Applied Geography 43 (2013) 36–44

Contents lists available at SciVerse ScienceDirect

Applied Geography

journal homepage: www.elsevier.com/locate/apgeog

Automated geographic context analysis for volunteered information

Laura Spinsanti [1], Frank Ostermann*

European Commission Joint Research Centre, Institute for the Environment and Sustainability, Digital Earth and Reference Data Unit, Via E. Fermi 2749, 21027 Ispra, Va, Italy

Keywords:
Volunteered geographic information
Social media
Spatial data infrastructures
Spatio-temporal clustering
Geographic context analysis

A B S T R A C T

Several studies show the impacts of (geo)social media and Volunteered Geographic Information (VGI) during crisis events, and have found intrinsic value for rescue teams, relief workers and humanitarian assistance coordinators, as well as the affected population. The main challenge is how emergency management and the public can capitalize on the abundance of this new source of information by reducing the volume to credible and relevant content.

In this paper, we present the GeoCONAVI (Geographic CONtext Analysis for Volunteered Information) approach and a prototype system, designed to retrieve, process, analyze and evaluate social media content on forest fires, producing relevant, credible and actionable VGI usable for crisis events. The novelty of the approach lies in the enrichment of the content with additional geographic context information, and use of spatio-temporal clustering to support scoring and validation. Thus, the system is focusing on integrating authoritative data sources with VGI. Evaluation in case studies shows that the prototype system can handle large amounts of data with common-off-the-shelf hardware, produces valuable results, and is adaptable to other types of crisis events.

Introduction and related work

In this paper, we present a proof-of-concept system to extract, to process and to analyze volunteered information on forest fires from social media networks. The novelty of the approach lies in its focus on using location to filter and verify the information circulating in social media networks. By first geo-referencing the retrieved content and turning it into volunteered geographic information (VGI), it can subsequently enrich it with additional geographic context information from authoritative data sources, and cluster it spatio-temporally to support filtering and verification.

We have argued elsewhere for the importance of clearly distinguishing terms such as "crowd-sourced geo-information", "citizen science", and VGI (Craglia, Ostermann, & Spinsanti, 2012). The GeoCONAVI system currently consumes all types of user-generated content: Volunteered as well as contributed (Harvey, 2012), on geographic locations or not. We have decided to use the established term VGI throughout the paper, because the output of the GeoCONAVI system is just that: Information on geographic features and processes related to a crisis event, which has been volunteered with the purpose to be consumed and acted upon by others who are affected by that same event.

Our motivations are the opportunities provided by changing ways in which environmental information is collected, distributed and used. Public authorities traditionally generate, manage, update and distribute information in accordance to established rules and procedures in closed systems to ensure reliability and trustworthiness. Information flowed from "top" to "bottom": Public authorities informed the public on notable risks or events through traditional broadcasting media (e.g. newspapers, radio, and television). The limited reach of available horizontal media (e.g. word-of-mouth, letters, and telephones) restricted any peer-to-peer communication. In the past decade, new information and communication technologies have greatly increased opportunities for collaboration and participation: Nowadays, many citizens use mobile device with wireless internet access to share freely various media through social networks[2] or more focused platforms for text messages,[3] images,[4] videos,[5] and maps.[6] Thereby, citizens have become providers of environmental information during crisis

* Corresponding author. Tel.: +39 033278 9242.
E-mail addresses: laura.spinsanti@jrc.ec.europa.eu, jrc.vgi.ff@gmail.com (L. Spinsanti), frank.ostermann@jrc.ec.europa.eu (F. Ostermann).
[1] Tel.: +39 033278 5544.

[2] e.g. Facebook, Google+.
[3] e.g. Twitter, Blogspot, Wordpress.
[4] e.g. Flickr, Picasa, Panoramio.
[5] e.g. YouTube, Vimeo.
[6] e.g. GoogleMaps, GeoCommons, MapBox.

0143-6228/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.apgeog.2013.05.005

events, as several studies have observed (De Longueville, Annoni, Schade, Ostlaender, & Whitmore, 2010; Palen & Liu, 2007; Puras & Iglesias, 2009; Roche, Propeck-Zimmermann, & Mericskay, 2011).

Other studies (Al-Khudhairy, 2010; De Longueville, Smith, & Luraschi, 2009; Hughes & Palen, 2009; Liu & Palen, 2010; Schade et al., 2013) indicate when and how VGI can be most useful in a crisis context: During the response phase, affected citizens are on the ground before the first emergency response teams arrive. Instead of remaining passive victims, they can become now active collectors and distributors of information, thus acting as in-situ sensors (Goodchild, 2007). Their local knowledge could be especially valuable in providing near real-time localized and accurate updates, increasing situational awareness for emergency managers and peers searching for information. During the damage assessment in the recovery phase of a disaster, photographs and reports add to information from remotely sensed images and professional surveyors.

From the perspective of public authorities, the main challenge to using VGI is the lack of managerial control over the lineage of the information, and thereby an unknown reliability and trustworthiness (Jennex, 2010). As a result, most public agencies have been more reluctant in adopting social media information than non-governmental or volunteer organizations. The latter, however, face the issues of sustainability and scalability: There is no guarantee that for a given event there is a sufficient volunteer force.

Therefore, we propose a novel approach to filter VGI, evaluate its quality, and thereby improve its utility. It focuses less on the analysis of the source and content of an individual piece of information. Instead, it relies on geographic location and geographic context information to emulate two heuristics that humans use to deal with new information (Flanagin & Metzger, 2008): Expectancies ("What do I already know?") and social confirmation ("What do others say?"). To this end, the GeoCONAVI system queries authoritative geographic context information and clusters VGI in space and time. The system is implemented as a fully operational prototype based on prior research (Ostermann & Spinsanti, 2011).

This paper investigates the following main research questions:

1. How can geographic location improve the filtering and quality assessment of social media content?
2. What are the specific design requirements to utilize geographic context information?
3. What is the performance of the prototype system, and how well is adaptable to other crisis events?

