“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

Bertrand De Longueville
European Commission
Joint Research Centre
Institute for Environment & Sustainability
T.P 262, 21020 Ispra, Italy
+39 0332785860
bertrand.de-longueville@jrc.ec.europa.eu

Robin S. Smith
European Commission
Joint Research Centre
Institute for Environment & Sustainability
T.P 262, 21020 Ispra, Italy
+39 0332786363
robin.smith@jrc.ec.europa.eu

Gianluca Luraschi
European Commission
Joint Research Centre
Institute for Environment & Sustainability
T.P 262, 21020 Ispra, Italy
+39 0332786364
gianluca.luraschi@jrc.ec.europa.eu

ABSTRACT

The emergence of innovative 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. LBSN can thus be seen by Geographic Information specialists as a timely and cost-effective source of spatio-temporal information for many fields of application, provided that they can set up workflows to retrieve, validate and organise such information. This paper aims to improve the understanding on how LBSN can be used as a reliable source of spatio-temporal 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.

Categories and Subject Descriptors

H.3.1 [**Information Systems**]: Content Analysis and Indexing

General Terms

Measurement, Documentation, Experimentation, Human Factors, Verification.

Keywords

VGI, social networking services, social media, twitter, forest fires, data mining, spatio-temporal data

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 [15]. When such user-generated contents have a geographical dimension, these are now 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 [6] . 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 paper aims to contribute to this growing body of literature by studying how Twitter[1] can be used as a source of spatio-temporal 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 generalisable outcomes to a range of scenarios and related LBSNs.

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

The remainder of this paper is structured around four main sections covering previous work 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). Finally, our main conclusion points can be found in section 5. In order to provide 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 [13]. Twitter allows users to post very short 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, a specific syntax for short messages has been created by users. 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 our 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.

Twitter currently generates a lot of hyperbole in the media, as the most prominent example of the 'social' trend the Web 2.0 features [1], that "will change the way we live" [11]. 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]"[5]. 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 [18] or the recent troubles surrounding the Iranian elections [2].

Although, the body of scientific literature about Twitter is growing[2], its potential for spatio-temporal information has still to be assessed. Most studies have focused on its social dimension by studying users motivations [10], interactions [8] or collaboration [7],with a recent article from the crisis informatics field examining Twitter adoption during mass convergence events [9]. This paper aims 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.

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 from the Twitter API with a level of accuracy of 1 second. The spatial dimension of Twitter is more complex, where georeferencing takes two 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 smart phones allow users to automatically update this location *field* each time a tweet is posted, thus converting Twitter into a genuine LBSN. Secondly, tweets themselves can contain geographic coordinates or more often place names that can be geocoded using a gazetteer service. This direct form of 'geotweeting' has been noted as an area for development for Twitter by founder Biz Stone, where georeferenced tweets will be created more readily [17]. Like many forms of LSBN, an important aspect is not only that these forms of GI are easily produced but that they are also readily gathered and extracted from their source platforms to be reused in other applications.

2.2 Harvesting spatio-temporal information from the web

The idea of harvesting spatio-temporal information from the web has seen some activity in recent years. For example, it has been demonstrated that general purpose *Points of Interest* (POI) can be automatically derived from users' map annotations [14] and vague geographic regions (e.g., Midlands, or Middle West) delineated [12]. 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 [4]. In addition, a Geospatial Exploratory Data Mining Web Agent that retrieves geographic information from web pages (related to outdoor activities), has also recently been discussed [16]. Recent work by the authors has presented a workflow to retrieve, validate, and filter spatio-temporally referenced images from Flickr within the context of flooding in the UK [under review]. As such, this paper aims to explore the role of Twitter as another source of spatio-temporal 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. Data mining and web-crawling scripts, written in PHP, were then applied 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.. Having adopted this methodology, we now turn to the specific case of the Marseille forest fire.

