Metadata – both shallow and deep: the fraught key to big data mass state surveillance

A report prepared for the Key issues in Big Data Surveillance: Research Workshop
May 12-14 2016, Queen's University, Kingston, Ontario

Andrew Clement andrew.clement@utoronto.ca
Jillian Harkness jillianvictoriaharkness@gmail.com
George Raine george.w.raine@gmail.com

Faculty of Information
University of Toronto
May 11, 2016

This is a draft paper for workshop discussion. We invite feedback, but please don't cite without first contacting the authors.

Abstract

This report examines the significance of metadata for big data analytics, especially as conducted by the Five Eyes security surveillance agencies. We compare the definitions and operationalization of metadata in academic, professional, legal and popular discourses with both official public claims about metadata by the heads of these agencies and that revealed through the published secret documents leaked by Edward Snowden.

Metadata plays a central role in state surveillance operations because it is amenable to automated processing and when analysed in large volumes from a wide range of sources promises to be highly revealing about an individual's behaviour, movements, beliefs, associations and affiliations. We find that what these agencies regard in practice as metadata is much more expansive and potentially intrusive than is generally understood. In their eyes, metadata effectively includes any data that can be algorithmically derived from the content of communication or other forms of expression. This raises serious questions about whether these government agencies are adequately protecting citizens personal privacy as well as whether their public statements can be trusted. We recommend that the protection of metadata be raised to the same legal standard as that of communication content and that the Five Eyes state security surveillance agencies be more transparent and accountable in their handling of personal information.

Contents

Abstract 1
Introduction 2
Conventional meanings of metadata 3
Role of metadata in big data analytics 5
Metadata from the perspective of the Five Eyes 6
Metadata in the NSA 8
Metadata in GCHQ 12
Metadata in the CSE 12
Findings 14
'Deep' versus 'shallow' metadata 14
Researching with the Snowden Archive 14
Implications (for discussion) 15
For surveillance researchers 15
For those calling for state surveillance accountability and reform 16
References 16

Introduction

> "We kill people based on metadata,"
> former head of the National Security Agency Gen. Michael Hayden[1]

Until recently, metadata - literally "data about data" - was unfamiliar outside the information fields, where it is a term of art used in describing a wide variety of information objects to enable cataloguing and retrieval. However, since Edward Snowden's 2013 whistleblowing exposures of the extraordinary scope of secret mass state surveillance and the key role that metadata plays in such surveillance, the term has now entered public discourse. We are learning that with the application of big data collection and analysis techniques, metadata can be much more revealing of an individual's behaviour, movements, beliefs, associations and affiliations than previously understood. In the hands of state actors, metadata can provide the basis of highly intrusive intervention into people's lives. While the coming into prominence of metadata draws attention to a relatively new and potent mode of surveillance practice, distinct from the popular imaginary of eavesdropping or surreptitious reading of communications, it remains an ambiguous and contested term. In particular, officials who lead state security agencies often deploy the term publicly in ways evidently designed to reassure audiences skeptical of their behind-the-scenes surveillance activities but which given its ambiguity also raise serious questions about their actual practices and trustworthiness. This is significant because metadata enjoys lower legal and constitutional protection

[1] http://abcnews.go.com/blogs/headlines/2014/05/ex-nsa-chief-we-kill-people-based-on-metadata/

than communication content. Furthermore, recurring discrepancies between what security and law enforcement agencies say about metadata and what they actually do with it call into question the adequacy of democratic governance of these powerful and sometimes necessarily secret arms of the state.

This report seeks to clarify the multiple meanings of metadata as defined and operationalized in various settings. More specifically it seeks to determine whether the practices of the Five Eyes signals intelligence agencies, notably the U.S. National Security Agency (NSA), the U.K. Government Communications Headquarters (GCHQ) and Canada's Communications Security Establishment (CSE), in relation to metadata conform to their public statements as well as the conventional and legal understandings of the term. To set the stage, we summarize the various definitions of metadata (and its equivalents) as they have appeared in academic, professional, legal and popular discourses as well as in the emerging field of of big data analytics. The core of the report draws on our study of the more than 500 secret documents of the Five Eyes alliance that the media has published based on Edward Snowden's leak, as found in the Snowden Digital Surveillance Archive.[2] Using the Archive's various search and index features, we selected and analysed documents for what they reveal about how the security agencies actually generate metadata from their global communication interception apparatus and subsequently organize, access and use it in their intelligence operations. To this research authors Raine and Harkness bring their professional archives background and interest in Metadata as well as their experience in designing and building the Archive. They are both familiar with the documents individually and the corpus as a whole. Clement brings his longstanding research and advocacy interests in surveillance and privacy. The report concludes by identifying implications for action research and policy advocacy around big data surveillance, especially in relation to state security agencies.

