University of Nebraska - Lincoln
DigitalCommons@University of Nebraska - Lincoln

Library Philosophy and Practice (e-journal) | Libraries at University of Nebraska-Lincoln

2-27-2014

Effectiveness of Metadata Information and Tools Applied to National Security

Cassidy Pham
San Jose State University, cassidypham@hotmail.com

Follow this and additional works at: http://digitalcommons.unl.edu/libphilprac

Part of the Library and Information Science Commons

Pham, Cassidy, "Effectiveness of Metadata Information and Tools Applied to National Security" (2014). *Library Philosophy and Practice (e-journal)*. Paper 1077.
http://digitalcommons.unl.edu/libphilprac/1077

Library Philosophy and Practice ISSN 1522-0222

Effectiveness of Metadata Information and Tools Applied to National Security

Cassidy Pham

Introduction

Recent events, in particular, the ongoing Snowden affair, has increased debate about government surveillance within the country and abroad. Among the most common concerns is the infringement of privacy in terms of metadata information. Though there is ample debate concerning the legal and political issues, few, if any, discussions are concerned with the effectiveness of metadata used by the intelligence community (IC). From smartphones to Facebook profiles, the IC, with its latest tools, can collect a substantial amount of information in the form of metadata. With the application of metadata as part of the national security apparatus, it has greatly enhanced the capability of the IC to retrieve and analyze information in the 21st century.

Statement of the Problem

Research is needed to examine the effectiveness of the metadata-related tools and applications in order to determine its usefulness in national security. With the legal and political ramifications in the government's sweeping surveillance of metadata, and the roughly 67 billion dollars budgeted for U.S. intelligence in a period where Americans are less willing to pay for costly government programs, it is imperative for the public to better understand how metadata is used as an intelligence gathering tool as well as how effective it is. The limitation or complete absence of public awareness may unnecessarily limit the IC's ability to effectively use metadata for national security concerns.

Literature Review

Leaked government documents, concerning the software and metadata scheme of surveillance in the form of the National Security Agency's (NSA) XKeyscore will be explained. Various applications of the metadata generated by XKeyscore, and similar programs will be discussed using leaked government documents, declassified documents, scholarly journals, and books. In addition, the interoperability among local, state, federal, and international agencies within the IC using more or less the same type of sources.

Though the actual structure of the metadata has not been made available to the public, training materials in the form of a PowerPoint presentation released by The Guardian provides a rough overview of the possible metadata elements in the National Security Agency's (NSA) so called XKeyscore. This program is essentially a metadata generator that extracts and converts data into separate indexes, or metadata tags. Below is a table from one of the slides that showcases the type of information that XKeyscore extracts and indexes.

Plug-in	Description
E-mail Addresses	Indexes every E-mail address seen in a session by both username and domain.
Extracted Files	Indexes every file seen in a session by both filename and extension
Full Log	Indexes every file seen in a session by both filename and extension.
HTTP Parser	Indexes every DNI session collected. Data is indexed by the standard N-tupple (IP, Port, Casenotation, etc.)
Phone Number	Indexes every phone number seen in a session (e.g. address book entries or signature block)
User Activity	Indexes the Webmail and Chat activity to include username, buddylist, machine specific cookies, etc.

Figure 1: XKeyscore Presentation: Plug-ins (The Guardian, 2013).

It is apparent from the table above and commentary by Glen Greenwald that IP and e-mail addresses are among the items extracted and indexed by XKeyscore (Greenwald, 2013). Web files such as word documents and html files, and HTTP activities consisting of internet browsing and communications are also targeted (Greenwald, 2013). Though the government has yet to confirm the existence of XKeyscore, a recent declassified document does verify some aspects of the metadata collected by the government. According to a document by the Foreign Intelligence Surveillance Court, it is confirmed that the U.S. government does collect "telephony metadata", which according to the document, includes:

> comprehensive communications routing information, including but not limited to session identifying information (e.g., originating and terminating telephone number, International Mobile Subscriber Identity (IMSI) number, International Mobile station Equipment Identity (IMEI) number, etc.), trunk identifier, telephone calling card numbers, and time and duration of call (Foreign Intelligence Surveillance Court, 2013).

