International Journal of
Geo-Information

Article

Spatial Context from Open and Online Processing (SCOOP): Geographic, Temporal, and Thematic Analysis of Online Information Sources

Colin Robertson * and Kevin Horrocks

Department of Geography and Environmental Studies, Wilfrid Laurier University,
Waterloo, ON N2L 3C5, Canada; kevin.horrocks@greenanalytics.ca
* Correspondence:crobertson@wlu.ca; Tel.: +1-519-884-0710 (ext. 4757)

Academic Editors: Bert Veenendaal, Maria Antonia Brovelli, Serena Coetzee, Peter Mooney and Wolfgang Kainz
Received: 16 May 2017; Accepted: 22 June 2017; Published: 26 June 2017

Abstract: The Internet is increasingly a source of data for geographic information systems, as more data becomes linked, available through application programing interfaces (APIs), and more tools become available for handling unstructured web data. While many web data extraction and structuring methods exist, there are few examples of comprehensive data processing and analysis systems that link together these tools for geographic analyses. This paper develops a general approach to the development of spatial information context from unstructured and informal web data sources through the joint analysis of the data's thematic, spatial, and temporal properties. We explore the utility of this derived contextual information through a case study into maritime surveillance. Extraction and processing techniques such as toponym extraction, disambiguation, and temporal information extraction methods are used to construct a semi-structured maritime context database supporting global scale analysis. Geographic, temporal, and thematic content were analyzed, extracted and processed from a list of information sources. A geoweb interface is developed to allow user visualization of extracted information, as well as to support space-time database queries. Joint keyword clustering and spatial clustering methods are used to demonstrate extraction of documents that relate to real world events in official vessel information data. The quality of contextual geospatial information sources is evaluated in reference to known maritime anomalies obtained from authoritative sources. The feasibility of automated context extraction using the proposed framework and linkage to external data using standard clustering tools is demonstrated.

Keywords: geospatial data; data integration; surveillance; spatial analysis; VGI

1. Introduction

The proliferation of online and streaming spatial information sources has created new opportunities for social and natural sciences, and by extension, geographic information science [1]. The use of web-based data and information sources in geographic information systems (GIS) is still in its infancy, as classical database management system-based applications are adapted or augmented to handle unstructured data. The emerging fields of web science [2], linked data [3], and netnography [4], signal a sea-change in the acquisition of information for research. GIScience has figured prominently in these burgeoning fields, with its long history of data integration and conjoining disparate datasets through shared spatialities. In the research on volunteered geographic information (VGI) in recent years, advanced analysis has been mostly limited to platforms with a robust data model such as Open Street Maps [5] and highly focused on issues of data quality [6,7]). For more ephemeral sources of VGI such as geosocial data, analysis has been mostly limited to mapping distributions (e.g., [8,9]), tracking population-level trends (e.g., [10]), identifying points of interest (e.g., [11,12]), and extracting place-related contextual information from geocoded tweets (e.g., [13]), Flickr images (e.g., [14–16]), or other sources (e.g., [17]). The spatial analysis of web documents however, has not been considered fully within this literature, leaving the tools for extraction and data modelling somewhat disjointed from the analytical methods needed to understand these data. In this paper, we aim to develop an analytical approach for web documents obtained through geographic information extraction methods.

Geographic information retrieval (GIR) research provides tools for extracting and ranking information about locations and places from text documents [18]. Typically, the GIR task takes the form of a collection of documents, a query that has a geographical component, and a measure of relevance which assess the degree each document is associated with the query. GIR research therefore encompasses the techniques for identifying place names within bodies of text (e.g., geoparsing), assessing the interpretations of identified place names (e.g., disambiguation), and scoring and ranking based on thematic and/or spatial similarities. A rich literature has developed around the methodologies to support these GIR tasks [19,20]. Yet, comparatively little research has been reported utilizing these techniques in applied geographic analysis approaches generally or in case studies.

Recent research has emphasized the integration of multiple sources of VGI [21–23]. Cross-platform integration can also serve as a potential tool to facilitate quality assessment [24]. Exploiting geographic information obtained from web documents has the potential to scale this assessment exponentially, so long as geographic content can be accurately extracted and modelled from documents, whether through encoded coordinates available in geolocation APIs, mobile device location sensors, map interfaces, or natural language methods. For example, [25] used newspaper archive data to extract a geographic database of historical flood events. Here, GIR tools are used to generate a spatial database that supports a geoweb application for additional visualization and querying.

However, a key challenge in the use of spatial data obtained from web sources is determining the information context. In the field of information science, context can be defined as any information that can be used to characterize or improve interpretation of an entity, which may be a person, place, or a software object. Context provides the extra-information that enhances understanding of a unit of information—whether in a conversation, a statistical analysis, or a map. We consider the building of contextual information from web-scraped data through the use of spatial, temporal, and thematic identifiers, which are then used to provide interpretation to independent datasets. With increasing use of user-generated, transactional, and machine-generated geographical information, there is growing need for contextual data (e.g., [26]) to aid interpretation and support abductive reasoning. As more components of an information processing, analysis, and reporting pipeline are automated, developing tools that aid in identifying information context becomes critical.

Online surveillance systems offer an illustrative use-case to explore how spatial analysis of web document data can provide information context. A surveillance system can be broadly defined as a system designed for collection, collation, integration, analysis, and reporting of information in an ongoing setting, such that the information, maps, reports, and summaries produced by the system are useful to decision makers. The concept has been mostly defined for disease surveillance systems, many developed in the wake of terrorism attacks and bioterrorist fears (see [27]). These systems exhibit a variety of objectives, from trend detection, outbreak control planning, disease burden characterization, to situational awareness. Many systems have also focused on modelling and visualizing the spatial and temporal characteristics of information being tracked [28]. A good example is the HealthMap system, which tracks global disease outbreaks and public health threats through surveillance of online information sources [29,30]. The HealthMap system uses four stages of data processing in structuring information; acquisition, categorizing, clustering, and filtering of both official and informal information sources. Customized dictionaries for structuring disease keyword matching and location references are used in HealthMap. However, computational surveillance systems can be deployed in any scenario where information currency and granularity are critical needs.

One of the ways to formulate contextual information for unstructured content being consumed by an information system is through an ontology. Ontology-based models provide an explicit definition of key concepts, relationships, and processes to represent a particular domain of knowledge. These systems can represent complex processes and incorporate a range of information sources at multiple levels/scales [31]. However, ontological models are best indicative of a 'normal state' of a process, and are generally poorly equipped to handle anomalous or unforeseen perturbations to the system [32]. As well, exploiting web-based information sources for geographic context in a surveillance system contains significant heterogeneity in information provenance and relevance, as well as uncertainties introduced through processing of natural language data that complicate formalization. Secondly, ontologies formalize knowledge and encode reasoning schemes from a deductive perspective, yet in many circumstances, such as monitoring situational awareness and anomaly detection, objectives are more diffuse and dynamic. Abductive reasoning approaches advocated for Big Data analytics [33] instead aim at identifying plausible explanations for observed data, and are by nature evolutionary and flexible. We take this approach to context extraction for online surveillance in this paper, focusing on fusing thematic, temporal, and spatial properties of online data sources.

