SAND2011-8305 C
SAND2011-8305C

Advances in the Use of Open-source Information

Back to the Future with Open-source Information and Tools

Karl Horak, Denise Bleakly, and Michael McDaniel
Sandia National Laboratories[1]
P.O. Box 5800, MS-1379, Albuquerque, NM, 87185
kehorak@sandia.gov, drbleak@sandia.gov, mmcdani@sandia.gov

Abstract

Great advances have been made in the application of open-source information for Safeguards. However, the open-source "ecosystem" is rapidly evolving. Recently, the combination of powerful smartphones and the World Wide Web have led to novel developments for mapping the impact of natural and man-made disasters. Researchers are learning to harness online volunteers as citizen scientists. Online social media such as Facebook, LinkedIn, and Twitter are becoming accepted channels of credible information. Geospatially aware applications (augmented reality apps) are capable of overlaying data upon a smartphone's viewscreen.

These and other developments will have (in fact, already have had) a significant effect on the collection, evaluation, structuring, analysis, and dissemination of safeguards-relevant information. In addition, the modern, open Internet is posing new security threats—not just viruses and their ilk, but threats based on digital traces left by browsing the open Web.

The authors present a case study of the disasterous toxic flood near Devecser, Hungary in October 2010 and then give recommendations for the future use of open-source information and software tools.

Introduction

Safeguards open-source information once meant labor-intensive, manual information processing and sophisticated, expensive, one-of-a-kind software systems [1] to deal with:

- Newspaper clipping services
- Hardcopy academic journals
- Industry and trade publications
- Traditional library research
- Paper maps, diagrams, and photographs
- Lots of photocopies, or
- Scanned versions of the same.

---

1 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. SAND2011nnnnX.

Now with the Internet (plus handy phones, digital cameras, personal GPS[2], 3 and 4-G networks, and wi-fi), it is a new world for information technology.

By way of an example, here is the story about Ushahidi[2]. In Ruwanda in 1994 warnings about the genocide were transmitted by fax. It never spread far enough to generate international assistance from the United Nations. In Kenya 2007, post-election violence similarly seemed about to spiral out of control. This time eyewitness reports were transmitted by e-mail and text messages. An enterprising group called Usahidi aggregated these messages and placed them on a Google map layer for the entire Internet to see. Statesmen noticed and the likes of Kofi Anan responded. The violence abated.

What was the difference between 1994 and 2007? Text messaging and Google Maps had been around for years, but the software environment had matured to the point that people could make connections between disparate datasets and provide an online, map-based visualization.

Open Source for Detecting Undeclared Nuclear Activities

By way of an example more closely related to the IAEA's mission, consider the Fukushima Daiici disaster. Citizens were concerned about gaps in the official radiation data. On their own, they constructed personal Geiger counters (using instructions from the Web) and made use of their private automobiles to sample large areas. This data has been uploaded to websites for aggregation and the results can be visualized in online maps and Google Earth[3].

This ability to create online maps in response to an event is termed "crisismapping." It's becoming common practice wherever there is an emergency, whether man-made (fraudulent elections) or natural (recent flooding in Vermont, Figure 1)[4].

The use of ordinary but motivated citizens to collect data refers to "citizen sensors." The use of private citizens to make expert observations in a scientific endeavor is called "citizen science" and it is an example of crowdsourcing.

A typical smartphone has six sensors (digital compass, magnetometer, barometer, thermometer, accelerometer, and gyroscope), an 8 MP camera, hi-def video, GPS, 32 MB of storage. With ease of integrating hardware components such as credit card readers with cell phones, it's conceivable that someone will devise a small, individual radiation detector for airborne radiation measurements that integrates easily with a smartphone.

