SAND2011-3018 C

SAND2011-3018C

Open Source Geographic Information for Safeguards Analysis

Denise Bleakly, Karl Horak, Michael McDaniel

International Safeguards & Technical Systems
Sandia National Laboratories
Albuquerque, NM, USA
E-mail: drbleak@sandia.gov, kehorak@sandia.gov, mmcdani@sandia.gov

Abstract:

In this era of user-generated Web content, geographically referenced information is being published to open sources at an astounding rate. One might conceptually understand these data as the product of a distributed, decentralized sensor network capable of detecting the geographic signals of nuclear proliferation. Within an information-driven safeguards regime, these data, often created and shared by common citizens, can be invaluable to the detection of undeclared nuclear activity. Such information, however, is often overlooked and underutilized because, at present, no tools exist to systematically and efficiently extract and utilize these data. This paper describes an ongoing project that seeks to enable safeguards analysts to efficiently and effectively use open source geospatial information by leveraging web-based information technologies in novel ways.

While a great deal of geospatial data are published in well defined, easily detectable formats, most data are unstructured, heterogeneous and complex. Geospatial and domain-specific ontologies can be used to detect and convert these data into usable and semantically interoperable formats that can be effectively incorporated into an analyst's work. Working closely with safeguards analysts and other stakeholders to establish high-level requirements and derive use cases ensures that these tools are integrated into analysts' existing workflow for efficient use and high adoption.

Keywords: Open source, safeguards, geodata, geospatial, GIS, analysis, web

1 Introduction

Today's Internet is characterized by numerous interactive features that provide a plethora of avenues for user-contributed content. Ubiquitous cell and smart phones usually combine camera, web browser, global positioning system (GPS), and other tools that permit the person on the street to upload text, photographs, and video to any number of blogs, social media outlets, news networks, and other online repositories in near-real time. Growing numbers of tablet computers further swell the ranks of highly mobile, amateur and professional web authors

Much of this user-generated content is geospatially referenced. For example, modern digital cameras embed spatial coordinates by default, tweets (140-character Twitter messages) and Facebook updates allow users to "geotag" their location. Geospatial information can simply be entered manually, obtained from cell tower triangulation, or precisely derived from GPS.

And while data in social media streams are unstructured in the extreme, geospatial information can very often be systematically extracted. By way of an example, one of us gives his Twitter location as "Usually ABQ," reflecting an approximate location due to frequent travel for work. Twisst, a Twitter-based service that notifies users of daily flyover times of the International Space Station [1], manages to recognize "ABQ" as the airport code for the Albuquerque International Airport and derived the correct time zone and correct location to within 8 miles. Twisst then sends out a personalized tweet when an ISS flyover is imminent. Still others have demonstrated that an individual's location can be derived from implicit geographic information by analyzing regional colloquialisms and high-level topics within the text of Twitter posts [2].

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Importantly for the purpose of this study, some of this web content may accidentally, incidentally, or purposefully include information on nuclear facilities, materials, and perhaps even proliferation activities. However, such information is often overlooked or not used because the scripts to extract these data must be hand-crafted. Even when a well-defined Application Programming Interface (API) is available, a solution must be constructed for each data source. In order to make this information available to safeguards analysts, web-based technologies need to be leveraged in novel ways so that the end user need not rely on web programming expertise.

All of this points to a large and growing body of data that has the potential to contain geospatially referenced, safeguards-relevant information. Within an information-driven safeguards regime, these data, often created and shared by common citizens [3] can be invaluable to the detection of undeclared nuclear activity. One might conceptually understand social networks as a distributed, decentralized sensor network capable of detecting the signals of nuclear proliferation, often with geospatial metadata. Unfortunately, these data are hidden in a forest of innocuous information and as such, are often overlooked and underutilized because no tools currently exist to systematically and efficiently extract and make use of these data.

Recently, articles discussing nuclear proliferation detection using geospatial data have focused on the use of aerial and satellite imagery for change detection analysis [4], [5]. This has led the way to the systematic use of satellite and aerial imagery within the safeguards community as one way of identifying undeclared nuclear activity [6]. Other geospatially referenced open source information, such as ground level images from tourists and visitors, “crowdsourced” map data, and geospatial references in blogs or discussion wikis are a resource that have not been systematically analyzed to determine their usefulness in safeguards analysis.

