IDES project: a tool for safety and security in the environment, based on GIS and semantic technologies

F. Gargiulo[1], G. Persechino[1], M. Lega[2] & A. Errico[1]
[1]CIRA - Italian Aerospace Research Centre, Italy
[2]University of Naples Parthenope, Italy

Abstract

In the Campania region in south-west Italy, there is growing evidence, including from a World Health Organization (WHO) study, that the accumulation of waste, illegal and legal, urban and industrial, has contaminated soil, water, and the air with a range of toxic pollutants including dioxins. An effective environmental monitoring system represents an important tool for the early detection of environmental violations. The IDES (Intelligent Data Extraction System) Project is a geo-environmental intelligence system (GIS), developed by the CIRA (Italian Aerospace Research Centre) with the contribution of universities and other government bodies, which aims to implement an advanced software and hardware platform for image, data and document analysis in order to support law enforcement investigations. The main IDES modules are: imagery analysis module to monitor land-use and anthropogenic changes; environmental GIS module to fuse geographical and administrative information; epidemiological domain module; and semantic search module to discover information in public sources such as blogs, social network, forums and newspapers. This paper focuses on the semantic search module and aims to provide the greatest support to those extracting possible environmental crimes, through the collection and analysis of documents from online public sources. People rarely denounce criminal activity to the authorities. On the other hand, every day sees many people exposing the status of land degradation through blogs, forums and social networks. In addition, given the public interest, journalists often document critical environmental issues. All this unstructured information is often lost, due to the difficulty in collecting and analysing it. The IDES semantic search module is an innovative solution for

WIT Transactions on Ecology and The Environment, Vol 191,
www.witpress.com, ISSN 1743-3541 (on-line)

doi:10.2495/SC141122

aggregating the common uneasiness and thoughts of the people; it is able to transform and objectify public opinion in “human sensors” for safety environmental monitoring. In this paper we introduce methods and technologies used in some case studies and, finally, we report some representative results, highlighting innovative aspects of this applied research.

Keywords: Illegal dumping, Landfill monitoring, interoperability, Text semantic search, Information retrieval, Geographical Information Systems.

1 Introduction

A recent census of Italian illegal dumping sites estimates the presence of 4866 incidences of illegal dumping. Only 21% of surveyed landfill sites have been reclaimed and more than 700 contain hazardous waste [1].

Citizens' exposure to toxic waste is a major public health problem; therefore, the responsible public authorities must act as quickly as possible to identify such environmental issues. Identifying the problem, as soon as it occurs, would reduce reclamation costs. Therefore, more frequent and more targeted monitoring of large areas is necessary.

On the other hand, public authorities do not have large budgets, so the solution must also be economically feasible. Many employers and institutions are involved in territory control, necessitating high costs for this activity. But control of the territory is a necessary action for the prevention of such unlawful activity. It is a difficult challenge to perform long-term environmental monitoring with today’s manned aircraft because of limitations on vehicles, cost, and missions.

Moreover, where there are small sources of pollution and contamination over a wide area, illegal dumping is very difficult to detect. The use of satellite monitoring exceeds the current limits of traditional in situ methods of detection. The satellite images are able to continuously monitor, in terms of space and time, a large part of the territory. There are several other remote sensing data products that can be used in the environmental monitoring, including fusion of optical data with synthetic aperture radar data to detect cattle ranching and use of thermal imagery to monitor landfills [2] and detect illegal dumping [3]. In addition, remote sensing data can be strategically combined with other data layers in geographic information systems to monitor the vulnerability of cultural sites [4] and anticipate environmental violations [5][6] Early intervention in waste accumulation is the only way to prevent its transformation into an illegal land fill site. Continuous satellite scanning allows the detection of anomalies connected to the dumping area, while employing fewer operators. In this way, the operators are only called out to aimed interventions, leaving more time for other important police activities. Another advantage of early warning in the contaminated area is more effective remediation so that the territory can soon be returned to its natural state. The treatment of small accumulations of waste requiring disposal and reclamation entails much lower costs than those for the remediation of large volumes of waste, and it is safer:

protecting the people living in the vicinity from exposure to the risk of contamination.

Early warning also prevents the contamination of wider areas caused by the dispersion phenomena mediated by the trophic chain and atmospheric agents. Government bodies must ensure citizens' health by reducing their exposure to contaminants.