The objective of the paper is to show how the GeoCONAVI approach works and that it produces valid and actionable output. We then discuss the requirements that need to be met in order for GeoCONAVI to work, showing that the approach is feasible for many types of applications and environments. Finally, the paper discusses the performance of our specific prototype and its adaptability to other types of crisis events. The paper structure is as follows: In the next section, we describe the overall system, the modules and the methods used. In the third section, we briefly present results and evaluation of the case studies, while the fourth section discusses the system evaluation and we answer the research questions. The paper concludes with a synthesis of the results and knowledge gained.

Methods and system architecture

This section presents the modules that make up the GeoCONAVI workflow and the corresponding software implementation. The objective was to implement a proof-of-concept prototype that runs autonomously for a substantial amount of time and allows the evaluation of the approach. We decided to investigate forest fires because of their increasing importance in environmental domain conservation, their seasonality and the availability of the European Forest Fire Information Service (EFFIS). We limited the geographic scope to four countries (Italy, France, Portugal and Spain) because they have frequent seasonal forest fires, the authors are able to read and understand content in these languages, and to reduce the challenges of character encoding. The following Fig. 1 shows the main processing phases for the social media information in the left column, the corresponding state of the VGI in the middle column, and the respective GeoCONAVI modules in the right column.

Fig. 2 below gives a more detailed overview of the system architecture, detailing modules and sub modules. The first three phases of Fig. 1 aligns from left to right instead from top to bottom. The top layer in Fig. 2 shows external data sources, while the implemented GeoCONAVI (sub) modules form the middle layer, and the lower layer is the data storage, implemented in an Oracle DBMS. The Disseminator has been implemented as a web-mapping interface (upper right).

The following sections describe each module in more detail.

Sensor

The GeoCONAVI Sensor is an opportunistic sensor. This means that it listens to a specific "frequency" of broadcasted information by citizens, but it does not have a particular interface where someone could provide information directly. This distinguishes it from portals where citizens can provide information for a specific purpose in a participative effort. The rationale for this is that persons in distress are unlikely to use an interface or infrastructure that they are not used to, instead relying on known services and social networks. There is no reason why some stakeholder could

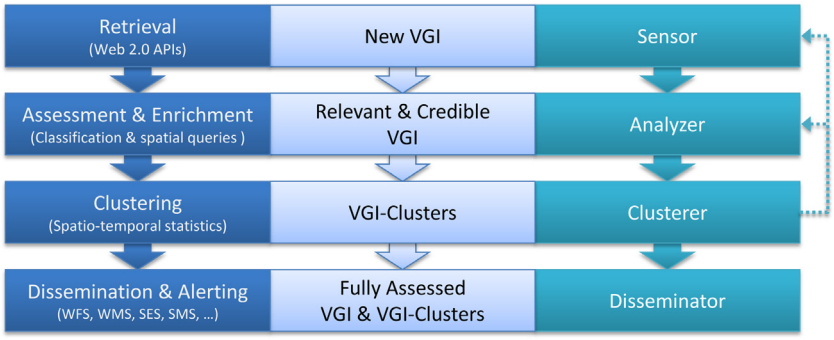


Fig. 1. Overview of Processing Steps.

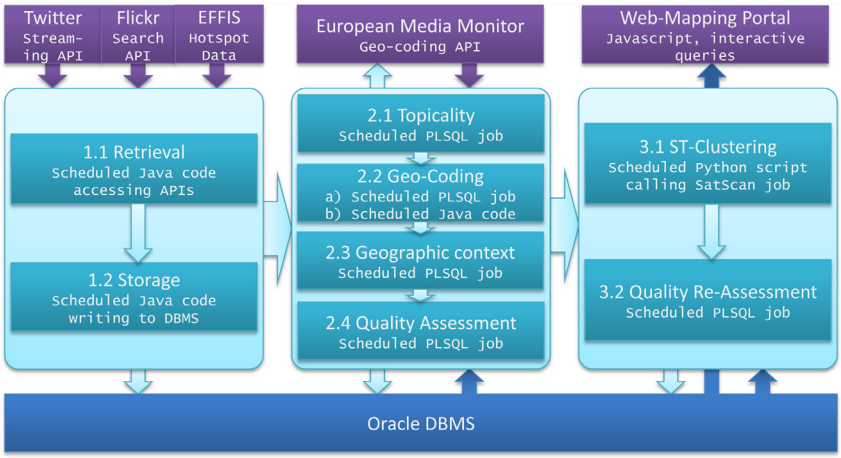


Fig. 2. Overall GeoCONAVI system design.

not develop a smartphone application or internet portal that serve as interface for more structured data from direct participation. The two approaches are not mutually exclusive and can be easily integrated, e.g. improving ("tweaking") social media communication (Starbird and Stamberger 2010).

The Sensor consists of two sub-modules, connecting to the micro-blogging platform Twitter and the image-sharing platform Flickr: Their APIs are well-documented and allow detailed queries with high rate limits. Further, their content base includes information on forest fires, they are popular with citizens, and they represent textual and visual VGI.

The two sub-modules are written in Java and query the APIs using a set of keywords. Similar to the (re)calibration of a sensor, we have calibrated our virtual sensor by refining this set of keywords over the course of two years. We first started out with an extensive set of keywords in common European languages that were developed in discussion with forest fire domain experts (Ostermann & Spinsanti, 2013). The set covered the five concepts of *fire*, *area*, *vegetation*, *actors*, and *actions*. The objective was to prioritize recall over precision, i.e. using a set of keywords that would not miss any relevant information, despite an increased level of noise. For example, the *action* keyword "evacuation" is used in many different situations, while the Romanic *fire* keywords "incendi*" can refer to any type of uncontrolled fire, and the Germanic *fire* keyword "fire" is also often used metaphorically. By analyzing the results (see Section 2.2.1), we were able to reduce the keywords significantly without losing valuable information. The current set of keywords is based on the three concepts of *fire*, *area*, and *vegetation*, using the four languages of the target countries: *Feu*, *fogo*, *fuego*, *fuoco*, *incendie*, *incendio*, *incêndio*, *ettari*, *hectáreas*, *hectareas*, *hectares*, *boschivo*, *forêt*, *florestais*, and *forestal*, plus applicable declinations.