---

2 Geographical Positioning System

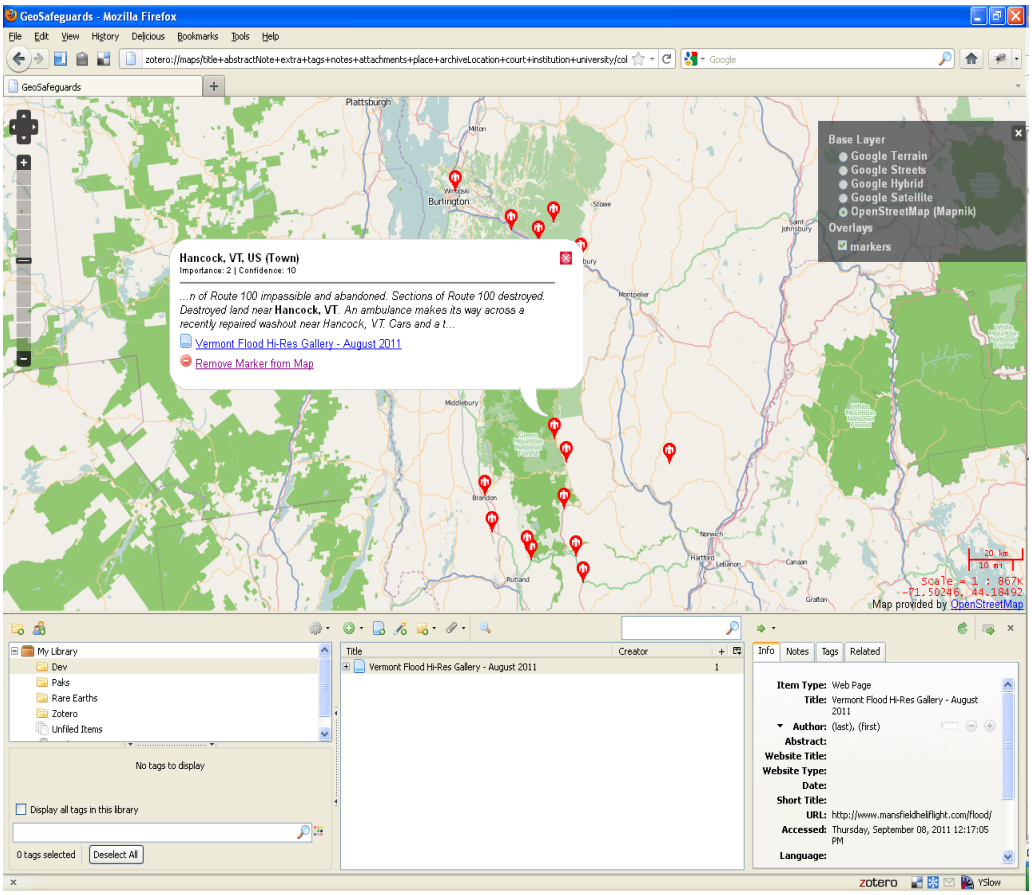

Figure 1: A Sandia-modified Zotero map based on a news article.

Crowdsourcing citizen scientists armed with mobile devices, formidable desktop computing power at home, and ubiquitous access to the modern Internet can now turn to services like Google Maps, Google Earth, Yahoo! Maps, or OpenStreetMaps to transmit and display their observations. This can be combined with datamining in publicly available online industry reports, journals, newspaper articles, and government documents.

In addition there is a large body of volunteered information on the Web in the form of social media like Facebook, Twitter, LinkedIn, Flickr, Foursquare, Delicious, and many others. These social media websites allow "geo-tagging" of content, for example, Flickr and Foursquare.

Because most modern cell phones have cameras and GPS, their photographs are automatically geo-referenced. There is now a large body of volunteered geospatial information in the open source. Some of these social media sites are providing safeguards-relevant information, either purposely or incidentally[5].

There is now more data than ever before and a faster pace of innovation than ever before. As Joe Lewis of Sandia National Laboratories has said, "I try to keep up with the pace of

technological change, but then I have to take a nap...." It's extremely difficult to hold one's position in the river of technology that's rushing by us.

