Integration of Volunteered Geographic Information into Spatial Data Infrastructures: a case study

Gianluca Luraschi

(European Commission – Joint Research Centre Institute for Environment & Sustainability Ispra, Italy email: gianluca.luraschi@jrc.ec.europa.eu), Bertrand De Longueville (European Commission – Joint Research Centre Institute for Environment & Sustainability Ispra, Italy email: bertrand.de-longueville@jrc.ec.europa.eu)

Riassunto
La recente evoluzione di Internet, nota con il termine Web 2.0, ha portato ad un aumento senza precedenti di contenuti creati da utenti non specializzati (Rinner et al., 2008). Quando tale informazione contiene una dimensione geografica si definisce come Volunteered Geographic Information (VGI), ed è potenzialmente una fonte efficace per conoscere l'ambiente che ci circonda (Goodchild, 2007) .
Tuttavia, l'integrazione VGI in Spatial Data Infrastructures (SDI) non è una sfida facile, dato che spesso l'informazione su base volontaria e` considerata come non adeguatamente strutturata, poco documentata, e non validata secondo gli standard scientifici (Flanagin & Metzger, 2008).
Con questo lavoro proponiamo una soluzione concreta a questa sfida. Senza essere esperti di incendi forestali, e senza conoscere la traduzione in tutte le lingue in cui potrebbero essere stati descritti degli incendi, siamo riusciti a costruire dataset validi, cioe` pertinenti rispetto all'argomento trattato, utilizzando come fonte d'informazione il noto photo sharing portal Flickr. Per questo scopo abbiamo integrato un thesaurus ambientale, GEMET, nell'approccio basato su workflow definito nei nostri lavori precedenti (De Longueville *et al*., 2009). Per validare i dataset generati, abbiamo fatto un confronto con dati provenienti da fonti giornalistiche certificate.

1. Introduction

Fire is typified by a situation of crisis, where responsible authorities need up-to-date situational awareness in order to effectively coordinate response. Much of the information traditionally gathered for such situations comes from official, trusted sources (e.g. emergency services, local authorities, mapping agencies). However, in most cases, citizens are present at the site of disaster and are providing VGI (Hughes & Palen, 2009). While this VGI can be timely, its value to the situational awareness is unproven. VGI can contain false information, interpretation, rumors and, in general, the information accuracy is unknown. Therefore, the crisis management community is wary of using it for decision making. We aim at rising awareness about the strengths and weaknesses of VGI as a novel information source. In this purpose, we developed a process to retrieve and validate dataset (De Longueville *et al*., 2009); in this work we have applied it in a c scenario involving multi-linguality and semantically complex topic. The information resulting from our VGI integration process is validated using journalistic sources.
The remainder of paper is structured as follows: In section 2 we describe the concepts and technologies that are the basis of our research. We then describe the use case on which we applied our works (section 3). In section 4, a description of the VGI to SDI data integration workflow is provided, taking in account the multi-lingual aspects of our case study. In section 5 we analyze the obtained results in the light of information obtained through independent media sources. This is followed by conclusion and future work items in section 6.

2. Previous works

In this section we describe previous works in the field of VGI initiatives, and its relation to Spatial Data Infrastructures. We also describe the Flickr platform that is used as source of VGI in the use case and the GEMET thesaurus.

Volunteered Geographic Information as a data source for SDIs

According to Goodchild (2007), the term Volunteered Geographic Information (VGI) is used to designate any user-generated content that has a relation to the surface of the earth.

Popular examples are GPS tracks of cars and points of interest such as look-outs, restaurants, coffee bars, etc. There are various VGI applications that allow users to upload and browse information in various media (text, pictures, videos, documents, etc.). The information is linked through a spatial reference to a location on a map.

In the other hand, a *Spatial Data Infrastructure* (SDI) is ”a distributed system that allows acquiring, process, distributing, using, maintaining, and preserving spatial data” (Maguire and Longley; 2005), where the core components of an SDI are people (including partnerships), data, policy network architecture and technical standards (Rajabifard *et al.*, 2002).

