Data integration and quality requirements in emergency services

Chiara Francalanci⋆ and Barbara Pernici

Politecnico di Milano, DEIB
piazza Leonardo da Vinci, 32, Milano, Italy
`{chiara.francalanci,barbara.pernici}@polimi.it`

Abstract. The requirements and an initial architecture for supporting the analysis of social media contents in emergency management tools are discussed. The use of social media to improve emergency maps and to support early warning is discussed, and the emerging requirements are outlined. The proposed architecture is designed to support enhanced Emergency Copernicus services, as proposed in the European project E^2mC (Evolution of Emergency Copernicus services). In the project a Witness component is going to be developed, to support social media monitoring and analysis and federated crowdsourcing.

Keywords: emergency services, social media, map improvement, data quality

1 Introduction

Emergency response activities are based on many support services. In this paper we focus on emergency systems providing tools that leverage social media in emergency management processes.

A classification of social media use in crisis response has been proposed in [6], where messages are classified in the following categories:

1. Requests for assistance by victims.
2. Distribute official warnings.
3. Establish situational awareness.
4. Sharing media (e.g., pictures) to assist in damage estimation projects.
5. Possibility of direct engagement among users, citizens, organizations, responders (not only to share information).

In the early phases of an emergency, in most cases it is important to exploit all possible information sources is a very short lapse of time, therefore an automatic analysis of these messages can provide further support in the emergency management activities. An initial proposal of a system for extracting information

---

from tweets in emergency situations was put forward in the TORCIA project [4, 5], funded by the Lombardy Region, to manage information originated by tweets and transforming it into useful information, such as, for instance, early warning signals for events and identifying consequences of events on the infrastructures, providing their geolocalization.

A further step in this direction is planned in the E^2mC H2020 European project, starting in November 2016, which aims to demonstrate the technical and operational feasibility of the integration of social media analysis and crowdsourcing within the full Copernicus EMS (Emergency Management System), which provides annotated maps as a product, starting from satellite and aerial images. The main goals of the project are focusing on exploiting social media and crowdsourcing to improve the Mapping and Early Warning service components of the system, as illustrated in Fig. 1.
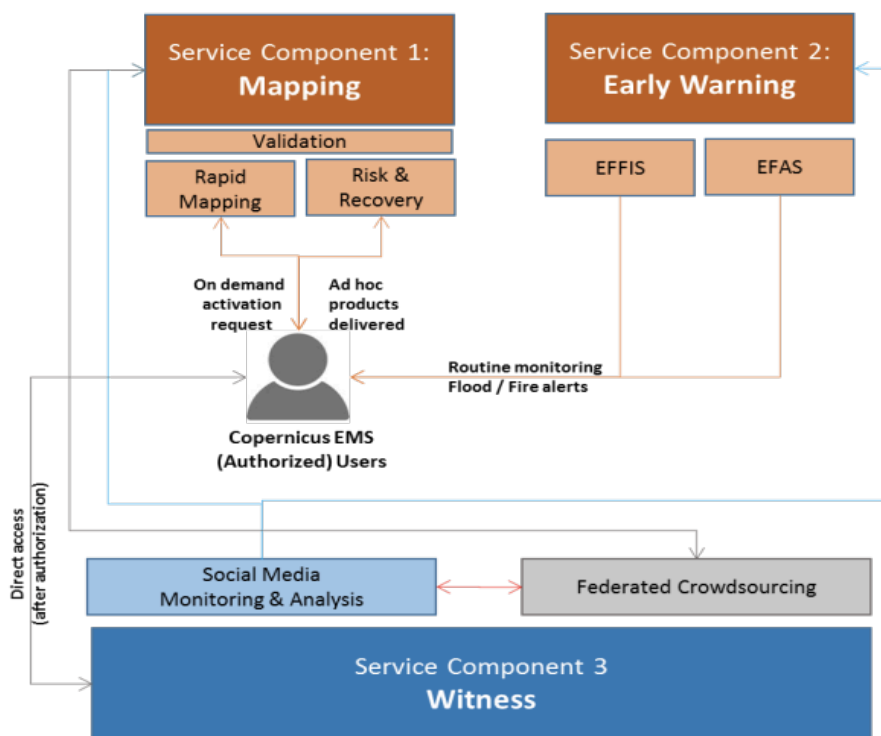
Fig. 1. E^2mC extended Copernicus Emergency Management Service

The architecture illustrated in the figure shows the three main service components of E^2mC:

- *Mapping*: to produce annotated maps in near real-time about areas of interest. An example of map is provided in Fig. 2, showing a grading map that annotates an aerial image with the severity of the damage to buildings and roads.
- *Early warning*: services more and more focusing on the so called impact-based forecast, where an estimate of the potential impact is provided together with the location and severity of the forecasted event.
- *Witness*: a new component to be developed in the project, to collect, integrate, and curate information from multiple social media sources and crowd-sourcing.
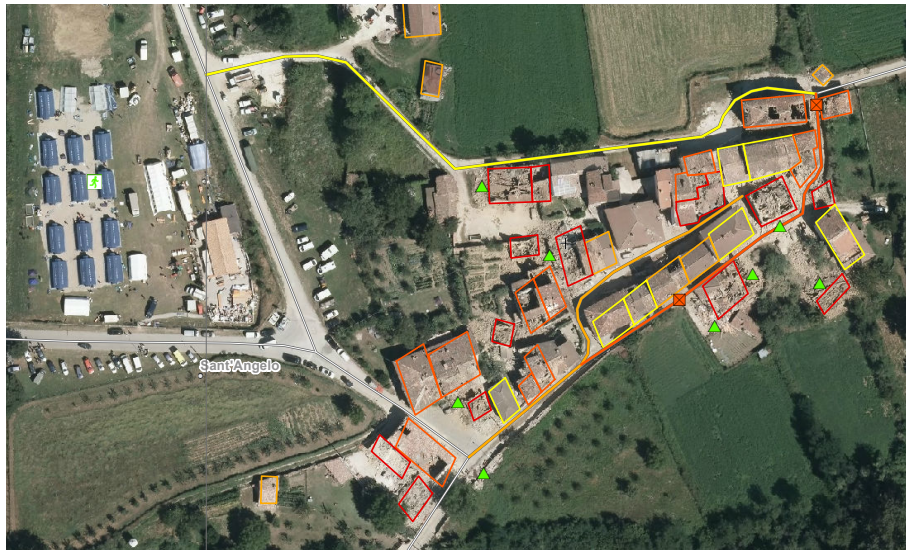
Fig. 2. Detail of an aerial grading map from emergency.copernicus.eu

The goal of this paper is to discuss the main components to be studied in E^2mC and the focus will be on illustrating research challenges of the Witness component.

The project is characterized by the need for integrating or at least relating several different sources of information (existing maps, satellite, social, crowd). The social media, in addition, provide a number of different types of information (text, sound, images, videos). As shown in a recent analysis of two emergency events described in [8], 50% of the tweets contain a link with a URLs. As a consequence, the analysis needs to go beyond the analysis of the text of the tweets, providing also tools to understand and possibly exploit also associated information. As a consequence, first there is the need to build an integrated knowledge base, and then to exploit its contents as a basis for supporting the improvement of the quality of the maps derived from satellite images, in a semiautomatic way.

The Witness service is going to be developed as an independent service, to be connected to the other existing services mentioned above, but also to any other service operating in this domain, as an independent component. In Section 3, we illustrate the main components for information extraction and analysis and underlying design choices for a flexible architecture for E^2mC.

While the quantity of information posted on Twitter is large, the quality of tweets is often low. In a situation of emergency, the institutions and the individuals involved need precise information that can be practically helpful and make their work more efficient and effective. In E2mC, information is useful if it can improve at least one the two fundamental services of Copernicus, that is early warning and rapid mapping. In both cases, the ultimate goal is to improve maps, by providing more timely and complete information. To this aim, the information collected from social media needs to be:

- recognized and interpreted, to guarantee that all required contextual information is available;
- validated, to guarantee dependability;
- integrated, to avoid redundancy and identify spamming.

For example, a picture has to be recognized and associated with a specific point of interest, has to be validated as a trustworthy representation of the emergency situation in the corresponding geographical position and, finally, it has to be grouped with other pictures providing the same type of evidence. These activities cannot be always and exclusively carried out by an automated tool. For example, a picture may represent a building, but the tool has no reference images of that building, or a picture may be posted without geotagging information and the tool fails to associate it with a position. In these cases, a crowd of individuals who are willing to cooperate to improve the quality of social media information can be extremely helpful. In E2mC, crowdsourcing will be aimed to:

- mitigate the risks associated with information integration;
- reduce error rates, interpreting and classifying information;
- improve the overall accuracy and dependability of technology's outputs.

2   Research challenges and potential developments

Several research challenges are posed by a service-based architecture in this specific domain. The main requirements are quantity of data (additional data), its quality (live maps combined with social multimedia contents) and its timeliness (first crisis maps available within few hours), while satellite data are not always available and costly services might need to be requested from the satellite platform: there is a need to complement satellite information to improve quality and the irregular provision of information from the satellites.