Fortunately, there are open-source tools designed to help with processing open-source information. Open-source software is non-proprietary in the sense that it is licensed such that the source code is visible to users and developers.

Some open-source geospatial tools include:

- Zotero and Zotero Maps
- OpenStreetMaps
- Google Maps sketchups
- GEOnet Names Server
- GNIS (Geographic Names Information Service)
- GeoHacks

Leveraging technologies and developments outside the safeguards arena saves costs and reduces deployment times. Already open-source software is commonly used for computer operating systems (Linux), webservers (Apache), programming languages (Python), databases (Postgress, MySQL), and many other aspects of the modern computing ecosystem. Open-source information is used widely as well and has long been recognized as important to safeguards and the IAEA. Together these support the entire analysis and reporting process: optimizing information utilization, expanding and diversifying sources, enhancing evaluation and analysis, and improving capabilities to detect undeclared nuclear materials and activities. With regard to the future, let's point out some important trends in each area of information collection, evaluation, structuring, analysis, security, and dissemination.

Collection

Finding open-source information is no longer news aggregators, internal databases, subscription databases, automated search tools, web crawlers, and Member-State supplied information, but new social media search resources, meta-search tools, and geospatial tools.

We have found that open-source research can be broken down into three phases: general searches, geo-enabled searches, and searches for structured geospatial data.

General searches

These are typical web-based searches using Google, Yahoo!, Google Scholar, and so on. Depending on the information domain being searched, query capabilities vary.

Geo-enable searches

Based upon a place name, geographic point, or geographically defined bounding box, it is

now possible to search for information tied to that location. One can find news items, scholarly articles, government reports, blogs, tweets, Facebook posts, photographs, and videos all based upon geospatial references.

Searches for structured geospatial data

Additional effort is required, but information in specialized formats can be obtained relative to a location of interest. This may be CAD drawings, shape files, digital terrain maps, or other files. Typically one needs specialized software and training to make best use of this material. That said, some open-source tools, such as uDig, are very adept at handling a wide variety of data inputs.

Evaluation

Information needs to be assessed based on its volume, relevance, reliability, timeliness, and veracity. One also has to be concerned with its interoperability (its ability to work and play well with other software systems) as well as its internationalization features (its ability to be machine translated or imported into other understandable forms).

Tools are being developed to automate the framing of documents and compute values of "trust" for web-based data. Control of analytical data can be made with reliability settings for the information sources. Statistical comparisons can be made use of as well as ground-truthing by experts. Systems can be set up to allow duplicate scoring of analyses by numerous independent evaluators. Crowdsourcing and even online games can be applied to difficult analytical problems[6].

Some metrics for evaluating the accuracy and utility of information from social media include:

Metric	Description
Blog/microblog biography	Self-supplied information regarding background, interests, and other experience. Could well be fabricated totally or in part.
Number of posts	How many and how frequently updates are made by this authority.
Number of followers	How many people subscribe to the blog. Does not guarantee that people are reading it.
Number followed	How many people the author subscribes to.
Tracebacks/retweets	Postings from other blogs/microblogs that refer back to the one in question.
Location	Physical location of the author. Sometimes automated and highly accurate. Other times self-supplied and possibly fabricated.
Timing	The time stamp on the post. Usually generated by a server and highly accurate.
Social authentication	Reputation within an online community. Sometimes generated by algorithms, such as Technorati.
Triangulation	Obtaining the same information from different

	sources. Comparing reports from different angles.
Origins	Cited origin of the information in the blog.
Language	Self-evident
Photographs	Are images included that document the posting?
Media authentication	Is there outside authentication by mainstream media?
Engage the source	Contact the author directly either by e-mail or by comments in the blog.
Follow up	Take action to have Agency resources used to verify observations and reports in a blog.

Structure

Data must have some digital structure in order for it to be useful. Four activities are needed to provide the structural elements that give data its enduring value.