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

[2] As an example, the online scientific literature database Scopus.com provides 9 results for 2007, 24 results for 2008 and 48 results for 2009 while searching for the keyword 'Twitter'.

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 Carpiagne 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 South-East 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, Twitter usage in France is different from other countries where studies have been undertaken, where most research has involved Twitter usage for incidents in the United States. Instead the focus is on a European country, where English is not the first language and where Twitter users are much lower. According to [3], Twitter has 11.5 million accounts, 62.14% of these are based in the United States, 7.87% in the United Kingdom and 5.69% in Canada. In comparison to these top 3 countries, France only has 0.9% of this total. In addition, the ratio of Twitter penetration between France and the US (in terms of total Internet users) is roughly 1:10[3]. However, such figures must be treated with care, as the figures for Twitter users are estimates and it is possible that individual users may have more than one Twitter account (potentially inflating any numerator in our brief analysis).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, allowing us to expect a relatively large number of tweets. Lastly, the event recieved a lot of attention from the media, allowing

3.2 Assumptions/hypothesis to verify

In order to provide focus to the study, the following hypotheses were set out:

H1: Twitter is an extremely fast information dissemination platform to report exceptional events.

H2: As an LBSN, Twitter will provide 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 [9]. The remainder of the paper discusses the material provided to assess these hypotheses and our findings.

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 felt to be 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 [9]. 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

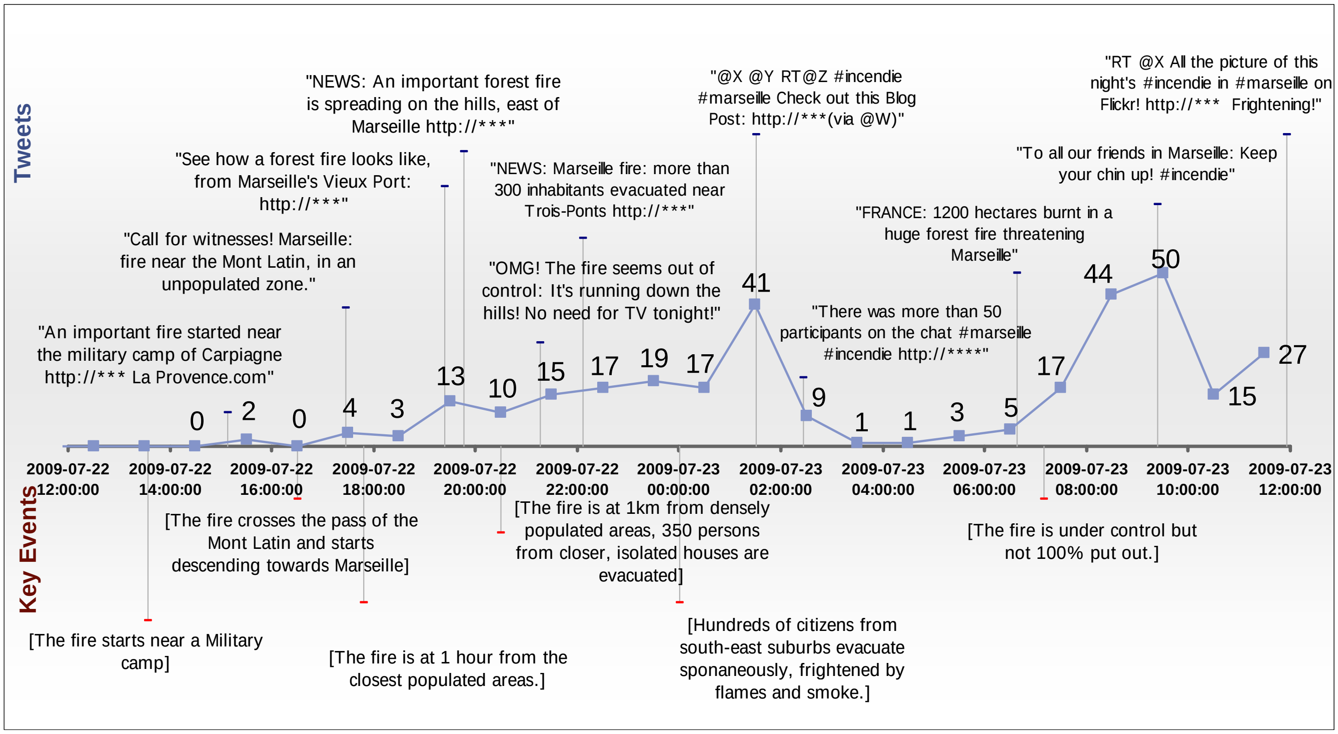
Figure 1 shows a time line with the major events related to the Marseille Fire and the number of relevant Tweets per hour, with examples of tweets (translated by the authors) taken at key moments in the event.

