SCIENCE & DIPLOMACY

A quarterly publication from the AAAS Center for Science Diplomacy

Maynard Holliday and Chris Holden, "Advanced Web-Based Temporal Analytics for Arms Control Verification and Compliance," Science & Diplomacy, Vol. 3, No. 3 (September 2014*). http://www.sciencediplomacy.org/article/2014/advanced-web-based-temporal-analytics-for-arms-control-verification-and-compliance.

*The complete issue will be posted in September 2014.

Advanced Web-Based Temporal Analytics for Arms Control Verification and Compliance

Maynard Holliday and Chris Holden

Traditional monitoring of arms control treaties, agreements, and commitments has required the use of National Technical Means (NTM)—large satellites, phased array radars, and other technological solutions. NTM was a good solution when the treaties focused on large items for observation, such as missile silos or nuclear test facilities. As the targets of interest have shrunk by orders of magnitude, the need for other, more ubiquitous, sensor capabilities has increased. The rise in web-based, or cloud-based, analytic capabilities will have a significant influence on the future of arms control monitoring and the role of citizen involvement.

Rose Gottemoeller, the U.S. under secretary of state for arms control and international security, noted that weapons of mass destruction verification and monitoring could be enhanced by the incorporation of new, information-age tools that can harness the power of the crowd to generate data that can then be analyzed. However, she stated explicitly that "the goal of using open source information technology and social networks should be to add to our existing arms control monitoring and verification capabilities, not to supersede them."[1]

Since 1999, the U.S. Department of State has had at its disposal the Key Verification Assets Fund (V Fund), which was established by Congress. The Fund helps preserve critical verification assets and promotes the development

Maynard Holliday is a researcher at Sandia National Laboratories.

Chris Holden is the manager of customer success at Recorded Future.

of new technologies that support the verification of and compliance with arms control, nonproliferation, and disarmament requirements. The V Fund program concentrates on the following arms control treaties and nonproliferation categories:

- missiles/warheads/space/strategic treaties
- comprehensive test ban and nuclear test ban treaties
- nuclear nonproliferation treaty/fissile material cutoff
- nuclear arms reduction treaties
- chemical weapons convention
- biological weapons and toxins convention
- open skies treaty
- arms control in the information age/social media

Sponsored by the V Fund to advance web-based analytic capabilities, Sandia National Laboratories, in collaboration with Recorded Future (RF), synthesized open-source data streams from a wide variety of traditional and nontraditional web sources in multiple languages along with topical texts and articles on national security policy to determine the efficacy of monitoring chemical and biological arms control agreements and compliance. The team used novel technology involving linguistic algorithms to extract temporal signals from unstructured text and organize that unstructured text into a multidimensional structure for analysis. In doing so, the algorithm identifies the underlying associations between entities and events across documents and sources over time. Using this capability, the team analyzed several events that could serve as analogs to treaty noncompliance, technical breakout, or an intentional attack. These events included the H7N9 bird flu outbreak in China,[2] the Shanghai pig die-off,[3] and the fungal meningitis outbreak in the United States last year.[4]

These analog events were all unique, and investigating each highlighted different issues that could also be broadly applied to disease outbreak tracking. Although all of these events were natural or accidental, the analysis was able to show East/West contrasts in social media and Internet reporting.

This research is not the first attempt to model or monitor for disease outbreaks using information available on the web. A scoping review published last year[5] reported thirty-two studies published between 2002 and 2011 that used web data for disease surveillance. Several of those studies leveraged patterns in web searches to anticipate disease outbreaks, which differ from the RF approach of analyzing human-reported content from news, blogs, and social media.

Organizations such as HealthMap and Crowdbreaks are doing related work in disease monitoring by combining news analytics on select sets of mainstream and social media with crowd sourced, geo-located disease reporting. These efforts primarily focus on aggregating reports and issuing alerts regarding the incidence of diseases during a recent timeframe.

The current work undertaken by Sandia and RF breaks ground in several areas: the scope of media sources being measured for indications of disease outbreaks; the temporal dimension of text analysis used to timeline key events during an outbreak; and the source metadata analysis allowing a novel understanding of patterns and anomalies in disease reporting by language, media type, and geography.

Introducing Capabilities for Large-Scale Analysis of the Web

Leveraging data from the web has long required hours of brute force information scraping and data wrangling or the use of niche information sets. Cobbling together a useful and workable data set from disparate media sources generally proves difficult and painstaking because of the information's unstructured nature. Policy analysts will likely recognize the unpleasantness of piecing together shifts in political statements or going back through historical records to timeline the series of small-scale events leading up to a major political event.

The current research process leverages linguistic algorithms tuned to recognize indications of disease outbreak to reduce the manual information gathering elements of web data analysis. This scalable technique provides researchers with an index of web information updated in real time and suitable for use in a variety of analytic platforms.

Unlike teams of consultants reading social media or article after article from scientific publications, this system mines quantifiable intelligence from the open web in real time. This includes, as of this writing, more than 720,000 disease outbreak events and more than 15,000 nuclear or radiological material transactions reported over more than five years.

