

E²mC: Improving Rapid Mapping with Social Network Information

Jose Luis Fernandez-Marquez**, Chiara Francalanci*, Sharada Mohanty***, Rosy Mondardini**,
Barbara Pernici*, Gabriele Scalia*

* Politecnico di Milano

** University of Geneva

*** EPFL

Contact author: chiara.francalanci@polimi.it

Abstract

E²mC aims to demonstrate the technical and operational feasibility of the integration of social media analysis and crowdsourced information within both the Rapid Mapping and Early Warning Components of Copernicus Emergency Management Service (EMS). Copernicus is a European Commission programme developing information services based on satellite earth observation. A fundamental innovation with E²mC is to combine the automated analysis of social media information with crowdsourcing, with the general goal of improving the quality and dependability of the information provided to professional users within the Copernicus network. The automated analyses will focus on multimedia information (mainly pictures), which is most useful for rapid mapping purposes. A fundamental challenge to enable the effective use of multimedia information is geolocation. The paper presents a methodology to extract, integrate and geolocate information from social media and leverage the crowd to clean, validate and complement this information. Preliminary results from testing the methodology are presented based on the analysis of tweets on the earthquake that struck Central Italy in August 2016.

1. Introduction

The availability of a number of information sources from social media makes it possible to use it as a basis for gathering information to support decision making and other activities. Social media have proven particularly valuable in collecting information in emergency situations, such as in earthquakes, in particular in very destructive ones such as Haiti and Nepal, and in large floods and hurricanes, such as Hurricane Matthew and floods in the Philippines, in the different phases of such events (E²mC Team, 2017). Several social media platforms are used in crisis management, including text analytics, event detection, image and video analysis, geospatial analysis; on the other hand, also crowdsourcing initiatives are often started in crisis situations, for mapping purposes such as in Humanitarian OpenStreet Map projects (<https://www.hotosm.org/>) or to gather information with the support of networks of volunteers, such as in the case of the Haiti earthquake, where four different crowdsourcing initiatives were used providing the information needed to coordinate the emergency response, namely Open Street Map, CrisisCamp Haiti, Ushahidi, and GeoCommons (Castillo, 2016; E²mC Team, 2017). In this paper, we propose to combine information gathering from social media and crowdsourcing for supporting a specific task that needs to be performed in such events with

stringent temporal constraints: Rapid Mapping. The production of updated maps helps rescue teams and field operators to understand how the area has been affected, the severity of damages, the grouping of persons in given areas, and they need to be produced within the first 24-72 hours after the events. Such activity is currently supported by earth observation activities such as Copernicus, with Emergency Management Services (EMS) (<http://emergency.copernicus.eu/mapping/>), with tools to produce maps based on satellite data. The goal of this paper is to present how rapid mapping activities can be supported by images extracted from social media and with crowdsourcing, and to delineate a path for improving the quality of additional information to the available satellite data to prepare more timely and more precise maps. The goal of reducing mapping times significantly and improving their precision is being studied within the E²mC (Evolution of Emergency Copernicus services) H2020 European Project (<https://www.e2mc-project.eu/>).

The paper is structured as follows. In Section 2., we briefly introduce the E²mC goals and architecture. In Section 3., the E²mC approach to image extraction and geolocation is illustrated, and in Section 4. we discuss how a process including crowdsourcing tasks can improve the quantity and quality of available information for rapid mapping.

2. E²MC: goals and architecture

E²MC aims at integrating social media analysis and crowdsourcing in a new *Copernicus Witness EMS (Emergency Management System) Service Component*, to improve the timeliness and accuracy of geo-spatial information provided during the crisis management cycle and, particularly, in the first hours immediately after the event. The focus of E²mC is on rapid mapping. To this aim, the project aims to leverage social media analysis and crowdsourcing techniques. These techniques should be used in combination, by exploiting the synergies of both within a unified Social&Crowd platform. Personnel involved in crisis management currently uses social media as a source of information that is considered useful at the beginning of the response phase. Information is searched for manually and classified by manual inspection or by contacting the author of the information (for example, geotagging information is often gathered by writing to the social media users who have posted interesting pictures). This manual process allows employees to make a fast, although rough, assessment of the impact of the crisis, in particular when satellite images are not yet available. On the other hand, when official information emerges after some time from the start of the crisis, they use social media information only to confirm official data when their interpretation is not straightforward. While crowdsourcing has been previously applied to mapping, the use of social media information (i.e., the information that is spontaneously published by social media users on general purpose social channels, such as Twitter) for mapping purposes is limited. In particular, to the best of our knowledge, multimedia information, that is pictures and videos, have not been previously collected, automatically processed and practically used to improve maps produced with satellite technologies in the context of emergency management.

