International Journal of Digital Earth

ISSN: 1753-8947 (Print) 1753-8955 (Online) Journal homepage: http://www.tandfonline.com/loi/tjde20

Digital Earth from vision to practice: making sense of citizen-generated content

M. Craglia , F. Ostermann & L. Spinsanti

To cite this article: M. Craglia , F. Ostermann & L. Spinsanti (2012) Digital Earth from vision to practice: making sense of citizen-generated content, International Journal of Digital Earth, 5:5, 398-416, DOI: 10.1080/17538947.2012.712273

To link to this article: http://dx.doi.org/10.1080/17538947.2012.712273

Accepted author version posted online: 30 Jul 2012.
Published online: 20 Aug 2012.

Submit your article to this journal

Article views: 577

View related articles

Citing articles: 28 View citing articles

Full Terms & Conditions of access and use can be found at
http://www.tandfonline.com/action/journalInformation?journalCode=tjde20

Download by: [Wageningen UR Library] **Date:** 14 May 2017, At: 22:44

International Journal of Digital Earth,
Vol. 5, No. 5, September 2012, 398–416

Taylor & Francis
Taylor & Francis Group

Digital Earth from vision to practice: making sense of citizen-generated content

M. Craglia*, F. Ostermann and L. Spinsanti

Digital Earth and Reference Data Unit, European Commission Joint Research Centre, Ispra, Italy

(*Received 10 May 2012; final version received 6 July 2012*)

The vision of Digital Earth (DE) put recently forward under the auspices of the International Society for DE extends the paradigm of spatial data infrastructures by advocating an interactive and dynamic framework based on near-to-real time information from sensors and citizens. This paper contributes to developing that vision and reports the results of a two-year research project exploring the extent to which it is possible to extract information useful for policy and science from the large volumes of messages and photos being posted daily through social networks. Given the noted concerns about the quality of such data in relation to that provided by authoritative sources, the research has developed a semi-automatic workflow to assess the fitness for purpose of data extracted from Twitter and Flickr, and compared them to that coming from official sources, using forest fires as a case study. The findings indicate that we were able to detect accurately six of eight major fires in France in the summer of 2011, with another four detected by the social networks but not reported by our official source, the European Forest Fire Information Service. These findings and the lessons learned in handling the very large volumes of unstructured data in multiple languages discussed in this study provide useful insights into the value of social network data for policy and science, and contribute to advancing the vision of DE.

Keywords: Digital Earth; social networks; volunteered geographic information; data quality

1. The evolution of the Digital Earth (DE) concept

Digital Earth is a very powerful metaphor conveying the concept of a multi-resolution, multi-dimensional digital replica of our planet helping us to develop a shared understanding of our changing environment and its consequences. First popularised by Vice President Al Gore in 1998, the concept gained some currency in the scientific literature and has recently seen a revival largely due to the efforts of the International Society for Digital Earth (ISDE) supported by the Chinese Academy of Science.

Figure 1 shows the most widely cited scientific papers on DE out of a search on Google Scholar. At the end of 2011, there were 8680 references on DE. The table includes only the 14 references cited 20 times or more (see Table 1).

As shown, we can see two clear clusters. The first around 1999–2000 builds on the Al Gore speech (Gore 1999) and the papers published as a result of the first

*Corresponding author. Email: massimo.craglia@jrc.ec.europa.eu

ISSN 1753-8947 print/ISSN 1753-8955 online
© 2012 European Union
http://dx.doi.org/10.1080/17538947.2012.712273
http://www.tandfonline.com

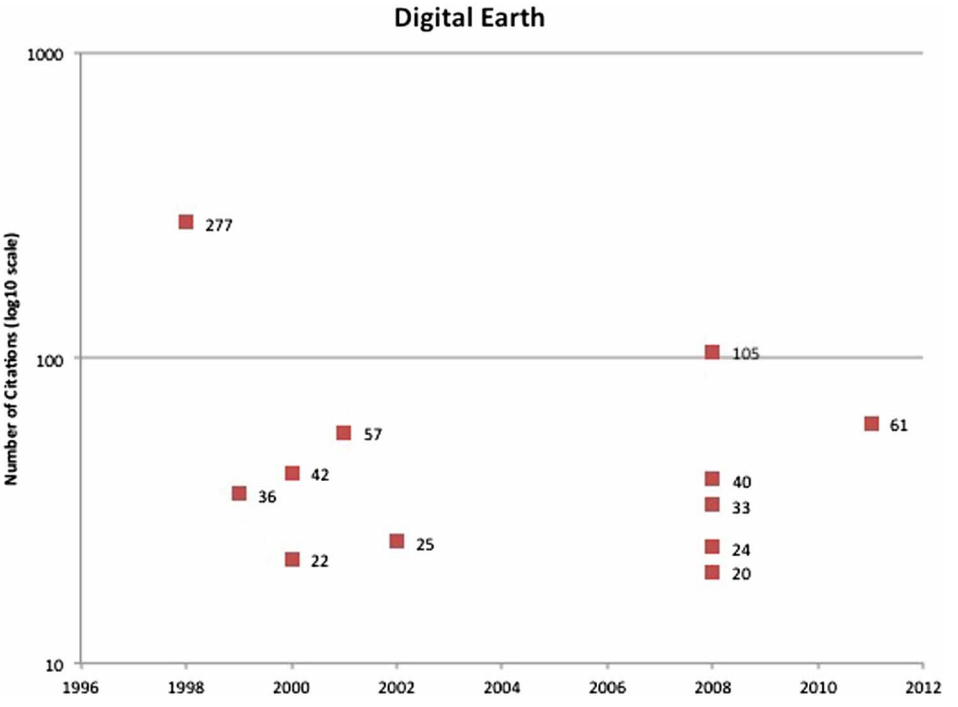

Figure 1. Scientific papers on Digital Earth cited 20 times or more in Google Scholar, accessed 29 December 2011.

international conference on DE organised by the Chinese Academy of Science in Beijing in 1999. The second cluster in 2008–2009 is the result of the first issue of the International Journal of DE, and the position paper on Digital Earth of the Vespucci Initiative for the Advancement of Geographic Information Science (Craglia *et al.* 2008).