This data set grows continuously as additional material is published on the web. As new information is published, a variety of linguistic analysis techniques are used to analyze, catalogue, and offer up the newly structured data visually for researchers or as a data feed for developers and statisticians.

How does all of this relate to weapons proliferation, bioterrorism, and disease monitoring?

Information about disease incidences is reported constantly across many publicly available sources on the web. Illness and reporting on chemical exposure show up particularly prominently in local news and social media channels, but information might also be posted in a blog by health researchers, summarized in a report by the Centers for Disease Control and Prevention, or detailed by organizations such as Healthmap, which is based at Boston Children's Hospital.

Currently, RF captures and analyzes the contents of more than 450,000 sources. Researchers can leverage this massive information set to quickly aggregate and query across real-time summaries on particular disease outbreaks, symptom recognitions, or successful treatments.

For example, an aid worker in Kuala Lumpur tweeting that she contracted H7N9 this week will place a disease outbreak event involving H7N9 in the capital city of Malaysia at that time point. Organizing reporting on such events at a massive scale results in data that is ideal for mapping open source reporting on disease outbreaks. Researchers can pose and answer questions: How did the Chinese media portray the H7N9 scare during 2013? How did fungal meningitis spread between U.S. states over time? What was the social media response to the Huangpu Shanghai pig die-off?

Data Structure

The software platform constantly mines text ingested from web sources for contextual clues about what is being reported. As new information is collected, the system automatically identifies *specific events* (natural disasters, disease outbreaks, arms trade, company acquisitions, political protests, etc.), *entities* (persons, organizations, places, disease types, malware, etc.) related to these events, and the *time point(s)* when that event is/was reported to happen. There are five major component blocks to generating usable web intelligence data:

1. Harvesting—in which text documents are retrieved from various sources on the web and stored in the text database.
2. Linguistic analysis—in which texts are analyzed to detect events and entities, time and location, etc. This is the step that takes us from the text domain to the data domain.
3. Refinement—in which data is analyzed to obtain more information; this includes calculating the momentum of entities, events, documents and sources, synonym detection, and ontology analysis.
4. Data analysis—in which different statistical and artificial intelligence based models are applied to the data to detect anomalies.
5. User experience—the different user interfaces to the system, including the web interface, the RF alert mechanism, and the application programming interface for other systems.

These results enable what is described above as temporal analytics. Linguistic algorithms detect what actual calendar time a text is referring to when an event, such as a reported disease incidence, is described. This can be both absolute times ("9:37AM, September 11, 2001," "at the end of 2012") and relative times ("three weeks ago," "tomorrow"). Information about the publication time of a document combined with linguistic analysis is used to map all events to a calendar time interval.

Combining temporal analysis with ontological information about the world (who is the leader of a certain country, in which country is a certain city located, etc.) and the contextual information about a given information source (where

is it being published, who is the author, etc.) gives a structured view of what is happening in the world at any given point in time.

Generic Hardware Configuration

Sandia provisioned a stand-alone instance of RF's analytical tool. This allowed an instance that analyzed and displayed data not only from the open web index but from documents that were proxies for internal documents. These proxy sources contained everything from long-ago published magazine articles to scientific studies to long-form books. Examples of proxy sources include the Kathleen Vogel book, "Phantom Menace or Looming Danger?: A New Framework for Assessing Bioweapons Threats," the David Hoffman book, "The Dead Hand: The Untold Story of the Cold War Arms Race and its Dangerous Legacy," and the J. C. Hymans *Foreign Affairs* article, "Botching the Bomb: Why Nuclear Weapons Programs Often Fail on Their Own—and Why Iran's Might, Too."

In terms of hardware, Sandia's stand-alone instance consisted of three virtual machines, each with eight central processing unit cores and 64GB of RAM per machine. It is important to note that this form factor can scale upwards depending on the speed required to harvest data and the size of the data to be indexed.

Tailored Source Material

The system complemented the already vast number of websites that RF indexes with subject specific sites and publications germane to the weapons of mass destruction proliferation queries that are of interest. For example, the system included Jeffrey Lewis's blog, *Arms Control Wonk*; the United Nations' Biological and Toxin Weapons Convention Implementation Support Unit disarmament page; the Center for Arms Control & Non-Proliferation website; the James Martin Center for Nonproliferation Studies; and many others.

Temporal Analytic Results

The Sandia RF system did queries on events that were proxies for potential chemical or biological attacks, such as the H7N9 flu outbreak in China, the fungal meningitis outbreak in the United States, and the Middle East Respiratory Syndrome (MERS) outbreak. It also looked at outlier events such as the Huangpu River pig die-off. The analysis revealed many interesting facts and insights that otherwise may not have been revealed using other techniques.