While there exists a vast body of literature either on social media and emergency management, or on crowdsourcing and emergency management, there is a lack of experiences and technologies that allow the conjunct exploitation of social media and crowdsourcing in the

context of emergency management in the early warning and rapid mapping phases. Technical research aimed to provide integrated Social&Crowd solutions has also been explicitly called for in the pivotal survey paper published in (Imran et al., 2015).

3. Automated extraction and geolocation of social media content

Rapid mapping can significantly benefit from associating a location to the extracted multimedia, especially images. However, this association is often challenging. Most social media allow users to *geotag* the items that they post (that is, attaching a geographical location to the items in the form of metadata), but, in practice, only a small percentage of the content posted on social media is geotagged. For example, it is estimated that, on average, only the 0.5%-2% of tweets are geotagged (Inkpen et al., 2015; Castillo, 2016), but could be lower than that in specific and particularly unfortunate emergency contexts (Francalanci et al., 2017). Moreover, the location associated to a content, such as an image, does not necessarily coincide with the location of the post, leading to interpretation errors. Many social media, including Twitter, process images when uploaded by removing the associated location-related metadata and, therefore, a common solution is to infer the location of an image based on the location associated with the related post (*tweet*) (Castillo, 2016).

Focusing on natively geotagged content would limit the number of images available for rapid mapping, potentially excluding the bulk of the useful multimedia content. Being able to increase the percentage of posts with geolocation information could be extremely helpful. Assigning a location to an item can be based on both implicit and explicit geographical references contained in the item. Indeed, “while explicit metadata about locations may be absent, many messages in social media do contain implicit references to names of places” (Imran et al., 2015).

In (Francalanci et al., 2017), authors discuss the feasibility of geolocating images shared on Twitter messages by extracting implicit geographical references from the text of the tweet, and highlight a correlation existing between text features and image features. A traditional pipeline setting has been employed, based on a *recognition* phase, where location names are recognized in the text, followed by a *disambiguation* (or *geocoding*) phase, where location names are geocoded according to the locations they refer to. Besides the challenges related to the usefulness of the extracted locations, like their precision, accuracy and credibility, the extraction phase is challenging and introduces errors (false positives and false negatives), due to ambiguities which exist between location names and other proper or common names. These are called *geo/non-geo* ambiguities, such as the Italian city called None, which coincides with a very common word in the English language. There are also ambiguities among location names themselves (*geo/geo* ambiguities, as in London, UK; London, ON, Canada; London, OH, USA; London, TX, USA; London, CA, USA) (Inkpen et al., 2015).

To overcome these problems, the idea of using implicit geographical references has been refined by taking into account the social network of a post to obtain an *additional context*

potentially useful in overcoming ambiguities and, at the same time, privilege the extraction of locations related to the target event (Scalia, 2017).