SDIs are usually providing an ‘expert’ perspective on Geographic Information (GI), while VGI concept refers to a more ‘casual’ approach (Buttler, 2006). There is nowadays a wide consensus about the fact that VGI can benefit Spatial Data Infrastructures (SDI) driven by specialists in many fields (Craglia, 2008). Natural hazards is one of these very relevant fields, as citizens that are directly affected can often provide valuable geospatial information for risk and impact assessment

Flickr

Flickr is an online application that allows uploading, store and organizing digital photos.

Since its creation in 2004, Flickr has been recognized as one of the most innovative user generated contents sharing platform (Terdiman, 2004) and as a reference implementation of the Web 2.0. principle (O'Reilly, 2005). Flickr offers numerous features that make it an interesting VGI platform. The first of them is the multiplicity of uploading options, which includes direct upload from camera-enabled mobile phones. Such devices are becoming always more accessible to the mass market and many of them also include built-in GPS sensor, so it is expected that Flickr will contain in the future a growing number of geo-referenced contents that will be available within seconds after a photo has been taken. The possibility for users to assign a location to pictures – ‘to geotag’ - is another important feature. Indeed, the wide majority of cameras do not include a GPS device that automatically inserts picture location in the image file metadata. Flickr users can thus manually add this information using an online map interface. Flickr allows users to associate keywords – called ‘tags’ - to their pictures. This feature, which is supported by many Web 2.0 portals, is known as ‘folksonomy’ because the indexing of contents is the consequence of end-users actions instead of a top-down classification process (Voss; 2007). Folksonomy can be criticized for a potential lack of coherence that it can generate, however tags are valuable information to perform queries in the vast amount of images that can be found on Flickr. Thanks to its Application Programming Interface (API), Flickr can be accessed, viewed, updated, retrieved and analyzed in many ways and for many purposes.

GEMET, a multi-lingual environmental thesaurus

GEMET, the GEneral Multilingual Environmental Thesaurus[1], has been developed as an indexing, retrieval and control tool for the European Topic Centre on Catalogue of Data Sources (ETC/CDS) and the European Environment Agency (EEA), Copenhagen. GEMET was conceived as a thematic, multi-purpose thesaurus, aimed to define a common general language, a core of general terminology for the environment.

GEMET is structured as follows: each descriptor is arranged in a hierarchical structure headed by a Top Term. As a complement to the hierarchical “vertical” relations, an exhaustive series of strong “horizontal” relations between terms (RT, Related Terms) have been introduced. The current

[1] http://www.eionet.europa.eu/gemet/about

version contains 5.298 descriptors, including 109 Top Terms, and 1.264 synonyms in English. It provides a complete numerical equivalence (all the descriptors have an equivalent) with the following languages: Basque, Bulgarian, Dutch, Finnish, French, German, Hungarian, Italian, Norwegian, Portuguese, Russian, Slovenian and Spanish. For Danish, Slovak, Swedish and Greek some few descriptors are still missing - this issue is presently under work.
We used GEMET as a multilingual reference for the VGI processing tasks involving natural languages analysis (e.g. retrieval of relevant pictures in Flickr).

3. Use Case: Fires in Mediterranean Area
The use case presented in this paper relates to the use of VGI from photo sharing portal Flickr to map fire events which took place in Mediterranean region between the 1st of January 2009 and the 31st of August 2009. There were several reasons for this choice. Firstly, fires are events that have a precise location (compared, for example, to a seismic event). Secondly, the choice of Mediterranean region allowed testing multiple-language in queries to Flick. The time period has been chosen to cover a recent period that was expected to include several fire events of various extents. And finally "fire" is a term that could have several different meaning, and it is difficult to distinguish them, for instance: forest fire, fireworks, quick fire, campfire, fire away, … This was thus providing an interesting use case to test how the thesaurus approach can help to deal with semantically complex scenario.