This research was designed to survey current geospatial resources on the open Internet and examines the feasibility of providing geospatial tools to analysts who do not have a high level of GIS or web programming fluency. Therefore, the hypothesis of this work is that, by enabling safeguards analysts to efficiently and effectively extract and utilize geospatially referenced information from the Internet, these analysts will more often use these data to produce more complete and context rich analyses.

2 Methods

Several facets to the research have been identified for this project: (1) The identification of current open source tools with the potential assist analysts in extracting and managing geospatial data; (2) an examination of the growing number of geospatial data types and the spatialization of typically non-geographic data like photographs; and (3) a test case to demonstrate of the usefulness of these types of open source geospatial information.

2.1 Tools assessment

To guide the assessment of the tools available to achieve these goals, several requirements were established for the development of a safeguards toolset. First, the tool should be available at low or no cost to the end user. This requirement led to the examination of existing open source software that could be used, modified, and re-distributed free of charge and copyright restriction. Moreover, by leveraging open source software, existing capabilities need not be reinvented, thus saving significant development time.

Second, the developed tools should integrate easily into analysts’ existing workflows. This will help to ensure that the tools can be used widely among analysts by lowering the threshold of adoption. This also requires that tools have an intuitive design and functionality, refined user interface, and that users are closely involved development. This further requires that these tools be compatible with existing software and data.

Third, these tools must be extensible and thus adaptable to different analysis requirements, individual preferences, and future changes in workflow.

One tool that has been identified as a candidate for further development as a safeguards-specific tool is Zotero, a reference management software used to collect, organize, cite and share electronic research sources [7]. Because the workflow of safeguards analysts has a clear correspondence to the research process in which references are collected, tracked, and organized, and reports are generated, Zotero appears to be a valuable starting point for development. Also, because Zotero is free, open source, and extensible, it meets all of the established requirements for this project. Currently, Zotero is deployed as an add-on for the Firefox web browser and therefore integrates into one of the primary tools of safeguards analysts—the web browser. (Zotero is also being developed for use with other browsers and as a standalone tool.) Additionally, several modules are available that extend the core functionality of Zotero including word processor integrators and, of particular interest to this work, mapping capabilities.

Zotero Maps, an extension to the core Zotero program, identifies and geocodes place names within documents managed by Zotero. Figure 1 was created by storing a web-based article [8] in Zotero and generating a map using Zotero Maps. This capability, therefore allows a non-geospatially trained individual to rapidly extract and visualize geospatial information buried in unstructured text in significantly less time than traditional GIS methods of generating geospatial data.