Organize	To enable the use of industry-standard meta-data
	To allow continued use of existing topic trees and ontologies
Store and Retrieve	To enable subsequent searches by the widest possible number of engines for the largest number of users
Re-purpose	To enable easy sharing of data between applications
	To allow conclusions to be drawn from the widest possible collections of datasets
Future-proof	To enable continuity of use in the face of inevitable upgrades
	To allow legacy data to be migrated to new applications

Analyze

The central activity undertaken to impart meaning to data is analysis. The nature of the data being analyzed, its geospatial context, the analyst's behavior, and the emergent methods of visualization will all impact how analysis is carried out in the future.

Data	Everyone will become a data producer as well as an on-site data consumer.
Geospatial Context	Information will be displayed relevant to the user's current time and place.
User Behavior	Systems will track user location and activity as the context for data interaction.
Visualization	Applications will be able to respond to the command, "Tell me about the surrounding area" with graphs, maps, tables, and augmented reality (AR).

Figure 2 is an example of the view from one of the authors' window looking out towards the Innovation Parkway Office Complex (IPOC) as seen using a Droid X with the Layar augmented reality browser. AR not only correctly identifies the building and its associated

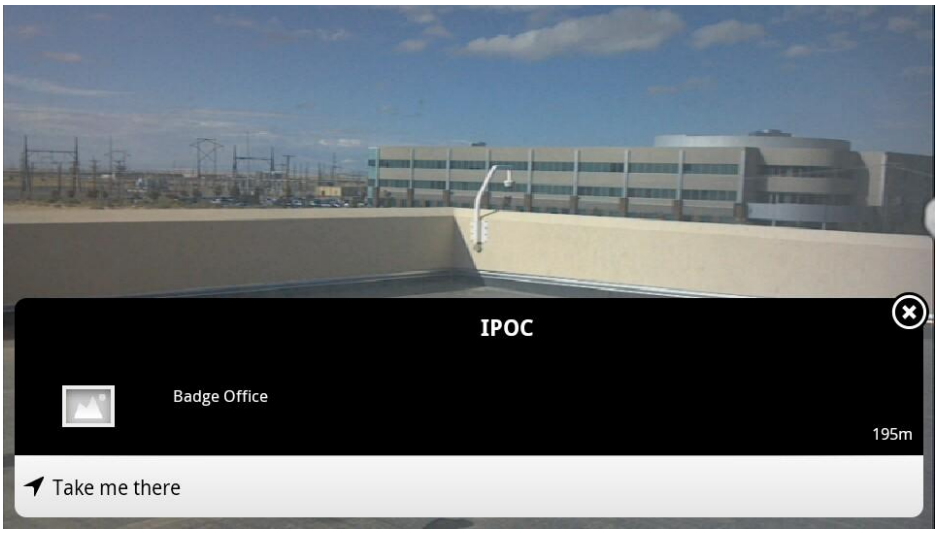

Figure 2: An example of augmented reality of a building at Sandia National Laboratories

function (Sandia's Badge Office), but also specifies the distance and can display an overhead map view as well.

Tools

Open-source analytical tools include geospatial analysis software such as:

- Yahoo!Placemaker
- GeoNames
- OpenLayers
- Zotero Maps
- uDig

Some geospatial information of particular interst to safeguards analysts might be:

Coordinates of the given location	GIS Data
Aerial / Satellite Images	Roads
Other maps of site	Fencelines
Other pictures of the site (Flickr, etc.)	Gates
Topographic or Elevation data	Utilities
CAD Data (or AutoCAD data)	Building Footprints / building names / builiding numbers
Building schematics	Environmental information
Site schematics	Other plannimetric information of interest
Other CAD data layers	What do the "anti-nuclear" sites say?