T

The wide-range of information that is extracted by XKeyscore has many useful applications in terms of national security. According to one of the slides, analysts can track a German-speaking person located in Pakistan by accessing documents with tags defined by the country of origin, the HTML language, and various other metadata under the "User Activity" plug-in (The Guardian, 2013). Though a good portion of the information is redacted, one particular slide showcases how the metadata is structured. Below is a metadata structure taken from a supposed "HTTP Parser" plug-in.

GET /search?hl=eng&q-islambad&meta= HTTP/1.0

Accept: image/gif, image/x-xbitmap, image/jpeg, image/pjpeg, application v/nds.ms-application/msword, application/x-shockwave-flash, */*

Referer: http://www.google.com/pk/

Accept-Language: en-US

User-Agent: Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1)

Host: www.google.com.pk (The Guardian, 2013).

It is apparent that XKeyscore has four notable metadata elements in this particular plug-in, "Accept-Language", "User-Agent", "Referer" and "Host"; though in this example, "Referer" and "Host" come from the same source. These elements tell the analyst that the item content of the information uses the standard U.S. English language code, and it was based on the Mozilla/4.0 browser scheme. In addition, the source of the item came from Google.com, specifically, the Pakistani host server.

Other sources suggest metadata information from XKeyscore and similar programs can be used on other relevant tools. For instance, in a leaked training manual from a PowerPoint slide provided by the New York Times, the SYANPSE Data Model creates visual graphics of metadata information (New York Times, 2013). In this case, the phone and e-mail metadata is used to display a social network diagram that depicts the relationships between people associated with the person of interest. In a research study, a software similar to SYANPSE Data Model called iMiner was used to graph the hierarchy of covert terrorist networks by determining the association and frequency of metadata, in particular, individuals, organizations, places and events (Memon & Larsen, n.d.). Metadata for this particular software is imported from other legacy systems or data sources formatted in XML or CVS (Memon & Larsen, n.d.).

In terms of interoperability, books provide substantial information in regards to intelligence sharing between domestic intelligence agencies. These sources tend to suggest inconsistencies and incoherence among the federal agencies, especially before 9/11. Differences in legal, policy, technology, process, and culture among each IC member stalled collaborative efforts (Lee, 2013, p. 25). After the events of 9/11, the intelligence agencies took great lengths to work closely with one another. The failure to predict the 9/11 attacks were due in large part by the lack of coordination between the IC (Johnson, 2012, p. 429). In response, the IC undertook an aggressive policy to address and overcome barriers, in particular, issues in technology and process. This transformation so to speak resulted in the acceleration of information sharing and the associated technologies between the various members of the IC (Lee, p. 25). As of today, the IC consists of 16 major members, such as the Central Intelligence Agency (CIA), the Defense

I

Intelligence Agency, the Federal Bureau of Investigation, the NSA, the NGA, and the intelligence units of the Army, Navy, Air Force, and Marines (Johnson, p. 8).

The integration of these various IC members has yielded successful results of interoperability in regards to data-sharing. In a case study by Keith Cozine (2013), the CIA, NGA, and NSA collaborated closely with one another in a clandestine operation to kill Osama Bin Laden. The NSA intercepted a phone call that possibly revealed the whereabouts of Bin Laden (Cozine, p. 83). How the call was intercepted remains undisclosed in this case study; nevertheless, tools similar to XKeyscore may have been used to collect and index the intercepted call, which suggest a possible role of metadata in terms of telephone surveillance. The CIA, for its part, setup a safe house near the Abbottabad compound, the supposed hideout of Bin Laden. They logged activities, and collected DNA samples from Bin Laden's children by organizing a fake vaccine drive (Cozine, p. 84-85). The NGA collected and analyzed geospatial imagery of the site as well as imagery of the person of interest (Cozine, p. 85). Metadata in the form of GIS certainly played a key role in this particular operation. Geospatial data collected by the three agencies, such as weather reports, maps of Abbottabad and the compound, and even information on the phases of the moon were shared and analyzed in order to determine the best method to launch an assault (Cozine, p. 86). The collaborative effort undertaken by these three agencies showcases a vast improvement in interoperability since the person of interest did turn out to be Bin Laden.