2 IDES – Intelligent Data Extraction System

The IDES project [7] aims to implement a software platform for data analysis within the domain of environmental criticalities.

IDES provides an integrated repository of information extracted from heterogeneous, physically distributed, unstructured sources (satellite and airborne data, web pages, etc.) by means of a capture, extraction and analysis process.

Based on the integrated information stored in a geographic information system and elaborated with the aid of advanced data analysis techniques and tools, IDES will be able to extract hidden information, that is information not immediately identifiable through a mere reading or a deeper analysis, even if performed by a domain expert. To this end, IDES will be able to uncover patterns and multi-disciplinary correlations not known a priori and to extract relevant information useful for government bodies.

A large amount of data describes anthropic activities with a heavy impact on environmental health: the Campania region of southern Italy has recently experienced an increase in the number of deaths from cancer and other diseases, exceeding the national average, caused by pollution from illegal waste disposal, landfill sites and dumps. The project team's focus is on an innovative environmental monitoring system, which considers dumping and landfill data related to specific industrial installations, urban solid waste temporary deposits/facilities and statistical data about urban people.

All the above mentioned data, originating from different sources available in the project, differ regarding semantics (chemical, emissions, county people, installation address, etc.) and structure (tabular, vector, raster, structured, unstructured). IDES is an environmental geographical information system (GIS) with the strategic mission to be a centralized and unified informative system, containing all data coming from different sources (satellite, airborne, terrain data) and tools enriched with geographical information, enabling detailed spatial analysis tasks.

In order to elaborate different types of data and to enable user analysis, the IDES software platform is based on three components:

- Semantic search module: text analysis component with semantic analysis features; its objective is to analyze text documents arising from intranet and web sources in order to automatically extract entities and relationships among them, with which to build a conceptual map useful for primary illegal crime detection;

- Imagery analysis module: image analysis component for analyzing multi-spectral and SAR (Synthetic Aperture Radar) images through advanced algorithmic features to extract patterns among data; this represents a powerful key to identify and geo-locate sites with potential illegal activity;
- Environmental GIS module: Geographical information system and geostatistical analysis component for spatial data exploration and analysis, with the capability to create statistically valid prediction information from a limited number of data measurements.

The applied methodology integrates content from various sources of text and images, ensuring consistency and semantic coherence, and obtains structured information, on which to perform innovative techniques of analysis and correlation of data to generate new knowledge by using natural language processing technologies and the semantic analysis of a large amount of documents written in natural language. This added scientific value correlates the data on the spatial distribution of productive activities, hydro-geographical patterns and roads, using geostatistical Kriging method. Results are structured into a geo-system reference, easily accessible by the end-user, highlighting the relationships between the geographical and non-geographical entities for mapping risks of illegal spillages.

The illegal burning of waste is known to release toxic substances into the atmosphere. Even if such fires are easily hidden among legitimate incineration resulting from the more general waste disposal problem, a useful instrument of public complaint is the social network distribution of information, capable of highlighting the common uneasiness and thoughts of the people.

This paper aims to develop an innovative solution, implemented in a decision support system for the identification of circumstantial evidence of environmental crimes through a geo-environmental system, realizing the extraction of intelligent information from unstructured data. It has been carried out within the IDES (Intelligent Data Extraction System) project. The contribution of CIRA (Centro Italiano di Ricerca Aerospaziale) to IDES provides application tools and infrastructure, developed for the identification and classification of environmental problems in relation to the province of Caserta.

Within the Campania Regional Operational Program FERS 2007-2013, the IDES Project responds to specific objective 2.a - Enhancement System for Research and Innovation and Implementation of Technology in Production Systems and, in particular, to operational objective 2.1 "Build and strengthen in the field of industrial research and experimental development of science, technology leadership that may lead to the placement of substantial shares of the productive fabric, including through joint development in the form of advanced services in industrial research and experimental development”.

The research and the prototype developed in IDES will integrate the contents of the various information sources of both text and images (satellite, Aerial, multi spectral, SAR, etc.), ensuring consistency and semantic coherence, to obtain structured information, on which to perform innovative techniques of analysis and correlation of data useful for generating new knowledge.

The use of tools implemented in IDES opens a window onto the development of multidisciplinary scientific research, arousing interest in the ability of the project to develop a comprehensive geo-information product for the user.