4. Description of the workflow
Principle and overview
The proposed data integration workflow aims at integrating VGI into SDI through several configurable steps. Its overall principle is based on the retrieval and processing of a subset of information contained in a wide scaled user-generated data repository. In other words, queries are sent to a web-based repository in order to extract relevant information, and then the result goes through several processing tasks that are designed to fit a particular need.
The workflow includes 6 steps: retrieval, formatting, validation, clustering, ranking and conversion. Each step is described in detail in previous works from the authors (De Longueville *et al*., 2009).
Step 1: retrieval
During the retrieval phase, queries are submitted through the Flickr API, and their results are saved locally for further processing. The Flickr API offers numerous options to submit queries using the *flickr.photos.search* method. The query parameters: the date picture has been taken (01/01/2009 to 31/08/2009); tags including "*fire"*, geographic bounding box of the Mediterranean area (-*7,33,26,46).*
We have performed the query in each of the languages managed by GEMET translating the word "*fire*". We included even languages that are not officially spoken in the Mediterranean area, as citizens that are speaking non-local languages (e.g., tourists) are assumed to be able to produce VGI as well.
By using a geographic bounding box, we excluded *de facto* all the pictures that have not been properly geo-tagged, even if they include geographic information in their metadata (e.g. a town name in the title or tags). This can look restrictive as only 3 to 4% of Flickr pictures are geotagged. However, we decided to focus on information provided on a voluntary basis, instead of generating it with more sophisticated procedures (e.g. by geocoding place names that appear in titles and tags).
Step 2: formatting
The number of pictures retrieved by language are Arabic = 0, Basque = 0, Czech = 0, Danish =15, German = 15, Dutch = 15, English = 741, Spanish = 3, Finnish = 0, Norwegian = 0; Swedish = 0; French = 73, Greek = 0, Hungarian = 0, Irish = 0, Italian = 60, Russian = 0, Portuguese = 0, Slovak = 0, Slovenian = 0.
Different languages uses words that syntactically are the same, i.e. in Danish , Dutch and German use "brand"; or a word in a language could be included in another language i.e.: "incendio" in

Italian, and "incendios" in Spanish. As a consequence, one single picture could have been retrieved several times. Using their unique identifier, it has been possible to take into account only one occurrence for each photo. In total, these queries returned a set of 838 unique pictures.

Step 3: Validation

The validation is a formal step to control if the minimal information required to process the data is available in the proper format. Validation tests highlighted that 34 of the 838 pictures had both latitude and longitude equal to 0. This was surprising, as all the result returned were expected to fit with a bounding box including excluding the location (0,0). No clear reason was found to explain why pictures that were supposedly correctly geo-tagged (as they were retrieved in response to a query that includes geographic criteria) had no latitude and longitude information. Possible reason is that a privacy setting prevented to retrieve the actual location even if the Flickr querying system was able to access the information. As a consequence, those 34 images have been removed from the VGI dataset, which will finally contain 804 records.

Step 4: Clustering

Clustering data consists in generating sets of implicit classes that describe the data (Jain *et al.*, 1999). In particular, a combination of simple spatial, temporal and relevance constraints have been evaluated.

The method applied to clustering the records retrieved is based on some hypothesis.

- H1 – Spatial Cluster: A fire is an event that occurs in a defined area.

H1.1: The pictures that recorded a single fire event should be georeferenced in a close area.

- H2 Temporal: A fire is an event that might be recorded in different close days.

H2.1 Fires that occurs in the same spatial cluster but in different period are referring to different events.

- H3: A fire is an event that should be documented by significant images.

H3.1: the more there are people affected by the fire, the more pictures will be uploaded on Flickr.

On this basis, we formulated 3 criteria we used to build relevant clusters:

- Criteria 1: Using the 'Haversine' formula[2] a distance between the two latitude/longitude points with a different in decimal degree in Mediterranean area is around 20 km. We have applied this criteria for generating spatial clusters.
- Criteria 2: A gap longer then 15 days between to consecutive date describes different events.
- Criteria 3: each photo in Flickr contains the identifier of the photo owner; the cluster has to contain at least 3 different photo owners.

Appling these Hypotheses sequentially at the 804 records, 2 clusters have been retained.

Cluster1: 5 different owners have documented 68 photos in a area around latitude:43.24 and longitude:5,44 from 18 to 22 on July 2009.