The gap in scientific papers on DE during the period 2003–2007 may be understood as a combination of factors: the lack of support for the DE concept by the Republican administration in the USA, the emergence of major private sector initiatives like Google Earth and Microsoft Virtual Earth (now Bing) which contributed enormously to the practical implementation of many features of the Gore vision under the new heading of 'geo-browsing', and the development of government-led initiatives like spatial data infrastructures (SDIs) and the Global Earth Observation System of Systems (GEOSS). This changed political and technological landscape diverted the attention of the scientific community from DE, even though a series of international conferences and symposia continued to be organised regularly by the ISDE. The turning point in 2008–2009 came about as the result of a number of factors:

(1) The recognition of the new opportunities created by open application interfaces for individuals to publish location-explicit information via social networks or group efforts (e.g. OpenStreetMap, Wikimapia, Ushahidi). The scale of these new developments blurred the traditional dichotomy between

Table 1. Analysis of 20 most cited references on Digital Earth from Google Scholar, 29 December 2011.

Title	Year	Citations
The Digital Earth: Understanding Our Planet In The 21st Century	1998	277
Next-Generation Digital Earth: A Position Paper From The Vespucci Initiative For The Advancement Of Geographic Information Science	2008	105
Evaluating Digital Libraries For Teaching And Learning In Undergraduate Education: A Case Study Of The Alexandria Digital Earth Prototype (ADEPT)	2000	78
Cartographic Futures On A Digital Earth	2011	61
Real-Time Global Data Model For The Digital Earth	2000	42
The Use Cases Of Digital Earth	2008	40
NII, NSDI And Digital Earth	1999	36
Distributed Geospatial Information Processing: Sharing Distributed Geospatial Resources To Support Digital Earth	2008	33
Visual Explorations For The Alexandria Digital Earth Prototype	2002	25
The Alexandria Digital Earth Prototype	2001	24
Defining A Digital Earth System	2008	24
Iterative Design And Evaluation Of A Geographic Digital Library For University Students: A Case Study Of The Alexandria Digital Earth Prototype (ADEPT)	2001	23
Discrete Global Grids For Digital Earth	2000	22
Digital Earth In Support Of Global Change Research	2008	20

producers and users of information, and challenged the official modes in the production of knowledge (Goodchild 2007).

(2) An increasing demand in the scientific community for multi-disciplinary efforts to address the global sustainability research (ICSU 2010) and at the same time a pressing need to communicate science better and more transparently. This became all the more urgent after the major controversy surrounding the release of the 2007 report of the Intergovernmental Panel on Climate Change (IPCC).

(3) In 2006, the International Symposia of Digital Earth culminated in the establishment of the International Society of Digital Earth, promoting and supporting international exchange, and leading to the launch of this journal in 2008.

(4) The acknowledgement that neither government-led initiatives like SDIs or GEOSS, nor private sector-led geo-browsers were sufficient to meet the needs of a more open and participatory governance and science. SDIs and GEOSS have made great strides in increasing visibility and to some extent accessibility to government and scientific data, but are still some way to go to support full data access and integration across multiple disciplines, and they are still cumbersome to use. Geo-browsers on the other hand provide limited functionality, but have been successful through ease of use and attractive visualisations.

With these considerations in mind, a reflection among scientists in the academic, government and private sector within the framework of the Vespucci Initiative for the Advancement of Geographic Information Science in 2008 considered how to update the vision of DE and reflect the changed technological and social landscape since 1998. The resulting position paper (Craglia *et al.* 2008) highlighted volunteered geographic information (VGI) and the emergence of ubiquitous sensor networks as the key new elements around which a revised vision of DE should be developed.

Annoni *et al.* (2011) considered DE in the context of the European policy challenges and initiatives. Whilst acknowledging that the term DE is rarely used in Europe, they recognised that there are multiple initiatives in Europe that could be harnessed to contribute to a collective vision of DE. These include, for example, Europe 2020 flagship initiatives on the Digital Agenda and Innovation Union, the development of the Infrastructure for Spatial Information in Europe (INSPIRE) (European Commission 2007), the Global Monitoring for Environment and Security initiative and the strong participation of Europe in the implementation of GEOSS. From a European perspective therefore, the key challenge to contribute to the development of DE is one of governance, that is to try and harness the multiple initiatives in Europe and channel them effectively into the global endeavour. From a scientific and social perspective, however, the key issue identified by Annoni *et al.* (2011) is the need to integrate effectively the top-down approaches of SDIs, GEOSS and related government-led initiatives, with the bottom-up information flows coming from citizens and sensor networks.

The crucial role of individuals was also highlighted in the revised vision for DE that emerged from the brainstorming workshop hosted in Beijing in March 2011 by the Centre for Earth Observation and Digital Earth (CEODE) of the Chinese Academy of Sciences. In this vision, DE in 2020:

- Will be dynamic and interactive exploiting the full range of information flows from sensors and people.
- Will synthesise heterogeneous information and provide metrics of quality and trust of both data inputs and outputs.
- Will be more participative as people will have a greater say in providing data but also in interpreting the data.
- Will be easy and fun to use with different levels of functionality available to different audiences.
- Will be a ubiquitous frame of reference as people and an increasing number of everyday objects will be on-line at all times. (Extract from Craglia *et al.* 2012, pp. 13–14)

As shown, the role of citizen-provided information is central to this vision both in terms of creating more dynamic flows of information, and in terms of a fostering a participative approach, which is crucial to narrow the gap between citizens, government and science.

Given this important role, there is a strong need to develop methods to integrate heterogeneous data and assess their quality and reliability (Craglia *et al.* 2008).This paper offers a unique insight into these research questions, reporting on an exploratory research undertaken at the Joint Research Centre on the use of citizens' data to complement official data on forest fires in Europe.