The algorithm selects a set of candidate locations for each tweet, identified as the n -grams which potentially refer to a location. In the current implementation, they are obtained using high-recall NER (Named-Entity Recognition) with multi-language capabilities (Al-Rfou et al., 2015; Sasaki et al., 2013), but, in general, they could also include other kinds of candidates, such as n -grams matching predefined patterns. Then, the algorithm tries to disambiguate candidate locations by building a *local* context, that is searching for geographical correlations among candidate locations. Only the locations for which it is possible to find a correlation are disambiguated, while the others are considered ambiguous. This technique is traditionally used for longer and context-rich documents, like web pages, but the short and decontextualized nature of tweets tends to reduce its effectiveness. Working on a case study (the earthquake that struck central Italy in August 2016), we have found that less than 4% of tweets have locations that can be disambiguated by building a local context. We have extended the algorithm to connect tweets in a *behavioral social network*, that is a social network based on implicit interactions among messages rather than explicit relationships among users (Castillo, 2016). In particular, tweets are connected if they share a similar content or belong to the same conversation. The social network provides an additional context by connecting tweets with both implicit and explicit relationships, thus building a *global* context. The global context overcomes the limits of the local contexts, allowing the disambiguation of their location or, alternatively, the inference of a related location if no locations are explicitly mentioned in the tweet's text. The idea of leveraging a behavioral social network is very recent, but has been used successfully in other research areas, such as topic identification (Nugroho et al., 2017). Using the global context, the algorithm is able to disambiguate locations in more than 20% of tweets with a precision > 90% in the Italian earthquake case study. As an example, let us consider the tweet in Fig. 3.2: the global context, below, is built using the neighbors of the tweet and allows not only the identification of "Saletta" as a location, but also the correct disambiguation of "Saletta" as a small town close to Amatrice, rather than one of the other more populated locations named "Saletta" that exist in Italy.

presunta vittima a **#Saletta** estratta viva dalle macerie poco fa! Ci sono ancora persone da tirare fuori, avanti così! #terremoto

- Saletta, Ferrara: population 1469
- **Saletta, Amatrice**: population 33
- Global context

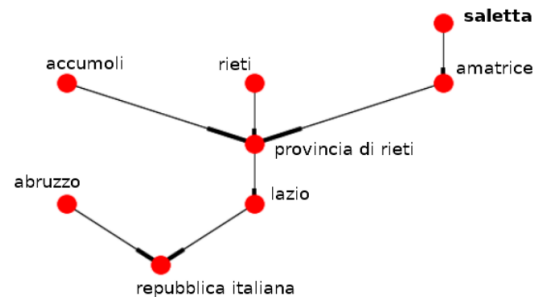


Figure 3.2 - Example of disambiguation of the location “Saletta” using the *global* context provided by other tweets in relationship

The algorithm is currently under development to further improve performance. Even if improving recognition, disambiguation and inference reduce geolocation errors, false positives/negatives cannot be totally avoided and represent an issue that can be successfully tackled with crowdsourcing (see Section 4).

Challenges are not only limited to the location extraction, but involve the extracted information itself: there exist a series of *imprecisions*, *uncertainties* and *ambiguities* in the information that can be extracted from Twitter, which evolves over time during an event. A content item (e.g., an image) can be re-posted at different points in time with a location information that can vary in *precision* (the granularity of the location) and *accuracy* (the likelihood that the data reflect the truth). Moreover, an item is not necessarily linked to a single user and, therefore, there could exist several versions of the same item which bring complementary or contradictory information. Also the correlation among the different items is not guaranteed: the image attached to a message could describe a different location with respect to the text, or could describe the right location but be related to another previous event. Let us consider the example shown in Fig. 4.2: the same image is posted by several different users in the first few hours after the earthquake with different descriptions. If, on one side, the different descriptions may allow us to increase the precision and accuracy of geolocation information, they can also create ambiguities that are difficult to be solved automatically when there is a conflict in the information provided by different authors. In Section 4, we discuss how crowdsourcing can help make decisions to solve the ambiguities in time-related information.

4. A Social&Crowd methodology

This section describes a methodology combining social media analysis and crowdsourcing for both Rapid Mapping and Early Warning components of the EMS. Additionally, this section describes the crowdsourcing contribution to overcome the limitations of automated analysis of social media content.

The Social&Crowd methodology is composed of the following processes:

1. **Data gathering** is the process of extracting relevant information for a given crisis event. Information can be obtained by the automatic analysis of social media, or by the crowd using browser plugins and/or mobile phone app.
2. **Automatic data validation and geolocalization** is an automatic process for establishing the relevance of the information gathered previously, and extract potential localization of the content as mentioned in Section 3.
3. **Crowdsourcing data validation and geolocalization** is a crowdsourcing process to validate the information extracted automatically for the social media analysis and solve disambiguations regarding potential locations.
4. **Information aggregation and visualization** merges the information coming from different sources and visualize it. Information gets ready to be used on the crisis management.

A more detailed information about the interaction between the different processes is presented in Figure 4.1.