Emerging open-source technologies also permit time-series animation of maps, geo-fencing, and geo-rectification. Time series analyis is best known today by animated weather radar

maps, but other subject matter (for example, a spreading power failure) is amenable to the technique. Geo-fencing is the ability of a map-based system to provide e-mail or text message alerts when an event is reported within a specified geographical perimeter. Geo-rectification aligns, rotates, and skews an oblique overhead image to align with a geographically correct coordinate system.

Crowdsourcing

"Crowdsourcing is the act of taking a job traditionally performed by a designated agent (usually an employee) and outsourcing it to an undefined, generally large group of people in the form of an open call."

National Geographic, in their "Valley of the Khans project," is using Internet-based volunteers to classify features in Mongolian aerial photography[7]. NASA's Zooniverse[8] provides amateur scientists the opportunity to make genuine discoveries in a variety of areas. An online game, FoldIt, harnesses the power of Internet volunteers to find solutions to computationally intractable problems in protein folding[9]. They recently made headlines when a team of FoldIt workers solved the structure of a key enzyme involved with HIV infection[6]. The solution had eluded scientists for ten years but the "winning" group found an optimal structure in just three weeks.

Even the U.S. Government posts frequent online challenges to encourage private groups and companies to work on some of the most vexing problems of our time[10]. Perhaps the best known example is DARPA's Grand Challenge for autonomous vehicles.

A well-known crowdsourcing engine is Amazon's "Mechanical Turk," named after the 19th Century chess-playing device. When finally debunked, it was learned that a human chess player was hidden inside the Mechanical Turk box. Amazon uses its network reach and extensive computing power to provide a means for human workers anywhere on the Web to carry out tasks posed by paying customers, likewise anywhere on the Web[11].

As one can imagine, this amorphous crowd needs expert guidance and a goal-oriented perspective in order to be productive and useful. Careful consideration needs to be given for means of maintaining quality in such a environment.

Secure

Quite frankly, security needs to be considered throughout the entire ope-source life cycle. A large number of open-source tools exist that cover the entire realm of computer and network cyber-security. The table below lists ten commonly used open-source software packages that could play a role in an information architecture.

Software	**Purpose**

Nessus	Vulnerability scanner
Snort	Intrusion detection
Nagios	Network monitor
SpamAssassin	Spam filter
ClamAV	Anti-viral tool
Open SSL	Secure sockets
OpenSSH	Secure shell
Nmap	Network mapper
Ossec HIDS	Host-based network intrusion detection
Wireshark	Network protocol analyzer

However, even more important in the current context is the issue of “digital footprints” and “browser fingerprints.” Each browser request to the World Wide Web includes a large amount of information in the request header that tells the responding website how to reply. Enough information is present to uniquely identify a particular computer's browser among millions of others[12].

This browser fingerprint can be discovered in web servers' logs and analytical software can then piece together the search terms, websites, and other aspects of online behavior. This information comprises one's digital footprint and conceivably it could tip off an adversary that the Agency is researching a particular facet of a State's nuclear profile. Particularly for online research involving sensitive issues, it may be prudent to use computers that are specially configured for anonymous browsing.

Disseminate

After one has collected, structured, evaluated, and analyzed a group of meaningful data, it becomes essential to share that information with others so that they can look at it through the lenses of their experience and apply expert judgment to make important decisions. Dissemination turns out to be a tricky problem when it comes to opening information channels while at the same time enforcing need-to-know restrictions.

Safeguards has a robust and well-functioning system for processing public information. SharePoint is their document management system and GES is soon to be their channel for sharing aerial and satellite imagery[13]. However, what is missing is an information “ecosystem” that can add value to data that is collected and analyzed from open sources.

One solution is to implement social media systems behind Agency firewalls; systems that permit sharing of annotations, comments, and insights among authorized staff. These “organic,” self-organizing systems might be in-house wikis, blogs, microblogs, or something else. At any event they foster an environment that enhances a number of key activities.

