.1.0.

O'Neill, 2015. Enabling the transition towards Earth Observation Science 2.0. In Keynote speech. ESA conference EO Science 2.0. ESRIN, Frascati, Italy.

OpenStreetMap, 2009. The OpenStreetMap stats report. http://www.openstreetmap.org/stats/data_stats.html, p.last access: 14.02.2009.

Opgenhaffen, M. & Smets, W., 2012. Twitter as emergency response tool during the Pukkelpop disaster: An analysis of# ppok and# HasseltHelpt. status: published. Available at: https://lirias.kuleuven.be/handle/123456789/396327 [Accessed May 21, 2015].

ORCHESTRA, 2008. ORCHESTRA, an open service architecture for risk management M. K. Klopfer, ed., ORCHESTRA consortium.

O'Reilly, T., 2005. What Is Web 2.0 Design Patterns and Business Models for the Next Generation of Software, Available at: http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html.

Orwell, G., 1948. Nineteen Eighty Four, Houghton Mifflin Harcourt.

Oscarson, D.B. & Calhoun, A.J.K., 2007. Developing vernal pool conservation plans at the local level using citizen-scientists. Wetlands, 27(1), pp.80–95.

Palen, L. et al., 2009. Crisis in a networked world: Features of computer-mediated communication in the April 16, 2007, Virginia Tech event. Social Science Computer Review, 27(4), pp.467–480.

Palen, L. & Liu, S.B., 2007. Citizen communications in crisis: anticipating a future of ICT-supported public participation. In Proceedings of the SIGCHI conference on Human factors in computing systems. San Jose, California, USA: ACM, pp. 727–736.

Pang, B. & Lee, L., 2008. Opinion Mining and Sentiment Analysis. Found. Trends Inf. Retr., 2(1-2), pp.1–135.

Paulson, L.D., 2005. Building rich web applications with Ajax. Computer, 38(10), pp.14–17.

Perez, S., 2010. This is What a Tweet Looks Like. *ReadWrite.com*. Available at: http://readwrite.com/2010/04/19/this_is_what_a_tweet_looks_like [Accessed January 16, 2016].

Piketty, T., 2013. Le capital au XXIe siècle, Paris: Seuil.

Port, R.F. & Van Gelder, T., 1995. Mind as motion: Explorations in the dynamics of cognition, The MIT Press.

Press, G., 2013. A Very Short History Of Data Science. Forbes. Available at: http://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/ [Accessed May 13, 2015].

Pultar, E., Raubal, M. & Goodchild, M.F., 2008. GEDMWA: Geospatial Exploratory Data Mining Web Agent. In W. Aref et al., eds. Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM GIS 2008). 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM GIS 2008). Irvine, CA, USA.

Quesnot, T. & Roche, S., 2015. Platial or Locational Data? Toward the Characterization of Social Location Sharing. In 2015 48th Hawaii International Conference on System Sciences (HICSS). 2015 48th Hawaii International Conference on System Sciences (HICSS). pp. 1973–1982.

Radanne, F., 2015. Attentats à Paris: Le twitto qui a lancé le hashtag «#PorteOuverte» raconte. *20minutes.fr*. Available at: http://www.20minutes.fr/web/1731183-20151115-attentats-paris-twitto-lance-hashtag-porteouverte-raconte [Accessed January 19, 2016].

Rajabifard, A. et al., 2000. From Local to Global SDI initiatives: a pyramid building blocks. In 4th Global Spatial Data Infrastructure Conference, Cape Town, South Africa. pp. 13–15.

Rehrl, K. et al., 2013. A conceptual model for analyzing contribution patterns in the context of VGI. In Progress in Location-Based Services. Springer, pp. 373–388.

Resch, B., 2013. People as Sensors and Collective Sensing-Contextual Observations Complementing Geo-Sensor Network Measurements. In J. M. Krisp, ed. Progress in Location-Based Services. Lecture Notes in Geoinformation and Cartography. Springer Berlin Heidelberg, pp. 391–406.

Rhind, D.W., 1999. National and international geospatial data policies. Geographical information systems: principles, techniques, management and applications, pp.767–787.

Rinner, C.K., 2008. The use of Web 2.0 concepts to support deliberation in spatial decision-making. Computers, Environment and Urban Systems, 32, pp.386–395.

Roche, S. et al., 2012. Are 'smart cities' smart enough. Global geospatial conference, pp.215–235.

Roche, S. & Kiene, B., 2008. L'intelligence collective géospatiale au service du diagnostic de territoire : GEOdoc. Revue des Nouvelles Technologues de l'Information, 13, pp.43–62.