Fusion of web-extracted data can take advantage of similarity measures that estimate the degree to which two or more vectors of information are alike. Pooling similarity measures can then be used to discriminate distinct groups of documents. Such an approach mimics human inference in filtering out irrelevant information and in arriving at consensus on the quality and relevance of information. Figure 1 outlines a conceptual model for integrating these information dimensions into a surveillance system, based loosely on Sinton's formal model of geographical information [34] as well as the triad model of Peuquet [35]; and the related pyramid model of Mennis et al. [36] that distinguishes between storage of data and synthesizing knowledge from that data. We propose the conceptual model in Figure 1 for handling web document data sources in the context of surveillance. Moving from left to right, raw data acquired from online APIs, web crawlers, or scripts that download and store webpage content are obtained in each time step. From each data source input into the pipeline, data extraction routines identify temporal, spatial, and thematic entities (themes derived from a pre-defined keyword list that can evolve over time). Note that the nature of these thematic entities can be either time-stamps and locational coordinates or durations and regions. Similarity measures in space, time, and theme can then be constructed to fuse 'similar' content-objects and present information outputs in an auxiliary application or method. Our research goal was to develop this framework through an implementation to a maritime surveillance system.

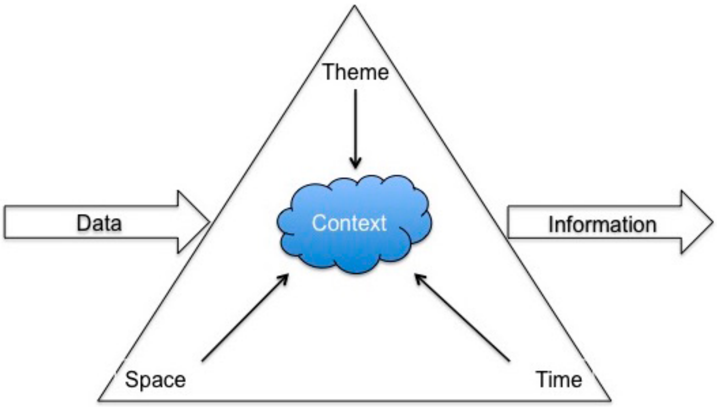


Figure 1. Simple conceptual model for learning context from integration of spatial, temporal, and thematic properties of Internet data sources within a surveillance system.

This paper is organized as follows. The following section outlines the architecture and technical details of an implementation of Figure 1, which we have denoted spatial context from open and online processing (SCOOP), as we are focused on identifying spatial outliers (areas on the map 'of interest' to a system user). We describe in detail the software components, implementation decisions, and overall processing chain for automated extraction, structuring, and analysis of web-based data. Section 3 outlines a detailed case study of the system deployed for maritime surveillance, with sample analyses and results from preliminary runs. Various results are outlined to demonstrate the abilities and potential of exploitation of open data and processing and analysis tools within the maritime domain using the SCOOP framework. Finally, we conclude with a brief discussion of the limitations of our current system and highlight research opportunities.

2. Materials and Methods

2.1. Project Architecture

The goals of the project methodology were to build an information processing system which could extract and structure web-based information in order to support spatial, temporal, and thematic analysis that could provide value to surveillance systems where early warning was a core requirement. To handle each class of entity, both custom and existing open-source Python tools were used. The information processing chain architecture is outlined in Figure 2.

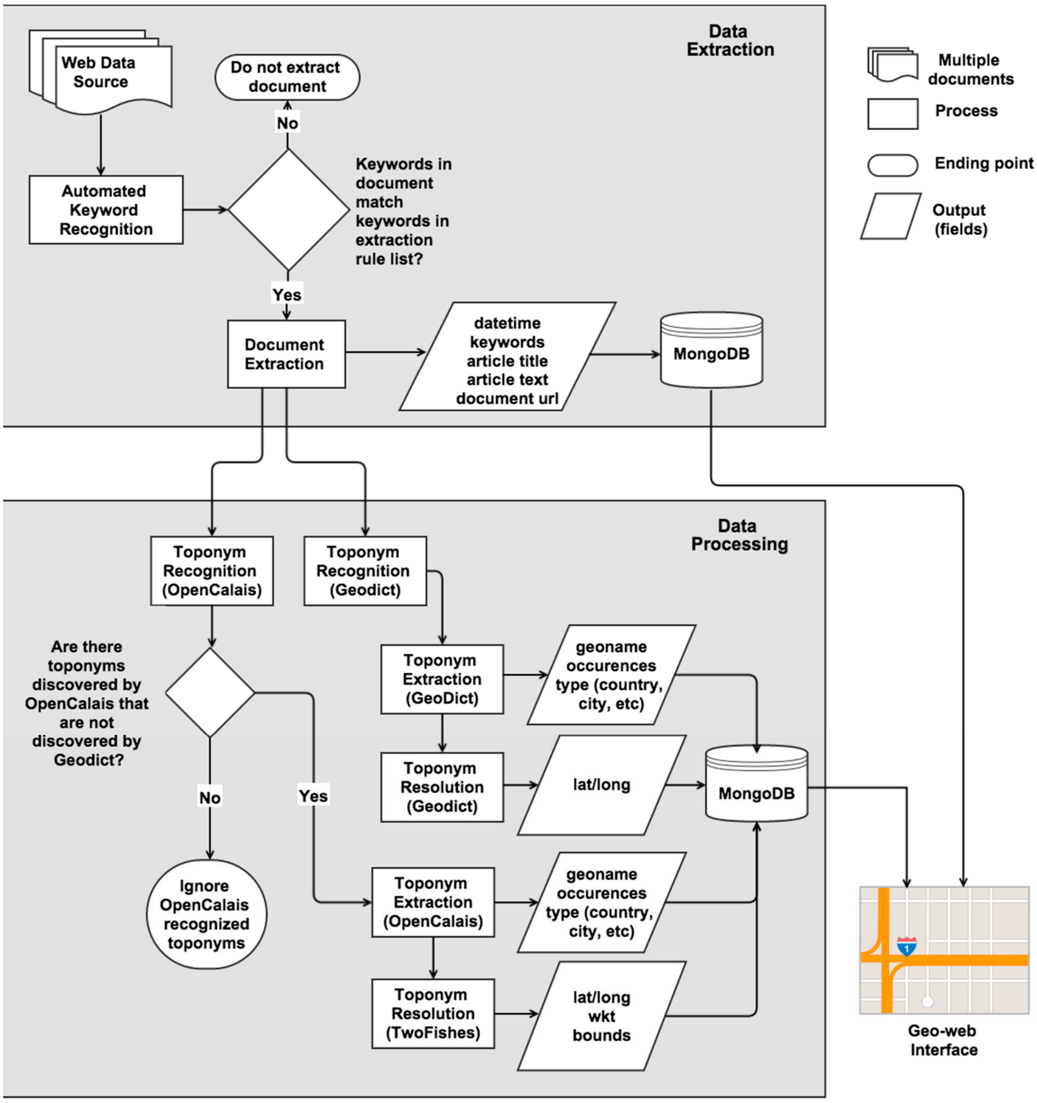

Data Extraction
Web Data Source
Automated Keyword Recognition
Do not extract document
No
Keywords in document match keywords in extraction rule list?
Yes
Document Extraction
datetime keywords article title article text document url
MongoDB
Multiple documents
Process
Ending point
Output (fields)
Data Processing
Toponym Recognition (OpenCalais)
Toponym Recognition (Geodict)
Are there toponyms discovered by OpenCalais that are not discovered by Geodict?
No
Yes
Ignore OpenCalais recognized toponyms
Toponym Extraction (GeoDict)
geoname occurences type (country, city, etc)
Toponym Resolution (Geodict)
lat/long
MongoDB
Toponym Extraction (OpenCalais)
geoname occurences type (country, city, etc)
Toponym Resolution (TwoFishes)
lat/long wkt bounds
Geo-web Interface

Figure 2. Process diagram outlining flow of data through the automated data extraction and processing stages.