2.1   Data requirements and quality requirements

A fundamental goal of the E2mC project is to integrate multimedia information from multiple sources. This goal raises a number of technical challenges that

are both related to specific technical issues (such as the identification of a tiny portion of useful information among large volumes of generic data) and to the need of considering multiple technical issues in combination to provide a usable result (such as the integration of text processing, image processing, geotagging and crowdsourcing functionalities to prepare a database of images that can aid Copernicus services). These issues are made even more challenging by the near real-time requirements and critical situations in which the system operates. The proposed approach is to adopt a keyword-based approach, which facilitates a multilingual annotation, and to support automatic and semi-automatic annotation by operators by micro-tasks for data quality checking supported by the crowdsourcing platform. The service-based approach will be adopted, providing operators with services to rapidly define new keywords and their associations on controlled vocabularies, and to automatically rank data wrt relevance and quality, e.g., from "full trusted" class (input for live notification and early warning process) to "to be checked" one (input for the crowdsourcing task). In particular, quality evaluation will be performed as a two-step process:

1. Authentication of source as reliable: for example, institutional sources representing operators that have emergency management responsibilities and official sources of news will be recognized and analyzed separately.
2. Triangulation of content as valid: this will be performed as a combined effort of software and crowdsourcing.

2.2   Ad hoc workflows with source evaluation

Service ownership and management have to be fully understood, to coordinate the different operators, for access control, and to allow the Copernicus EMS Rapid Mapping Team to build appropriate workflows to streamline the exploitation of data flows provided by the Copernicus Witness. New challenges are raised by the need of preserving privacy of subjects, of considering the quality of the data gathered from the different sources, and to control the output in terms of timeliness and quality fit for use in a given situation.

3   Towards a service-based architecture for E^{2}mC

The core component of the E^{2}mC software architecture is the platform. The platform is composed by a service portal and shared social media data. The service portal provides a set of core REST data management services and a set of REST micro-services that have local data and can access shared data through the data management services. Both types of services are used by the E^{2}mC applications: the Web application, the mobile application, the federated crowdsourcing application and the crawlers.

The main advantage of this architecture is the right balance between integration and flexibility, which are both needed to accommodate for changing requirements throughout the life span of the project. While core data and data

management services are shared, that is, fully integrated, micro-services are independent of each other and can evolve according to different research and experimentation plans. For a same micro-service, different instances or versions can be simultaneously active and operate on different data for either software testing or field experimentation purposes. This trade-off between integration and flexibility is key to emergency management. A core set of data and related data management services is necessary to guarantee that Copernicus rapid mapping and early warning activities have access to integrated information that over time is enriched by different applications, such as crowdsourcing, in a consistent way. Given that social media information has a value, but is not always dependable, it is important that all applications work towards building a common set of consistent data. On the other hand, the micro-service approach gives applications the degree of flexibility that is important to support incremental development and experimentation by independent teams, towards separate goals, along different schedules. At runtime, micro-services can also be beneficial to integrate new functionalities as they emerge.

It should be noted that crawlers are considered an application, even if they are not a user application. This allows designing data management services that accommodate both the insertion of raw posts as well as partly or fully tagged posts.
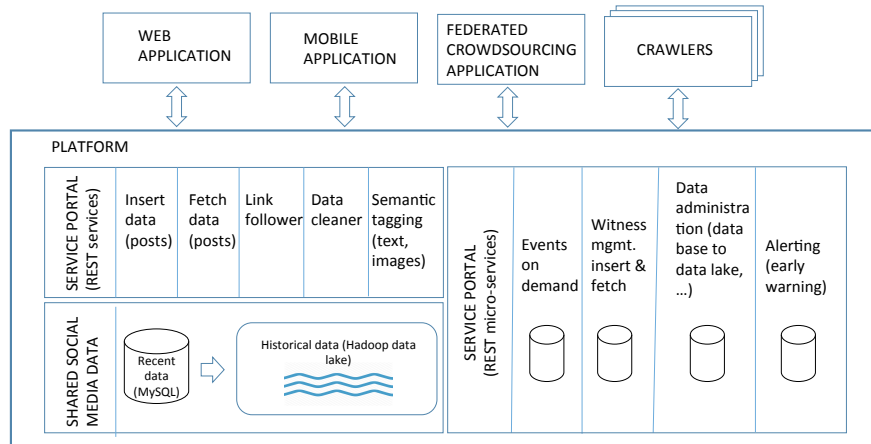
Fig. 3. E^2mC proposed software architecture