The paper is articulated as follows: after this initial introduction on the role of citizens' data for the new vision of DE, Section 2 reviews the literature focusing in particular on VGI, Section 3 describes the project and its outcomes and Section 4 concludes with a critical reflection on the added value of VGI for policy and science. As many talk about the value of VGI but few have undertaken rigorous research with very large volumes of VGI, we believe this paper provides a real contribution to the scientific debate on VGI and DE.

2. Citizens as producers of information

Participatory approaches in research and governance are not new (Weiner and Harris 2008). There is long tradition of calling on volunteers to provide information relevant to science, as for example the Birdcount in the USA (http://www.birdsource.org/gbbc/) and Spring Watch in UK (http://www.bbc.co.uk/nature/uk/). What has made a significant difference recently is the diffusion of Internet-based and social networks as media to increase the participation of the public in reporting news, providing information on natural disasters, traffic, tourist information and so on. New technologies, for example Web 2.0 platforms, mobile Internet and social networking access through smartphones, enable the public to contribute and participate on an unprecedented scale and have led to many and diverse initiatives using information by citizens (Elwood *et al.* 2011).

To understand the potential of this change it is worth noticing the amount of social information produced daily in the last years and now. In August 2006, geo-tagging facilities at Flickr started to operate; and by the year of 2007, more than 20 million geo-tagged photos had been uploaded to Flickr. In August 2011, Flickr announced its 6 billionth photo, with an increasing 20% year-over-year, over the last 5 years (http://blog.flickr.net/en/2011/08/04/6000000000/). Similarly, Twitter was launched in 2006. The increase in number of message is impressive: In 2010, the average number of Tweets sent per day was 50 million (http://blog.twitter.com/2011/03/numbers.html), while in March 2012 it was 340 million (http://blog.twitter.com/2012/03/twitter-turns-six.html). Also in 2010, the geo-tagging feature of the tweets was added. Although the amount of geo-enabled messages is around 1% of all messages, this still means one million geo-tagged information messages per day.

The universe of VGI (Goodchild 2007), or neo-geography (Turner 2006) is not uniform and understanding the different components and perspectives is important to develop the strategies necessary to assess the quality of the VGI provided. Through an extensive review of the literature, Coleman *et al.* (2009), p. 341) characterise the type of contributors of VGI (see Table 2).

Although the table provides a good overview of the different facets of VGI, it fails to capture adequately the different modes through which individuals or communities contribute such information. In order to arrive at a basic typology of VGI, we propose to consider two dimensions: first, the way the information was made available, and second, the way geographic information forms part of it.

Each of these two dimensions can be 'explicit' or 'implicit', with explicit denoting that the dimension is of primary concern to the piece of VGI, while implicit denotes that the dimension was not originally an integral part, and is only of secondary concern or derived. So, if a piece of information is about the characteristics of a place, it is explicitly geographic. On the other hand, information that is not about a place

Table 2. Examples of VGI contributors in each category along the spectrum.

	Mapping and navigation *(e.g. GPS-based Car Navigation)*	Social networks *(e.g. OpenStreetMap)*	Civic/governmental *(e.g. PPGIS)*	Emergency reporting *(e.g. Disaster Reporting)*
Neophyte	Relies on unit to provide directions and follows instructions to add basic point information using the Unit.	Identified gaps in map coverage, familiar with the locale, and has obtained the requisite GPS equipment. Interested in making a first contribution.	Views a GIS map in a town hall meeting around the siting of a power plant in the town	May use cell phone to add basic information detailing location of a potential new wildfire outbreak.
Interested amateur	Owns a personal system, uses it extensively, and has made several contributions. Is aware of both technology strengths & limitations and procedures required to make reliable contributions.	Owns the equipment; familiar with data editing software & processes. Regular contributor of edited map data and may assess other contributions.	Citizen fashions a map to present a counter claim in a town hall meeting around the siting of a power plant in the town.	May drive from place to place shooting geo-tagged photos showing extent of floodwaters.
Expert amateur	Familiar with the strengths and weaknesses of multiple systems, has owned more than one. May assess and occasionally amend the contributions of others.	Expert with the requisite equipment. Regularly assesses & edits contributions from others. Participates in specification development & decision-making.	Individual familiar with conditions in a given neighborhood and with the operation of the Web-based PPGIS system in use.	Familiar with requirements for data useful to emergency response personnel and may voluntarily travel to sites to provide such information on an 'on-call' basis.

Table 2 (*Continued*)

	Mapping and navigation *(e.g. GPS-based Car Navigation)*	Social networks *(e.g. OpenStreetMap)*	Civic/governmental *(e.g. PPGIS)*	Emergency reporting *(e.g. Disaster Reporting)*
Expert professional	Mapping or Location-Based Services professional.	Mapping or Location-Based Services professional.	Practicing Urban Planner.	Emergency planning and/or response personnel tasked with mapping the position and geographic extent of a given flood or wildfire.
Expert authority	Specialist consulted by other professionals re: specific problems and/or new developments.		City Planner with extensive knowledge of developments in the area of interest.	Specialist consulted by other professionals re: specific problems and/or new developments.

but can still be geo-coded is implicitly geographic. Likewise, if a piece of information is explicitly volunteered, it was made public by the author and contributed with a specific purpose in mind. Implicitly volunteered information on the other hand has been made publicly available by the author, but was not provided with a specific purpose. This gives us a matrix of four types of VGI as shown in Table 3.

The typology shown in Table 3 has impact on the sensing of VGI. We propose to differentiate between active and passive sensing, which correspond to explicitly volunteered and implicitly volunteered information. Other possible terms would be 'participatory' sensing and 'opportunistic' sensing (Jiang and McGill 2010). The former provides a framework for the citizen participation and includes mentioned examples such as counting birds. The latter approach provides no a priori guidelines, and aims to tap into the abundance of VGI offered on a day-to-day basis. An example would be the information about routing provided by users of TomTom navigation systems as they drive in their daily business.