H7N9

For H7N9 we found that open source social media were first to report the outbreak and give ongoing updates. Chinese sources noted that H7N9 was different from H1N1 because mortality was over a larger age distribution, (i.e., young and

old contracted the virus and died). The Sandia RF system was able to roughly estimate lethality based on temporal hospitalization and fatality reporting.

Additional reporting showed that humans did not contract it as easily as H1N1, and there was no evidence of human-to-human transmission. The system also monitored domestic Chinese as well as international preparedness measures taken as the virus spread and was more widely reported. Japan was noted as having the most transparent response.

These observations lent additional support to the idea that tracking social media in closed non-democratic societies can provide insight into sensitive issues (disease, political unrest, disaster response, etc.) that would otherwise go unreported or underreported through government sources.

U.S. Fungal Meningitis Outbreak

The analysis of the U.S. health system response to this outbreak demonstrated that if this was not a reportable disease and had no central agency tracking the cases, a social media analysis would show the distribution and rapid increase in deaths to initiate a national investigation (e.g., skin rashes, eye irritations, gastrointestinal distress). This reporting is a digital analog to the canary in a coalmine. We were also able to track public health messaging and actually pinpoint cases at the city and county levels through local blogs and postings about support group formation.

MERS Open Source Analysis

A retrospective open source analysis showed that authorities were unable to identify MERS as a syndrome until about a year after its outbreak, which raises questions concerning why it was not identified earlier. Possible reasons could have been that linking of like symptoms was lacking and that the availability of a laboratory test to type the coronavirus was slow to be developed. Further research in the region is required to confirm any of this. Reporting showed that MERS was much less infectious than SARS and that it was localized to the Middle East unlike SARS, which spread worldwide. Unverified concerns about exposure and transmission were observed through blog posts. Authorities localized the virus reservoir to the indigenous bat population, but only after a long investigation. They are still looking for other host species (e.g., domesticated animals). Further studies show that some variant of MERS has been prevalent in camels for about twenty years. This raises questions concerning the jump to the local human population: Why now? Did the virus mutate to infect humans? What changes in practice occurred that may have precipitated this?

Huangpu River Pig Die-Off

The analysis tracked the rapid assessment by Chinese authorities that H7N9 was not the cause of the pig die-off as had been originally speculated. Open source reporting highlighted a reduced market for pork in China due to the very public

dead pig display in Shanghai. Possible downstream health effects were predicted (e.g., contaminated water supply and other overall food ecosystem concerns). In addition, legitimate U.S. food security concerns were raised based on the Chinese purchase of the largest U.S. pork producer (Smithfield) because of a fear of potential import of tainted pork into the United States.

Improving Web Intelligence for Arms Control Compliance

The research detailed in this paper describes an initial effort at applying emerging web intelligence technology to monitoring and analysis of arms control compliance. This analysis included bio surveillance through open source data streams. Given the initial results, several currently available capabilities for researchers can enhance the available dataset.

The source set being analyzed can be significantly expanded. This is an ongoing process often driven by a particular research scope. However, this remains an area that can result in dramatic improvement of information relevancy by locale or subject.

Additional materials for collection would include more public social media that matches a particular set of key terms related to diseases within a scope of study. It is likely that inclusion of more local discussion boards or public health communities that have not been analyzed yet would also significantly enhance the local data set that could be investigated for early signals of outbreaks.

Sandia and RF are continuing to refine and improve their analyses by constantly adding more web sources and upgrading the user interfaces that display results. The hope is to increase the user base in the analyst communities within Sandia and elsewhere by making the data displays more intuitive. Further socialization of this capability with the greater analyst community is something being done both internally, at Sandia, and externally.

Work is also being done to enhance researchers' abilities to add their own documents and mash-up more structured data sets alongside the purely, public data set on a rolling basis. This would enrich the customized analytic capability.

Additionally, Sandia and RF are creating a custom event extractor for disease symptoms that would serve as a complement to the existing disease outbreak data structure. This would likely prove useful for early identification of emerging or resurgent diseases as well as serve as early warning of possible chemical and biological weapons activity. SD

Endnotes

1. Rose Gottemoeller, "Mobilizing Ingenuity To Strengthen Global Security," (remarks, South by Southwest Conference, Austin, TX, March 8, 2013), http://www.state.gov/t/us/197056.htm.

2. "H7N9 Timeline Based on Early Reporting," Recorded Future, https://www.recordedfuture.com/live/sc/6HSsFFBOxrP5.
3. "2013 Huangpu River Incident," Recorded Future, https://www.recordedfuture.com/live/sc/3qeOwFAWW0P7.
4. "Fungal Meningitis Outbreak Based on Early Reporting," Recorded Future, https://www.recordedfuture.com/live/sc/g13sjQe2v5Cw.
5. Theresa Marie Bernardo, Andrijana Rajic, Ian Young, Katie Robiadek, Mai T Pham, Julie A Funk, "Scoping Review on Search Queries and Social Media for Disease Surveillance: A Chronology of Innovation," *Journal of Medical Internet Research* 15, no. 7 (2013): e147, http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3785982.

New Mexico's 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.