The Sensor modules query the real-time Streaming API of Twitter continuously and the Flickr API in regular intervals. The retrieved information is parsed, validated, and attributes regarding content, location, and user are put into an Oracle DBMS. In 2012 we had a minimum of 7.5 million Tweets and a maximum of 9.5million Tweets per month during the main forest fire season (June to October). The number of retrieved images from Flickr is much lower, not exceeding 200,000 per month.

Analyzer

The Analyzer module has three sequential tasks: Topicality assessment, geo-coding, geo-context scoring. Topicality assessment is performed first, because manual evaluation of samples showed a high level of noise despite careful Sensor calibration. This is not surprising, since keywords can have several meanings (homonymy or polysemy) or can be used figuratively (metaphoric), as shown in Section 2.1. By filtering out noise early, the computation cost of the remaining modules is reduced significantly. The individual tasks are explained below.

Topicality

Established natural language processing tools need adaptation to deal with the short and unstructured nature of VGI content and annotations/tags (Cheng, Caverlee, & Lee, 2010). Since the GeoCONAVI system needs to perform a binary classification into "on-topic" or "off-topic", we choose a simple approach that tries to contextualize found keywords by searching for particular keyword (co-)occurrences and assigns scores based on a few rules. For example, the *fire* keyword "incendie" is most likely about a forest fire and not a buildings fire when used in conjunction with the *area* keyword "hectares" or the *vegetation* keyword "forêt". In order to establish those rules, we extracted a set of around 6000 Tweets with a purposeful (non-random) sampling method based on the temporal co-occurrence of a Tweet with major reported forest fires in Europe in August 2010. We manually classified these Tweets as being on-topic (i.e. about a forest fire) or off-topic (i.e. about anything else, including all other types of fires). Then we programmatically extracted all keyword occurrences, and created language specific rule sets to classify the content into four categories (Table 1 shows the French one): A ("Probably about a forest fire"), B ("Possibly about a forest fire"), C ("Possibly not about a forest fire") and D ("Probably not about a forest fire") .

Later tests with machine-learning algorithms (J48 and Naive Bayes) using the Weka software suite supported this rule set. For verification/error estimation, we used a standard stratified 10-fold cross validation. The results all showed roughly a 90% accuracy. The large number of false negatives introduced by the combination of "*incendie <= 0 AND hectar <= 0 AND forest <= 0*" is an issue, i.e. Tweets that the machine learning algorithms classified as not being about forest fires when in fact they were. We tried to adjust the costs by using a cost matrix on the results, and with the MetaCost classifier. Unfortunately, neither improved the results significantly - either the noise was not reduced, or many True Positives were misclassified. Another annotated data set from a different case study showed similar results. In any case, given the data and the attributes used, the results from the machine learning support our

Table 1
Topicality rule set for VGI in French.

```
FOR EACH VGI
        CASE = D
        IF feu OR feux THEN
                CASE = C
                IF hectares THEN CASE = A
                IF foret OR forets OR forêt OR forêts THEN CASE = A
                NEXT
        IF incendie THEN
                CASE = B
                IF hectares THEN CASE = A
                IF foret OR forets OR forêt OR forêts THEN CASE = A
        IF foret OR forets OR forêt OR forêts THEN
                CASE = C
                NEXT
        IF hectares THEN
                CASE = C
                NEXT
```

initial, hand-crafted rule set. Only VGI likely to be about forest fires (cases A and B) is sent to the geo-coder (2–3% of the retrieved data).

Toponyms and geo-coding

A typical geo-coder searches for place names (toponyms), looks them up in a gazetteer of place names, and assigns a pair of coordinates. Despite this simple premise geo-coding is greatly complicated by variations of the same toponym in different languages, variations in spelling due to special characters, and the same toponym often belonging to many distinct places, requiring disambiguation. The geo-coding is a necessary step in our workflow, because an analysis of our dataset shows that the share of geo-coded information is low: Around 20% for Flickr, and less than 1% for Twitter. However, even for already geo-coded information, an additional geo-coding seems necessary, since any location information already present has an unknown uncertainty, because it depends on a number of factors including hardware specifications, software settings, and user preferences. Then there is also the semantic uncertainty whether the topic of the VGI is located at or near the origin, e.g. a Tweet may be geocoded with the location from which it was sent, but the information maybe about another place. Similarly, a Flickr image may depict a scene that is far away from the location of the photographer. There are several potential sources for location information in our data, e.g. toponyms in the text body or in tags, toponyms in the source's user profile, and any actual geo-coding done by the device, application, or platform. Due to the low count of pre-geo-coded VGI, the limited utility of the user location field (Hecht, Hong, Suh, & Chi, 2011), and the high computational cost of the geo-coding, the Analyzer attempts to geocode the VGI only from toponyms in the text body (Tweets), or the title and tags (Flickr images). We experimented with several third-party web-based geo-coding services, but the results were not satisfactory. One particular challenge was the multilingualism of our data. In the end, we implemented and evaluated a simple approach based on string matching unigrams of the VGI content with a gazetteer from the Geographical Information System at the European Commission (GISCO) database. The administrative level of communes contains more than 57,000 toponyms in 292 provinces for the four considered countries (Italy, France, Spain and Portugal). The geo-coder uses the commune as well as Province names (if no commune was found) with a regular expression to avoid partial matches (Equation (1)).

To test the geo-coding, we performed the GISCO geo-coding on the manually annotated set of Tweets from the previous section, resulting in over 1500 geo-coded Tweets. This set was sent to a geo-coder from the European Media Monitor (EMM), which has a smaller gazetteer but more sophisticated geo-coding algorithms. We then manually evaluated both geo-coder results and categorizing them into the following classes (Table 2):

Since the geographic scope of our gazetteer is limited, we also annotated whether the Tweet contained a valid toponym within our geographic scope (CS = Y), a valid toponym outside of our geographic scope (CS = N), or no valid toponym at all (CS = X).

The following table (Table 3) shows the performance of each geo-coder for the possible combinations of categories for the whole set ($n = 1539$):

```
regexp_like(text, ' '|| label ||'[ |:|,|;|\.|\?|\!|#|(|)]' )
```
(1)

Table 2
Categories from the manual geo-coding evaluation.