The combination of crowdsourcing with the automated analysis of social media information allows to:

1. Reduce the amount of data gathered by the automated analysis of social media by filtering non relevant data.
2. Increase the data quality by contributing on the geolocalization process and solving disambiguation.
3. Increase the relevant media information by allowing the crowd to inject media content in the system.

4.2. Reducing amount of data

Automated analysis of social media information usually provide false positives, i.e., information not related for the crisis management. In this methodology we propose the use of crowdsourcing for the validation of the social information gathered from automatic analysis. Figure 4.2. shows an example of the crowdsourcing project where images gathered from social media and the associated texts are presented to the volunteers.

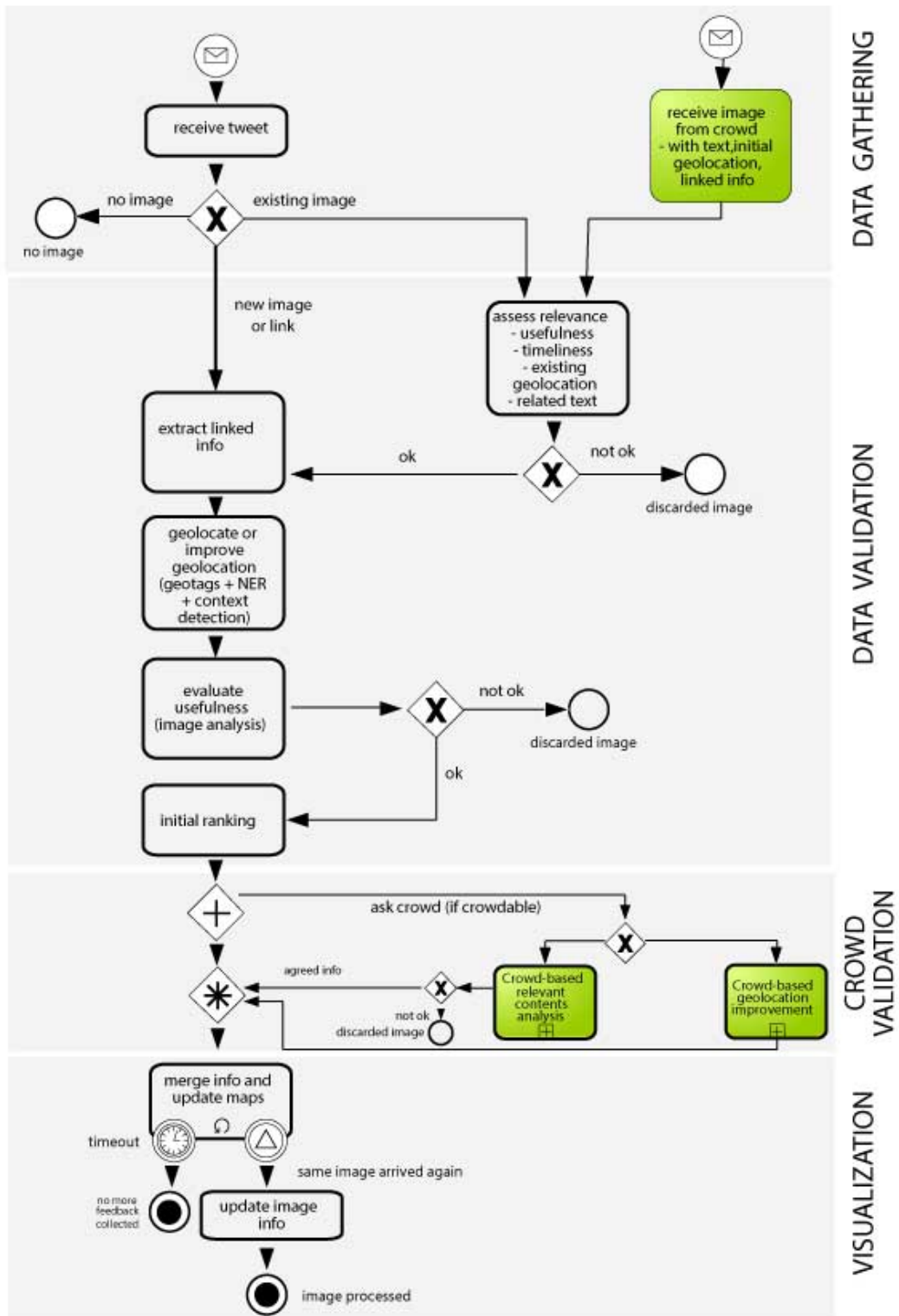
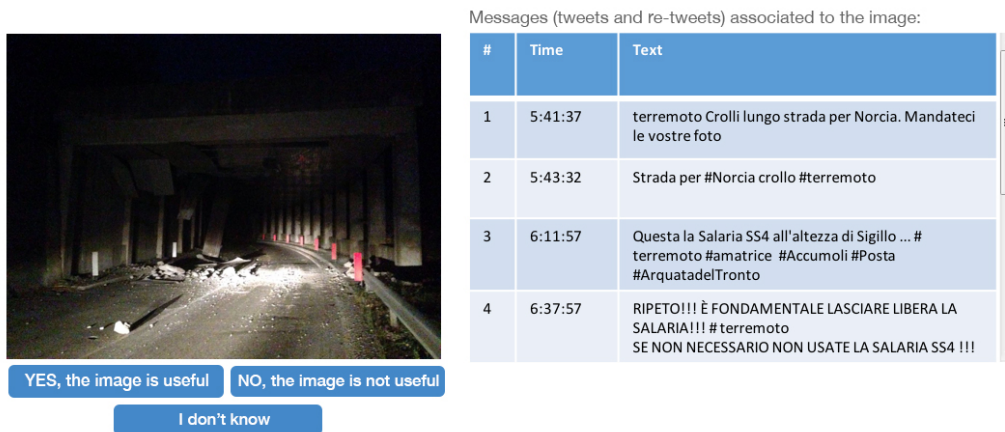