Cluster2: 6 different owners have documented 77 photos in a area around latitude: 38 and longitude:23,8 from 24 to 26 on August 2009.

The remainder, 659 pictures, where distributed in small space-time clusters, with only 1 or 2 different contributors.

Step 5: Ranking

The Ranking step aims to quantify to relevance of each cluster, using automatic means. In other words, the ranking score reflects the likeliness the cluster represents a *forest fire* (and not any other type of fire). The ranking value can be used to reduce noise (i.e. by eliminating clusters that are most likely not corresponding to fire event) or evaluate the probability that a cluster can be confused with another type of fire.

The ranking score of each cluster has been calculated by retrieving all the tags associated to each picture, and by summing the number of occurrences of words included in a set of ranking words terms that comes from GEMET. This choice has been motivated by the fact that the use of several relevant keywords in the tags is a stronger signal than a single keyword. In GEMET the 'fire' is

[2] http://en.wikipedia.org/wiki/Haversine_formula

defined as "The state of combustion in which inflammable material burns, producing heat, flames and often smoke". In GEMET the broader terms of "fire" is "disaster" and the narrowed terms are: "forest fire" and "grass fire". All those terms have been included in the set of ranking words. Following the broader terms link "disaster " other narrows and broader terms have been included in the ranking set of word: "human-made disaster", "natural disaster", "accident", "disaster zone". These set of ranking words have been translated in all the GEMET languages.
In 68 photos for the first cluster we have found 340 tags and the ranking score is 83, in 77 photos for the second cluster we have found 246 tags, and the ranking score is 95. The proportion of tags matching to a GEMET term related to forest fire is thus of 1,22 (83/68) for cluster 1 and 1,23 (95/77). If we consider that every picture has at least 'fire' or its translation among its tags (as it was a retrieval parameter), this means that on average, less that 1 picture out of 4 has a second term that can be found in GEMET. The proportion of tags that match to terms from GEMET other than 'fire' or its translations in tags is rather low : 5,5% for cluster 1, and 10,6% for cluster 2.
Step 6: Conversion
The conversion step is a formal process with limited added-value. It consists in uploading the final dataset in our SDI database and to create corresponding metadata. The final dataset contains 145 point features classified in 2 ranked clusters.

5. Comparison of output with independent data sources and discussion

In this section, we compare the output of our workflow with information from an independent data source: the European Media Monitoring system (EMM). EMM is developed by the European Commission's Joint Research Centre; it harvests on a daily basis articles and news from thousands of online media sources (including news agencies, major national journals and television networks) and then associates automatically all those news items to persons, organizations, themes (e.g. floods, ecology, armed conflicts, immigration), and places. Using EMM, we found that the two clusters created are referring to the July fires close to Marseille (FR) and the end of August fires occurred close to the Athens (GR), which both attracted a lot of focus form the media because they affected densely populated areas.

6. Conclusions and future works

In this paper, we provided an example on how VGI can be turned in a valuable source of information for SDIs, in a particular context characterized by multilingual (Mediterranean area) and semantic complexity (the word '*fire'* not only being associated to natural hazards). By integrating multilingual thesaurus in the described method, we converted pictures uploaded by users on the photo-sharing website Flickr into a dataset describing fires that took place in Mediterranean region between the 1st of January 2009 and the 31st of August 2009.
Our most prominent result is the confirmation that VGI might be used to detect major fires affecting an important number of citizens. It is important to highlight that, among the thousands of forest fires that takes place every summer in the Mediterranean area, the two ones our method clearly detected were featured in the news as the most important in terms of persons concerned (both Athens and Marseille have more than 1.000.000 inhabitants). This confirms the assumption that the VGI production is more important where there is a conjunction of damage and affected citizens, and provides interesting research directions for future works, as, oppositely, finer means are required to collect few but highly relevant VGI produced in scarcely populated zones.
One other achievement of this research is the demonstration that VGI can be retrieved without specific expertise in the field of interest and without knowledge of all the languages spoken in the studied geographic area, thanks to the use of a multi-lingual thematic thesaurus. In particular we provided an example on how GEMET can be used to automatically retrieve and rank multi-lingual VGI. While the multi-lingual retrieval of information clearly benefitted from the use of GEMET, it is less clear for the ranking process. As a matter of fact, we found that excepted the generic term 'fire', non-experts users tend to not use terms referred in GEMET (e.g. damage, accident, natural