Category	Description
True Positive (TP)	The Tweet has a valid place name that was correctly identified.
False Positive (FP)	The Tweet has no valid place name, but one was reported.
False Negative (FN)	The Tweet has a valid place name, but none OR a wrong one was reported.
True Negative (TN)	The Tweet has no valid place name, and none was reported (only applicable for EMM results; since the set consist of Tweets where GISCO reported a toponym, this would be a FP for GISCO).

Table 3
Geo-coders results and comparison.

GISCO	EMM	CS	Performance (number of Tweets)
TP	TP	Y	52
TP	FN	Y	1293
FP	FP	X	16
FP	TN	X	163
FN	TP	Y or N	10
FN	FN	Y or N	5

Regarding the EMM geo-coder, the results were as expected because it is designed for longer texts, has more discriminating geo-coding algorithms, and a smaller gazetteer. A positive surprise was the high precision of the GISCO geo-coder, which was deemed sufficient to use it for GeoCONAVI.

Social media content that does not contain any information on location has no immediate value as environmental information, so only VGI that was on-topic and geo-coded was passed on to the next module.

Geographic context scoring

The two previous modules are the pre-requisites for this important sub-module to work correctly. It is the first step in GeoCONAVI's novel approach to use geographic context information for assessing information quality (see Section 1). The aim is to enrich the VGI with geographic context by looking up characteristics of the locations identified. In principle, these characteristics could be any characteristics found in Spatial Data Infrastructures (SDI) or other databases with a geographic component. In practice, it is sensible to restrict the enrichment to domain- or application-specific information. In the case of forest fires, relevant context information includes distances to known remotely sensed hot spots or forest fires, the population density, and the predominant vegetation type of the area. The rationale is that geographic and temporal proximity to a known hotspot or fire increases the credibility considerably, while the absence of any combustible vegetation would decrease it (e.g., a Tweet about a forest fire originating from a desert might still be true, but it is unlikely to contain first-hand or helpful information). The population density influences credibility and relevance indirectly: A single Tweet on forest fires from a densely populated area is likely to be misinformation. The risk of human casualties or substantial property damage is highest in sufficiently forested terrain with a moderate population density, e.g. urban-rural border. Other context information could include smartphone penetration, socio–demographic parameters, history of forest fire events, infrastructure conditions, and others. Since our geo-coding is based on the communes in the GISCO dataset, we aggregated raster data sets on population density (DGUR-Degree of Urbanization) and land cover (CORINE data 2006) through zonal spatial analysis. The distance to hot spots uses the latest MODIS data from the EFFIS, downloaded at regular intervals and used in a spatial query searching for the nearest neighbor within the same time window.

In order to arrive at a single score for ranking and filtering, the Analyzer converts each result into an individual score for information i (equations (2)–(4)), before calculating a single Integrated Context Score (ICS) from them (equation (5)):

$$FS_{ik} = \begin{cases} 2 \times FC_k & \{FC_k \in \mathbb{R} : 0 \le FC_k \le 0.5\} \\ 1 & \{FC_k \in \mathbb{R} : 0.5 < FC_k \le 1\} \end{cases} \quad (2)$$

with FC_k being the ratio of forest cover for commune k where the information i was located,

$$HS_{ih} = \begin{cases} 1 - (D_h \div 50) & \{D_h \in R : 0 \le D_h < 50\} \\ 0 & \{D_h \in R : D_h > 50\} \end{cases} \quad (3)$$

with D_h being the distance in kilometers between the location of information i and the nearest hotspot,

$$PS_{ik} = \begin{cases} PD_k \div 200 & \{PD_k \in R : 0 \le PD_k \le 200\} \\ 2 - (PD_k \div 200) & \{PD_k \in R : 200 < PD_k \le 400\} \\ 0 & \{PD_k \in R : PD_k > 400\} \end{cases} \quad (4)$$

with PD_k being the number of inhabitants per square kilometer of commune k where the information i was located,

$$ICS_i = (FS_{ik} + HS_{ih} + PS_{ik}) \div 3 \quad (5)$$

The scoring functions are the result of discussions with domain experts and literature studies. However, we are aware that they are (yet) arbitrary and very likely have a lot of room for improvement.

Clusterer

The Clusterer emulates the social confirmation heuristic and searches for patterns and confirmation in the VGI. Additionally, the Clusterer can feed back detected patterns into the workflow and thereby improve the Sensor calibration (e.g. by focusing on a particular geographic area) or the scoring of the Analyzer.

We had to rely on external software for the Clusterer, since the available Oracle DBMS did not provide sufficient support for spatio-temporal clustering. After testing various software (CrimeStatIII, packages of R, ArcGIS, QGIS), we settled on using SatScan.[7] There is a substantial body of literature on it and it offers the widest variety of possible scan methods, including Space-Time Scan Statistics, Bernoulli and Discrete Poisson Models.

The choice of parameters is crucial. A first experiment used several settings on the data of a pilot case study. The input was 680 VGI items that had been filtered for French keywords and locations (Ostermann & Spinsanti, 2013). We used SatScan's space-time permutation model, which only needs a case file, i.e. the VGI data itself. A crucial decision is where to base the spatial location of the scanning windows (Kulldorff, Heffernan, Hartman, Assunção, & Mostashari, 2005): In our case, either on the VGI cases or on other locations, such as known locations of forest fires. From a conceptual viewpoint, this resembles a choice between detecting clusters in the data without prior knowledge about possible events, and trying to find the representation of known events in the data. We opted for both methods to compare them, running each twice with two parameter sets: Once with default parameter values (unrestricted cluster size but no spatial cluster overlap), and a second time with modified parameters (maximum cluster radius 50 km and spatial overlap possible). Thus, in total there are four sets of results. Each result set consists of a number of likely clusters ($p <= 0.0001$ estimated from 9999 Monte Carlo simulation runs).

An analysis of the clusters reveals that three clusters are identical in all four cases. We compared the clusters detected with the official fires registered in the EFFIS system. The results show that without prior knowledge, the system detects 50% of the known fires when restricting spatial overlap of clusters, and all of them (and a large number of potentially False Positives) when allowing for spatial overlap. A manual text analysis of the False Positive clusters leads to the hypothesis that there may be fires not reported

[7] http://www.satscan.org/.

by EFFIS, offering a potential improvement for alert systems based on remote sensing. With prior knowledge, the system is able to detect all known forest fires when we allow for spatial overlap of clusters thus confirming the social network activity around disaster reports.