2.2. Data Extraction

Extraction of data is focused primarily on data sources dedicated to providing information regarding events occurring within the domain of interest such as news articles or authoritative broadcasts. A list of all data sources was produced for usage in development of web crawlers. For each web data source, a custom crawler was written in Python to allow for site crawling, extraction of documents, and recognition of duplicate documents. Scrapy (http://scrapy.org/) was chosen as the primary Python-based framework for data extraction in this research due to its high-level scraping framework, recursive logic, and ability to extract comprehensive datasets. In addition to the custom crawlers, a list of keywords was created as a reference to be used during data extraction. This list can be developed using words from a structured lexicon, literature review, or expert surveys and can be automatically edited based on system feedback. To limit extraction of irrelevant documents, potential documents are scanned by the crawler for the presence of matching keywords. Should no keywords match within a candidate document, the document will not be extracted and the crawler will move on to the next candidate document located within the data source. As needed, the keyword list can be modified to fine-tune the extraction process. During the extraction of a document, the web crawlers are configured to extract data into various fields (e.g., date, text, title, url). Only English language sources were investigated in this research, and it is unclear how these tools would perform with other languages. Following the completion of a document extraction, the field-separated data is stored in a MongoDB database collection. MongoDB was used for data storage in this research as it allows for semi-structured storage of data and includes geo-spatial query and indexing natively. Given that the design of MongoDB is schema-less, empty fields are permitted, as well as the addition of future fields in a collection without past data being affected.

Potential new documents are scanned for their URL and document title prior to extraction and transferred into the database. To avoid the instance of duplicate extraction of documents, the 'title' and 'URL' fields of each document are considered to be a unique identifier. Should a unique identifier match with any identifiers already present in the database, the potential document will be considered a duplicate and be ignored.

2.3. Named Entity Recognition

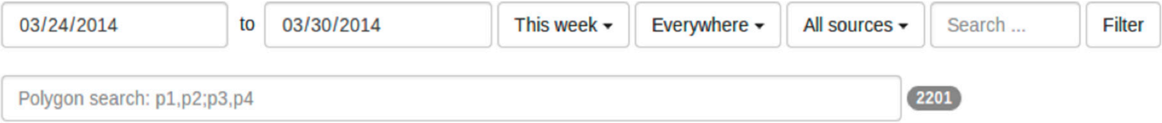
Following extraction and insertion into the database, a document's text content is then placed in a python-based task queue for data processing, which includes named entity recognition and concept mapping. Geographic entities are recognized, extracted, structured, and processed. This is accomplished through usage of open source tools for toponym recognition and resolution. For geographic entities within documents, the recognition of geographic entities is completed using two tools: Geodict (https://github.com/petewarden/geodict), an open source Python geographic entity recognition tool, as well as OpenCalais, a free-to-use multi-entity and relationship recognition tool with open-source Python wrappers.

2.4. Concept Mapping

Recognized geographic entities are mapped to meaningful concepts by matching them to an existing knowledge base or ontology. The GeoNames.org and MaxMind world cities databases are freely available gazetteers used as the primary geographic gazetteers in this research. The purpose of this step is to associate discovered named geographic entities with meaningful structured information (e.g., coordinates, polygonal area etc.). Extracted geographical entities stored in the MongoDB database are resolved to geographical location data (i.e., latitude and longitude), using two open source tools, Geodict and TwoFishes. Both point location and polygonal location data are resolved. This facilitates use of polygonal location data in geographic overlay analysis and queries for facilitating toponymn disambiguation. TwoFishes is an open-source coarse geocoder for translating toponyms to geographic coordinates, which also returns location data as well-known-text (WKT) geometries. TwoFishes is used to resolve toponyms recognized by OpenCalais, as well as provide WKT geometry and spatial extents when possible across all geographic entities.

2.5. Geo-Web Interface

A geo-web interface was developed as a tool to visualize geocoded documents stored within the database. The user interface offers a database query tool (Figure 3), interactive map and data explorer (Figures 4 and 5), JSON query exporter (Figure 6), and scraping statistics. The database query tool is designed for user-friendly extraction of data. It allows users who may not be familiar with command line interfaces to create their own datasets for analysis. Queried data is served up to the mapping interface, and each document is mapped to its stored coordinate data. This geo-web interface supports interactive exploration of new data sources and facilitates development of algorithms for exploiting these information sources.



03/24/2014 to 03/30/2014

Figure 3. Geo-web interface query tool. Allows users to set custom temporal, geographic, and thematic parameters to query, display, and extract data for analysis. Possible selections include custom or pre-set date ranges, geographic regions, specific data sources, and specific keyword and phrases. Users also have the option to draw a polygon on the mapping interface and display and extract all data within the polygon.

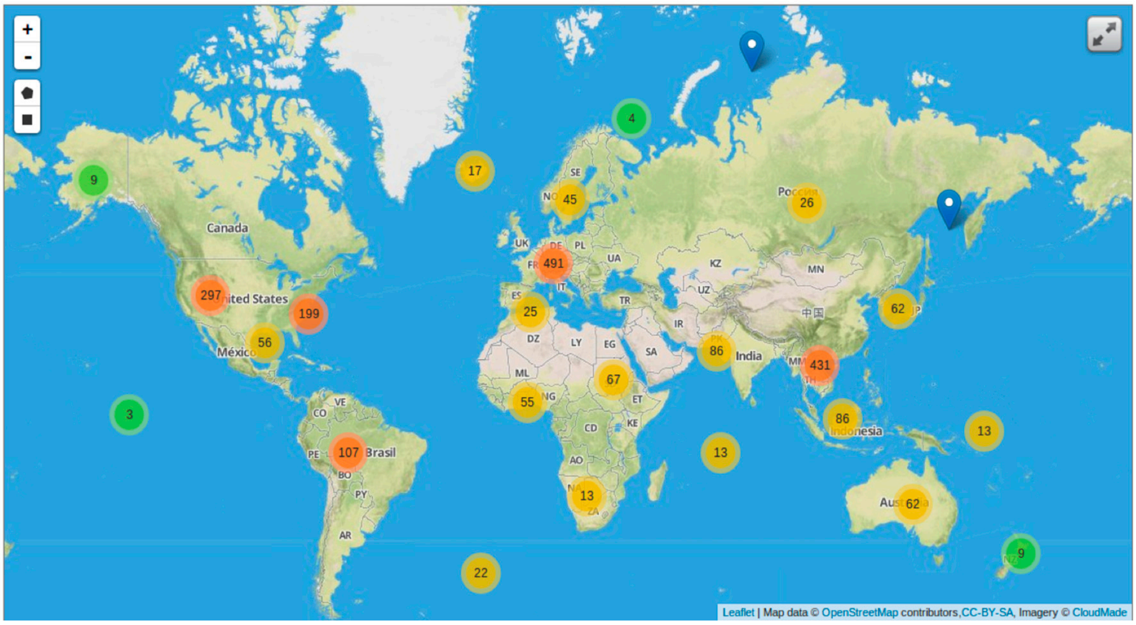
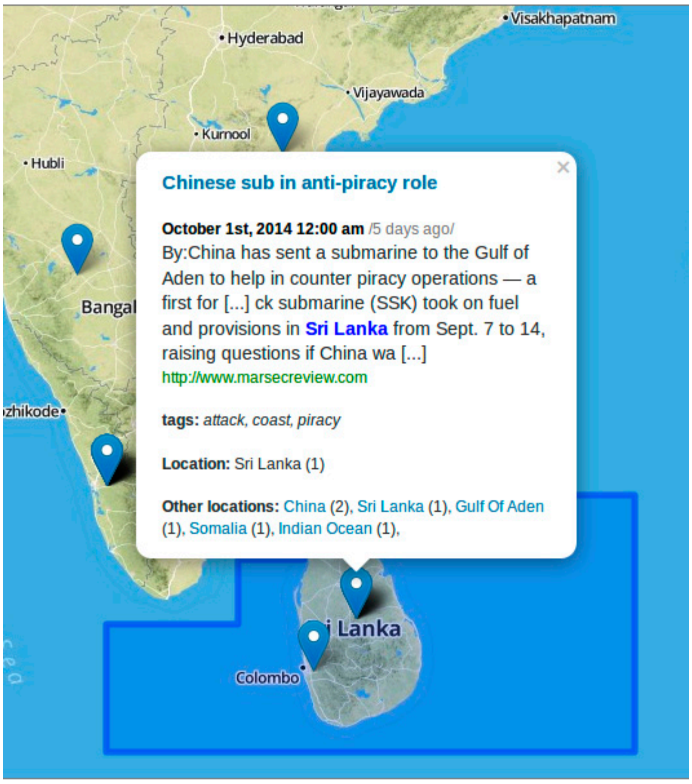