Roche, S., Propeck-Zimmermann, E. & Mericsay, B., 2011. GeoWeb and crisis management: issues and perspectives of volunteered geographic information. Geojournal, 78(1), pp.1–20.

Rogerson, P.A., 2001. Monitoring Point Patterns for the Development of Space-Time Clusters. Journal of the Royal Statistical Society. Series A (Statistics in Society), 164(1), pp.87–96.

Roick, O., Loos, L. & Zipf, A., 2012. A technical framework for visualizing spatio-temporal quality metrics of volunteered geographic information. Proceedings of the GEOINFORMATIK 2012—Mobility and Environment.

Ross, J.-M., 2009. The Rise Of The Social Nervous System. Forbes, (march 2009). Available at: http://www.forbes.com/2009/03/09/internet-innovations-hive-technology-breakthroughs-innovations.html?feed=rss_technology.

Rush, M., Holguin, A. & Vernon, S., 1976. Potential role of remote sensing in disaster relief management. Available at: http://ntrs.nasa.gov/search.jsp?R=19770004559 [Accessed May 21, 2015].

Salk, C.F. et al., 2015. Assessing quality of volunteer crowdsourcing contributions: lessons from the Cropland Capture game. *International Journal of Digital Earth*, 0(0), pp.1–17.

Sakaki, T., Okazaki, M. & Matsuo, Y., 2010. Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors. In Proceedings of the 19th International Conference on World Wide Web. WWW '10. New York, NY, USA: ACM, pp. 851–860.

Sankaranarayanan, J. et al., 2009. TwitterStand: News in Tweets. In Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. GIS '09. New York, NY, USA: ACM, pp. 42–51.

San-Miguel-Ayanz, J. et al., 2002. Towards a coherent forest fire information system in Europe: The European Forest Fire Information System (EFFIS). In X. Viegas, ed. Forest Fire Research & Wildland Fire Safety. Millpress, Rotterdam.

Schade, S. et al., 2013. Citizen-based sensing of crisis events: sensor web enablement for volunteered geographic information. Applied Geomatics, 5(1), pp.3–18.

Schnebele, E., Cervone, G. & Waters, N., 2014. Road assessment after flood events using non-authoritative data. Natural Hazards and Earth System Science, 14(4), pp.1007–1015.

Schneier, B., 2015. What's Next in Government Surveillance. The Atlantic. Available at: http://www.theatlantic.com/international/archive/2015/03/whats-next-in-government-surveillance/385667/ [Accessed August 24, 2015].

See, L. et al., 2015. Harnessing the power of volunteers, the internet and Google Earth to collect and validate global spatial information using Geo-Wiki. *Technological Forecasting and Social Change*, 98, pp.324–335.

Shah, A.R. et al., 2010. Applications in Data-Intensive Computing. Advances in Computers, 79, pp.1–70.

Sharl, A.T. (Eds.), 2007. The Geospatial Web. How Geobrowsers, Social Software and the Web 2.0 are Shaping the Network Society A. T. Sharl, ed., Springer.

Shubber, K., 2013. Mapping websites reveal just how stupid it is to geotag your tweets. Wired. Available at:

http://www.wired.co.uk/news/archive/2013-09/04/twitter-geotagging [Accessed August 24, 2015].

Sikder, I. & Woodside, J., 2007. Detection of Space-Time Cluster. In Information and Communication Technology, 2007. ICICT '07. International Conference on. Information and Communication Technology, 2007. ICICT '07. International Conference on. pp. 139–143.

Smith, R.B. & Woodgate, P.W., 1985. Appraisal of fire damage and inventory for timber salvage by remote sensing in mountain ash forests in Victoria. Australian Forestry, 48(4), pp.252–263.

Snow, C.P., 1971. Technology. New York Times.

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

Stamp, L.D., 1937. The Land of Britain: the report of the land utilization survey of Britain.

Steiger, E., de Albuquerque, J.P. & Zipf, A., 2015. An Advanced Systematic Literature Review on Spatiotemporal Analyses of Twitter Data. Transactions in GIS, p.n/a–n/a.

Stigler, S.M., 1986. The history of statistics: The measurement of uncertainty before 1900, Harvard University Press.

Stone, B., 2009. Location, Location, Location. Official Twitter Blog. Available at: http://blog.twitter.com/2009/08/location-location-location.html [Accessed August 24, 2009].

Sui, D., Elwood, S. & Goodchild, M., 2012. Crowdsourcing Geographic Knowledge: Volunteered Geographic Information (VGI) in Theory and Practice, Springer Science & Business Media.