For the implementation, we chose to restrict the maximum temporal extent of a cluster 10% of study period (i.e. 9 days), to prohibit geographic overlap, and unrestricted maximum cluster size. The Clusterer module exports data daily from the Oracle DBMS, feeds it into SatScan, parses the outputs, and uploads the results back into the database. Each cluster is assigned a preliminary score based on the confidence reported by SatScan. It can be used to rank likely events, which can be investigated further by a human domain expert. However, as a brief investigation has shown, the number of false positive clusters can be high, and further ranking measures are needed.

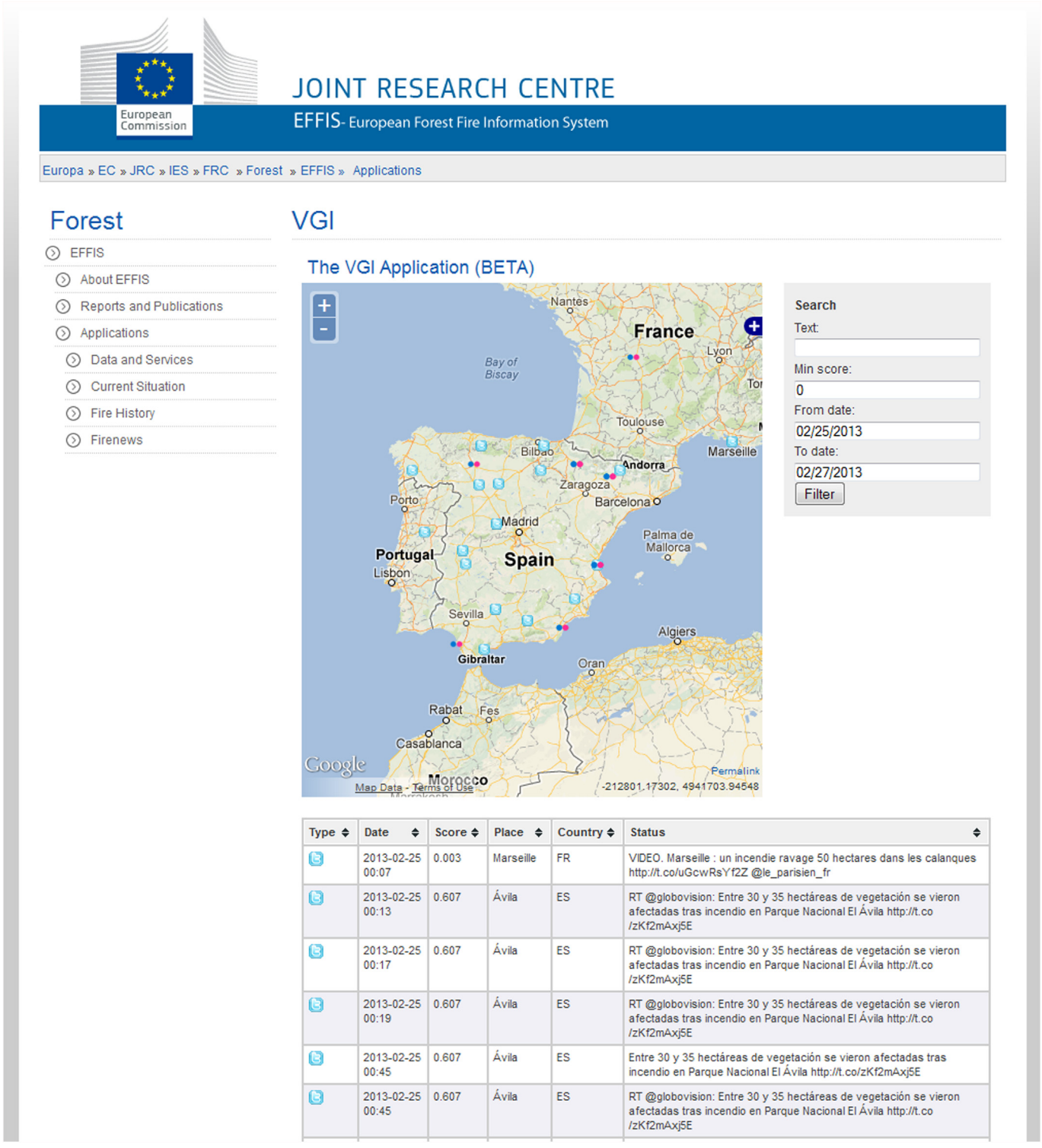


Fig. 3. A screenshot of the GeoCONAVI web interface

Web query interface

The final module in the GeoCONAVI workflow, as shown in Fig. 1, is the Disseminator, which aims to make the results available in various formats. Options include fully interactive web mapping interfaces using Web Mapping Service (WMS), Sensor Observation Services (SOS), subscriptions to plain text information broadcasts such as Short Messaging Service (SMS) or micro-blogging (Twitter account). Currently, the Disseminator is implemented as a basic web interface (Fig. 3).[8] It allows the user to query the DBMS and view the resulting dataset displayed over a basic web map and in tabular form, showing VGI type (Tweet or Flickr image), exact time and date, place and country, score and text (either the Tweet's message body, or the Flickr image's title and tags).

The query parameters are restricted to selecting a period, a search text that can be a keyword and/or a location, and a threshold for the score. Due to licensing restrictions on the content, the interface offers only querying and viewing functionality, but no download service.

Results and evaluation of methods

In this section, we present a brief overview of the results, and the evaluation of the GeoCONAVI system with respect to the first research question. We evaluated the GeoCONAVI system formally through a case study in on the data collected and processed in 2011. We reported on the case study in detail elsewhere (Ostermann & Spinsanti, 2013) and present here only the most relevant parts.

The case study is confined geographically to mainland France to reduce linguistic ambiguities and temporally to the fire season ranging from July to September. The following Table 4 shows the processing phase and corresponding data volumes.