Due to the lack of detail in the data-sharing apparatuses between the IC community, it is difficult to ascertain how exactly information is shared. Nevertheless, interoperability between the federal, state, and local agencies is quite clear in a number of sources. In a PowerPoint

presentation shown at the GIS for Local Government Conference, a slide outlines the Department of Homeland Security's (DHS) adoption of the Federal Geographic Committee (FDGC) and the National Spatial Data Infrastructure (NSDI) standards (Vanderheyden, 2005). These standards, according to the FDGC, "promote the coordinated development, use, sharing, and dissemination of geospatial data on a national basis" (Vanderheyden, 2005). The DHS consists of 22 different federal departments and agencies, such as the Transportation Security Administration, Nuclear Incident Response Team, and the Federal Emergency Management Agency (U.S. Department of Homeland Security, n.d.). As part of the DHS, these federal departments and agencies likely use the same GIS standards in their data collection and sharing processes. At the state and local level, a slide reveals that the DHS is connected to 87,000 entities (Vanderheyden, 2005). It is expected that these entities use the same GIS standards, or at the minimum, a metadata structure that can be applied to the standards of the FGDC and NSDI. The DHS is also a major intelligence IC member, which has its own intelligence analyst unit (Johnson, p. 7). The DHS's integration in the IC community may arguably suggest that FGDC standards are used to collect and share geospatial data among the 16 major members.

Information shared between the U.S. IC and its foreign counterparts is well-known by now, especially after recent leaks that put such activities in the public spotlight. Revelations, such as the recent news of U.S.-Australian surveillance operations in Asian countries, provide a glimpse to the interoperability within the IC at the international level (British Broadcasting Corporation, 2013). U.S. intelligence sharing with foreign counterparts was largely due out of necessity during the Second World War (National Security Agency, n.d.). The signing of the BRUSA Agreement in 1946, known today as the UKUSA, signaled increase partnership between

t

the intelligence apparatuses of the U.S. and U.K. throughout the 20th century, and well into the 21st century (National Security Agency, n.d.). As a result of 9/11, additional measures were taken to enhance standardization practices between the U.S. and U.K. by adopting an "integrated collaborative planning, based on the maintenance of a common operating picture and common intelligence inputs" (Svendsen, 2013, p. 36). This partnership cultivated in unified coordination between intelligence services that monitors and translates large quantities of foreign media and newswire as well as news agency output, which is arranged in geographical and thematic products (Svendsen, p. 25).

Leaked documents concerning the XKeyscore provide an example to the interoperability between between the U.S. and U.K. intelligence agencies in terms of metadata. As mentioned earlier, XKeyscore is a possible metadata extraction and indexing tool for the NSA. Below each PowerPoint presentation slide of XKeyscore provided by The Guardian is a notation labeled "TOP SECRET//COMINT//REL TO USA, AUS, CAN, GBR, NZL" (The Guardian, 2013). It is apparent from the notation that the U.S. is in close collaboration with the U.K. and various other countries. In this case, the countries of Australia, Canada, and New Zealand. A declassified NSA document confirms the working relationship between the intelligence apparatuses of all five nations mentioned above as part of the UKUSA Agreement (NSA, 1955). Since the PowerPoint slide appears to be intended as an instructional presentation, it is plausible that the mentioned countries use the same software for their intelligence gathering services. Ultimately, this may translate into a high level of interoperability on metadata standards and tools in association with national security.

Research Question

This research paper will propose a single question:

1. How effective is metadata applied to national security?

Methodology

A qualitative research method will be applied using two different experiments: determining the success rate by identifying associates of a participant; and determining the success rate of identifying hobbies, personal background, and preferences of a participant. These experiments will involve a total of ten (10) participants. A metadata social network diagram called Wolfram Alpha will be extensively used throughout the experiment. This metadata tool is capable of extracting metadata information from an individual's Facebook profile. The data, such as a person's social network, commentary, imagery, and various other online activities are translated and visualized in a graphical diagram.

The evidence is not an exact representation since the tools, such as XKeyscore and SYANPSE, are not used in the research. The government has yet to declassify and publically release these tools. Nonetheless, a close representation from commercial and academic sourced tools, such as the Wolfram Alpha, can be achieved based on the descriptions and capabilities of the tools identified in the leaked documents as shown in the literature review.