Sui, D.Z., 2008. The wikification of GIS and its consequences: Or Angelina Jolie's new tattoo and the future of GIS. Computers, Environment and Urban Systems, 32(1), pp.1–5.

Terdiman, D., 2004. Photo Site a Hit With Bloggers. Wired, http://www.wired.com, retrieved on 08/05/2009.

Terpstra, T. et al., 2012. Towards a realtime Twitter analysis during crises for operational crisis management. In ISCRAM 2012 Conference Proceedings. 9th International Conference on Information Systems for Crisis Response and Management. Vancouver.

Thatcher, J., 2013. From Volunteered Geographic Information to Volunteered Geographic Services. In D. Sui, S. Elwood, & M. Goodchild, eds. Crowdsourcing Geographic Knowledge. Springer Netherlands, pp. 161–173.

Tomlinson, R., 1963. Feasability Report on Computer Mapping System Agricultural Rehabilitation and Development Administration, Department of Agriculture, Government of Canada.

Tresch, J., 2012. The Romantic Machine: Utopian Science and Technology after Napoleon, University of Chicago Press.

Tuia, D. et al., 2009. Clustering and Hot Spot Detection in Socio-economic Spatio-temporal Data. In Transactions on Computational Science VI. pp. 234–250

Tulloch, D.L., 2008. Is VGI participation? From vernal pools to video games. GeoJournal, 72(3-4), pp.161–171.

Turner, A., 2006. Introduction to Neogeography, O'Reilly Media, Inc.

Ulrich, C., 2008. Twitter, média de l'ère Obama. Le Monde 2 - special Hi-Tech. Available at: http://www.lemonde.fr/archives/article/2008/11/14/twitter-media-de-l-ere-obama_1118891_0_1.html [Accessed August 25, 2009].

Van de Merckt, T. & Chevalier, J.-F., 2008. PaKDD-2007: A Near-Linear Model for the Cross-Selling Problem. International Journal of Data Warehousing and Mining, 4(2), pp.46–54.

Van Exel, M. & Dias, E., 2011. Towards a methodology for trust stratification in VGI. In VGI Pre-Conference at AAG.

Venkataramanan, M., 2014. My identity for sale. Wired. Available at: http://www.wired.co.uk/magazine/archive/2014/11/features/my-identity-for-sale [Accessed August 24, 2015].

Vidal, D., 2015. Vous avez dit terrorisme ? Le Monde Diplomatique, Manière de Voir(140), pp.2–5.

Vogt, J.V. et al., 2007. Developing a pan-European Data Base of Drainage Networks and Catchment Boundaries from a 100 Metre DEM. In AGILE International Conference. Aalborg, Denmark.

Völkel, M. et al., 2006. Semantic Wikipedia. In Proceedings of the 15th International Conference on World Wide Web. WWW '06. New York, NY, USA: ACM, pp. 585–594.

Wald, D.J.; Q., 1999. Utilization of the Internet for Rapid Community Intensity Maps. Seismological Research Letters, 70/6, pp.680–697.

Wallace, D.P., 2007. Knowledge management: Historical and cross-disciplinary themes, Libraries unlimited.

Walsh, J., 2008. The beginning and end of neogeography. GEO: connexion, 7(4), pp.28–30.

Walther, M. & Kaisser, M., 2013. Geo-spatial event detection in the twitter stream. In Advances in Information Retrieval. Springer, pp. 356–367.

Wilkin, S. & Balali, M., 2015. Iran's Guards increase monitoring of social media - state TV. Reuters newswire. Available at: http://uk.reuters.com/article/2015/03/02/uk-iran-internet-idUKKBN0LY1XT20150302 [Accessed August 24, 2015].

Wilson, T.V. & Fenlon, W., 2009. How the iPhone works. Retrieved from: http://electronics.howstuffworks.com/iphone3.htm on May 2015, 24(2009), p.9.

Wolf, G., 2010. The Data-Driven Life. The New York Times. Available at: http://www.nytimes.com/2010/05/02/magazine/02self-measurement-t.html [Accessed August 24, 2015].

Yang, B., Zhang, Y. & Lu, F., 2014. Geometric-based approach for integrating VGI POIs and road networks. International Journal of Geographical Information Science, 28(1), pp.126–147.

Zhang, C., Zhao, T. & Li, W., 2014. Towards an interoperable online volunteered geographic information system for disaster response. Journal of Spatial Science, (ahead-of-print), pp.1–19.

Zhao, L. et al., 2014. Unsupervised Spatial Event Detection in Targeted Domains with Applications to Civil Unrest Modeling. PLoS ONE, 9(10), p.e110206.