Figure 4. Geo-web user interface. By default, displays all extracted and processed data within the past week. As there is often many thousands of documents spread geographically, documents are clustered at lower zoom levels, with clusters breaking up and becoming more location specific as the user zooms in.



Chinese sub in anti-piracy role

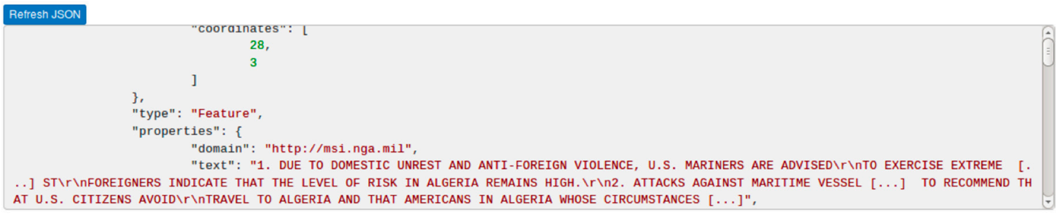
October 1st, 2014 12:00 am /5 days ago/
By:China has sent a submarine to the Gulf of Aden to help in counter piracy operations — a first for [...] ck submarine (SSK) took on fuel and provisions in **Sri Lanka** from Sept. 7 to 14, raising questions if China wa [...]
http://www.marsecreview.com

tags: *attack, coast, piracy*

Location: Sri Lanka (1)

Other locations: China (2), Sri Lanka (1), Gulf Of Aden (1), Somalia (1), Indian Ocean (1),

Figure 5. Geo-web interface mapping display. When a user clicks on a point, information regarding the respective document is displayed. Temporal (date and time), thematic (article text and keywords recognized for extraction), and geographic is displayed in a convenient popup window. This window also shows all other geographic locations that were mentioned (recognized and processed) within the article.



```
                    "coordinates": [
                            28,
                            3
                    ]
            },
            "type": "Feature",
            "properties": {
                    "domain": "http://msi.nga.mil",
                    "text": "1. DUE TO DOMESTIC UNREST AND ANTI-FOREIGN VIOLENCE, U.S. MARINERS ARE ADVISED\r\nTO EXERCISE EXTREME  [.
..] ST\r\nFOREIGNERS INDICATE THAT THE LEVEL OF RISK IN ALGERIA REMAINS HIGH.\r\n2. ATTACKS AGAINST MARITIME VESSEL [...]  TO RECOMMEND TH
AT U.S. CITIZENS AVOID\r\nTRAVEL TO ALGERIA AND THAT AMERICANS IN ALGERIA WHOSE CIRCUMSTANCES [...]",
```

Figure 6. JSON exporter tool. Once a user is satisfied with their query and information displayed, an export of this data can be generated using this tool. JSON data can easily be brought into a statistical analysis environment such as R, or converted to other formats.

2.6. Case Study: Harvesting Web Information for Maritime Surveillance

2.6.1. Introduction

Over the last two decades, globalization has increased flows of goods between countries which has intensified international shipping traffic. Additionally, technological advances in maritime traffic monitoring have led to decreased human resources for monitoring maritime safety, security and compliance. Anomalous events and activities in the maritime domain of interest include maritime pollution, terrorism, piracy, illegal trade, and human trafficking—and both legal and nefarious maritime activity is increasing [37]. Maritime domain awareness is defined as the effective understanding of all maritime activities that could impact the security, safety, economy, or environment and is achieved through the development and implementation of tools that compile timely information about maritime infrastructure, cargo, people, and vessels that contribute to enhanced security. The purpose here was to develop tools for exploiting online, publicly available information such that data can be structured in a way that will support spatial, temporal, and space-time analysis of web-based documents relating to events occurring in the maritime domain.

One of the key motivations to exploit online sources of information is the perceived information value within these online sources that have the potential to provide early-warning capabilities for nefarious maritime activity [32,38]. In many fields where early detection is critical, from tracking emerging epidemics [29] to identifying shifts in public sentiment [39], novel sources of information are being explored for the ability to provide context-specific, nuanced information often with explicit geographic and temporal references. While this is relatively simple for structured information sources (e.g., geocoded Tweets accessed through a public API), it is more difficult for unstructured information (e.g., online news articles, images). The role of informal sources of online information in maritime anomaly detection is that of context. While core data sources such as satellite-based systems and coastal radar provide highly detailed vessel location and track data, supporting kinetic anomaly detection methods, it was hypothesized that fusion with contextual information could greatly improves these methods, lowering false alarms and false negatives.

2.6.2. Methods

Data were extracted from six web sources including maritime websites, an official maritime alert broadcast, and blogs. These six sources were selected to represent a variety of web document sources, from generic news websites, to highly focused maritime websites and reports. Our goal in this selection was to mimic the variety of document data in a real-world application while keeping the total number of sources relatively low. A set of keywords was identified with expert input and used to extract documents.

A random sample of ten documents was taken from the extracted web documents and manually analyzed to determine accuracy of toponym resolution in comparison to the geographic locations mentioned within the extracted thematic text of the document (Table 1). A toponym is defined as correctly recognized if the toponym determined in the geoprocessing phase matches an implied toponym in the document text.

Table 1. Sample analysis of ten randomly sampled documents.