A key part of the selection process for participants was to avoid predetermine knowledge between the participant and observer. Two observers, one for each experiment, participated in to insure the avoidance of predetermine knowledge. These participants have no relationship with

t

the observers since the study requires that the observer analyze the data without any background knowledge of the participant. With permission from participants, the observer uses the Wolfram Alpha to extract and visualize the metadata from their Facebook profile.

In the first experiment, the observer will identify and categorize the participants' contacts into groups using the social network diagram. The observer designates different social groups by examining the connection between contacts. As shown in Figure 2, four distinct groups are discovered: high school friends, college friends and friends with similar hobbies, co-workers, and family. Each colored circle represents an individual contact, and each line between two contacts indicates they are associates. Separate groups can be identified based on the color and the lack of a connection with other groups as seen in the circle labeled "Family" in Figure 2.

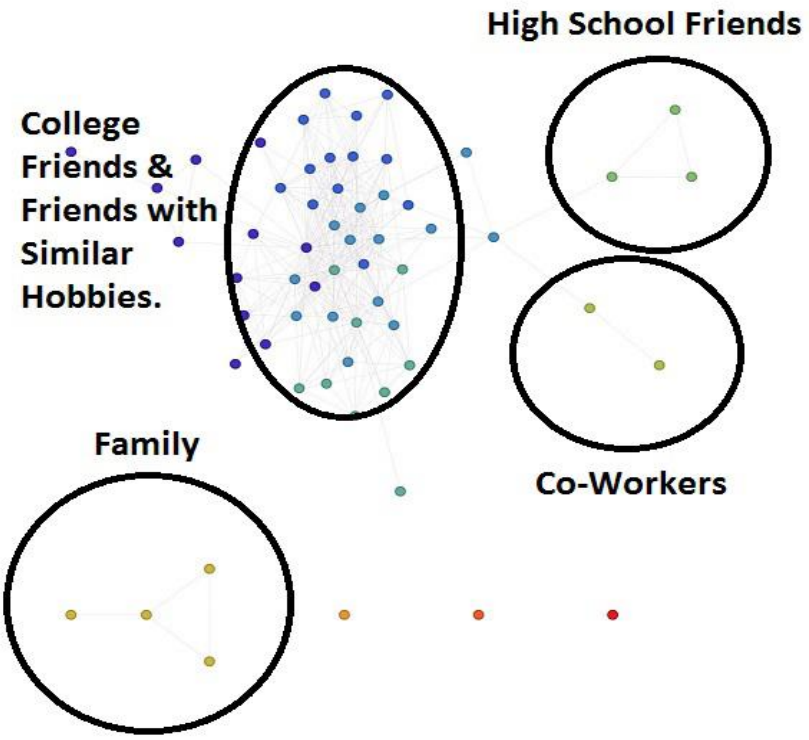

Figure 2. Wolfram Alpha: Social Network Diagram (Wolfram Alpha, 2013).

E

examination of metadata text of participant's contacts is also used to determine social groups. This information, as shown on Figure 3, is procured by clicking on any colored circle, which displays basic information of associated contacts, such as age, gender, first and last name, where the person goes to school, and where the person lives.

Figure 3. Wolfram Alpha: Contact Profile (Wolfram Alpha, 2013). [Note: Name of contact omitted for privacy].

For the second experiment, the observer will identify participants' preferences and/or hobbies using the "Word Cloud" function, which visually displays the most used words posted by the participant in Facebook. As shown on Figure 4, the size of the text depicts how often the word is used compared to other words. The observer will analyze the frequency and relationship of each word to determine the participant's background information, such as hobbies and beliefs. For example, the words "representatives" is a frequently used word based on its size. Its association with words, such as "advocate" and "vote" suggests to the observer that the participant is politically active. Unlike the previous experiment, the observer will not have access to the profile of the participants' contacts. This method prevents the observer from learning in advance any background information of the participant.

Figure 4. Wolfram Alpha: Word Cloud (Wolfram Alpha, 2013).

The experiments are followed by online interviews between the observer and the participants. The interviews confirm if the analysis undertaken by the observer is correct, and if not, participants provide clarification for any discrepancies. These discrepancies and other factors not considered during the experiment will be presented in the discussions segment of this research paper.

The result, which determines the effectiveness of metadata in a national security platform, is based on whether social contact groups and personal background information can be successfully identified by the observer through these two experiments.