These different dimensions should be added to the taxonomy in Table 3 to develop a more articulated perspective of VGI. For each row and column, there is potentially a different methodological approach to assess the quality of the information provided. Simplistic notions that only experts actively providing VGI would be quality assured need to be re-assessed. In the first instance, one needs to be clear about who is an 'expert' in what. Professional expertise has recognised standards to adhere to but in many instances local knowledge may be just as valuable (Goodchild 2009). In other cases the cumulative knowledge of a community can achieve high quality results as in the case of Wikipedia or approaches in which volunteered contribute within a rigorous methodological framework as was the case for Seti@Home, or the Bird Count, or organised efforts such as OpenStreetMap. Research on the volunteers contributing to Wikipedia has been undertaken, amongst others, by Anthony *et al.* (2005), and reported by Coleman *et al*. (2009). However, more research needs to be done in relation to VGI and its quality assessment. Currently, the quality assessment of VGI is relying heavily on human volunteers, such as the Stand-By-Task-Force in times of crisis events. This approach has proven to

Table 3. Typology of VGI.

	Geographic	
	Explicit	Implicit
Explicitly volunteered	This is 'True' VGI in the strictest sense. Examples include Open Street Map.	Volunteered (geo)spatial information (VSI). Examples would include Wikipedia articles about non-geographic topics, which contain place names
Implicitly volunteered	Citizen-generated geographic content (CGGC). Examples would include any public Tweet referring to the properties of an Identifiable place.	Citizen-generated (geo)spatial content (CGSC) such as a Tweet simply mentioning a place in the context of another (non-geographic) topic.

work well on several occasions. However, the increase in VGI from all over the globe will lead to problems of sustaining and scaling these efforts, calling for an automated approach that lets volunteers handle the difficult cases only. The research we report in the following sections deals with passively provided VGI in the context of emergency reporting, specifically forest fires. Part of the challenge is that a priori there is no way of knowing in which category of volunteers the providers fit. The next Section 3 describes the research project, and its outcomes.

3. 'Next generation DE: engaging the citizens in forest fire risk and impact assessment' project

We describe in this section the outcomes of an exploratory research project undertaken by the European Commission Joint Research Centre in partnership with Google. The project aimed to address the problems of sustainability which manual processing of social media VGI faces. Therefore, the project developed and tested a methodology to harvest and analyse VGI automatically. As an application, it used the domain of crisis management of forest fires. This domain was chosen as the European Commission's Joint Research Center (JRC) has an institutional responsibility for monitoring forest fires in Europe, and has developed the European Forest Fires Information System (EFFIS, http://effis.jrc.ec.europa.eu). The concrete purpose of the research was to assess the value of VGI at different stages of a fire event (detection, spread, post-fire assessment) and compare this contribution with the flow of information from authoritative sources as used by EFFIS (Spinsanti and Ostermann 2010).

3.1. Project research objectives

We identified three major challenges to the utility of VGI: First, the sheer amount of VGI available, with millions of messages sent and photos uploaded each day. Second, the unstructured nature of VGI, with citizens exploiting the possibilities of the various platforms in creative and sometimes unintended ways. Third, the unknown quality of VGI, since platform specific tools for quality control are only emerging, and cross-platform tools are non-existent. Consequently, to date collecting, filtering, formatting, enriching, assessing and visualising VGI have been mostly in the hands of human volunteers.

As we have argued before, this approach faces issues of scalability and sustainability. Even if we restrict the sources to explicitly volunteered information and disregard all other publicly available information, the amount can be expected to increase substantially over the next years, and therefore it is imperative to develop some automatic or semi-automatic method for assessing the quality and usability of this data. The obvious starting point is to automate as much as possible filtering, validation and quality assessment, and let human analysts deal only with the highly dubious cases. This idea has already been taken up by the community and is expressed in efforts like the Swiftriver (http://ushahidi.com/products/swiftriver-platform) project. However, we propose an approach that goes beyond it. We argue that a crucial component for assessing the quality of short messages is geographic context, as we show in the following section.

3.2. Project approach

The project developed a semi-automatic workflow to retrieve information from various sources and evaluate them against multiple criteria. The aim is to provide a set of geo-referenced, annotated and evaluated information about wildfires, which can be used to enrich official authoritative data and accessed on the Web (Ostermann and Spinsanti 2011; Schade *et al.* in press).

The two key criteria for which we developed automated scoring functions are *relevance* and *credibility*. While there is already a substantial amount of research on assessing relevance and credibility from an information retrieval point of view, several issues make a new approach based on spatial context necessary and promising to pursue. First, in the context of crisis management, location is of paramount importance. Decision-makers and emergency workers need to know the location of an event to respond to it. However, knowledge about the location can also be used to create the context that is necessary to evaluate the credibility and the relevance. This is important as context may be otherwise very difficult to establish, considering that the information retrieved is often very short, and only little is known about its source.

Our process model approaches this task by adapting three heuristics humans use to deal with new information (Metzger *et al.* 2010): What do others know and say (social confirmation), what do we know about the situation the information is referring to (expectancies) and how does the information relate to my current needs? In particular, we focus on the second aspect by attempting to place any VGI within its geographic context.

The measure *relevance* is a well-established quality metric in the information (retrieval) sciences. It denotes how well a document meets the information needs of a particular user or use case. This means it is highly context dependent and will likely profit from our location-based approach. The measure *credibility* has two main characteristics, each with two aspects: The trustworthiness of the source, and expertise on the subject matter of the information. For example, my best friend is completely trustworthy, but he has no expertise on molecular biology, and any information coming from him on this topic is probably relayed, possibly distorted in the process, thus leading to a low accuracy. The aspects are a subjective component on the part of the information recipient, and an objective component of the information quality itself. Flanagin and Metzger (2008) term these credibility-as-perception and credibility-as-accuracy. Our approach of geographically contextualising unknown information clearly addresses the second, objective, component. Any assessment of credibility relies on the first and second of the heuristics we described: What do others say, and what do we already know about the source and the situation it addresses?

To operationalise relevance and credibility, we built on the work by De Sabbata and Reichenbacher (2012) and Friberg *et al.* (2011). Table 4 summarises, the criteria developed on relevance or credibility, and possible ways of measuring them.