ID	Keywords	Toponym Recognized	Correct (Yes/No)	Relevant (Yes/No)
554	chemical, freight	China	yes	yes
446	accident, alarm, death	Black Sea	yes	yes
1213	delay, oil	Panama	yes	yes
457	Cruise	Antarctica	yes	yes
1016	death, incident, pirate	United States	no	no
380	attack, cruise, incident	Fort Lauderdale, FL, USA	yes	yes
808	cargo, chemical, crash, oil, spill	Houston, TX, USA	yes	yes
1169	harbour, oil	United States	no	no
877	Cruise	New Jersey, USA	yes	yes
183	explosion, fire, oil, spill	Florida, USA	yes	yes

Keywords recognized within documents during the extraction phase typically reflect the article theme or event being described within the document. K-means clustering was conducted on extracted documents based on their keyword content to group the documents around their thematic content. Since analysis is conducted in an ongoing basis in a surveillance system, keyword clusters identified in extracted documents would reflect distinct events being reported in the web document sources. In order to conduct this cluster analysis, a document-term matrix was created to determine keyword frequencies within each extracted document and abstract text data to numeric form which is used as the basis of clustering. The number of clusters (k) parameter is critical for k-means clustering in real applications. We assessed the percentage of explained variance based on several candidate values for k.

Plotting the sum of squared errors with respect to values of k, we selected the lowest number k where any higher number resulted in only slight reductions in error (i.e., 'the elbow method'). K-means was used as it is a simple and widely used approach for keyword clustering, though in principle any clustering technique could be employed here. This was also a fairly simple case because the keyword constraint set by the list used for document extraction reduced the parameter space and potential for sparsity significantly.

Spatial analysis was conducted to identify significant spatial clusters of thematic clusters using spatial scanning methods. The spatial scan statistic is a cluster detection methodology that supports identifying the location and size of clusters in space, time, and space-time, originally formulated by Kulldorff and Nagarwalla [40]. Circles are drawn around the centroid of each region/case (or arbitrary locations), up to some maximum radius (at default defined as 50% of the 'population at risk'). The likelihood ratio test statistic is calculated for every candidate cluster location, defined generally as

$$D(s) = P(\text{data} \mid H_1), \quad (1)$$

where the probability model can be Poisson, Binomial, Gaussian, or others. The likelihood ratio defines the probability of the number of events inside a circle given the alternative hypothesis that the circle is a cluster (area of elevated risk) divided by the probability of the data given the null hypothesis that the circle is not a cluster (equal risk inside and outside the circle). The maximum likelihood ratio of all tested circles is the most likely cluster. Significance can then be tested against a distribution of maximum scores calculated from data generated from randomizations where no spatial clusters are present. Ranking the observed likelihood ratio within this distribution yields the *p*-value.

Using a Bernouilli probability model, we can treat documents in thematic clusters as cases, and complement documents as controls, to determine the probability of finding the observed spatial pattern of thematic clusters, given the hypothesis that they are not spatially structured. If we find spatial clustering in thematic clusters, we may have greater confidence in the 'signal' from an operational surveillance perspective. This is just one candidate method for identifying interesting patterns of extracted information, and other methods for space-time surveillance will be evaluated for their utility in the system. By applying this model to the previously created thematic clusters, spatial outliers in thematically clustered events can be visualized and located. It is expected that, generally, events that are similar in theme and 'talking about the same event', will also display as clusters in space. Thematic clusters can also be generated by defining temporal criteria, thus we can jointly exploit thematic, temporal, and spatial information. Here, concurrency in time determined by the timing of when data were extracted.

3. Results

3.1. Toponym Resolution

A total of 12,789 articles were extracted and inserted into the database based on keyword matching. In the small sample used for validation, eight of the ten tested toponyms in the sample were determined to be correct. The incorrect instances were investigated further, and it was discovered that instances of "US" appearing in document text were automatically recognized as a toponym for "United States", which may not have been implied in the document. Further, recognized toponyms were analyzed to determine whether they held any relevance to the thematic content being discussed in the document text. As a document may have multiple recognitions and resolutions of toponyms, some place names that are mentioned may not have any strong relevance to the theme of the article. Issues surrounding the problem of geographic relevance is outlined in the discussion.

3.2. Sample Analysis

A subset of data with a temporal range from November 2013 to May 2014 was extracted utilizing a query on the geoweb interface. The purpose of this case study is to analyze the spatial clusters generated from previously determined thematic clusters. Differences in geographic location and thematic content of various clusters can be explored in a spatio-temporal visualization (Figure 7). In this case study, eight thematic clusters are generated via k-means clustering, two of which are further examined. A unique identification number was assigned to all documents within each of the two clusters, and further passed through the Bernoulli spatial scan statistic. The following constraints were put in place for the spatial scan: a minimum temporal cluster size of one month, and a maximum temporal cluster size 50% of study period. Results from the spatial scan analysis of Cluster 2 of 8 are outlined in Figure 8. Results from the spatial scan analysis of Cluster 3 of 8 are outlined in Figure 9.

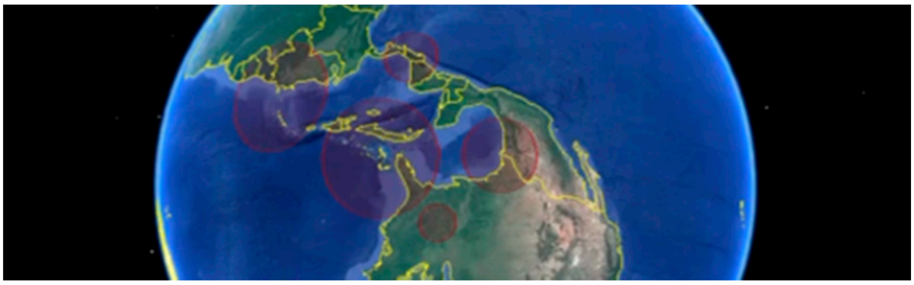


Figure 7. Generated spatial clusters are translated into a .kml file, and visualized in Google Earth software.

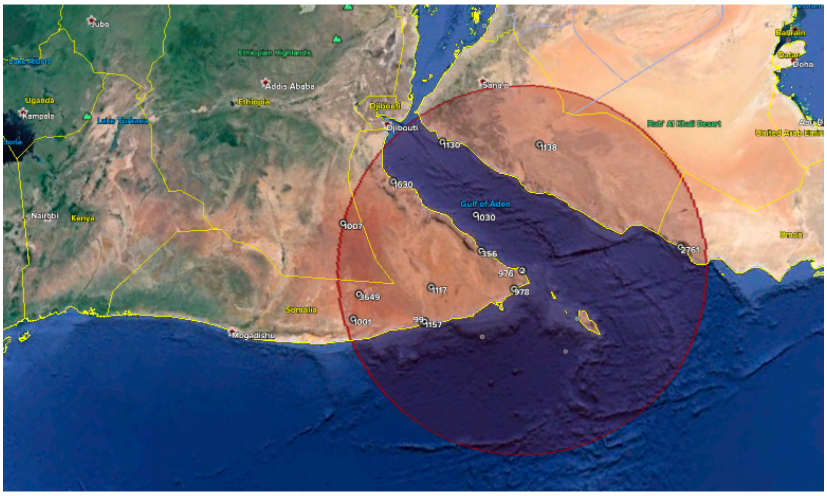


Figure 8. Results of spatial clustering of thematic cluster #2. An event cluster is generated surrounding Somalia and the Gulf of Aden. This cluster contains 17 location IDs, an observed/expected ratio of 3.41, a relative risk of 3.67, and a p-value of 0.00003.