There were eight major fires reported by EFFIS in the considered time window, with burned areas varying between 40 and 240 ha. Comparing cluster locations, time extent and content, we could establish that GeoCONAVI detected six out of the eight fires. The first undetected fire was in a military training area without public access, while the other overlapped with another cluster. Allowing geographic overlap, however, would introduce much more noise in the form of smaller clusters (see Section 2.3). The remaining five clusters cannot be associated with known fires from EFFIS, but a manual check of the text revealed that in fact they refer to forest fires, with four of them reporting burned areas greater than 40 ha.

A manual evaluation of the content through random samples taken during most processing phases shows that the actionable or informative content is overall rather low, and increases significantly only during the last processing step, i.e. the iterated filtering by keywords.

The processed case study data (without the raw Tweets for licensing reasons) can be accessed via a GeoCommons web map (Fig. 4)[9]:

In summary, the case study performs well indicating that forest fires are represented in the social media content and detectable by automated methods with a 92% of filtered VGI resulting on-topic. The GeoCONAVI system is able to detect the spatio-temporal "reality" of forest fires in social media content. The spatio-temporal clustering finds many clusters of content on fire events (not only forest fires), which can have an enormous echo in social media VGI if they occur in centers of major cities, or cause several fatalities. The implications are that although geographic context can improve situational awareness and assessment of remaining clusters, topicality scoring and filtering contributes most to overall information quality, either before and/or after the spatio-temporal clustering phase.

[8] http://forest.jrc.ec.europa.eu/effis/applications/vgi/.
[9] http://geocommons.com/maps/183605.

Table 4
Processing steps and corresponding data volume.

Processing phase	Data volume
(0) Keyword filtered retrieval from API	21.9 million Tweets, 54,000 Flickr images
(1) Filtering for French keywords	659,676 Tweets, 39,016 Flickr images
(2) Calculating topicality for each VGI and filtering for high scores	25,684 VGI items,
(3) Successfully enriched VGI	5770 VGI items
(4) Spatio-Temporal Clustering	129 clusters containing 2682 VGI
(5) Excluding smaller clusters (<6 items)	75 clusters containing 2565 VGI
(6) Filtering for keywords in clusters	11 clusters containing 469 VGI

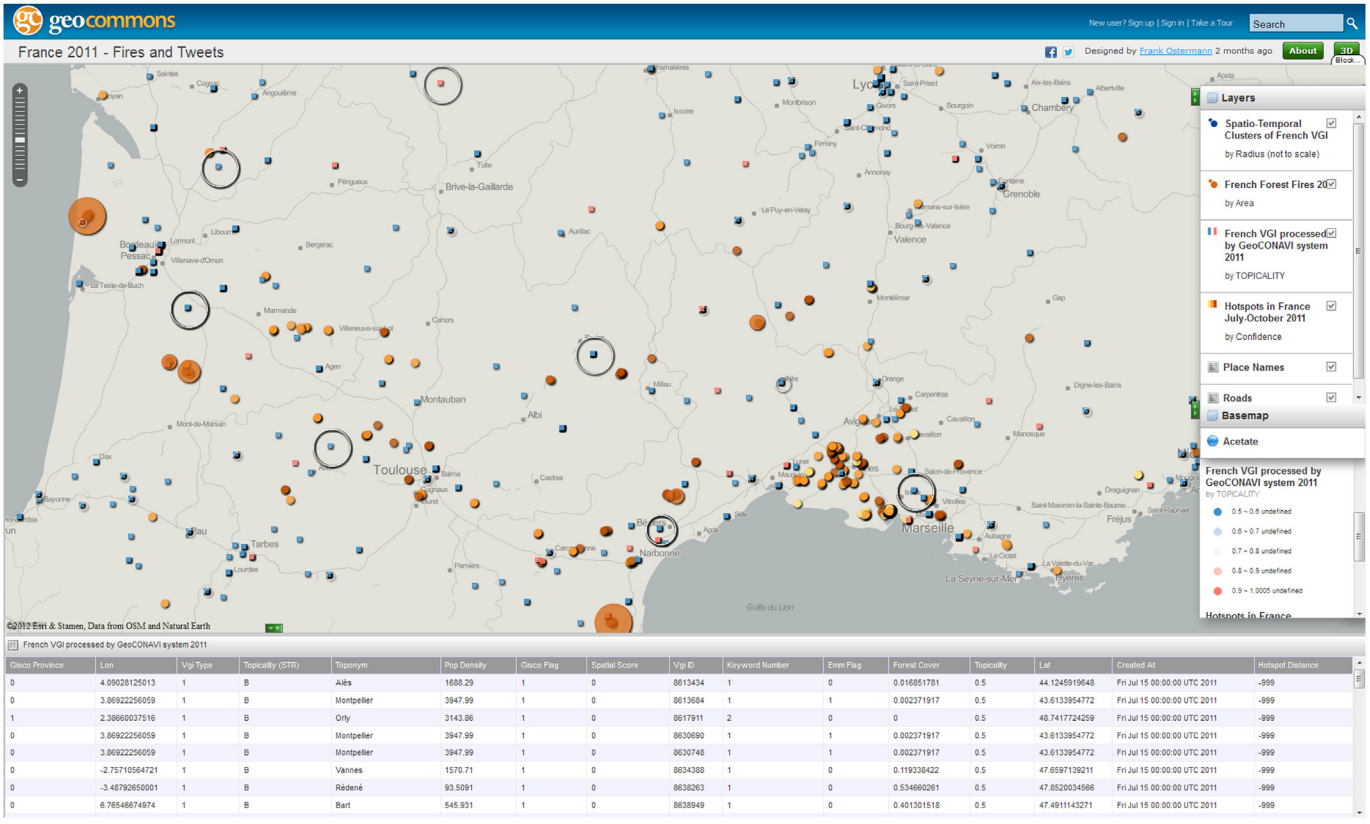
Discussion and evaluation of the system

In this section, we cover the two remaining research questions on data integration and system adaptability in separate sections.

What are the specific design requirements to utilize social media and geographic context information?

An important requirement is flexibility of the platform - technology is a moving target, and during the development and deployment of GeoCONAVI, there have been several changes to the used APIs. Even more critical are changes in the user activity of the platforms. User preference of social networks varies geographically and over time. For example, the Flickr platform has seen a decrease in use during recent years, with other photo-sharing platforms and social networks like Facebook surpassing it (Douglas 2011; Offer 2011). Accordingly, the Sensor module needs constant fine-tuning or calibration, suggesting a brokering approach (Díaz, Granell, Huerta, & Gould, 2012) and we have in fact already explored options in this direction (Schade et al., 2013).