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

[3] The proportion of Twitter users in terms of Internet users (by country) can be estimated by multiplying the proportion of Twitter users from a given country by the 11.5 million total (for all Twitter users) and then dividing this by the total number of Internet users in this country. Using national Internet usage figures based on the World Factbook 2009 https://www.cia.gov/library/publications/the-world-factbook/, it can be estimated that Twitter usage expressed by Internet usage is: US= 3.20%, Canada= 2.34%, UK= 2.25% and France= 0.33%. The ratio between France and the USA is, therefore, approximately 1:10.

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.



(Sources: information provided during and after the fire by AFP and La Provence, and selected Twitter contents)

Figure 1: Chronology of the Marseille Fire, number of related tweets per hour and selected tweets’ contents

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. We 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 [17]. 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).

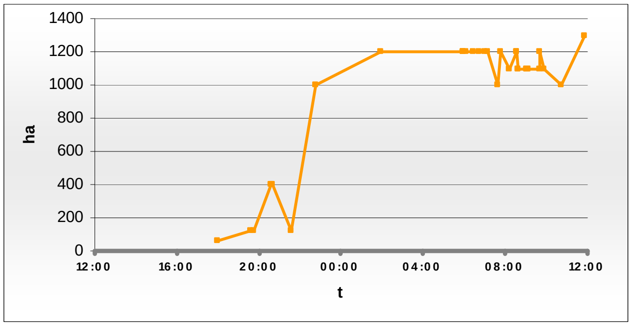


Figure 2: Hectares of burnt area reported in tweets over time

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.

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).