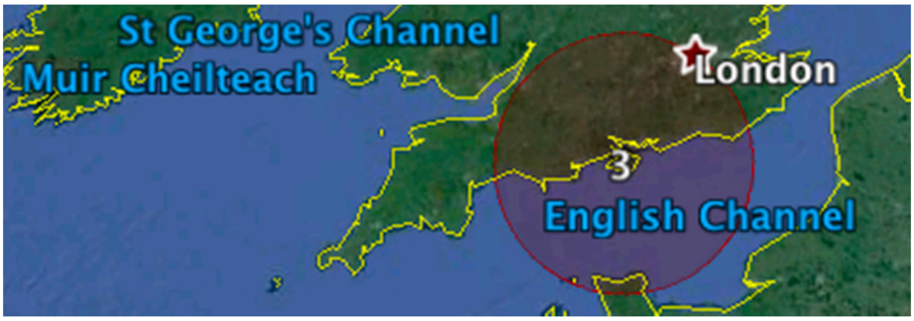


Figure 9. An event cluster encasing southern UK and the English Channel. This cluster contains 31 location IDs, an observed/expected ratio of 9.06, a relative risk of 9.24, and a p-value of 0.012.

To assess the thematic content of documents lying in these generated clusters, a subset was performed using the unique document IDs to separate clustered events into a new dataset. Wordclouds generated from the clusters keywords provided insight into the nature of the events identified in the clusters in Somalia (piracy) and the UK (fire).

As the spatial scan statistic allows for temporal constraints as well as on-line processing [41], the characteristics of clusters can be tracked over time, and newly emergent clusters discovered. Clusters can be generated and analyzed by day, month, or year. For example, the clustering of piracy events in the Malacca strait may shift to a different geographic location, or become smaller due to tighter clustering of events in a specific area. By performing automated data extraction, processing, and analysis, we can visualize maritime events by theme over time, and by utilizing multiple open data sources, we are gathering information about the maritime domain that may not be reported in official, authoritative channels.

3.3. Fusion with Authoritative Data

The results of information processing and data clustering techniques were assessed to determine quality and accuracy of the events within spatial clusters. By validating the SCOOP methodology of data processing, strength is given to the concept of using open web data sources for building intelligent surveillance systems. Multiple web data sources were explored without the aid of an information extraction system in order to gain an understanding of recent events occurring within the maritime domain. The purpose of this step was to obtain a list of suitable events to help formulate a query that could be performed within the database. An anomalous event regarding a vessel explosion aboard the Chinese 'Tian Xiang 69' cargo vessel, occurring on 4 November 2014, was chosen to perform this analysis. As nearly all toponym references within the documents surrounding this event focused in specific locations in China, a spatio-temporal query was performed by drawing a bounding box covering these locations as well as an additional 50 km spatial allowance. A three-day allowance was given for the temporal aspect of the query with the assumption that non-authoritative news outlets often do not report on events until the following days. A query was deemed successful if three or more documents deemed relevant to the specific event were returned. The results of this query were further passed through the clustering step of SCOOP, in order to determine spatial outliers of documents gathered from the query. AIS (automated identification system) is a system primarily used for tracking, identifying, and data exchange. Broadcast transmitters are commonplace aboard a vast number of commercial and non-commercial vessels due to international standards, laws, and guidelines. Information transmitted in a typical AIS broadcast signal includes vessel identification, position, course, and speed. By utilizing the live AIS-based vessel tracking and record keeping service 'MarineTraffic.com', records of AIS broadcasts from the Tian Xiang 69 were located at the time of the anomalous event. All AIS data broadcast by the vessel on 4 November 2014 was exported and further assessed by message time-stamps in order to isolate the exact time period where the anomalous event occurred. Each of these isolated records was further plotted to a map by their coordinate data. It is expected that, generally, the plotting of coordinate data of AIS transmissions at the time of anomalous events will either overlap or lie relatively close in space to its respective spatial cluster that was generated by the SCOOP system. Data fusion was performed by integrating the results of the authoritative AIS information into the results of the spatial cluster analysis (Figure 10). By validating the accuracy of open-data focused spatial clusters of anomalous events, the ability of the SCOOP system to harvest and process thematic, temporal, and geographic data from open web sources is highlighted in its ability to detect anomalous events that are confirmed in an authoritative capacity, as well as events that may be lacking in authoritative confirmation. The exploitation of data from open web sources provides valuable information—useful as a complement to traditional methods of maritime domain awareness.

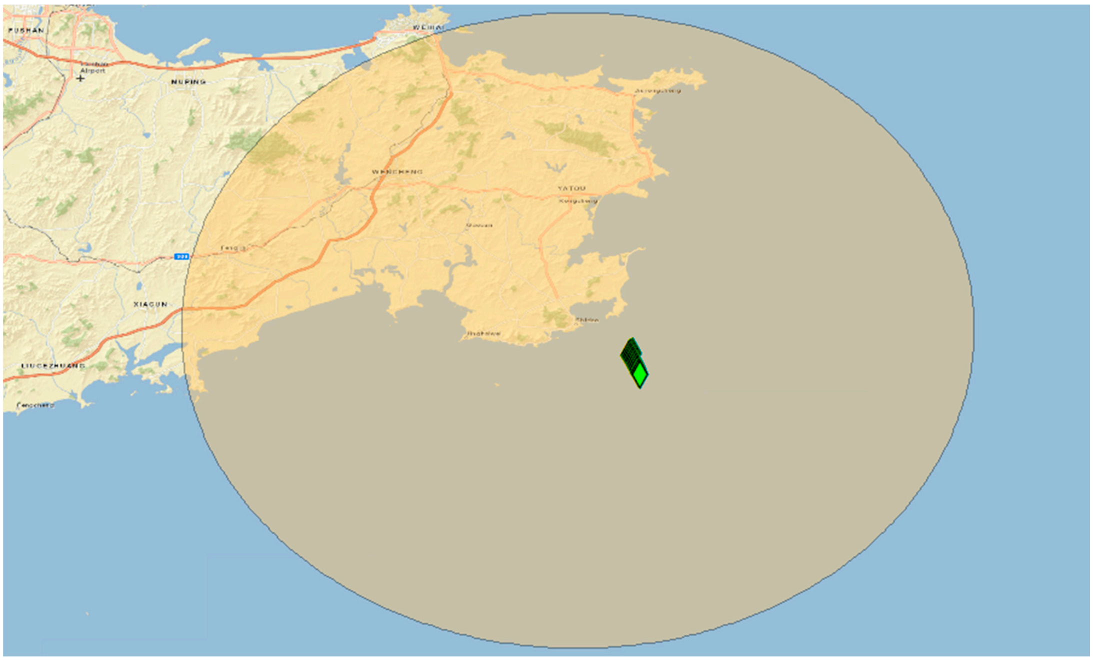


Figure 10. Fusion of data is performed by overlaying historical vessel automated identification system (AIS) data from time of incident over top of spatial cluster which was generated using the SCOOP methodology. AIS broadcasts from the Tian Xiang 69 vessel are represented by small diamonds and shown located directly within the generated cluster. Its initial isolated AIS broadcast is located 58 km to the closest cluster outer edge.