Equally crucial for a successful processing with GeoCONAVI are the sequence of processing and the choice of parameters. Currently, social media content is rarely geo-coded, and even if, the uncertain quality and semantics of this geo-coding recommends additional geo-coding. For a singular piece of information, an optimal solution would be a triangulation of all available location information of some content, e.g. places mentioned in the text or in tags, places mentioned in the source's user profile, and finally any geo-coding already done by the hardware/software/network platform. Additionally, especially for photographs, ancillary information about the angle and focal length, combined with the generation of a view shed based on a digital terrain model, can help to determine what the photograph is about without actually analyzing the image data (Ostermann, Tomko, and Purves 2013). When we consider a cluster of geographically related content, we can also use thematically related information, such as in our case the forest cover and the road network. The case studies have also shown that at least for forest fires, the Sensor picks up a large amount of noise due to the ambiguity of search terms. It makes sense to filter out probable noise before further processing it. A repeated noise filtering of the processed clusters eliminates most noise and increases useful content significantly. In the case studies, relatively simple parameters and approaches worked already well. We can expect an even better performance with improved geo-coding methods that are more error-tolerant and disambiguate better. Further improvements include the addition of data on the sources, and with more geographic context data (e.g. forest fire risk indices). With the inclusion of geographic information, "traditional"



Cisco Province	Lon	Vgi Type	Topicality (STR)	Toponym	Pop Density	Cisco Flag	Spatial Score	Vgi ID	Keyword Number	Emm Flag	Forest Cover	Topicality	Lat	Created At	Hotspot Distance
0	4.09028125013	1	B	Alès	1688.29	1	0	8613434	1	0	0.016851781	0.5	44.1245919648	Fri Jul 15 00:00:00 UTC 2011	-999
0	3.86922256059	1	B	Montpellier	3947.99	1	0	8613684	1	1	0.002371917	0.5	43.6133954772	Fri Jul 15 00:00:00 UTC 2011	-999
1	2.38660037516	1	B	Orly	3143.86	1	0	8617911	2	0	0	0.5	48.7417724259	Fri Jul 15 00:00:00 UTC 2011	-999
0	3.86922256059	1	B	Montpellier	3947.99	1	0	8630690	1	1	0.002371917	0.5	43.6133954772	Fri Jul 15 00:00:00 UTC 2011	-999
0	3.86922256059	1	B	Montpellier	3947.99	1	0	8630748	1	1	0.002371917	0.5	43.6133954772	Fri Jul 15 00:00:00 UTC 2011	-999
0	-2.75710564721	1	B	Vannes	1570.71	1	0	8634388	1	0	0.119338422	0.5	47.6597139211	Fri Jul 15 00:00:00 UTC 2011	-999
0	-3.48792650001	1	B	Rédené	93.5091	1	0	8638263	1	0	0.534660261	0.5	47.8520034566	Fri Jul 15 00:00:00 UTC 2011	-999
0	6.76546874974	1	B	Bart	545.931	1	0	8638949	1	0	0.401301518	0.5	47.4911143271	Fri Jul 15 00:00:00 UTC 2011	-999

Fig. 4. Results from the 2011 case study

challenges enter the analysis, including questions of scale and areas for aggregating data (i.e. modifiable areal unit problem).

There are several avenues for disseminating results from systems such as GeoCONAVI, including broadcasting via dedicated social media channels, SMS, and web maps. Another option would be to send alerts during the response phase through the Sensor Event Service (SES) (Bröring et al. 2011), which can be used to push sensor data (including detected events) to subscribed clients. Similar to the input, the output can easily be adapted to suit the needs of different user groups such as decision makers or citizens on the ground. While decision makers work collaboratively using large desktop screens or wall-projections to coordinate crisis response, citizens are likely to be mobile, employing devices with a small screen, and looking for concrete information on evacuation routes, shelters, or the whereabouts of friends and family members.

There are also a number of unresolved legal issues arising from VGI. Two main problems could be copyright infringements and incorrect information. In both cases, the provider of raw VGI and derived information could be liable in negligence. Although a clearly visible disclaimer and the ability to prove that the provider adhered to basic standards of care should be sufficient (Hickling Arthurs Low Corporation, 2012), administrations are still likely to be reluctant to agree on a tight integration of VGI and official data because of these liability issues.

What is the performance of the prototype system, and how well is adaptable to other crisis events?

The system is intended to work in near real-time, with the Clusterer having to work with batches of VGI. The information available, the type of event and the intended use determine the exact frequency.

GeoCONAVI has shown that it is possible to analyze a multilingual single-topic crisis event type with standard off-the-shelf hardware and medium-sized enterprise DBMS: Looking at Fig. 2, every module that is not a PLSQL job runs on a single Intel Xeon X5550 machine with 12 GB RAM without significant CPU usage. A computational bottleneck is the search for toponyms within the Oracle DBMS, but only because GeoCONAVI implements the simplest approach possible, i.e. brute-force string matching. We expect a fully optimized and indexed DBMS to perform much better, without even including the implementation of advanced techniques such as map-reduce. Thus, from a resource point of view, the GeoCONAVI system provides valuable information with few investments necessary compared to traditional remote sensing and emergency response infrastructure.

The GeoCONAVI system has been implemented for the specific use case of forest fires, but the system architecture was designed to be easily adaptable to different types of environmental events. The input of domain experts is indispensable to provide information on useful keywords, valuable context information and parameters for the spatio-temporal clustering. Since the full original data is retained, users of the system can easily adjust the displayed information by narrowing it down through keyword and geographic queries, or weighting and filtering of the topicality and geographic context scores. Currently, this is only possible through directly manipulating the DBMS and not through a graphic or web user interface, but this limitation is easy to overcome. There are number of portals providing inspiration, such as Twitcident.[10]

[10] http://twitcident.com/.