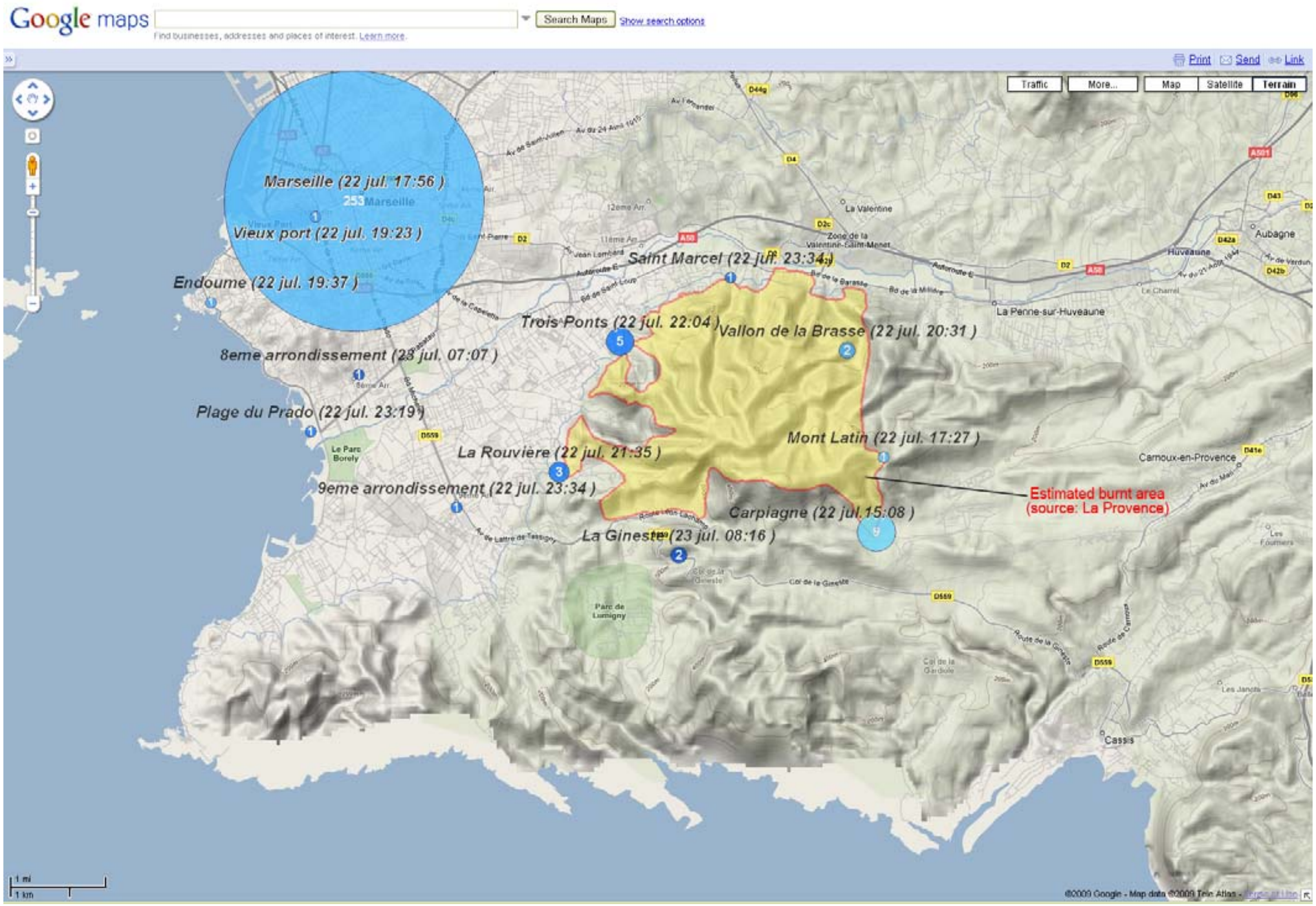

Google maps
Search Maps
Show search options
Find businesses, addresses and places of interest. Learn more.
Print
Send
Link
Traffic
More...
Map
Satellite
Terrain
Marseille (22 jul. 17:56)
253
Vieux port (22 jul. 19:23)
Endoume (22 jul. 19:37)
Saint Marcel (22 jul. 23:34)
Trois Ponts (22 jul. 22:04)
Vallon de la Brasse (22 jul. 20:31)
8eme arrondissement (23 jul. 07:07)
Plage du Prado (22 jul. 23:19)
La Rouvière (22 jul. 21:35)
Mont Latin (22 jul. 17:27)
9eme arrondissement (22 jul. 23:34)
Carpiagne (22 jul.15:08)
La Gineste (23 jul. 08:16)
Estimated burnt area
(source: La Provence)
1
5
2
3
2
9
La Valentine
Aubagne
La Penne-sur-Huveaune
Carnoux-en-Provence
Cassis
Parc de Lumigny
Le Parc Borely
1 mi
1 km
©2009 Google - Map data ©2009 Tele Atlas

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 spatio-temporal 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 5).