2.1 Benefits and limitations of semantic search method with respect to remote sensing technology

In recent years, many innovative and technological projects have been developed in order to prove that remote sensing images have an important role to play in law enforcement and the prosecution of environmental crime, such as illegal waste management, illicit burning, abandonment, dumping or uncontrolled disposal of waste, etc. Also, improvements to satellite technology are expanding through the processing of low-resolution imagery (e.g. Moderate Resolution Imaging Spectroradiometer - MODIS), and high-resolution imagery (e.g. IKONOS, GeoEye), and the use of this resource as a tool for environmental compliance and enforcement will increase. Remote sensing, developed in the GIS domain (by Esri ArcGis software), is used to detect many characteristics of landfill sites. If these characteristics indicate that landfill sites have a negative impact on the environment, then these sites could be potentially illegal.

The remote sensing approach demands a large budget for the acquisition and processing of imaging. As described in the introduction, the IDES Project also merges remote sensing analysis and data acquisition for the preparation of sample datasets, collected at different heights (with the help of aerial work platforms, drones, etc.). In this study, IDES develops the semantic search module in order to highlight the advantages of this approach with respect to monitoring by satellite technology.

Remote sensing is not an alternative to ground surveys for monitoring contamination, but it may be used to identify vegetation stress, high soil temperature, etc. as a symptom of soil contamination. The processing is developed by means of a training site and requires supervised methods ~~(it is a technical term)~~ for the classification and analysis of data. Supervised methods are methods that attempt to discover the relationship between input attributes (sometimes called independent variables) and a target attribute (sometimes referred to as a dependent variable). The relationship discovered is represented in a structure referred to as a model. Usually models describe and explain phenomena, which are hidden in the dataset and can be used for predicting the value of the target attribute knowing the values of the input attributes. The supervised methods can be implemented in a variety of domains such as marketing, finance and manufacturing. From this, the creation of thematic maps of land use is very important for representing terrain expansion and the forms of degradation, such as derelict land, buildings, landfill sites and abusive quarries.

The first limitation is that no universal method has been developed: the vegetation stress may be an indicator of a landfill site, but there are other factors which may induce vegetation stress, for example trampling of vegetation, such as at building sites, deposit areas or sport fields. In other words, if changes can be detected, it cannot be determined whether they are caused by illegal dumping.

As a second limitation, the geographical area, in which this phenomenon develops, is often very wide; so the use of airborne or satellite platforms leads to too-high costs and time consumption.

Generally, the activities of IDES are based on applications involving analysis of the planning and monitoring of a phenomenon or event and estimation of damage as a consequence of an event. This general process concerns phenomena at a global and local scale, such that it introduces spectral and radiometric parameters, based on an analysis of elements that respond differently to different wavelengths of electromagnetic radiation (spectral signature). These methods should be applied throughout the whole cycle of the event, from pre-event conditions to post-event conditions. This leads to consideration of the time parameter, meaning the period of time that elapses between two successive shots of the same area. The practical use of this parameter depends on orbit or flight level altitude, as well as the characteristics of the mission. As the high-resolution weather sensors cannot be used for applications on a local scale, the panchromatic commercial sensors, although not having the same temporal resolution, are considered high temporal resolution systems, recording the same scene from 1.5 to four days.

Although the localization of the event is well determined thanks to the high level of accuracy of image resolution, the temporal parameter cannot be continuously explicated because the research applications could not sustain such a cost of nonstop image acquisition. In particular, the use of multispectral images can discriminate between the different spectral responses of objects, but it is necessary to integrate high spatial resolution and radiometric images with data geo-archives for the constructing of multi-temporal investigations. The Cogito software approach solves the problem of the geo-localization of data in a well-defined temporal space, proved by the human sensors (e.g. via social networks).

2.2 IDES – Semantic search module

The IDES semantic search module aims to demonstrate that the use of natural language processing technologies and the semantic analysis of the text allows you to extract structured or semi-structured information specific to the application domain of IDES from a large number of documents written in natural language. The approach integrates information obtained from the analysis of the text with the data made available by the institutions participating in the "table" of the project available in GIS. Then the use of statistical and geostatistical analysis may provide support to the process of characterization of possible environmental crimes, from the point of view of both scenario analysis and analysis of specific cases.

The IDES project stems from the fact that extracting knowledge from structured databases is now a well-established and profitable cognitive process, but often the majority of the available information from text files and images is unstructured and only human-readable.