Conclusions

The results from our case studies show that social media content encloses potentially useful information, and can act as additional communication channel for citizens who have been affected by a disaster. An obstacle to an efficient use is the content's sheer volume and its unstructured nature. Very often, neither the available hardware nor software allows citizens to search social media content efficiently, and make sure all important information is received and read. Therefore, the integration and dissemination of social media content is an important and valuable contribution to the overall disaster management effort. The results presented in this paper show that focusing on the geographic context of the VGI provides a useful approach to deal with the information overload by filtering and assessing the social media content based on credible, authoritative spatial information.

We are convinced that it is in any case necessary to monitor and analyze social media streams during disasters. While some argue that social media is a self-correcting medium and the "wisdom of the crowds" will eventually single out and delete false information, this may happen too late in a time-critical situation like a crisis event. Therefore, the authorities charged with managing the disaster can try to counter false information - if they know about it. With the raw data still being available, the sources of misinformation can be found. This, however, directly points to important ethical questions of privacy and consent to use. Unless the user-content has been volunteered for a specific purpose, either through a portal or the use of certain (hash-) tags, all social media content could be considered "contributed" and not "volunteered" (Harvey, 2012).

Acknowledgments

This work profited greatly from the expertise of a number of colleagues at the JRC, most notably the EFFIS and EMM research groups. We would like to thank Dr. Massimo Craglia from the Digital Earth and Reference Data Unit in particular for invaluable input and support. The research was funded by an internal JRC grant for exploratory research.

References

Al-Khudhairy, D. H. A. (2010). Geo-spatial information and technologies in support of EU – crisis management. *International Journal of Digital Earth, 3*(1), 1–16.

Bröring, A., Echterhoff, J., Jirka, S., Simonis, I., Everding, T., Stasch, C., et al. (2011). New generation sensor web enablement. *Sensors, 11*(3), 2652–2699.

Cheng, Z., Caverlee, J., & Lee, K. (2010). You are where you Tweet: a content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management* (pp. 759–768). Toronto, ON, Canada: ACM.

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), 398–416.

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), 242–259.

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, USA.

Díaz, L., Granell, C., Huerta, J., & Gould, M. (2012). Web 2.0 Broker: a standards-based service for spatio-temporal search of crowd-sourced information. *Applied Geography, 35*(1–2), 448–459.

Douglas, N. (2011). *Flickr is dying. Here's the graph*. Slacktory. http://slacktory.com/2011/08/flickr-is-dying-graph/ Accessed 31.01.13.

Flanagin, A., & Metzger, M. (2008). The credibility of volunteered geographic information. *GeoJournal, 72*(3), 137–148.

Goodchild, M. F. (2007). Citizens as voluntary sensors: spatial data infrastructure in the world of Web 2.0. *International Journal of Spatial Data Infrastructures Research, 2*(1), 24–32.

Harvey, F. (2012). To volunteer or to contribute locational information? Towards truth in labeling for crowdsourced geographic information. In D. Sui, S. Elwood, & M. Goodchild (Eds.), *Crowdsourcing geographic knowledge: Volunteered Geographic Information (VGI) in theory and practice* (pp. 31–42). Berlin: Springer.

Hecht, B., Hong, L., Suh, B., & Chi, E. H. (2011). Tweets from Justin Bieber's heart: the dynamics of the location field in user profiles. In *Proceedings of the 2011 Annual Conference on Human Factors in Computing Systems* (pp. 237–246). Vancouver, BC, Canada: ACM.

Hickling Arthurs Low Corporation. (2012). *Canadian Geospatial Data Infrastructure volunteered geographic information (VGI) primer (No. 21e)*. In *Canadian Geospatial Data Infrastructure*.

Hughes, A. L., & Palen, L. (2009). Twitter adoption and use in mass convergence and emergency events. In *Proceedings of the 6th International ISCRAM Conference*. Gothenburg, Sweden.

Jennex, M. E. (2010). Implementing social media in crisis response using knowledge management. *International Journal of Information Systems for Crisis Response and Management, 2*(4), 20–32.

Kulldorff, M., Heffernan, R., Hartman, J., Assunção, R., & Mostashari, F. (2005). A space–time permutation scan statistic for disease outbreak detection. *PLoS Medicine, 2*(3), e59.

Liu, S. B., & Palen, L. (2010). The new cartographers: crisis map mashups and the emergence of neogeographic practice. *Cartography and Geographic Information Science, 37*(1), 69–90.

Offer, J. (2011). *The slow decline of flickr*. Codehesive. http://www.codehesive.com/index.php/archive/the-slow-decline-of-flickr/ Accessed 31.01.13.

Ostermann, F. O., & Spinsanti, L. (2011). A conceptual workflow for automatically assessing the quality of volunteered geographic information for crisis management. In *Proceedings of AGILE 2011*. Utrecht, Netherlands.

Ostermann, F. O., & Spinsanti, L. (2013). Context analysis of volunteered geographic information from social media networks to support disaster management: a case study on forest fires. *International Journal of Information Systems for Crisis Response and Management, 4*(4), 16–37.

Ostermann, F. O., Tomko, M., & Purves, R. (2013). User evaluation of automatically generated keywords and toponyms for geo-referenced images. *Journal of the American Society for Information Science and Technology, 64*(3), 480–499.

Palen, L., & Liu, S. B. (2007). Citizen communications in crisis: anticipating a future of ICT-supported public participation. In *CHI 2007 proceedings* (pp. 727–736). San Jose, USA.

Puras, J. C., & Iglesias, C. A. (2009). Disasters2.0. Application of Web2.0 technologies in emergency situations. In *Proceedings of the 6th International ISCRAM Conference*. Gothenburg, Sweden.

Roche, S., Propeck-Zimmermann, E., & Mericskay, B. (2011). GeoWeb and crisis management: issues and perspectives of volunteered geographic information. *GeoJournal, Online First*, 1–20.

Schade, S., Díaz, L., Ostermann, F. O., Spinsanti, L., Luraschi, G., Cox, S., et al. (2013). Citizen-based sensing of crisis events: sensor web enablement for volunteered geographic information. *Applied Geomatics, 5*(1), 3–18.

Starbird, K., & Stamberger, J. (2010). Tweak the tweet: leveraging proliferation with a prescriptive syntax to support citizen reporting. In *Proceedings of the 7th International ISCRAM Conference*. Seattle, USA.