4. Discusssion

The case study into marine surveillance using the SCOOP framework demonstrated how web data extraction using scraping along with geographic, temporal, and thematic structuring can be used to build a database for operational spatial analysis within a surveillance system. Through fusion of SCOOP data with independent vessel locations in AIS data we showed that the clusters discovered through thematic and spatial analysis coincided with actual events. The web document data with sufficient structuring provide contextual information about events detected in authoritative sources such as the AIS data. Such data fusion approaches whereby space, time, and theme (Figure 1) are co-aligned to provide contextual information may be an important approach for building intelligent online surveillance systems.

The role of human interaction in the synthesis of big data is an area of active research, with advocates in visual-analytics calling for a human-centric role, evaluating and comparing information outputs [42,43]. On the other side of the spectrum, machine-intelligence researchers aim to automate as much of the data-analysis and decision-support work flow as possible, even if the understanding is mired within blackbox algorithms [44,45]. This approach is the norm in applications with well-defined problems that lend themselves to optimization schemes such as online recommendation engines. In applications that centre on geographic information and the development of decision-support tools, human interpretation is still paramount; as the ability of humans to synthesize and reason about spatial patterns across multiple dimensions is still considered superior to machines. We identify information context as a key aspect to building intelligent applications with big geographic datsets obtained from hetergeneous sources. Our approach in a limited case study demonstrated the ability to identity real events and support simple visualization and querying using a processing chain of open and freely available tools and APIs.

In the analysis presented here, context is provided by spatial and thematic clustering. First, thematic clusters were derived from a k-means analysis of terms in extracted web documents which were run through the data processing pipeline in Figure 2. While k-means clustering was used in this example, any clustering method could be used here. Next, spatial clustering of the thematic clusters was performed to provide further evidence that thematic clusters pertained to an event of interest. These results were visualized on a web-map and were queryable across space, theme, and time.

Several areas of future research are opened up by this research. Concept mapping is typically ambiguous, meaning multiple concepts are matched to a given entity, and a step is required to choose the most likely and/or correct concept for each recognized entity. In the geographic case, there are several strategies for disambiguating geographic place references. One approach is using a relation to compare candidate concepts for each entity, for example, for geographic entities. Many spatial concepts have been utilized to develop heuristics or rules to aid in geographic disambiguation [46,47]. Spatial relationships (i.e., proximity and intersection) as candidate concepts closer in physical space are more likely to be correct. Finally, in the current architecture, the development of crawlers is an implementation bottleneck. We are currently exploring the development of crawler templates and social media data mining for a more robust and dynamic data collection system.

The technical details for optimizing GIR methods were not the objective of this research. Rather, we aimed to develop an applied example of our SCOOP model in the context of maritime surveillance. The context database developed from GIR processing of web document data using open source tools provided data sufficient for standard analysis tools within an automated system.

Acknowledgments: Funds for this research came from the National Sciences and Engineering Research Council of Canada.

Author Contributions: C.R. planned and conceived of the project. C.R. and K.H. designed the overall conceptual architecture; K.H. implemented the system and analyzed the data; K.H and C.R. wrote the paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Warf, B.; Arias, S. *The Spatial Turn: Interdisciplinary Perspectives*; Taylor & Francis: New York, NY, USA, 2008.
2. Berners-Lee, T.; Hall, W.; Hendler, J.; Shadbolt, N.; Weitzner, D.J. Creating a Science of the Web. *Science* **2006**, *313*, 769–771. [CrossRef] [PubMed]
3. Bizer, C. The Emerging Web of Linked Data. *IEEE Intell. Syst.* **2009**, *24*, 87–92. [CrossRef]
4. Kozinets, R.V. *Netnography: Doing Ethnographic Research Online*; SAGE Publications: London, UK, 2010.
5. Brovelli, M.A.; Minghini, M.; Molinari, M.; Mooney, P. Towards an Automated Comparison of OpenStreetMap with Authoritative Road Datasets. *Trans. GIS* **2017**, *21*, 191–206. [CrossRef]
6. Haklay, M. How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environ. Plan. B Plan. Des.* **2010**, *37*, 682–703. [CrossRef]
7. Goodchild, M.F. The quality of big (geo)data. *Dialog. Human Geogr.* **2013**, *3*, 280–284. [CrossRef]
8. Li, L.; Goodchild, M.F.; Xu, B. Spatial, temporal, and socioeconomic patterns in the use of Twitter and Flickr. *Cartogr. Geogr. Inf. Sci.* **2013**, *40*, 61–77. [CrossRef]
9. Hawelka, B.; Sitko, I.; Beinat, E.; Sobolevsky, S.; Kazakopoulos, P.; Ratti, C. Geo-located Twitter as proxy for global mobility patterns. *Cartogr. Geogr. Inf. Sci.* **2014**, *41*, 260–271. [CrossRef] [PubMed]
10. Quercia, D.; Ellis, J.; Capra, L.; Crowcroft, J. Tracking "Gross Community Happiness" from Tweets. In Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work, New York, NY, USA, 11–15 February 2012; ACM: New York, NY, USA, 2012; pp. 965–968.
11. Mummidi, L.N.; Krumm, J. Discovering points of interest from users' map annotations. *GeoJournal* **2008**, *72*, 215–227. [CrossRef]
12. Mülligann, C.; Janowicz, K.; Ye, M.; Lee, W.-C. Analyzing the Spatial-Semantic Interaction of Points of Interest in Volunteered Geographic Information. In *Spatial Information Theory*; Springer: Berlin/Heidelberg, Germany, 2011; pp. 350–370.
13. MacEachren, A.; Jaiswal, A.; Robinson, A.; Pezanowski, S.; Savelyev, A.; Mitra, P.; Blanford, J. SensePlace2: GeoTwitter analytics support for situational awareness. In Proceedings of the 2011 IEEE Conference on Visual Analytics Science and Technology (VAST), Providence, RI, USA, 23–28 October 2011; pp. 181–190.