- Finding rich data sources
- Working with large volumes of data despite hardware, software, and bandwidth constraints
- Cleaning the data and making sure that data is consistent
- Melding multiple datasets together
- Visualizing that data
- Building rich tooling that enables others to work with data effectively

In addition, data itself must become "smart." Systems must use metadata, ontologies, and the semantic web to be able to deal with:

- Vastness
- Vagueness
- Uncertainty
- Inconsistency
- Deceit

A Case Study

By way of an example[3], let's look at the disastrous toxic flood in Devecser, Hungary last year. Following our protocol, first we carry out a general web search for "Hungarian toxic flood 2010" in which we learn that there were:

- 621,000 results for everything related to the search terms,
- 68,000 blog posts on the topic,
- 533 homepages that mentioned the event,
- 569 twitter mentions that were found, and
- 43,300 images available.

In the second step, a geospatial search using GeoHack, Zotero Maps, and Google Maps with the Photographs layer turned on brings to light a tremendous number of relevant resources, maps, and images. Figure 3 shows how one can drill down from a news item found by online searching, to a Zotero Map, to a Google Maps view with Panaramino images, to an individual photograph.

Finally, in the third step, we search for specialized, structured geospatial data and find that large scale, structured GIS data is not common for this area. However, this varies widely from country to country. The authors' research did located the Hungarian national geospatial archives. Figure 4 shows an aerial image of the flooding geo-rectified atop a Google Maps street layer. Normally this is an advanced capability of commercial GIS tools, but it is now available as a free online service via MetaCarta.[4]

---

3 A non-nuclear one, just to avoid sensitivity issues.

4 http://labs.metacarta.com/rectifier/rectify/

Conclusions

Information, technologies, and other developments external to the safeguards domain increase available data, have the potential to significantly reduce costs, and cut deployment times.

The open-source information and software ecosystem is self-sustaining and growing[14]. With an estimated 5 billion smartphones expected to be in use within the next decade, open-source information will increase exponentially (at least in the short term). The numbers of mobile Internet users continues to rise, the inter-connectivity of the globe increases, and the

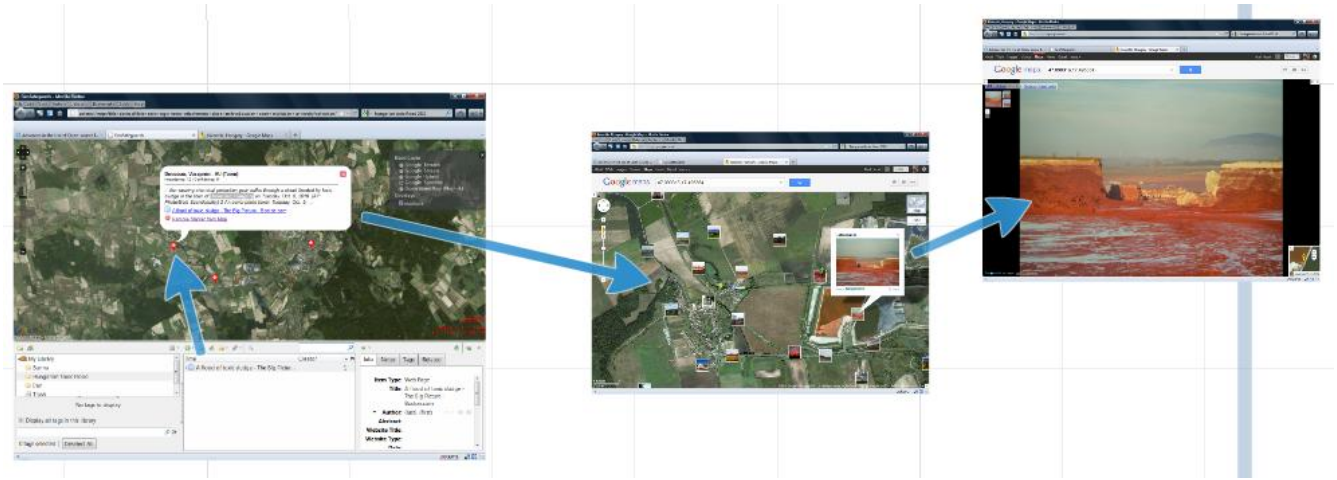

Figure 3: Drilling down through open-source geospatial information.

ability to share information continues to be enhanced.