3.3. System workflow and implementation

The workflow we developed to process VGI in this project is called CONAVI (CONtextual Analysis of Volunteered Information) and includes four main tasks that need to be carried out semi-automatically or automatically:

Table 4. Criteria to measure relevance and credibility.

Criterion	Description	Credibility (C) Relevance (R)	Measure
Topicality or specificity	Content topic	R	(Co-)occurrences of use case specific keywords (e.g. fire AND forest).
Coverage or completeness	Information content	R	(Co-)occurrences of use case specific keywords (e.g. fire AND forest AND xy hectares).
Novelty or timeliness	Information publishing date	R, C	n/a for (near-) real time use.
Spatial proximity	Geographic location relative to other information, or known event, or own location	R, C	Distance to own location, to high risk areas or known events.
Temporal proximity	Temporal location relative to other information, or known event, or own location	R, C	Distance to own location, or known events.
Clusters or Co-location or redundancy	Other VGI in spatio-temporal proximity	C	Probability of belonging to a cluster.
Accuracy or granularity or clarity	Geographic precision of geo-location	C, R	Level of detail (Country, Region, Town), vagueness of boundaries (administrative region vs. vernacular or landscapes), conflicting or confirming location from source profile, message origin, message content.
Source type	Public authority or private user; identifiable or anonymous	C	Authoritative, certified, uncertified; Level of authority with respect to use case; strength of certification mechanism.
Source behavior	Characteristics of generated content, reception in the community	C	Posting history (frequency), ratings, followers or friends.
Geographic context	Characteristics of the location near or around the VGI origin or content	R, C	Land cover, population density, weather, soil type, points of interest, infrastructure.

Source: Adapted from De Sabbata and Reichenbacher (2012), Friberg *et al.* (2011), and own contributions.

(1) Retrieving and storing VGI from various sources.
(2) Enriching the data retrieved with information about source, content, location and geographic context turning it into explicit information.
(3) Clustering of the VGI in space and time.
(4) Disseminating the results.

These phases are shown in Figure 2, and will be explained in more detail in the corresponding sections below.

3.3.1. Retrieval and storage

In the era of Web 2.0, the various geo-referenced media are mostly socially generated, collaboratively authored and community-contributed. The time- and geo-references, together with text metadata, reflect where and when the media was collected or authored, or the locations and times described by the media content. The enriched online multimedia resources open up a new world of opportunities to discover geographic related knowledge and information of our human society. In general, there are several types of media with time- and geo-references on the Internet: (1) geo-tagged photos on photo-sharing websites like Flickr, (2) geo-referenced videos on websites like YouTube, (3) geo-referenced web documents, like articles in Wikipedia and blogs, (4) geo-referenced microblogging websites like Twitter and (5) 'check-in' services (users can post their location at a venue and connect with friends) such as Foursquare. We did not consider any sites that have a national focus, since our regional focus is on South-West Europe (i.e. France, Italy, Spain, Portugal).

For the prototype, we decided on Twitter and Flickr, because they have well-documented application programming interfaces (API) that allow detailed queries without any rate-limits, they have a large potential content base on forest fires requiring automatic processing, and represent textual and visual VGI. The other sites did not have a sufficiently developed or maintained API, would provide a redundant type of VGI (images), have a comparatively small content base, or have more privacy restriction on shared content. We implemented the retrieval using Java scripts scheduled to run at regular intervals and query the Search API of Flickr (http://www.flickr.com/services/api/) and the Streaming API of Twitter (https://dev.twitter.com/docs/streaming-api). Concerning the storage of the retrieved VGI, testing

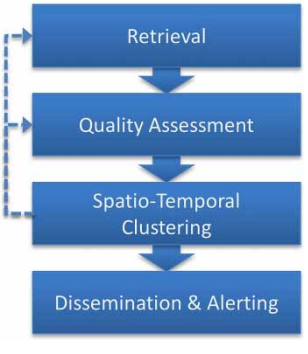

Figure 2. Main workflow phases of CONAVI system.

showed that the large volume to be expected (5 GB/day) excluded the use of local database management system (DBMS), while the lack of publicly accessible spatial analysis capabilities excluded the use of cloud storage services. At the end, we decided on a relational DBMS solution implemented with Oracle.

Regarding the search parameters, we refrained from using geographic parameters, since a quick investigation revealed that only about 1% of the Tweets and 20% of the Flickr images were already geo-coded. Instead, we relied on an extensive set of keywords so that we would not miss any relevant VGI. We chose the keywords with input from domain experts at the EFFIS, deriving them from key concepts pertaining to forest fires, that is fires, area, vegetation, actors and action. The set was composed by 77 words in 7 languages and covers the following concepts: fire (wildfire, Feuerwehrmann, foc, fogo, fuego, fuoco, incendi, incendie, incendio, . . .), area (hectares, ettari, . . .), vegetation (forest, brushwood, forêt, florestais, forestale, forêt, Gestrüpp, shrub, sterpaglie, Unterholz, vegetação, . . .), actors (firefighter, bombeiro, Canadair, helicopter, pompier, . . .) and actions (evacuation, alarm, alert, . . .).

3.3.2. Classification, geo-coding, context enrichment and quality assessment

As argued in the previous section, the enrichment of the retrieved VGI with additional information serves to assess both its credibility and relevance.

First, it is necessary to establish how likely the VGI is actually about a forest fire, that is classifying the content and the topicality. Since the difficulties of existing document classification and natural language processing systems with short, unstructured text such as found in Tweets and Flickr images descriptions are well-known (Gelernter and Mushegian 2011), the well-defined application case (forest fires) led us to another approach based on keyword occurrence: A search using a regular expression detects all specific fire-related keywords in a piece of VGI. The findings are compared to the occurrence of keywords found in a manually annotated ground truth sample (i.e. a sample of Tweets referring to known forest fires). An advantage of this approach is its simplicity, speed and reliability: Any string matching to find keywords is fast because of the short message length and the limited set of keywords to look for. The search for keywords and subsequent assignment of a topicality score is accomplished from within Oracle by a scheduled job to take advantage of speed.