Although 80% of the information for many organizations is unstructured [1], historically, the analysis has always been limited to the structured part of the available data. Losing such a high percentage of useful knowledge, especially in the areas of investigation, is currently a very high price to pay for the lack of efficient tools.

The main objective of the semantic search module is to provide support for the identification of potential environmental wrongdoing related to "illegal disposal of waste". Illegal waste disposal can take many forms, including: burning of toxic substances, illegal discharge in rivers, abandonment of waste, fly-tipping, etc.

The semantic search module relies on the expert system, Cogito SEE Suite .

The first step then is to identify the documentary sources in which to execute the semantic searches. For this purpose, in addition to the national and local press, some groups on Facebook, which were particularly active in reporting fires and the illegal discharge of wastes, were taken into account. Persons resident in the province of Caserta make daily postings in natural language on these groups; they report what they see and hear (fire, smoke, burning smell, etc.), the place where this happens (the road, the municipality, etc.) and often other information. The problem is that the descriptions of these events may be inaccurate or incomplete because they are written in natural language.

The semantic search module has been configured to extract from the posts of the Facebook group and from the national and local press the potential "illegal disposal of waste" events. The tool has been configured by expert linguists, who have analyzed the documentary sources and, in collaboration with the IDES project domain expert, have contributed to the definition of a taxonomy of specific domains and a list of events (named entities) to be extracted from documents.

The extensions of the semantic search tools focused mainly on the implementation of the taxonomy and the classification and extraction rules.

The table below shows an extract from the response list of defined events. There are five classes of events: waste disposal, burning, discharge of waste, waste traffic and bad smell and altogether 55 subclasses of events (named events).

For example, the class ‘burning’ contains the events: burning of special waste, burning of hazardous waste, burning of discarded tires, illegal disposal of asbestos waste, burning of waste electrical and electronic equipment, illegal disposal of hospital waste, etc.

An event is represented as in Fig. 1.

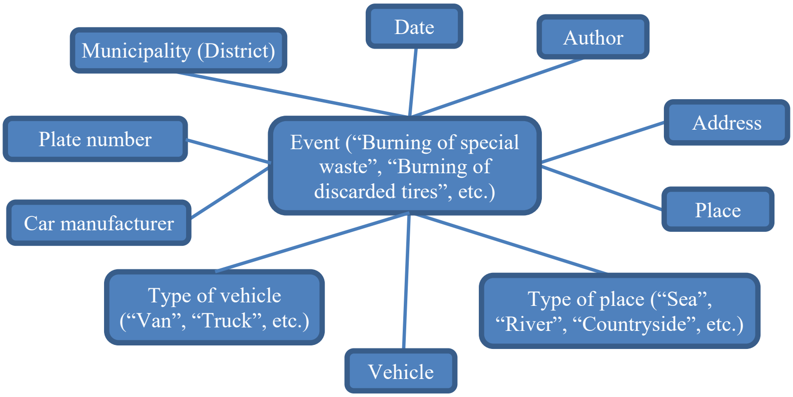

Date
Municipality (District)
Author
Plate number
Event ("Burning of special waste", "Burning of discarded tires", etc.)
Address
Car manufacturer
Place
Type of vehicle ("Van", "Truck", etc.)
Type of place ("Sea", "River", "Countryside", etc.)
Vehicle

Figure 1: Attributes of the events.

For each event retrieved, the tool tries to locate values for the attributes of the event:

- *Author*: name and surname, if specified in the document, the offender (people, but also company names, etc.).
- *Place*: proper name of the place where the event occurs.
- *Type of place*: the place where the offence occurs (e.g. marine waters, surface waters, soil, underground in the case of spillages, but also places like railway stations, airports, etc.).
- *Address*: address of the place or event venue type when specified.
- *Date*: the date of the event.
- *Type of vehicle*: vehicles, such as cars, motorcycles, trucks used for illegal trafficking, spills etc.
- *Vehicle*: the model of the vehicle (Grande Punto, Classe A, etc.), if specified.
- *Car manufacturer*: the manufacturer of the vehicle (Fiat, Alfa Romeo, etc.), if specified.
- *Plate number*: the number plate of the vehicle, if specified.

Of course, it is possible that not all the information appears in the text.

Information about this event is not organized into a rigid structure. For example, the address of the event is an attribute of the event and not a common attribute, as in a relational schema. This structure reflects the actual cases in which disclosures and complaints are incomplete.