Figure 4: Geo-rectified aerial image of Hungarian flooding atop a Google Maps layer, courtesy of MetaCarta.

Recommendations

1. Where proprietary, commercial systems are already in place, open-source solutions are probably not the optimal fit, at least in the short term.
2. But innovative and novel features can not be overlooked.
3. Dedicated teams need to be assigned to monitor new and emergent information sources, software systems, and tools.
4. But don't overlook the synergies to be generated by serendipitous events.

Open-source software continues to play an important role in 21st Century information technology. Recall that key components of the Internet are open source, that many motivational factors make open source a viable, long-term solution, and open-source solutions are out-pacing our ability to stay abreast of developments.

Open-source information continues to grow at a remarkable rate, largely due to the popularity of mobile devices: smartphones and tablet computers. Armed with personal GPS, high quality video and still cameras, almost ubiquitous Internet access, and many other features, private citizens and their smartphones will generate an expanding body of information available online.

References

[1] "Inspector_and_Analyst_in_Headquarters.pdf." Zaruki, R. and C. Norman, Safeguards Symposium 2010.

[2] "Erik Hersman on reporting crisis via texting | Video on TED.com." [Online]. Available: http://www.ted.com/talks/erik_hersman_on_reporting_crisis_via_texting.html. [Accessed: 28-Sep-2011].

[3] "Pachube?:: blog: Real-Time Radiation Monitoring in Japan - Internet of Things in Action." [Online]. Available: http://blog.pachube.com/2011/03/real-time-radiation-monitoring-in-japan.html. [Accessed: 28-Sep-2011].

[4] "Open sourcing the post-Irene Vermont flood relief effort." [Online]. Available: http://vtdigger.org/2011/09/16/open-sourcing-the-post-irene-vermont-flood-relief-effort/. [Accessed: 28-Sep-2011].

[5] "The Spy Who Tweeted Me: Intelligence Community Wants to Monitor Social Media | Danger Room | Wired.com." [Online]. Available: http://www.wired.com/dangerroom/2011/09/social-media-spies/. [Accessed: 28-Sep-2011].

[6] "Protein-Folding as a Computer Game - Brainiac." [Online]. Available: http://www.boston.com/bostonglobe/ideas/brainiac/2011/09/protein-folding.html. [Accessed: 28-Sep-2011].

[7] "Map Explorer | Field Expedition: Mongolia, National Geographic." [Online]. Available: http://exploration.nationalgeographic.com/mongolia. [Accessed: 28-Sep-2011].

[8] "Zooniverse - Projects." [Online]. Available: http://www.zooniverse.org/projects. [Accessed: 28-Sep-2011].

[9] "The Science Behind Foldit | Foldit." [Online]. Available: http://fold.it/portal/info/science. [Accessed: 28-Sep-2011].

[10] "The central platform for crowdsourcing US Government challenges, contests, competitions and open innovation prizes | Challenge.gov." [Online]. Available: http://challenge.gov/. [Accessed: 28-Sep-2011].

[11] "Amazon Mechanical Turk - Welcome." [Online]. Available: https://www.mturk.com/mturk/welcome. [Accessed: 28-Sep-2011].

[12] "Who am I?" [Online]. Available: http://proxify.com/whoami/. [Accessed: 28-Sep-2011].

[13] "Research and Development Programme for Nuclear Verification 2010–2011." IAEA.

[14] "Software and Community in the Early 21st Century - YouTube." [Online]. Available: http://www.youtube.com/watch?v=NorfgQlEJv8. [Accessed: 28-Sep-2011].