14. Kennedy, L.; Naaman, M.; Ahern, S.; Nair, R.; Rattenbury, T. How flickr helps us make sense of the world: Context and content in community-contributed media collections. In Proceedings of the 15th International Conference on Multimedia, Augsburg, Germany, 23–28 September 2007; ACM: New York, NY, USA, 2007; pp. 631–640.
15. Rattenbury, T.; Naaman, M. Methods for extracting place semantics from Flickr tags. *ACM Trans. Web* **2009**, *3*, 1:1–1:30. [CrossRef]
16. Feick, R.; Robertson, C. A multi-scale approach to exploring urban places in geotagged photographs. *Comput. Environ. Urban Syst.* **2014**, *53*, 96–109. [CrossRef]
17. Noulas, A.; Mascolo, C.; Frias-Martinez, E. Exploiting Foursquare and Cellular Data to Infer User Activity in Urban Environments. In Proceedings of the 2013 IEEE 14th International Conference on Mobile Data Management, Milan, Italy, 3–6 June 2013; Volume 1, pp. 167–176.
18. Jones, C.B.; Purves, R.S. Geographical information retrieval. *Int. J. Geogr. Inf. Sci.* **2008**, *22*, 219–228. [CrossRef]
19. Purves, R.S.; Clough, P.; Jones, C.B.; Arampatzis, A.; Bucher, B.; Finch, D.; Yang, B. The design and implementation of SPIRIT: A spatially aware search engine for information retrieval on the Internet. *Int. J. Geogr. Inf. Sci.* **2007**, *21*, 717–745. [CrossRef]
20. Derungs, C.; Palacio, D.; Purves, R.S. Resolving fine granularity toponyms: Evaluation of a disambiguation approach. In Proceedings of the 7th International Conference on Geographic Information Science (GIScience), Columbus, OH, USA, 18–21 September 2012; pp. 1–5.
21. Croitoru, A.; Crooks, A.; Radzikowski, J.; Stefanidis, A. Geosocial gauge: A system prototype for knowledge discovery from social media. *Int. J. Geogr. Inf. Sci.* **2013**, *27*, 2483–2508. [CrossRef]
22. Holderness, T. Geosocial intelligence. *SMART Infrastruct. Facil. Pap.* **2014**, *33*, 17–18.
23. Stefanidis, A.; Crooks, A.; Radzikowski, J. Harvesting ambient geospatial information from social media feeds. *GeoJournal* **2013**, *78*, 319–338. [CrossRef]
24. De Albuquerque, J.P.; Fan, H.; Zipf, A. A conceptual model for quality assessment of VGI for the purpose of flood management. In Proceedings of the 19th AGILE Conference on Geographic Information Science, Helsinki, Finland, 14–17 June 2016.
25. Yzaguirre, A.; Smit, M.; Warren, R. Newspaper archives + text mining = rich sources of historical geo-spatial data. In Proceedings of the IOP Conference Series: Earth and Environmental Science, Halifax, NS, Canada, 5–9 October 2015; 2016; Volume 34, p. 12043.
26. Crampton, J.W.; Graham, M.; Poorthuis, A.; Shelton, T.; Stephens, M.; Wilson, M.W.; Zook, M. Beyond the geotag: Situating "big data" and leveraging the potential of the geoweb. *Cartogr. Geogr. Inf. Sci.* **2013**, *40*, 130–139. [CrossRef]
27. Wagner, M.M.; Moore, A.W.; Aryel, R.M. *Handbook of Biosurveillance*; Elsevier: London, UK, 2006.
28. Robertson, C.; Nelson, T.A.; MacNab, Y.C.; Lawson, A.B. Review of methods for space-time disease surveillance. *Spat. Spatio-Temporal Epidemiol.* **2010**, *1*, 105–116. [CrossRef] [PubMed]
29. Brownstein, J.S.; Freifeld, C.C.; Reis, B.Y.; Mandl, K.D. Surveillance Sans Frontières: Internet-Based Emerging Infectious Disease Intelligence and the HealthMap Project. *PLoS Med.* **2008**, *5*, e151. [CrossRef] [PubMed]
30. Freifeld, C.C.; Mandl, K.D.; Reis, B.Y.; Brownstein, J.S. HealthMap: Global infectious disease monitoring through automated classification and visualization of Internet media reports. *J. Am. Med. Inf. Assoc.* **2008**, *15*, 150. [CrossRef] [PubMed]
31. Little, E.G.; Rogova, G.L. Designing ontologies for higher level fusion. *Inf. Fusion* **2009**, *10*, 70–82. [CrossRef]
32. Garcia, J.; Rogova, G. Contextual Knowledge and Information Fusion for Maritime Piracy Surveillance. In *NATO Advanced Study Institute (ASI) on Prediction and Recognition of Piracy Efforts Using Collaborative Human-Centric Information Systems*; Salamanca: NATO Science for Peace and Security, Sub-Series—E: Human and Societal Dynamics; Ios Press: Amsterdam, The Netherlands, 2013; Volume 109, pp. 80–88.
33. Miller, H.J.; Goodchild, M.F. Data-driven geography. *GeoJournal* **2015**, *80*, 449–461. [CrossRef]
34. Sinton, D. The inherent structure of information as a constraint to analysis: Mapped thematic data as a case study. *Harv. Pap. Geogr. Inf. Syst.* **1978**, *6*, 1–17.
35. Peuquet, D. It's about time: A conceptual framework for the representation of temporal dynamics in geographic information systems. *Ann. Assoc. Am. Geogr.* **1994**, *84*, 441–461. [CrossRef]
36. Mennis, J.L.; Peuquet, D.J.; Qian, L. A conceptual framework for incorporating cognitive principles into geographical database representation. *Int. J. Geogr. Inf. Sci.* **2000**, *14*, 501–520. [CrossRef]

37. United Nations Conference on Trade and Development (UNCTAD). *Review of Maritime Transport*; United Nations: Geneva, Switzerland, 2011; p. 229.
38. Kazemi, S.; Abghari, S.; Lavesson, N.; Johnson, H.; Ryman, P. Open data for anomaly detection in maritime surveillance. *Expert Syst. Appl.* **2013**, *40*, 5719–5729. [CrossRef]
39. Cheong, M.; Lee, V.C.S. A microblogging-based approach to terrorism informatics: Exploration and chronicling civilian sentiment and response to terrorism events via Twitter. *Inf. Syst. Front.* **2011**, *13*, 45–59. [CrossRef]
40. Kulldorff, M.; Nagarwalla, N. Spatial disease clusters: detection and inference. *Stat. Med.* **1995**, *14*, 799–810. [CrossRef] [PubMed]
41. Kulldorff, M.; Heffernan, R.; Hartman, J.; Assuncao, R.M.; Mostashari, F. A space-time permutation scan statistic for disease outbreak detection. *PLoS Med.* **2005**, *2*, e59. [CrossRef] [PubMed]
42. MacEachren, A.M.; Kraak, M.-J. Research Challenges in Geovisualization. *Cartogr. Geogr. Inf. Sci.* **2001**, *28*, 3–12. [CrossRef]
43. Andrienko, G.; Andrienko, N.; Demsar, U.; Dransch, D.; Dykes, J.; Fabrikant, S.I.; Tominski, C. Space, time and visual analytics. *Int. J. Geogr. Inf. Sci.* **2010**, *24*, 1577–1600. [CrossRef]
44. Peng, P.; Chen, H.; Shou, L.; Chen, K.; Chen, G.; Xu, C. DeepCamera: A Unified Framework for Recognizing Places-of-Interest Based on Deep ConvNets. In Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, Melbourne, Australia, 19–23 October 2015; ACM: New York, NY, USA, 2015; pp. 1891–1894.
45. Zhu, Y.; Newsam, S. Land Use Classification Using Convolutional Neural Networks Applied to Ground-level Images. In Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems, Seattle, WA, USA, 3–6 November 2015; ACM: New York, NY, USA, 2015; pp. 61:1–61:4.
46. Rauch, E.; Bukatin, M.; Baker, K. A Confidence-based Framework for Disambiguating Geographic Terms. In Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References, Edmonton, AB, Canada, 31 May 2003; Association for Computational Linguistics: Stroudsburg, PA, USA, 2003; Volume 1, pp. 50–54.
47. Smith, D.A.; Crane, G. Disambiguating Geographic Names in a Historical Digital Library. In *Research and Advanced Technology for Digital Libraries*; Constantopoulos, P., Sølvberg, I.T., Eds.; Springer: Berlin/Heidelberg, Germany, 2001; pp. 127–136.