Conventional meanings of metadata

Ever since the Snowden revelations prompted journalists to explain the technical aspects behind mass surveillance, an understanding of the term metadata has, for the average citizen, been largely shaped by the news media. These definitions are varied; often they are simplified accounts echoing those put forth by governmental offices, such as a recent definition from the *CBC*, where metadata is defined as "information associated with communication that is used to identify, describe or route information" and more specially as possibly including "personal information, including phone numbers or email addresses, but not the content of emails or recordings of phone calls".[3] More elaborate and insightful definitions have also sometimes been offered; in 2013, *The Guardian* explained that "[m]etadata provides a record of almost anything a user does online, from browsing history – such as map searches and websites visited – to account

[2] https://snowdenarchive.cjfe.org
[3] Burke, "'Difficult to determine' scope of privacy beach...," 2016

details, email activity, and even some account passwords. This can be used to build a detailed picture of an individual's life."[4]

Within legal frameworks, at least in the UK, US and Canada, the term metadata is rarely, if ever used. The concept of metadata within the law developed historically out of the ways in which the law has approached communications technology, in so far as this technology facilitates the production of personal information that may or may not be subject to privacy laws. How metadata is viewed in Canadian law is still being interpreted in court.[5]

Part VI of the *Criminal Code* states that:

> private communication means any oral communication, or any telecommunication, that is made by an originator who is in Canada or is intended by the originator to be received by a person who is in Canada and that is made under circumstances in which it is reasonable for the originator to expect that it will not be intercepted by any person other than the person intended by the originator to receive it…"[6]

Whether or not metadata can be defined as private communications has been discussed in Canadian courts. Commentators such as Craig Forcese and the Office of the Privacy Commissioner refer to recent court cases as well as the Supreme Court interpretation of the Criminal Code, which protects "any derivative of that communication that would convey its substance or meaning."[7] As a derivative of a communication, metadata has been shown to provide significant details about the meaning of a communication[8] and "may permit the drawing of inferences about an individual's conduct or activities."[9] This definition focuses, as in the media definitions, on the surface or contextual information about a communication but acknowledges the potential sensitivity of this information when analysed and linked with other available data. Forcese reads the 2014 Supreme Court decision *R v. Spencer* as the "clear authority" that some forms of metadata, in this case "the name, address, and telephone number of a customer associated with an IP address," can be used to reveal details about privates lives and should be protected as personal information.[10]

Laws in the UK and US do not use the term metadata either, but focus more specifically on the technical aspects of the type of communications information. In the UK, the *Regulation of Investigatory Powers Act (RIPA)* in part defines "communications data" as including "(a) any traffic data comprised in or attached to a communication (whether by the sender or otherwise) for the purposes of any postal service or telecommunication system by means of which it is being or may be transmitted... (b) any information which

[4] Ball, "NSA stored metadata of millions…," 2013
[5] Office of the Privacy Commissioner, 2014, p. 9
[6] Criminal Code. R.S.C., 1985, c. C-46. VI.183
[7] As quoted in Forcese, 2015, p. 137
[8] Ibid. p. 137, 148
[9] Office of the Privacy Commissioner, 2014, p. 10
[10] Ibid. p. 148

includes none of the contents of a communication (apart from any information falling within paragraph (a)).”11 In the US, one example of metadata, “Call Detail Records” is legally defined “as session identifying information (including an originating or terminating telephone number, an International Mobile Subscriber Identity number, or an International Mobile Station Equipment Identity number), a telephone calling card number, or the time or duration of a call” and “*[e]xcludes from such definition*: (1) the contents of any communication; (2) the name, address, or financial information of a subscriber or customer; or (3) cell site location or global positioning system information.”[12] (*emphasis added)*