For determining any geographic context, finding the location is a prerequisite. Although it may seem paradoxical, the number of fully geo-referenced VGI is still low. Even with geo-referenced VGI, the coordinates can be unreliable, since the geo-referencing depends on the hardware specifications of the device, on the software used to report the VGI, on the option settings of the user, and on any geo-coding done by the social media platform. Furthermore, they do not often represent the location that the VGI is about, but the location of the device or of the residence of the user. For these reasons, extracting place names in the content and geo-coding is indispensable.

However, the detection and disambiguation of place names in short, unstructured text like free-form image metadata or microblogs faces particular challenges. We tested several freely available online services, but most services need as input well-formatted place names, cannot return more than one result and do not support

multiple languages. One application that can deal with several languages and that accepts unstructured text is Yahoo!Placemaker. A major drawback besides the limited number of queries per hour is that this application requires the language identification of the input text (experiments using the wrong language led to wrong results). This could be solved in two different ways, both unsatisfactory from our point of view: The first approach is to use the Tweets metadata about the language, but our manual annotation of data found that many users living in not English-speaking countries have never changed the default setting that is English. The second approach is to add a computational layer to detect automatically the language (a possible parser was Google Translate application) but due to the unstructured nature of the Tweets (grammar is often ignored by the user due to the limited number of possible words) this approach produced unacceptable results and was discarded.

For these reasons, we implemented a simple search for place names based on string matching the words of VGI content with the Geographical Information System at the European Commission (GISCO) database of place names for our area of interest (Spain, Portugal, France, Italy) at the most detailed resolution, which is the commune level, again implemented as a scheduled job in Oracle.

Having geo-referenced the VGI, the central key step is to provide a geographic context to the VGI by looking up characteristics of the location identified, and use this new knowledge to increase or decrease the VGI's credibility and relevance. In principle, these could be any characteristics found in relevant spatial databases. In the case of forest fires, of primary interest are distances to known hot spots or forest fires, the population density and predominant vegetation type. Since our geo-coding is based on the GISCO data-set, we used the latter as base to aggregate raster data-sets on population density and land cover from 2006 through zonal spatial analysis. While this aggregation could pose some problems, at the level of municipality hierarchy used, we noticed few inaccuracies. The distance to hot spots was implemented using the latest data from the EFFIS, downloaded at regular intervals, uploaded to Oracle and used in a spatial query.

We defined an integrated quality score (IQS). The aim of assigning a score to each VGI has to deal with several facets, each of which is contributing to the final value. These values are combined in a weighted sum:

$$\mathrm{IQS}(\mathrm{VGI_j}) = \sum_{\mathrm{i=1}}^{\mathrm{N}} \mathrm{w}_i \mathrm{v}_i(\mathrm{s}_{ji}) \tag{1}$$

where w being weight for criterion *i*, v being the value function or rule for criterion *i* and s being the score for the VGI item j. In our specific case

$$\mathrm{IQS} = (\text{topicality} \times \text{weight } 1) + (\text{geo-coding} \times \text{weight } 2) + (\text{context} \times \text{weight } 3).$$

3.3.3. Spatio-temporal clustering

The next phase of clustering the VGI emulates the search for social confirmation (or rebuttal), and is crucial for detecting events. We want to aggregate VGI close in space and time. Each cluster then represent a potential forest fire event on the Social side. As for physical sensors, the cluster can then be scored using a combined measure that

gives the 'temperature' of the clusters. If the quality assessed measure is over a certain threshold the cluster is likely to report about a forest fire.

We had to rely on an external software since Oracle does not provide sufficient support for spatio-temporal clustering. After testing various software (CrimeStatIII [http://www.icpsr.umich.edu/CrimeStat/], packages of R [http://www.r-project.org/], ArcGIS/ArcInfo [www.esri.com] QGIS [http://www.qgis.org/]), we settled on using SatScan [http://www.satscan.org/] since it has been well-published, is in use, and offers the widest variety of possible scan methods, including Space-Time Scan Statistics, Bernoulli and Discrete Poisson Models. In regular intervals, data are exported from Oracle and fed into SatScan using a command line call with various parameters. The outputs are parsed and uploaded into Oracle via a Python script. As a final step a script computes the IQS score for each cluster.

3.3.4. Dissemination and alerting

The detection of events in a near-to-real time stream of information is a challenging task that needs further investigation. For the moment, we consider the detection of a cluster a likely event which can be investigated further by a human domain expert.

For the dissemination of results, various avenues are open, including broadcasting via social media, SMS and web maps. For the latter, we submit the highly likely candidates and clusters to a web application developed by the EFFIS, which is currently only available inside the intranet of the JRC as the public release would need to be discussed with the competent authorities in the European Union member states.

3.4. Project results – case study: France 2011

In this section, we report the outcome of the CONAVI workflow in a case study of forest fires in France in the summer of 2011 (July to September). During that period there were eight major forest fires reported by EFFIS. Using the steps previously described in Table 5 represents the data narrowing selection.

As indicated in Table 5, we start with almost 22 million pieces of information and end up with less than 500 in 11 clusters. All of these are Tweets. Analysing the images in Flickr containing we found that those in clusters (which were discarded subsequently) were about non-forest fires. From a human driven search we found some images related to 2 out of the 8 EFFIS fires considered. A deeper analysis of

Table 5. Data volume at various processing steps.

Step 0	Initial data-set	21.8 million Tweets + 57 thousands Flicks
Step 1	French keywords	802,260 Tweets + 49,067 Flicks
Step 2	Topicality score single VGI	35,420 VGI
Step 3	Toponyms	9051 VGI
Step 4	S/T cluster (with pop density)	129 clusters – 2682 VGI
Step 5	Excluding small cluster (≤ 5)	75 clusters – 2565 VGI
Step 6	Topicality of each cluster (fire-related keyword + 'hectare')	11 clusters – 469 VGI

Flickr use of textual metadata should be performed as these results are not very promising. The distribution of the fires detected by EFFIS and CONAVI is shown in Figures 3 and 4.