disaster). In future works, we plan to investigate the use of alternative multi-lingual thesauri that reflects better the natural language used by non-experts.
The study of Volunteered Geographic Information is a recent field of Geographic Information Science, and is widely influenced by rapid evolutions of the Information Technologies. A growing number of scientists is now considering such user-generated georeferenced contents as a genuine source of information that can enrich our knowledge of spatio-temporal phenomenon that take place on the Earth surface. In this context, the works presented in this paper can be used as a scientific basis supporting the development of the next generation of VGI-enabled Spatial Data Infrastructures that will rely on numerous VGI and traditional sources and will be applied to a wide variety of thematic fields.

References

Butler, B. (2006) "Virtual globes: The web-wide world," *Nature* 439: 776-778.
Craglia, M., Goodchild, M. F., Annoni, A., Camara, G., Gould, M., Kuhn, W., Mark, D., Masser, I., Maguire, D., Liang, S. & Parsons, E. (2008). Next-Generation Digital Earth. A position paper from the Vespucci Initiative for the Advancement of Geographic Information Science. *International Journal of Spatial Data Infrastructures Research*, 3, 146-167.
De Longueville, B., Luraschi, G., Smits, P., Peedell, S. & De Groeve, T. (2009) From Volunteered Geographic Information to Spatial Data Infrastructures: a data integration workflow based on a case study. *(Paper under review)*
De Rubeis, V., Sbarra, P. and Tosi, P. (2009) "Web based macroseismic survey: fast information exchange and elaboration of seismic intensity effects in Italy," in *Proceeding of the 6th International ISCRAM Conference*, ed. Jonas Landgren and Bartel Van de Walle (Gothenburg, Sweden), http://www.iscram.org/ISCRAM2009/papers/
Flanagin, A. J. & Metzger, M.J. (2008). The credibility of volunteered geographic information. *GeoJournal*, 72, 137-148
Goodchild, M. F. (2007). Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69/4, 211-221.
Hughes A.L. and Palen, L. (2009) "Twitter Adoption and Use in Mass Convergence and Emergency Events," in *Proceeding of the 6th International ISCRAM Conference*, ed. Jonas Landgren and Bartel Van de Walle (Gothenburg, Sweden), http://www.iscram.org/ISCRAM2009/papers/.
Jain, A.K., Murty, M.N. and Flynn, P.J. (1999) "Data clustering: A review," *ACM Computing Surveys* 31, no. 3 (1999): 316-323.
Maguire, D. J. & Longley, P.A. (2005). The emergence of geoportals and their role in spatial data infrastructures. Computers, Environment and Urban Systems, 29, 3-14.
O'Reilly, T. (2005). What Is Web 2.0 Design Patterns and Business Models for the Next Generation of Software. *http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html*,
Rajabifard, A., Feeney, M. F. & Williamson, I.P. (2002). Future directions for SDI development. International Journal of Applied Earth Observation and Geoinformation, 4, 11-22
Rinner, C., Kessler, C. & Andrulis, S. (2008). The use of Web 2.0 concepts to support deliberation in spatial decision-making. Computers, Environment and Urban Systems, 32, 386-395.
Terdiman, D. (2004). Photo Site a Hit With Bloggers. Wired, http://www.wired.com, retieved on 08/05/200.
Voss, J. (2007). Tagging, Folksonomy & Co – Renaissance of Manual Indexing?. Proceedings of the 10th International Symposium of Information Science (Cologne), , 234-25
Wald, D. J., Quitoriano, V., Dengler, L. A. & Dewey J. W. (1999). Utilization of the Internet for Rapid Community Intensity Maps . Seismological Research Letters, 70/6, 680-697.