The differences in these varying approaches reveals how metadata can imply very different things across varying communities of practice. In her *Introduction to Metadata 3.0*, Anne J. Gilliland introduces the “widely used but frequently underspecified term” within the framework of the archival discipline.[13] She notes that the term originated with data management and today in practice metadata is generally, “the sum total of what one can say about any information object at any level of aggregation.” An information object can vary from a film or book to email or phone call; therefore metadata in this definition suggests anything one could say about these items, from a title to any salient feature of the contents. For archivists and information managers, metadata reflects an information object’s content, context and structure, and enables preservation as well as “intellectual and physical access.”[14] Despite at times recognizing that communications metadata may reveal a significant amount of personal information, media and legal definitions of metadata tend to limit their focus to the specific types of information that can be read from the ‘surface’ of the information object without delving into the object’s content. By contrast, the archival definition of metadata, as put forth by Gilliland, acknowledges that varying levels of aggregation and detail, as well as relationships between information objects and systems, may impact how one defines metadata as opposed to data, or context as opposed to content.[15] Gilliland notes that these “distinctions… can often be very fluid and may depend on how one wishes to use a certain information object.”[16]

Role of metadata in big data analytics

Questions of how metadata is used, and by whom, are often missing from the more simplified account of metadata collection. Critics of these popular and governmental definitions note the dismissal of metadata’s importance to big data analysis,[17] which takes advantage of the ease with which this kind of structured data can be processed

[11] *RIPA*, Section 21.4
[12] *US Freedom Act*, Summary. Sec. 101
[13] Gilliland, 2008, p. 1
[14] Gilliland, 2008, p. 2
[15] Gilliland, 2008. p.14
[16] Gilliland, 2008, p. 14-15
[17] Lyon, 2014, p. 3, 10

automatically to link very large collections of data elements and find patterns in order to create meaning and draw conclusions.[18] Proponents of big data analysis claim metadata enables efficient "distill[ing] terabytes of low-value data…to… a single bit of high-value data."[19] Through the development of mobile communications technology, an environment has emerged in which ordinary users, often without realizing, produce large amounts of metadata on a daily basis.[20] Access to this mass of personal data, when analysed through big data analytical techniques and software, allows for broad and deep access to personal information. Arguably, this access has been downplayed through the conventional meanings of metadata summarized above, to the benefit of both corporate business practices and surveillance agencies.[21]

Metadata from the perspective of the Five Eyes

The Five Eyes, frequently abbreviated FVEY or 5VEY in their internal documents, is an intelligence alliance among the U.S., U.K., Canada, Australia and New Zealand dating back to World War II. Each of these countries maintain several intelligence agencies that participate in the alliance, but it is those focused on signals intelligence (SIGINT) that concern us here, respectively: National Security Agency (NSA), Government Communications Headquarters (GCHQ), Communications Security Establishment (CSE),[22] Australian Signals Directorate (ASD), and Government Communications Security Bureau (GCSB). During the Cold War, FVEY developed a globe spanning signal interception capability known as ECHELON.[23] Initially targeted at the Soviet Union, whistleblowers[24] and journalists[25] have recently revealed that alliance, led by the NSA, has greatly expanded this network into a comprehensive surveillance apparatus of extraordinary scope and domestic penetration, albeit of highly questionable efficacy in its principal stated mission to aid in the 'War on Terror.'[26]

The classified internal documents that Edward Snowden leaked to journalists in June 2013 and subsequently published in major news media[27] have offered an unprecedented

[18] Sagiroglu and Sinanc, 2013, p. 43; Fisher et. al ,2012, p. 53
[19] Fisher et. al, 2012, p. 50
[20] Lyon, 2014, p.3; Gilliland, 2008, p. 8
[21] Lyon, 2014, p. 10; Laprise, 2016, p. 208, 214
[22] Also formerly referred to as the Communications Security Establishment Canada (CSEC). It is this now unofficial name and acronym that appears most frequently in the Snowden documents.
[23] In the mid-1970s, "the very existence of GCHQ and the [worldwide US/UK] Sigint network were then closely guarded secrets." https://theintercept.com/2015/08/03/life-unmasking-british-eavesdroppers/
[24] Notably Mark Klein, William Binney, Thomas Drake, Edward Snowden.
[25] Notably James Bamford, James Risen, Eric Lichtblau, Glenn Greenwald, Laura Poitras, Barton Gellman, Ryan Gallagher,
[26] http://www.zdnet.com/article/nsa-whistleblower-overwhelmed-with-data-ineffective/
[27] Notably *The Guardian, The Washington Post, der Speigel, The Intercept, The New York Times*, ... See: https://snowdenarchive.cjfe.org/greenstone/cgi-bin/library.cgi?e=q-00100-00---off-0snowden1--00-2----0-10-0---0---0direct-10-and%2cand%2cand-TE%2cTT%2cDE%2cSU--4--