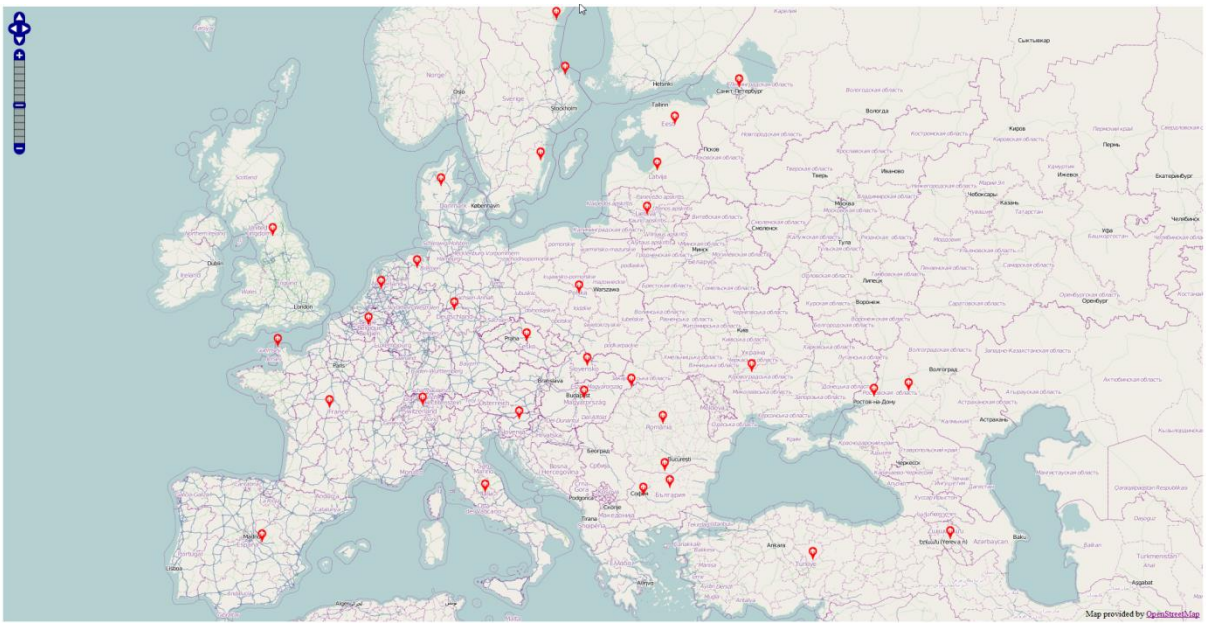

Figure 1 Map generated with Zotero Maps from the web-based article *Plans for New Nuclear Reactors Worldwide* [8].

While dozens of tools exists with potential benefit to this work, two open source software projects for handling geospatial data specifically are OpenLayers and GeoServer. OpenLayers is a pure JavaScript library for displaying and manipulating geospatial data within a web browser [9]. With the ability to display published geospatial resources from anywhere on the internet, OpenLayers is an excellent candidate for the front-end interface of a geospatial safeguards tool. GeoServer is a software server used to organize, edit and publish geospatial data [10]. Written in Java, GeoServer is a potential geospatial back-end candidate for storing and sharing geospatial data within an organization. Both OpenLayers and GeoServer are compliant with Open Geospatial Consortium standards [11] and are therefore interoperable with a wide variety of other geospatial software.

2.2 Geospatial data

Traditionally, geospatial data have come in two primary types, each with relatively few but well defined file formats: vector (or geometric) and raster (or image) data types. While this distinction still holds in general, the emergence of collaborative web technologies has supported the rise of dozens of new ways of encoding geospatial data, as well as a shift in the ways in which geospatial data is produced and conceptualized. Because high quality geospatial data is no longer created and published only by

large government agencies, as has generally been the case in the past, traditional methods of geospatial data discovery fail to lead to many of these new data types and formats.

To guide the development of these tools and to determine how such tools might assist in the geospatial data discovery process, three different search strategies have been devised to discover geospatial data in all formats. First, a general Internet search strategy using search engines such as Google, Google Scholar, and Wikipedia leads to the discovery of unstructured geospatial data in text and images. These data require additional computational procedures to transform them into geospatial data types useable in a mapping context. Second, a geographically enabled search strategy using specific geospatial filters such as coordinate pairs, bounding box coordinates, or administrative boundary names, leads to the discovery of geotagged data and geospatial web services. These data are generally unstructured but have associated geospatial metadata. Examples are geotagged images or blog posts. Third, structured geospatial data, such as ESRI shapefiles and GeoTIFF images, are discovered through geospatial data portals and clearing houses. In general these outlets are run by government or not-for-profit agencies.

Note that these search strategies are not mutually exclusive and one search strategy can lead to the discovery of different types of data.

2.3 Test Case: Paks Nuclear Power Plant, Paks, Hungary

To demonstrate the wide variety of open source geospatial referenced information available using this phased search strategy, a theoretical test case was developed based on the need to collect information helpful to a safeguards and security analyst understand Paks Nuclear Power Plant (NPP) near Paks, Hungary.

2.3.1 General Internet search

The first search included sites such as Google, Google Scholar and Wikipedia. Over 580,000 results were received on Google by searching for “Paks Nuclear Power Plant”. The first entry returned was that of Paks NPP on Wikipedia. The second site listed was the home page of the power plant.