It is preferred, in the first instance, to achieve this semi-structured organization and to perform, at a later stage, the appropriate processing events based on the analysis to be carried out. The analysis of the events involves both the quality and the quantity of extracted data.

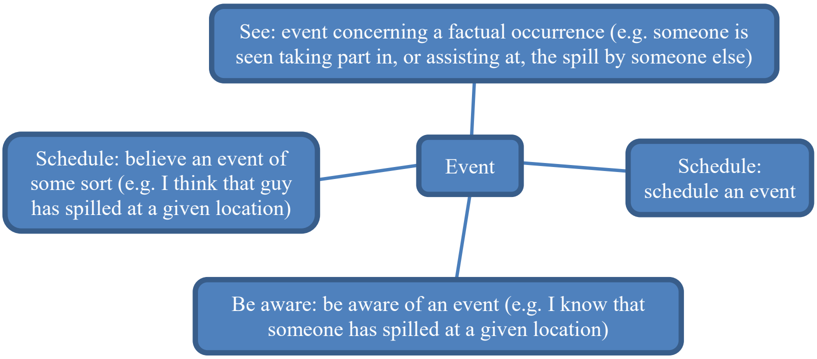

See: event concerning a factual occurrence (e.g. someone is seen taking part in, or assisting at, the spill by someone else)
Schedule: believe an event of some sort (e.g. I think that guy has spilled at a given location)
Event
Schedule: schedule an event
Be aware: be aware of an event (e.g. I know that someone has spilled at a given location)

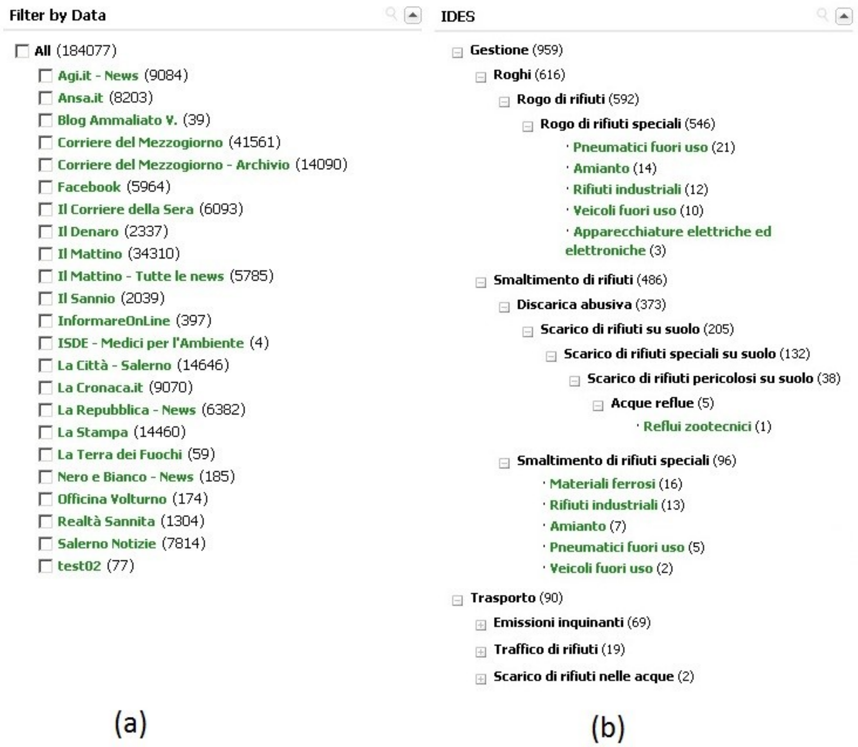
Figure 2:Meta-events.

In addition to the events, the definition of meta-events was proposed, with meta-event meaning an event category that has, as an object, other events. Meta-events are depicted in Fig. 2.

The second extension of the search tool has been the design and implementation of the domain taxonomy. The taxonomy is the hierarchical structure used to classify the documents. In this case, the classification is multidimensional, i.e. each document may be classified into zero, one, or more nodes. Nodes (or concepts) are linked together by IS-A relations. The taxonomy contains 79 concepts; at the first level there are two main concepts (waste management and waste transport) and the maximum depth of the tree is five. The taxonomy is designed to support information searches into abusive waste disposal in unstructured documents and is very dependent on the definition of events.