glimpse into these highly secretive surveillance agencies. The Snowden trove reveals for the first time in fascinating detail their inner workings, identifying hundreds of individual surveillance programs.[28] But the view offered is at best a sliver of the full picture. So far only about 500 out of the more than 50,000 Snowden leaked have been made public. Some vast surveillance programs are mentioned only in extremely abbreviated summaries. Many of the documents have been released to the public only in fragmented form, with heavy redactions by both government officials and newspaper editors. They are full of obscure, cryptic acronyms, code words and arcane technical details that call for security expertise and organizational experience to properly decipher.

Nevertheless, while keeping these limitations in mind, there is already such an abundance of material in the Snowden trove dealing with metadata that we have a good basis for painting a reliable, if preliminary, picture of how these agencies discuss and operationalize our topic at hand.

For our study of metadata within the FVEY we relied extensively on the Snowden Digital Surveillance Archive, a publically accessible finding aid to the full corpus of published Snowden documents and related media articles that we designed, built and is now hosted by the Canadian Journalists for Free Expression.[29] From working with the documents in building the archive, we developed strong suspicions that internally, the FVEY agencies take a much more expansive view of metadata than suggested by their public statements and the popular media definitions discussed above. In switching to a research role, we sought to test our suspicions, while being open to possible disconfirming evidence. We initially made use of the Archive's full text search, indexing and document description features to locate documents and stories relatively dense in details about metadata and then pursued thematic linkages between documents, such as by surveillance program, to amplify the contexts in aid of interpretation.

Metadata is evidently an important topic within the FVEY. A search in the Archive on 'metadata' produces 1644 word hits. Fourteen documents contained 'metadata' in their title, which we examined these first. The surveillance programs, legal justifications and internal policies mentioned therein informed further archival searches. The domain knowledge we had gained from arranging and describing the Snowden documents also greatly aided our initial searches for identifying fertile points for research as well as in understanding the documents we selected.

---

echelon%2c%2c%2c-----0-1l--00-en-50---50-about-TE%3a%28echelon%29--01-3-1-00-00--4--0--0-0-01-10-0utfZz-8-00&a=d&cl=CL4

[28] See https://snowdenarchive.cjfe.org/greenstone/cgi-bin/library.cgi?e=d-00100-00---off-0snowden1--00-2----0-10-0---0---0direct-10-and%2cand%2cand-TE%2cTT%2cDE%2cSU--4--echelon%2c%2c%2c-----0-1l--00-en-50---50-about-TE%3a%28echelon%29--01-3-1-00-00--4--0--0-0-01-10-0utfZz-8-00&a=d&cl=CL6.15

[29] https://snowdenarchive.cjfe.org

While we expected that exploring such a conceptually vague and varied phenomenon as metadata would yield a mix of results, we were struck by just how heterogenous the results were. For example, disparate GCHQ surveillance programs harvest hotel room bookings (ROYAL CONCIERGE), social media activity (STELLARWIND) and text message geolocation data (in partnership with the NSA under DISHFIRE). Each different program generates different types of metadata, making classification of the surveillance agency's handling of metadata difficult. To facilitate comparing the various agency interpretations of metadata with each other's, with their public statements and with the various legal definitions in their respective jurisdictions, we looked at the three most relevant agencies in turn - NSA, GCHQ and CSE.[30] We also focused on those surveillance programs in which metadata plays a particularly prominent role, notably XKEYSCORE, which provides front-end interface to many signals intelligence databases around the globe and is accessed by all members of the alliance, as well as by selected third party agencies.

Metadata in the NSA

Several documents in the Snowden trove across a span of several years use identical language that appears the NSA has adopted as a standardized definition of metadata:

> *"Communications metadata refers to structured "data about data": it includes all information associated with, but not including content, and includes any data used by a network, service or application to facilitate routing or handling of a communication or to render content in the intended format; it includes,* **but is not limited to**, *dialing, routing, addressing, or signaling information and data in support of various network management activities (e.g. billing, authentication or tracking of communicants)"*[31] *(emphasis added)*

On the surface, this definition appears very similar to the understandings of metadata in archival theory, described above. However, beyond the obligatory '*but not including content',* such an unrestricted definition opens the door for the NSA to justify the unfettered algorithmic analysis of communications content as metadata extraction, a stretch by any conventional understanding of metadata. This is illustrated by the NSA's most prominent analysis engine, XKEYSCORE. Compared with many other tools and surveillance programs mentioned in the archive, important aspects of XKEYSCORE are

[30] We exclude Australia's DSO and New Zealand's GCSB from our treatment here as there are relatively few Snowden documents that relate to these partners, nor does metadata appear prominently among them.