The Wikipedia site for “Paks Nuclear Power Plant” had a wide variety of geographically referenced data [12]:

- A location Map
- Latitude and longitude coordinates
- Multiple current and historic images of the site
- Links to the Paks NPP website and many other related sites
- Links to other papers and references

The Paks NPP home page [13] in English also provided some geospatially referenced information:

- An address of the facility
- A location map
- Images of the facility in their “Virtual Tour and Gallery Links”

From these two websites alone, a substantial geospatial reference to the site can be built, note that these data are in text and image formats and as such cannot be easily utilized by traditional geographic information systems (GIS). However, by applying additional computational procedures such as natural language processors to extract place names, these data can be formatted for use in geospatial applications.

2.3.2 Geo-enabled search

From the information gained during the general internet search, specifically the coordinates of the Paks NPP (46.5725N, 18.854167E), Google Earth was used to get an aerial image of the site, dated 20 December 2006. Also available in Google Earth are 3-dimensional building renderings, including a photorealistic rendering of the main buildings at the Paks plant, including those housing the reactors, turbines and control rooms (Figure 2). Google Earth, which has become the lay person’s geographic

information system (GIS) of choice, has the ability to overlay data from dozens of already defined sources including Web Mapping Service (WMS) layers from any external source.

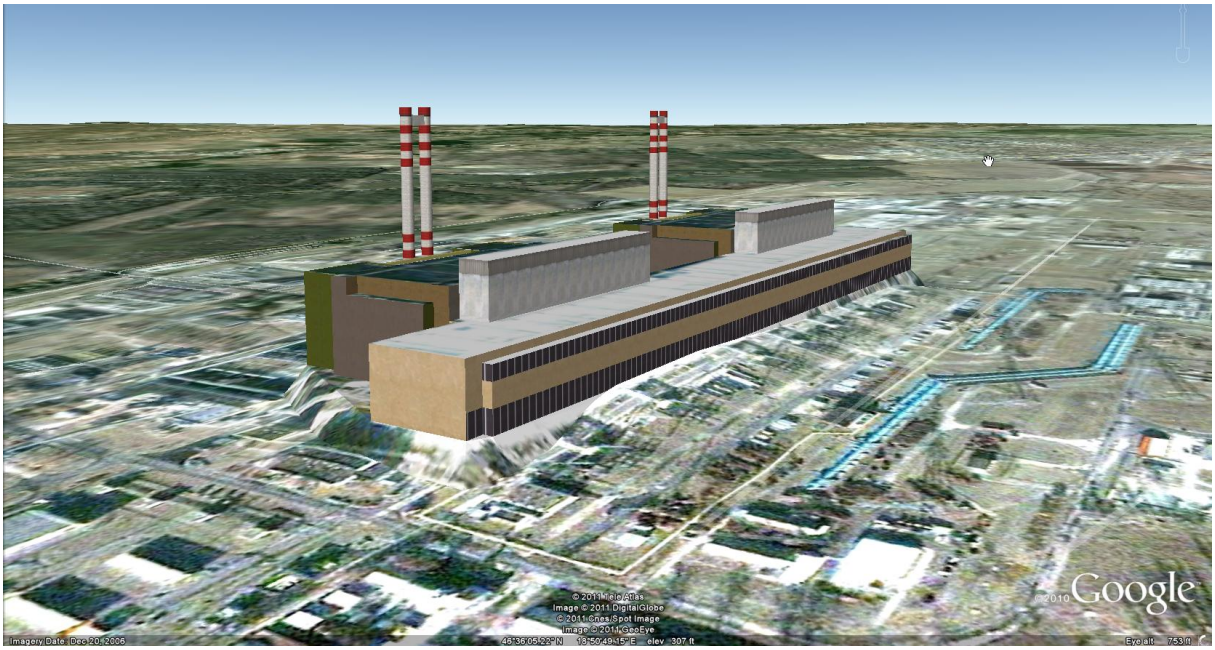

Figure 2: Photorealistic 3D rendering of Paks Nuclear Power Plant in Google Earth

Next, Wikimapia [14], a crowdsourced mapping service that allows users to digitize and annotate geographic features, was examined. Users have digitized building as infrastructure at the NPP site, including reactor housings 1 through 4, cooling water input and output systems, switchyard, control room building, visitors center, fire station, meteorological tower, and bus station, among others (Figure 3). This information can be extracted thorough Wikimapia's API in XML, JSON, KML, and binary formats.