Figure 4.1 - S&C methodology

Volunteers answer whether the information is relevant for a given crisis situation or is not. When an agreement between the answers from the crowd is reached (e.g., 8 out of ten answers agreed that the content is relevant), the information is moved to the aggregation and visualization phase or discarded.

Is this image providing useful information about the disaster ?



#	Time	Text
1	5:41:37	terremoto Crolli lungo strada per Norcia. Mandateci le vostre foto
2	5:43:32	Strada per #Norcia crollo #terremoto
3	6:11:57	Questa la Salaria SS4 all'altezza di Sigillo ... # terremoto #amatrice #Accumoli #Posta #ArquatadelTronto
4	6:37:57	RIPETO!!! È FONDAMENTALE LASCIARE LIBERA LA SALARIA!!! # terremoto SE NON NECESSARIO NON USATE LA SALARIA SS4 !!!

Figure 4.2 - The crowd helps eliminating content that is not relevant to the emergency.

4.3. Increasing data quality by contributing on the geolocation process

Many of the social media content such as images related with a crisis are not geolocated. The automatic extraction of the location can produce different options, or an inaccurate location (e.g., region, or city). Figure 4.3. shows an example of crowdsourcing project where different candidate locations for a given image are provided to the volunteer. In this project the volunteers would contribute to reduce the uncertainty regarding the localization of the image. Additionally, volunteers can contribute to add a more accurate location of the image than those one extracted automatically.

4.4. Adding new data from the crowd.

Automatic extraction of social media information related with a crisis is usually limited to the main social media platforms such as Twitter or Instagram, while many other channels remain unexplored. This unexplored channels would include blogs, online newspapers, or public email list. As already implemented by the GeoTag-X platform (<https://geotagx.org/>), a simple firefox plugin allows volunteers to quickly add information that they considering relevant for a crisis management. Just with one click on the plugin, the link to the source of the information is sent to our database and transferred to the validation process (see Figure 4.5.)

Where is this image located?



ADD LOCATION

I don't know

System suggestions:

Norcia, Perugia, Umbria (Italy)



SELECT

Sigillo, Posta, Rieti, Lazio (Italy)



SELECT

Figure 4.3 - The crowd helps geolocating content by selecting among possible locations proposed automatically by the geolocation algorithm.

Insert the location of the image:



SELECT

Figure 4.4 - The crowd geolocates content by proposing a location that is more precise than that proposed by the geolocation algorithm.