[31] This definition is found in several different documents. e.g.: https://snowdenarchive.cjfe.org/greenstone/collect/snowden1/index/assoc/HASH1aaa.dir/doc.pdf and https://snowdenarchive.cjfe.org/greenstone/collect/snowden1/index/assoc/HASH011c/8cb4f95b.dir/doc.pdf

extensively described, allowing a relatively comprehensive understanding of its capabilities, scope and use.

XKEYSCORE is one of the NSA's most powerful tools, and is often in demand among its trusted "2nd party" (i.e. other members of the Five Eyes) and "3rd party" partners.[32] Access to the tool has been shared with GCHQ, the Australian Signals Directorate (ASD) (Australia), CSE (Canada), GCSB (New Zealand), the *Bundesnachrichtendienst* (Germany), and the National Defence Radio Establishment (FRA) (Sweden). Described in the *Intercept* as "the NSA's Google,"[33] this tool allows analysts unprecedented access to communications metadata largely harvested by the NSA from fibre-optic cables and cached in over 700 servers at 150 storage sites scattered across throughout the world. An unofficial "users guide" to XKEYSCORE developed by Booz Allen Hamilton gives a technical description of the operation of the tool aimed at surveillance analysts. It includes detailed illustrations of its user interface, as well as specifications of how different metadata fields can be used to query the NSA's vast archives of intercepted communications. These fields are very valuable in identifying the NSA's operational interpretation of metadata.

Conventional metadata fields are well represented - unsurprisingly analysts have the ability to query communications by IP address, phone number, email account, etc. However, metadata query capabilities extend far beyond the obvious to include accessing anything contained in email messages, web searches, text chats and file attachments - in short, the full range of communications content.

The users guide describes various metadata 'extractors', including for phone number and email address. These tools appear to scan digital communications traffic for any mention of email addresses and phone numbers, not just in the routing data but in the message bodies, and retrieve these as distinct metadata fields:

> *"The phone number extractor query looks through the **content** of an email for phone numbers. This is very similar to a PINWALE DoPhone query except the traffic that XKEYSCORE finds may survey (i.e. unselected, non-tasked data) and might not be in PINWALE. XKEYSCORE may be your only hope at finding an email address for a target where you only have their phone number as lead information." (emphasis added)*[34]

Another document further describes the email address extractor query:

---

32 https://snowdenarchive.cjfe.org/greenstone/collect/snowden1/index/assoc/HASH01d9/b8e19abe.dir/doc.pdf

33 https://theintercept.com/2015/07/01/nsas-google-worlds-private-communications/

34 https://snowdenarchive.cjfe.org/greenstone/collect/snowden1/index/assoc/HASH012e/6318a2b1.dir/doc.pdf p.10

> *"The query searches within bodies of emails, webpages and documents for....(you guessed it)...Email Addresses."*[35]

The NSA and its intelligence partners, chief among them GCHQ, use these content-derived metadata extraction capabilities not only for counter-terrorism, but also in pursuit of diplomatic advantage. Targeting the offices of at least 35 world leaders, FVEYs agencies intercepted both phone and email communications from officials and staffers,[36] extracting phone numbers and email addresses to be used in identifying further targets.[37] [38]

Beyond phone numbers and email addresses, other XKEYSCORE documents make clear its ability to query the content of email messages and their attachments for arbitrary text strings:

> *"Allow queries like.... [s]how me all documents that reference Osama Bin Laden"* [39]

Interestingly, the XKEYSTORE documents refer to this kind of database querying via metadata as 'contextual' search. Arguably 'context' is an even more ambiguous term than 'metadata,' providing for wide interpretative flexibility. The conflation of metadata with context is illustrated in the following excerpts from another XKS training document, *Guide to using Contexts in XKS*. (See Figure 1)

These examples provide clear evidence that in practice, the NSA subscribes to a view that anything it can derive algorithmically from communications content can be regarded as metadata, including information extracted from the communications content itself. Short of a human analyst actually reading a message or listening in on a phone call, any distinction between content and metadata has been erased. As XKEYSCORE is shared among all Five Eyes SIGINT partners, not withstanding any jurisdictional differences between them, they all have in effect adopted this expansive view of metadata.