The third geo-enabled search was through the GeoHack [15] website. GeoHack is a tool developed by of the Wikimedia community's Toolserver project that aggregates mapping services that are capable of displaying georeferenced content from many different sources. By querying a latitude and longitude coordinate pair, GeoHack returns links to various mapping services that display data centered on these coordinates as well as links to other web-based resources related to these coordinates and thus serves as a valuable jumping off point to a large amount of geospatial data. From here, a large number other websites containing geo-tagged information were discovered to include:

- 28 global map services sites (Google Maps, Wikimapia, OpenStreetMap, etc)
- 12 Wikipedia links
- 10 photo hosting websites
- 19 "other sites"
- Over 100 regional map services

While each of these sites do not necessarily represent unique data points as some links are coincident or contain identical data, this does illustrate the relative ease with which recent aerial and satellite imagery and geographic data visualizations are obtainable.

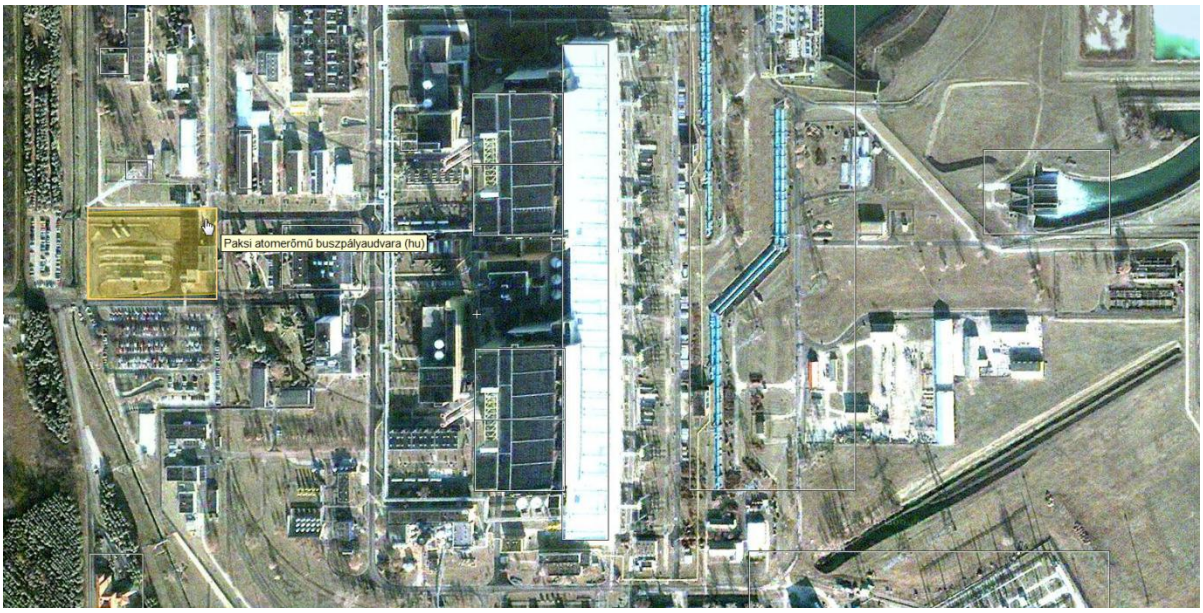

Figure 3: User digitized and annotated features of Paks NPP on Wikimapia.

2.3.3 Structured geospatial data search

Finally, an on-line search was conducted for standard structured geospatial data such as ESRI Shapefiles, digital elevation models (DEM), and GeoTIFF images. Effective use of data in most of these formats requires specialized GIS software (for example, ArcGIS or MapInfo) and a trained geospatial specialist. However, several XML-based geodata formats (for example, KML and GML) have emerged in recent years that allow these data to be used within a web-based computing framework and thus available to a larger number of analysts.