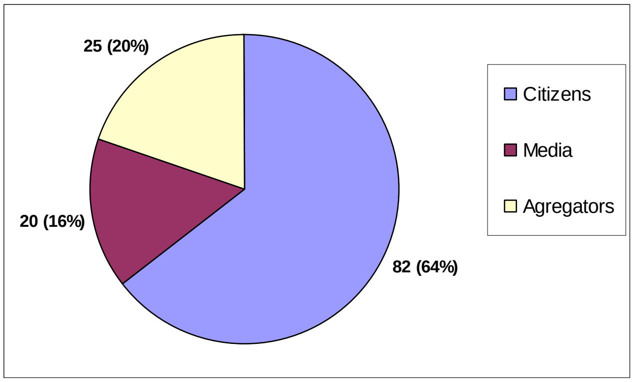


Figure 4: Number of users that published tweets by type

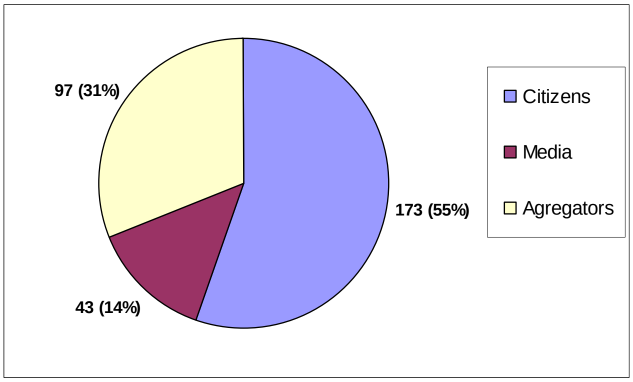


Figure 5: Number of tweets published by user type

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. 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. *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[4] to automatically re-publish the contents a RSS feeds into tweets, and thus very easily reproduce information contents on Twitter without human intervention. In our sample, nearly 1 tweet out of every 3 (31%) has been published by an aggregator. This finding is important to understand the possible redundancy of any 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 our sample 18.8% contained the "*RT*" code), where 91% of these tweets containing RT have been published by citizens. However, the fact that the media do not use the RT syntax does not means they act solely as a primary source of information. The aggregators are by their own nature secondary providers of information, whilst most of the media are expected to forward information from news agencies, often with limited added value. This can create problems to set up quantitative quality filters on top of Twitter: the fact that information is tweeted numerous times cannot be interpreted as a proof of veracity, or other sense of 'truth'. The use of Twitter as a crisis information source as suggested by H4 is thus conditioned by the capacity to easily discriminate such primary and secondary sources..

It is also interesting to underline that hash tags were present in 22% of the tweets. A consensus rapidly emerged about which hash tags to use: 68 out of these 69 tweets contained a combination of the hash tags #marseille and #incendie, as initially suggested by two citizen-users[5]. The use of a pair of hash tags including entire words contrasts with instances where short hash tags designate particular events without ambiguity within the Twitter users community (e.g.: #obamainaug, #bneflood, #grfires). Provided they are used consistently, hash tags appears to be means to support the comprehensive collection of event-related information on Twitter.

89% of the Tweets containing at least one hash tag were published by citizens. This tends to partially support H3: citizens use Twitter for their open online conversations, where other user types may act in different ways.

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% [10] and 25% [9]) and provides evidence for accepting H4 in this use case. It is a common practice on Twitter to use abbreviated URLs[6]; 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

[5] On 22/07/09 at 21:51: "@X nobody on twitter with hashtags like #canadair #incendie or #marseille?" and on 23/07/09 at 01:02 "Maseille fire: use the hash #marseille and #incendie for simpler searches" (translation by the authors)

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

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:

- *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 non-web based media and, again, typically does not act as primary sources of information.