The semantic search module indexes all documentary sources on a daily basis, although it is possible to define a different refresh rate for each document source if necessary. During the indexing process for each new document published, the following steps shall be carried out:

1. Semantic analysis: all the concepts present in the document are identified and disambiguated using the semantic network present in the module.
2. Classification: the document is classified in no, one or more classes of the domain taxonomy, using the result of the previous analysis and classification rules.
3. Extraction and meta-events: events and meta-events are extracted using the result of semantic analysis and extraction rules. Events and meta-events extracted on a certain date can be obtained in RDF (Resource Description Framework) format. In the RDF model, in addition to the already described attributes of the event, these are also included: the source documents from which the data were extracted (e.g. Facebook), the document (e.g. the post), the phrase, the date of acquisition, the unique ID of the event, the subclass of the event, the event class and the event destination.



Filter by Data

- □ All (184077)
  - □ Agi.it - News (9084)
  - □ Ansa.it (8203)
  - □ Blog Ammaliato V. (39)
  - □ Corriere del Mezzogiorno (41561)
  - □ Corriere del Mezzogiorno - Archivio (14090)
  - □ Facebook (5964)
  - □ Il Corriere della Sera (6093)
  - □ Il Denaro (2337)
  - □ Il Mattino (34310)
  - □ Il Mattino - Tutte le news (5785)
  - □ Il Sannio (2039)
  - □ InformareOnLine (397)
  - □ ISDE - Medici per l'Ambiente (4)
  - □ La Città - Salerno (14646)
  - □ La Cronaca.it (9070)
  - □ La Repubblica - News (6382)
  - □ La Stampa (14460)
  - □ La Terra dei Fuochi (59)
  - □ Nero e Bianco - News (185)
  - □ Officina Volturno (174)
  - □ Realtà Sannita (1304)
  - □ Salerno Notizie (7814)
  - □ test02 (77)

(a)

IDES

- ⊟ Gestione (959)
  - ⊟ Roghi (616)
    - ⊟ Rogo di rifiuti (592)
      - ⊟ Rogo di rifiuti speciali (546)
        - · Pneumatici fuori uso (21)
        - · Amianto (14)
        - · Rifiuti industriali (12)
        - · Veicoli fuori uso (10)
        - · Apparecchiature elettriche ed elettroniche (3)
  - ⊟ Smaltimento di rifiuti (486)
    - ⊟ Discarica abusiva (373)
      - ⊟ Scarico di rifiuti su suolo (205)
        - ⊟ Scarico di rifiuti speciali su suolo (132)
          - ⊟ Scarico di rifiuti pericolosi su suolo (38)
            - ⊟ Acque reflue (5)
              - · Reflui zootecnici (1)
    - ⊟ Smaltimento di rifiuti speciali (96)
      - · Materiali ferrosi (16)
      - · Rifiuti industriali (13)
      - · Amianto (7)
      - · Pneumatici fuori uso (5)
      - · Veicoli fuori uso (2)
- ⊟ Trasporto (90)
  - ⊞ Emissioni inquinanti (69)
  - ⊞ Traffico di rifiuti (19)
  - ⊞ Scarico di rifiuti nelle acque (2)

(b)

Figure 3: (a) Documentary sources; (b) Classification of Facebook posts.

Documents have been indexed in Italian and in English, although most are written in Italian.

3 Results

During the first six months of 2013, approximately 184,000 documents were indexed; Fig. 3(a). The total number of Facebook groups analyzed is about 5,964, from which 2175 reports of events were extracted. An example of Facebook document (posts) classification is shown in Fig. 3(b).The first RDF data processing filters events that have at least one of these attributes: address, place, town, type of place. If there are multiple values, they are used to attribute greater precision; indeed, we use as primary source the data field “address”, that it is more accurate and then the type of place and so on. The result of this elaboration can be available in the GIS and therefore can be used for geostatistical and spatial analysis. For example, Fig. 5 shows the Kriging model describing the distribution of the events in the province of Caserta.