Results

The results from the first experiments show that social groups can be successfully identified with metadata, in the form of a social network diagram. These unique social groups are often isolated or have few connections with other groups, such as how the "Family" group in Figure 2 is completely isolated from the other three social groups. Thus, it is relatively straightforward when identifying social groups with the aide of metadata text from the profiles of the participants' contacts. However, the contact profiles often lack information related to hobbies and interests. As a result, some groups were misidentified. As shown from Table 4 below, for participant 3, 4, and 6, a social group was incorrectly identified; however, for the rest of the participants, all social groups were correctly identified by the observer.

Table 1

Social Network Table

Participant #	# of Identified Social Groups	# of Social Groups Correctly Identified
1	4	4
2	3	3
3	4	3
4	4	3
5	3	3
6	3	2
7	3	3
8	2	2
9	4	4
10	3	3

Table 1: Social Network Table

There are mixed results from the second experiment. Out of the 10 participants, only 6 participants had identifiable social backgrounds. Factors, such as low posting activity, and minimal use of nouns made it difficult to determine a participant's background information. Nevertheless, from the 6 participants, assumptions of their personal background could be made with a relatively high degree of accuracy based on the frequency and pattern of the words. Two incorrect assumptions were made with participant 5 as it became evident after the interview that the participant works for public transit, and owns a vehicle. And it became apparent that participant 3 was not an alcoholic. The social backgrounds for 4 of the participants were correctly identified out of the 6 participants.

Table 2

Word Cloud Table

Participant #	**Words**	**Background(s)**	**# of Correctly Identified Background(s)**
1	PS3, computer, games, XP, iPhone, iTunes	(1) Tech-savvy (2) Gamer	2
3	Barcardi, rum, vodka, liquor, class, school, professor, assignment	(1) Alcoholic (2) College student	1
4	Representative, vote, advocate, school, teacher, faculty, academic	(1) Politically active (2) Student	2
5	Bus, light rail, VTA, station, school, student, university, SJSU	(1) Uses public transit (2) Does not own a car (3) College student (4) Goes to San Jose State University	2
7	Eat, food, restaurant, brother, UCSC	(1) Frequently eats out (2) Has a brother (3) Goes to UC Santa Cruz	3
9	Sequester, government, Republican, tea, party, Obama, Ted, Cruz, president, Obamacare, support, vote	(1) Politically active	1

Note: Participant #s reflect the same participants from Table 1.

Discussion

As pointed out earlier, the data from the experiments do not show an exact representation of the metadata tools used by the IC. Nonetheless, Wolfram Alpha is much like the IC's SYANPSE, which displays a social network diagram based on metadata information, such as contact information from cellular devices and e-mail. The similar features and functions of the two metadata tools do suggest a level of credibility. The observers were largely inexperienced, especially in comparison to fully-trained intelligence analysts belonging to the IC. However, their ability to successfully—to a certain degree—identify social network groups and personal background information without any professional training is impressive.

It should be noted that the observer for the second experiment did not have the luxury of using the profiles of the participants' contacts. Normally, analysts examine in conjunction with other sources to affirm the validity of the information. In other words, they cross-reference between multiple sources. In addition, second and third degree contacts are typically examined to better understand the context surrounding the target in question and his or her associates (Johnson, 2012). This study only examined second degree contacts.

With this wealth of metadata, and the tools to organize and graphically visualize the information, the IC can apply sophisticated analytics techniques to identify persons of interests and associates. The first experiment is applicable as a national security tool since it efficiently connects contacts, and isolates social groups. This is particularly useful in disassociating contacts that are unaware or uninvolved with the person interest. It is also valuable in finding unknown associates of targets he or she already knows about. The second experiments provide additional

contextual information. Naturally, metadata explains the who, what, where and when. It does not explain the why and how. These questions are left for analysts to fill in by using contextual information, such world cloud experiment, to develop a so-called picture. Though the experiments were innately limited in scope, the relative success in the application of metadata shows its effectiveness as a national security tool.