The map shown in Figure 3 illustrates (1) MODIS hotspots, (2) geo-coded VGI and (3) forest cover. Concerning MODIS hotspots, it shows only the MODIS hotspots that were registered from July to September, and which according to EFFIS estimates, were likely to be caused by burning vegetation. Concerning geo-coded VGI, the VGI is already pre-filtered according to keyword occurrences, and shows only those likely to be about forest fires, with the dots sized being proportional to the number of VGI from that commune. Finally, forest cover shows the percentage land cover classified as forest under the GlobeCover 2009 classification scheme (classes 20–120), on a commune basis.

The map shown in Figure 4 illustrates (1) forest fires, (2) processed and filtered VGI clusters and (3) forest cover. The forest fires are divided into two classes: Those reported by EFFIS, and those not reported, but detected through clusters of VGI. The spatio-temporal clustering results were filtered for keyword occurrence, with the remaining 11 clusters all being about forest fires. The dots are sized proportional to the number of VGI in that cluster originating from that commune, with locations belonging to the same cluster sharing the same colour.

An overall assessment of the case study shows that the system is able to retrieve mostly of the fire registered by EFFIS (6 or 7 out of 8 depending on the parameters). Moreover it points out 4 other relevant event (bigger that 40 ha) that could have been missed.

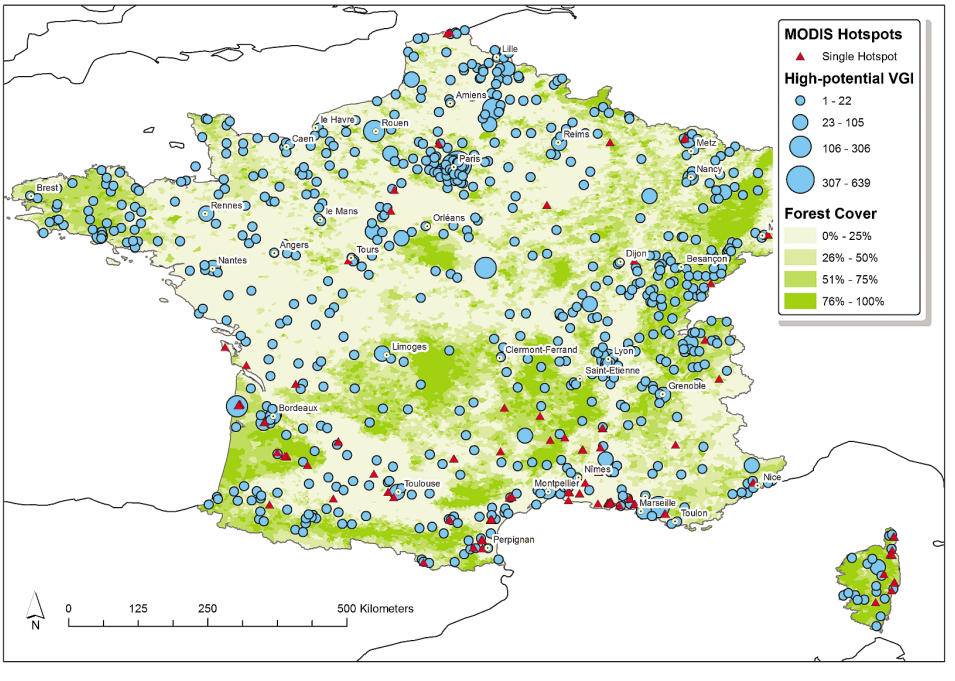

Figure 3. Map showing raw data for France.

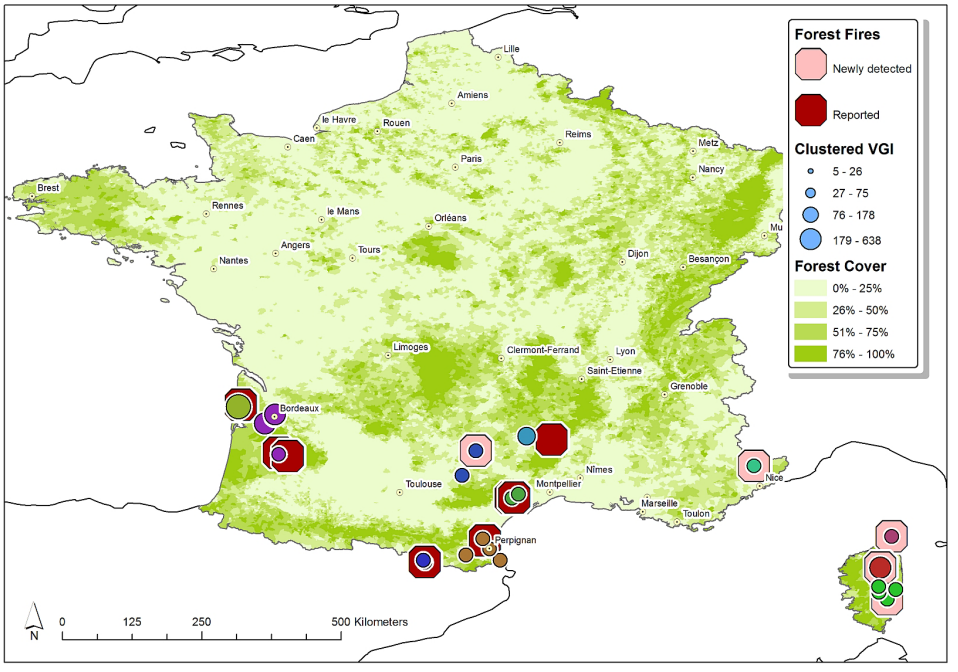

Figure 4. Map showing processed data for France.

Summing up the results, for the moment people do not use extensively the automatic geographic references for their information, but this could be encouraged especially in an emergency situation. When reporting an event such as a forest fire toponyms are mostly used in the text. The most common behaviour on Twitter is to spread around information as it arrives without any further editing, hash-tags are used but they are not yet a good practice. Finally, different social media have different rules to retrieve, extract and assess information. A fine-tuning of the system has to consider these differences.

We conclude that the echo of forest fire events in the social web has enough volume to be listened and the system can be tuned to receive only the signals we are interested in.