The quality and resolution of the GIS data discovered for the Paks NPP site ranged from very little to extremely high. While a large amount of data were discovered at state and regional scales, very little data were found at local and site-specific scales. For example, geospatial data for Hungary and Hungarian counties were abundant, while data for municipal scales and the Paks NPP site in particular were more difficult to come by. However, what one might consider “micro-level” geodata, such as geotagged photographs, were widely available. This trend might indicate the need to and benefit of examining other sources of geographically referenced data to supplement this mid-scale data void.

2.3.4 Results

Based on discussions with analysts, desirable geospatial information for safeguards and security analysis includes:

- Overhead aerial or satellite imagery
- Reference maps and images to provide context
- Reference information such as roads and other nearby geographic features
- Ground-based photographs
- Detailed site information
- GIS/map data to use in analysis

Each of these data types were discovered on the open Internet with relative ease. However, effective use of these data for analysis requires specialized training and expensive software tools that may not be widely available to analysts. Moreover, no tools exist (to the knowledge or the authors) that allow for the systematic detection, extraction and utilization of these data within a system that can be easily incorporated into the analysis workflow. Also, notably missing from this list are unstructured data (such as text data) containing geospatial references. Because these data are not easily used in a geospatial framework they are often ignored or overlooked. Future work will seek to enable analysts to detect,

extract, and utilize these data for use within a geospatial system in addition to these other data types by developing tools that are accessible to analysts and that integrate into workflows with minimal divergence from proven methods.

3 Ontology development

In order to systematically discover and integrate heterogeneous and unstructured data it is necessary to develop and apply a standardized definition of terms and relationships. Ontologies, formal specifications of terms and their relationships within a given knowledge domain [16], allow for the standardization of heterogeneous and unstructured data by defining the spatial, temporal, and thematic dimensions of the data. By applying computer-readable metadata based on these ontologies (semantic mark-up), it is possible to further automate detection and processing of these data by allowing computational reasoning about the geospatial and thematic relationships among data.

For this project, several ontologies are likely needed. First, a geospatial ontology will probably be necessary to define types of geospatial entities (for example, administrative districts, natural features, etc.) as well as geospatial relationships (for example, adjacency, proximity, containment, etc.). Second, a place names ontology will be necessary to standardize the identification of named geographic entities. Several existing place name ontologies exist in the form of online gazetteers, including GeoNames [17] and Yahoo! GeoPlanet [18]. Finally, thematic or domain ontologies are needed to define safeguards relevant terms and relationships. Examples might include ontologies for the nuclear fuel cycle and nuclear reactor or centrifuge facility operations. Domain ontologies are often specified using Web Ontology Language (OWL) definitions [19].

Existing semantic web specifications can also support the standardized detection and extraction of domain specific geospatial data by. Existing implementations of the semantic web, which seeks to supplement web resources with computer readable metadata, can support these functions, though the adaptation of these standards has only occurred within relatively narrow parts of the Internet.

The utilization of ontologies to increase the effectiveness of safeguards activities has been explored [20], although not within a geospatial context.

4 Conclusions

Using a phased search strategy that included general Internet search, geospatially-enabled search, and structured GIS data search, it is possible to assemble a basic geographically referenced set of data without a specialized GIS analyst or expensive GIS software.

Because of the ease of use and low life-cycle costs, the use of these open source tools to create a basic geospatially referenced data set has the potential to increase the use of geospatially referenced data in future safeguards analysis. When configured to work within an existing safeguards analysis workflow, these tools can allow analysts to efficiently and effectively utilize both structured and unstructured geospatial data from the open Internet, a capability that generally is available only to those with specialized training and expensive, proprietary tools.

Several no-cost, open source tools exist that lend themselves further development for use in safeguards specific analysis. Zotero, an open source reference management tool, for example, has the ability to produce a map from place names in a web-based document or collection of documents in seconds rather than the hours or days when using traditional GIS tools, all the while extracting, storing, and organizing references in a structured manner. OpenLayers and GeoServer are open source projects that allow for the handling, visualization, and sharing of geospatial data.