Additionally to the data gathered using the browser plugin, local volunteers can contribute by submitting media content such as images and video relevant to the crisis. This media content can be submitted in two ways: (1) Simply tweeting it with a specific hashtag, or (2) using a specific mobile application. Using Twitter simplifies the process, but images metadata is automatically removed when the image is submitted. Moreover, the text related with the images is unstructured requiring text processing techniques to classify the media content. Using a mobile phone tools, such as Epicollect (<http://www.epicollect.net/>), allows sending the media content with metadata, attached position gathered from the mobile phone GPS and add a form attached with image where volunteers can help to provide extra information regarding the situation in that location. Figure 4.6. shows an example of Epicollect project for school assessment.

6. Conclusions

Images shared on social media can represent a fundamental aid to rapid mapping activities. Their relevance to mapping is dramatically increased by the availability of geolocation information. Unfortunately, this type of information is rarely available. Starting from the assumption that, at the current state of the art, algorithms can help geolocate multimedia content, but are not error free, a fundamental goal of the E²mC project is to combine algorithms and crowdsourcing in a unified methodological approach. In turn, this approach will be supported by an integrated platform.

In this paper, we have presented the first step that the research team has taken to create a unified methodology that combined a software-based and a crowdsourcing approach to filter, disambiguate, classify, and geotag social media information for rapid mapping purposes. Our approach is innovative in that it is based on a novel geolocation algorithm that leverages the real-time, content-based connections among tweets. This algorithm can propose to the crowd a geolocation for a significant percentage of images. Working on the algorithm's output, the crowd can more easily validate and, if necessary, complement the algorithm's results. Even individuals who have no direct, personal knowledge of the location hit by the emergency can contribute to the improvement of the quality of the information that is then fed to the professionals performing the mapping tasks. In this paper, we have highlighted the questions to be forwarded to the crowd depending on the algorithm's output and we have organized the combined work of the algorithm and crowd in a unified information management process.

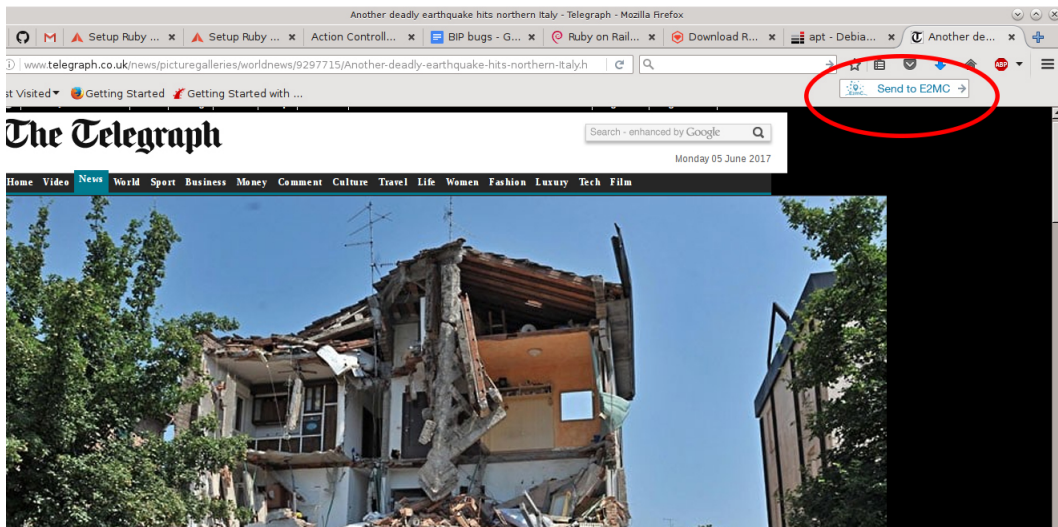


Figure 4.5. Firefox plugin (as developed by GeoTag-X).