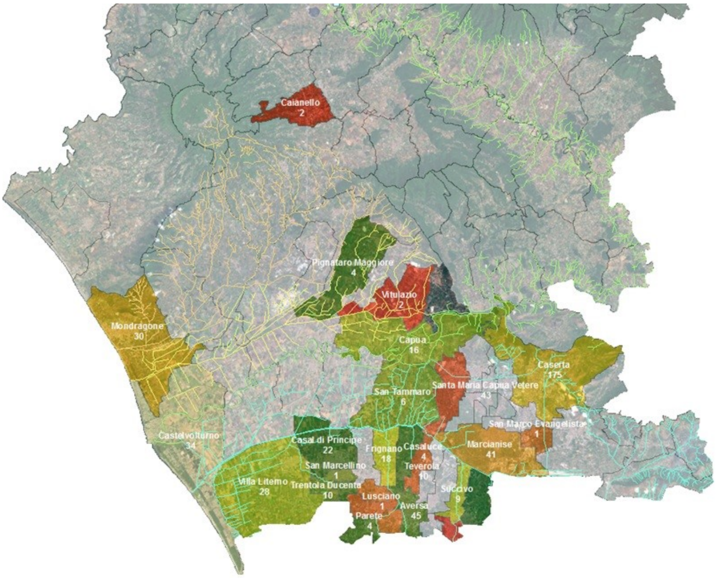


Figure 4: Map showing distribution of events in the province of Caserta.

4 Conclusion and future work

In this paper, the semantic search approach has been developed. In Table 1, the advantages and limitations are shown for each approach.

Table 1: Advantages and limitations of different approaches.

	Advantages	Limitations
Satellite approach (High Resolution Imaging)	High radiometric sensor resolution	High costs for acquisition database
	High spectral sensor resolution	High costs for data interpretation
	High spatial sensor resolution	Low temporal sensor resolution
	Georeferencing processing	
Airborne approach	High spatial sensor resolution	High costs of database acquisition
	Detailed information on the composition and related physical properties of detected objects	
Semantic search approach	High temporal sensor resolution	Uncertainty in the event time/date
	Mid/low costs for database acquisition	Uncertainty in the event positioning

The results obtained by the semantic search approach prove that the applied methodology provides a useful instrument at an inexpensive level in respect to satellite and airborne sensor systems for environmental monitoring to support the investigation of potential illicit crime. The human sensors allow an easier observation of local events. This approach strongly depends in turn on: the quality of the information provided by users and the correct extraction and classification of information from natural language texts. The biggest problem related to the quality of information is the possible presence of more posts that refer to the same event. In this case, the same event would be retrieved several times and this could have an effect on later analysis. In future work we intend to define a measure of similarity between events that allows us to identify potential duplicates (same event, same day, same place, etc.) with a degree of reliability. As regards the correctness of the classification and the extraction of information from text, to the best of our knowledge, there are no public datasets that are specific to this domain against which to measure the performance (precision, recall, etc.) of the natural language processing algorithms used. In future work, we plan to prepare specific domain datasets for performance evaluation.

References

[1] Persechino, G., Schiano, P., Lega, M., Napoli, R.M.A., Ferrara, C., Kosmatka, J., Aerospace-based support systems and interoperability: The solution to fight illegal dumping, *WIT Transactions on Ecology and the Environment*, **140**, pp. 203-214, WIT Press, 2010

[2] Lega, M., Napoli, R.M.A., A new approach to solid waste landfills aerial monitoring, *WIT Transactions on Ecology and the Environment*, 109, pp. 193-199, WIT Press, 2008

[3] Lega, M., Ceglie, D., Persechino, G., Ferrara, C. & Napoli, R.M.A., Illegal dumping investigation: A new challenge for forensic environmental engineering, *WIT Transactions on Ecology and the Environment*, 163, pp. 3-11, WIT Press, 2012

[4] Lega, M., D'Antonio, L., Napoli, R.M.A., Cultural heritage and waste heritage: Advanced techniques to preserve cultural heritage, exploring just in time the ruins produced by disasters and natural calamities, *WIT Transactions on Ecology and the Environment*, 140, pp. 123-134, WIT Press, 2010

[5] Lega, M., Kosmatka, J., Ferrara, C., Russo, F., Napoli, R.M.A., Persechino, G., Using advanced aerial platforms and infrared thermography to track environmental contamination, *Environmental Forensics*, 13 (4), pp. 332-338, Taylor & Francis, 2012

[6] Lega, M., Napoli, R.M.A., Aerial infrared thermography in the surface waters contamination monitoring, *Desalination and Water Treatment*, 23 (1-3), pp. 141-151, Taylor & Francis, 2010

[7] Persechino, G., Lega, M., Romano, G., Gargiulo, F., Cicala, L., IDES project: An advanced tool to investigate illegal dumping, *WIT Transactions on Ecology and the Environment*, **173**, pp. 603-614, WIT Press, 2013