In order for the efficient use of this heterogeneous and unstructured geographically referenced material, more work needs to be done to create standardized and automated processes for discovering, integrating and organizing the data. Development and application of ontologies and semantic technologies will be necessary to achieve this goal.

Finally, while open-source data can be an important supply of new types of information for safeguards analysis, it must be approached with some caution. Open-source data, especially crowdsourced information can be inaccurate, incomplete, biased or even fabricated [21]. A future goal of this research is to develop tools and methodologies that provide safeguards analysts the ability to differentiate valid and reliable geospatial data from those data that cannot be trusted.

5 References

[1] "Twisst - tweeting ISS passes near you." http://www.twisst.nl/.

[2] J. Eisenstein, B. O'Connor, N. A. Smith, and E. P. Xing, "A latent variable model for geographic lexical variation," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pp. 1277–1287, 2010.

[3] M. F. Goodchild, "Citizens as sensors: the world of volunteered geography," *GeoJournal*, vol. 69, no. 4, pp. 211–221, 2007.

[4] R. R. Vatsavai et al., "Geospatial image mining for nuclear proliferation detection: Challenges and new opportunities," in *International Geoscience and Remote Sensing Symposium*, pp. 48–51, 2010.

[5] I. Niemeyer, S. Nussbaum, and M. J. Canty, "Automation of change detection procedures for nuclear safeguards-related monitoring purposes," in *International Geoscience and Remote Sensing Symposium*, vol. 3, p. 2133, 2005.

[6] R. Wallace, G. Anzelon, and J. Essner, "Safeguards information from open sources," *Journal of Nuclear Materials Management*, vol. 37, no. 4, pp. 30-40, 2009.

[7] "Zotero." http://www.zotero.org/.

[8] World Nuclear Assosication, "Plans for New Nuclear Reactors Worldwide." [Online]. Available: http://www.world-nuclear.org/info/inf17.html. [Accessed: 31-Mar-2011].

[9] "OpenLayers." http://openlayers.org/.

[10] "GeoServer." http://geoserver.org/display/GEOS/Welcome.

[11] "Open Geospatial Consortium." http://www.opengeospatial.org/

[12] "Paks Nuclear Power Plant," *Wikipeida*. [Online]. Available: http://en.wikipedia.org/wiki/Paks_Nuclear_Power_Plant. [Accessed: 04-Apr-2011].

[13] "Paks Nuclear Power Plant Ltd." http://paksnuclearpowerplant.com.

[14] "Wikimapia." [Online]. Available: http://wikimapia.org/#lat=35.1025&lon=-106.6117&z=10&l=0&m=b. [Accessed: 13-Apr-2011].

[15] "GeoHack - Paks Nuclear Power Plant." [Online]. Available: http://toolserver.org/~geohack/geohack.php?pagename=Paks_Nuclear_Power_Plant&params=46_34_21_N_18_51_15_E_region:HU_type:landmark. [Accessed: 13-Apr-2011].

[16] T. R. Gruber and others, "A translation approach to portable ontology specifications," *Knowledge acquisition*, vol. 5, pp. 199–199, 1993.

[17] "GeoNames." http://www.geonames.org/.

[18] "Yahoo! GeoPlanet™ - YDN." http://developer.yahoo.com/geo/geoplanet/.

[19] "OWL 2 Web Ontology Language Document Overview," *W3C Recommendation*. [Online]. Available: http://www.w3.org/TR/owl2-overview/. [Accessed: 21-Apr-2011].

[20] S. Labana, A. I. ElDesouky, A. S. ElHefnawy, and A. F. El-Gebaly, "Ontology Potentials in Increasing Effectiveness of Safeguards and Nuclear Weapon Non-Proliferation Activities," presented at the Symposium on International Safeguards: Preparing for Future Verification Challenges, Vienna, 2010.

[21] A. Berriman, R. Leslie, and J. Carlson, "Information Analysis for IAEA Safeguards," in *INMM 2004 symposium, Orlando*, 2004.