The Lambda Architecture [7] shown in Fig. 4 will be deployed as the software infrastructure supporting data storage, data processing and service/micro-service provisioning. The main advantage of the Lambda Architecture is to support the deployment of Hadoop in a context requiring both big data storage&analytics and real-time querying of streaming data. An obvious advantage

of the Hadoop framework is its high performance with data analytics (write once and read many paradigm). On the other hand, E2mC requires micro-services that are based on the (quasi) real-time availability of information, especially for Early Warning. The Lambda Architecture allows the project to benefit from Hadoop performance, while providing the infrastructure for the real-time processing of streaming data from social media.
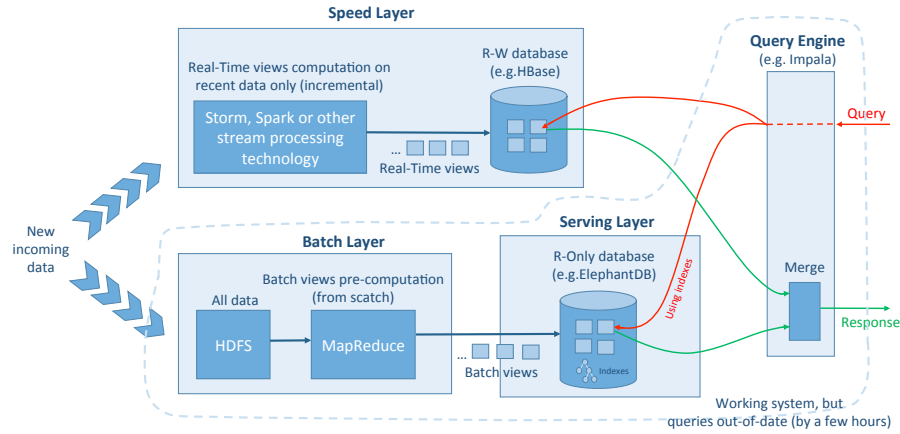

Fig. 4. Lambda architecture

4   Concluding remarks

In this paper, we introduced the main requirements and discussed architectural choices for additional services for Emergency Management Systems, based on social media and crowdsourcing. Several issues are still open and need further investigation in the project. We mention here two important ones which are critical for the development of the project. First, as mentioned above user involvement is planned, assigning users tasks for obtaining further information and human analysis of the data. Such information, as discussed above, is inherently of low quality. In addition, it is important to identify possible groups of reliable and available users to perform the assigned tasks, to devise strategies and workflows to assign tasks and collect and integrate information, and in general creating communities of supporting users during the emergency phases. Another important aspect is the identification of the interested stakeholders for the use of the results of the project, since many different roles and institutions are usually involved in emergency response. It is important to identify a technical solution that is providing also an adequate support to the management of the system and involvement of several actors. The key roles and owners of the in-

volved processes need to be identified and supported by an adequate distributed architecture, both in technical terms, and in ownership and responsibility.,

References

1. Copernicus, a European system for monitoring the Earth, http://www.copernicus.eu/
2. E^{2}mC Evolution of Emergency Copernicus services, Project H2020-EO-2016, `http://pernici.faculty.polimi.it/it/e2mc-evolution-of-emergency-copernicus-services/` (2016)
3. Flamini, A., Grandoni, D., Britti, F., Salvi, F.: e-GEOS capabilities in rapid emergency response. Two case studies: LAquila earthquake and Parma/Pepang typhoon. ISPRS Archive Vol. XXXVIII, Part 4-8-2-W9 (2010)
4. Francalanci, C., Giacomazzi, P.: TORCIA: una piattaforma collaborativa per la gestione delle emergenze, Mondo Digitale, n. 57 (2015)
5. Francalanci, C., Giacomazzi, P.: TORCIA - A decision-support collaborative platform for emergency management, DATE Conference (2015)
6. Lindsay, B.R.: Social Media and Disasters: Current Uses, Future Options, and Policy Considerations. Congressional Research Service 7-5700, www.crs.gov, R41987 (2011)
7. Marz, N., Warren, J.: Big Data: Principles and best practices of scalable realtime data systems, Manning Publications Co. Greenwich, CT, USA (2015)
8. K. Meesters, K., v. Beek, L., V. d. Walle, B.: #Help. The Reality of Social Media Use in Crisis Response: Lessons from a Realistic Crisis Exercise, 49th Hawaii International Conference on System Sciences (HICSS), Koloa, HI, pp. 116-125 (2016)