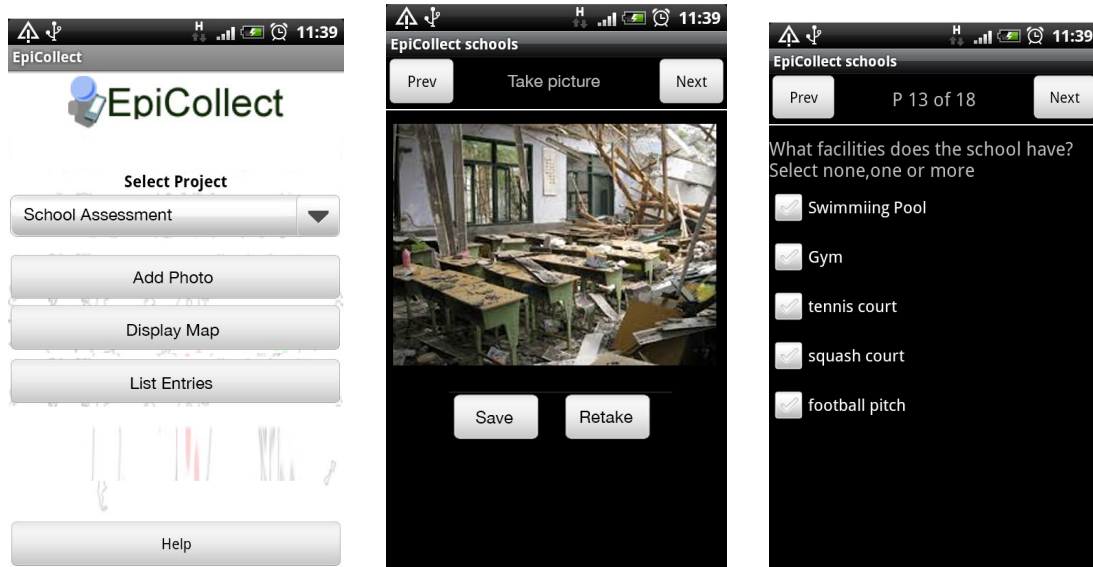


Figure 4.6. Epicollect example

Acknowledgments

This work has been partially funded by the European Commission H2020 project E²mC "Evolution of Emergency Copernicus services" under project No. 730082. This work expresses the opinions of the authors and not necessarily those of the European Commission. The European Commission is not liable for any use that may be made of the information contained in this work. The authors thank Paolo Ravanelli for his support in data management and software development.

References

- R. Al-Rfou, V. Kulkarni, B. Perozzi, and S. Skiena, POLYGLOT-NER: Massive Multilingual Named Entity Recognition, Proc. 2015 SIAM Intl Conf. on Data Mining, pp. 586–594, 2015.
- C. Castillo. Big Crisis Data. Social Media in Disasters and Time-Critical Situations. Cambridge University Press. New York, NY, USA, 2016.
- E²mC Team, Critical review of crowdsourcing and social media use associated with Copernicus EMS service evolution challenges, Deliverable D1.1, public, Jan. 2017
- C. Francalanci and B. Pernici, Data integration and quality requirements in emergency services, in Advances in service-oriented and Cloud computing, A. Lazovik, S. Schulte, Springer Communications in Computer and information Science, in press, 2017
- C. Francalanci, P. Guglielmino, M. Montalcini, G. Scalia, and B. Pernici, IMEXT: A method and system to extract geolocated images from Tweets - Analysis of a case study, IEEE RCIS'17, Brighton, UK, May 2017
- M. Imran, C. Castillo, F. Diaz, and S. Vieweg. Processing social media messages in mass emergency: A survey. ACM Computing Surveys (CSUR), pp. 47(4):67, 2015.
- D. Inkpen, J. Liu, A. Farzindar, F. Kazemi, and D. Ghazi. Detecting and disambiguating locations mentioned in twitter messages. In International Conference on Intelligent Text Processing and Computational Linguistics, pp. 321–332. Springer, 2015.
- R. Nugroho, J. Yang, W. Zhao, C. Paris, and S. Nepal, What and With Whom? Identifying Topics in Twitter Through Both Interactions and Text. IEEE TSC., in press, 2017
- T. Sakaki, M. Okazaki, and Y. Matsuo. Tweet analysis for real-time event detection and earthquake reporting system development. IEEE TKDE, 25(4), pp. 919–931, 2013.
- G. Scalia, Network-based content geolocation on social media for emergency management, Politecnico di Milano, Master's Thesis, April 2017