As noted in the literature review, metadata tools, such as the XKeyscore are truly invaluable as intelligence gathering tools. According to leaked documents, over 300 terrorists were captured using the XKeyscore (The Guardian, 2013). Also, examples from various case studies, such as the killing of Osama Bid Laden indicate successful use and application of metadata. With the added benefit of the interoperability of these various tools, the IC can lighten the burden and share information. Rather than having one agency find a needle in a haystack—a haystack of infinite size, it is far more efficient and effective to divide the hay into multiple stacks among multiple players.

Conclusion

Much of the information and sources are conjecture since they are based on leaked documents. And of course, the government has yet to fully disclose the information on the tools, which exasperates the problem. Nonetheless, declassified documents, journal articles, and metadata tools that are relatively similar to the ones used by the IC, insure legitimacy to the evidence. Overall, the results from the experiments indicate a high success rate since untrained observers were able to analyze the metadata diagrams, and accurately determine social groups and personal backgrounds for most of the participants. Evidence from various sources, such as

case studies, journal articles, and leaked government documents further support the effectiveness of metadata as part of a national security platform. As the country, and the rest of the world becomes more dependent on smart devices, social media sites, the internet, and other 21st century necessities, metadata and the associated tools are equally necessary for the IC to effectively face the threats of national security.

References

British Broadcasting Corporation. (2013, November 1). Australia ambassador summoned amid Asia US spying reports. Retrieved November 1, 2013 from http://www.bbc.co.uk/news

Cozine, Keith. (2013). Teaching the intelligence process: The killing of Bin Laden as a case study. *Journal of Strategic Security*, 6(5), 80-87. Retrieved from http://scholar.google.com

Foreign Intelligence Surveillance Court. (2013, April 25). *In re application of the Federal Bureau of Investigation for an order requiring the production of tangible things from* [PDF Document] (Docket No. BR 13-80). Retrieved October 26, 2013 from http://www.dni.gov/files/documents/PrimaryOrder_Collection_215.pdf

Greenwald, Glen. (2013, July 31). XKeyscore: NSA tool collects 'nearly everything a user does on the internet'. *The Guardian*. Retrieved October 26, 2013 from http://www.theguardian.com/world/2013/jul/31/nsa-top-secret-program-online-data/print

Johnson, L. K. (Eds.). (2012). *The Oxford handbook of national security intelligence*. New York, NY: Oxford University Press.

Lee

, Newton. (2013). *Counterterrorism and cybersecurity: Total information awareness*. New York, NY: Springer.

Memon, N., & Larsen H. L. (n.d.). Investigative data mining toolkit: A software prototype for visualizing, analyzing, and destabilizing terrorist networks. [PDF Document]. Retrieved October 25, 2013 from http://ftp.rta.nato.int/public/PubFullText/RTO/MP/RTO-MP-IST-063/MP-IST-063-14.pdf

National Security Agency. (n.d.). UKUSA Agreement release 1940-1956. Retrieved October 31, 2013 from http://www.nsa.gov/public_info/declass/ukusa.shtml

NSA (1955, May 10). *Amendment no. 4 to the appendices to the UKUSA Agreement*. [PDF Document]. Retrieved October 30, 2013 from http://www.nsa.gov/public_info/_files/ukusa/new_ukusa_agree_10may55.pdf

New York Times. (2013, September 28). Documents on N.S.A. efforts to diagram social networks of U.S. citizens. Retrieved October 25, 2013 from http://www.nytimes.com/interactive/2013/09/29/us/documents-on-nsa-efforts-to-diagram-social-networks-of-us-citizens.html

Svendsen, Adam. (2013). *Intelligence cooperation and the War on Terror: Anglo-American security relations after 9/11*. New York, NY: Routledge

The Guardian. (2013, July 31). XKeyscore presentation from 2008 – read in full. Retrieved October 24, 2013 from http://www.theguardian.com/world/interactive/2013/jul/31/nsa-xkeyscore-program-full-presentation

U. S. Department of Homeland Security. (n.d.). Who joined DHS. Retrieved October 30, 2013 from http://www.dhs.gov/who-joined-dhs

Vanderheyden, Chris. (2005, October 17). Local GIS for homeland security. [PDF Document]. Retrieved October 28, 2013 from http://gisconference.cas.psu.edu/2005/proceedings/1_mon_1100.pdf

Wolfram Alpha (2013). Face profiles and metadata diagrams. Retrieved November 26, 2013 from http://www.wolframalpha.com