4. Conclusions

The vision of DE articulated by Craglia *et al.* (2012) extends current international efforts on SDI by adding a dynamic dimension based on real-time or near-to-real time information coming from sensors and citizens. This paper focuses on the latter and in particular on VGI, which has gained over the last few years a great deal of attention with regular tracks at international conferences, and specialised workshops. Whilst many discuss the value of VGI, few have, yet, done large-scale experiments as those described in this paper, particularly in relation to what we have classified as information, which is implicitly volunteered and implicitly geographic in nature.

The proof-of-concept prototype described in the previous section has been operational for the second half of 2011 and collected a large volume of VGI.

This volume, which is increasing at an exceedingly rapid pace, suggested an approach, which focused on discarding as much VGI as possible with a low probability of being relevant and/or credible, and then enriching/clustering the remainder to extract the relevant information. This worked for the specific case study and the geographic and temporal window analysed but clearly any operational implementation would need a more powerful and scalable infrastructure capable of coping with hundreds of thousands or millions of VGI per hour, and provide sufficient fail safe measures.

The system described in this paper has shown that applying a sequence of relatively straightforward methods can lead to results that extract useful and timely information from the large volumes of VGI being exchanged daily. As more and more of this content will be geo-coded at source, the proposed system could be simplified. On the other hand, further improvements include using a machine-learning approach to classify the VGI, and a finer grained geographic context analysis. Finally, the approach presented here can be easily adapted to a wide variety of applications. For this reason, we believe that lessons learned are a useful addition towards the implementation of the vision for DE.

References

Annoni, A., *et al*., 2011. A European perspective on Digital Earth. *International Journal of Digital Earth*, 4 (4), 271–284.

Anthony, D., Smith, S., and Williamson, T., 2005. *Explaining quality in Internet collective goods: Zealots and good samaritans in the case of Wikipedia* [online]. Available from: http://web.mit.edu/iandeseminar/Papers/Fall2005/anthony.pdf [Accessed 30 January 2012].

Coleman, D., Georgiadou, Y., and Labonte, J., 2009. Volunteered Geographic Information: the nature and motivation of produsers. *International Journal of Spatial Data Infrastructures Research*, 3, 332–358.

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

Craglia, M., *et al*., 2012. Digital Earth 2020: towards the vision for the next decade. *International Journal of Digital Earth*, 5 (1), 4–21.

De Sabbata, S. and Reichenbacher, T., 2012. Criteria of geographic relevance: an experimental study. *International Journal of Geographical Information Science*, 26 (8), 1495–1520.

Elwood, S., Goodchild, M.F., and Sui, D.Z., 2011. Researching volunteered geographic information: spatial data, geographic research, and new social practice. *Annals of the Association of American Geographers*, 102 (3), 571–590.

European Commission, 2007. Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE), Luxembourg: Publications Office [online]. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32007L0002:EN:NOT [accessed 12 August 2011]

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

Friberg, T., Prödel, S., and Koch, R., 2011. Information quality criteria and their importance for experts in crisis situations. *In*: M.A. Santos, L. Sousa, and E. Portela, eds. *Proceedings of the 8th international ISCRAM conference*. Presented at the 8th international conference on information systems for crisis response and management, 8–11 May 2011, Lisbon, Portugal. Brussels: ISCRAM Association, 1–5.

Gelernter, J., and Mushegian, N., 2011. Geo-parsing messages from Microtext. *Transactions in GIS*, 15 (6), 753–773. doi:10.1111/j.1467-9671.2011.01294.x

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, 24–32.

Goodchild, M.F., 2009. NeoGeography and the nature of geographic expertise. *Journal of Location Based Services*, 3 (2), 82–96.

Gore, A., 1999. The Digital Earth: understanding our planet in the 21st century. *Photogrammetric Engineering and Remote Sensing*, 65 (5), 528.

International Council for Science (ICSU)., 2010. Grand challenges in global sustainability research: a systems approach to research priorities for the decade [online]. Available at: http://www.icsu-visioning.org/wp-content/uploads/Grand_Challenges_Nov2010.pdf [accessed 9 May 2012].

Jiang, M., and McGill, W.L., 2010. Human-centered sensing for crisis response and management analysis campaigns. *In*: S. French, B. Tomaszewski, and C. Zobel, eds. *Proceedings of the 7th international ISCRAM conference*. Presented at the 7th international conference on information systems for crisis response and management, 2–5 May 2010, Seattle. Brussels: ISCRAM Association, 1–11.

Metzger, M.J., Flanagin, A.J., and Medders, R.B., 2010. Social and heuristic approaches to credibility evaluation online. *Journal of Communication*, 60 (3), 413–439.

Ostermann, F.O., and Spinsanti, L., 2011. A conceptual workflow for automatically assessing the quality of volunteered geographic information for crisis management. *In*: S. Geertman, W. Reinhardt, and F. Toppen, eds. *Proceedings of 14th AGILE international conference on geographic information science 2011*. Presented at the 14th AGILE international conference on geographic information science, 18–21 April 2011, Utrecht, Netherlands. Association of Geographic Information Laboratories for Europe, 1–6.

Schade, S., *et al*., in press. Citizen-based sensing of crisis events: sensor web enablement for volunteered geographic information. *Applied Geomatics*.

Spinsanti, L., and Ostermann, F.O., 2010. Validation and relevance assessment of volunteered geographic information in the case of forest fires. *In*: C. Corbane, D. Carrion, M. Broglia, and M. Pesaresi, eds. *Proceedings of the 2nd international workshop on validation of geo-information products for crisis management*. Presented at the 2nd International Workshop On Validation Of Geo-Information Products For Crisis Management, 12–13 October 2010, Ispra, Italy. Luxembourg: Publications Office of the European Union, 101–108.

Turner, A., 2006. *Introduction to neogeography*. Sebastopol: O'Reilly.

Weiner, D. and Harris, T.M., 2008. Participatory geographic information systems. *In*: J.P. Wilson and A.S. Fotheringham, eds. *The handbook of geographic information science*, Chapter 26. Malden: Blackwell, 466–480.