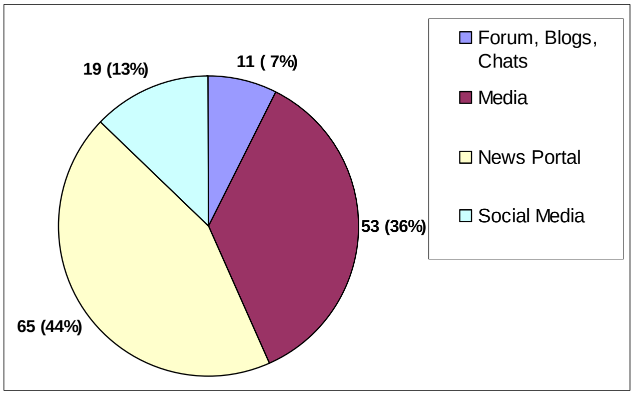


Figure 6: Number of unique cited URLs by domain type

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 portal sources, 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 paper has covered the application of Twitter as a source of spatial-temporal 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 was 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 chose to provide a geographic dimension (either directly on indirectly) to events they record, it appears to offer 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 did not produce primary content but compile existing sources into specific news-feeds that they broadcast to a targetted audience. This categorization is important to better understand the type of contents those re-using such content will be faced with. In particular, we highlighted that Twitter users provide a mix of primary and secondary information which cannot be easily distinguished by automatic means. This is a complex and important question to be addressed in further research.

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 came from blogs, chats, fora and other citizen generated information. Although few, 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 lag effects, 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 paper 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.

6. REFERENCES

[1] Twenty years of the world wide web: What's the score? The Economist, 2009-05-12. http://www.economist.com/sciencetechnology/displayStory.cfm?story_id=13277389

[2] Cardwell, S. A Twitter Timeline of the Iran Election. Newsweek Web Edition, 2009-06-25 http://www.newsweek.com/id/203953/

[3] Cheng, A., Evans, M., and Singh, H. Inside Twitter. An In-Depth Look Inside the Twitter World. Unpublished report by Sysomos, inc. June 2009. http://www.sysomos.com/insidetwitter/

[4] Crandall, D., Backstrom, L., Huttenlocher, D., and Kleinberg, J. Mapping the World's Photos. Proceedings of the 18th International World Wide Web Conference,, (2009), 761-761.

[5] Eudes, Y. Twitter, les pirates et les diplomates. Le Monde, 2009-08-24. http://www.lemonde.fr/technologies/article/2009/08/24/twitter-les-pirates-et-les-diplomates_1231380_651865.html

[6] Goodchild, M.F. Citizens as Voluntary Sensors: Spatial Data Infrastructure in the World of Web 2.0. International Journal of Spatial Data Infrastructures Research 2, (2007), 24-32.

[7] Honeycutt, C. and Herring, S. Beyond Microblogging: Conversation and Collaboration via Twitter. System Sciences, 2009. HICSS '09. 42nd Hawaii International Conference on, (2009), 1-10.

[8] Huberman, B., Romero, D., and Wu, F. Social networks that matter : Twitter under the microscope. First Monday 14, 1 (2009).

[9] Hughes, A.L. and Palen, L. Twitter Adoption and Use in Mass Convergence and Emergency Events. Proceeding of the 6th International ISCRAM Conference, (2009).

[10] Java, A., Song, X., Finin, T., and Tseng, B. Why We twitter: An analysis of a microblogging community, *in* Advances in Web Mining and Web Usage Analysis, Springer, 2009.

[11] Johnson, S. How Twitter Will Change the Way We Live. Time, 2009-06-05. http://www.time.com/time/business/article/0,8599,1902604,00.html.

[12] Jones, C.B.P. Modeling vague places with knowledge from the Web. International Journal of Geographical Information Science 22/10, (2008), 1045-1065.

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

[14] Mummidi, L. and Krumm, J. Discovering points of interest from users' map annotations. GeoJournal 72, 3-4 (2008), 215-227.

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

[16] Pultar, E., Raubal, M., and Goodchild, M.F. GEDMWA: Geospatial Exploratory Data Mining Web Agent. Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM GIS 2008), (2008).

[17] Stone, B. Location, Location, Location. Official Twitter Blog, 2009. http://blog.twitter.com/2009/08/location-location-location.html

[18] Ulrich, C. Twitter, média de l'ère Obama. Le Monde 2 - special Hi-Tech, 2008-11-14. http://www.lemonde.fr/archives/article/2008/11/14/twitter-media-de-l-ere-obama_1118891_0_1.html