---

35 https://snowdenarchive.cjfe.org/greenstone/collect/snowden1/index/assoc/HASH01f9/ed9d3833.dir/doc.pdf slide 2

36 http://www.theguardian.com/world/2013/oct/24/nsa-surveillance-world-leaders-calls

37 https://snowdenarchive.cjfe.org/greenstone/collect/snowden1/index/assoc/HASH3012.dir/doc.pdf

38 https://snowdenarchive.cjfe.org/greenstone/collect/snowden1/index/assoc/HASH1fcc.dir/doc.pdf

39 https://snowdenarchive.cjfe.org/greenstone/collect/snowden1/index/assoc/HASH01d9/b8e19abe.dir/doc.pdf slide 26

Figure 1 'Context' as metadata category in XKEYSCORE[40]

SECRET//COMINT//REL TO USA, AUS, CAN, GBR, NZL

Guide to using Contexts in XKS Fingerprints
Version 1.0

Example 1

```
$a = cc('pk') and (web_search('jihad') or document_body('planning
for jihad'))
```

Definition

(S//REL) Contextual expressions are those that restrict the search space for a particular expression. In example 1 above, we are looking for the string 'jihad' only in the normalized text of a web search, and 'planning for jihad' only in the context of the UTF-8 normalized text of an office document.

(S//REL) For example, web_search is a *scan* so web_search('jihad') will hit on web searches like:

"I want to participate in jihad"
"How do I avoid jihad"
"jihadi"
"bigjihad"

Communications Content

...

communication_body

Description:
(U//FOUO) The UTF-8 normalized text of all office document, email, and chat bodies.

Aliases:
doc_email_body

Context Type:
Scan
Aliased to email_body, chat_body and document_body.

Eample:
communication_body('how to' and 'build' and ('bomb' or 'weapon'))

40
https://snowdenarchive.cjfe.org/greenstone/collect/snowden1/index/assoc/HASHaff8.dir/doc.pdf
pp 1 & 1.8

Metadata in GCHQ

One particularly useful document in the Snowden archive, Content or Metadata? Categorisation of aspects of intercepted communications - GCHQ policy guideance,[41] clearly delineates GCHQ's official understanding of metadata. Similar such technical documents are not forthcoming for the other two major surveillance agencies. This document clearly specifies whether certain elements of an internet communication (such as an email) are to be classified as content or metadata, and which particular "class" under the *Regulation of Investigatory Powers Act* the communication falls under.

What is noticeable about this schema is how different it is compared to the NSA's. The contents of attachments to emails are labeled "content", in marked contrast to the NSA, which as we saw above in XKEYSCORE, operationalizes this element as a class of communications metadata.[42] Interestingly, GCHQ also identifies a class of "content derived data", including analysis of the language of the communication itself, presumably derived from analysis of the content of the message.[43] The document seems to imply that this class of data is considered conceptually distinct from both metadata and content.

However, GCHQ, along with the other surveillance agencies in the Five Eyes, has access to XKEYSCORE. It is unclear how they utilize this tool in accordance with their specific legal and policy requirements, as XKEYSCORE has been described as lacking any meaningful oversight and accountability features.[44]

Metadata in the CSE

In Canada, the CSE collects metadata as part of its mandate, through the National Defense Act, "to acquire and use information from the global information infrastructure for the purpose of providing foreign intelligence" (National Defence Act, S.273.64.1)). The act broadly defines 'global information structure as "electromagnetic emissions, communications systems, information technology systems and networks, and any data or technical information carried on, contained in or relating to those emissions, systems

---

41 https://snowdenarchive.cjfe.org/greenstone/collect/snowden1/index/assoc/HASHc296.dir/doc.pdf

42 https://snowdenarchive.cjfe.org/greenstone/collect/snowden1/index/assoc/HASHc296.dir/doc.pdf

43 https://snowdenarchive.cjfe.org/greenstone/collect/snowden1/index/assoc/HASHc296.dir/doc.pdf

44 http://abcnews.go.com/blogs/politics/2013/07/glenn-greenwald-low-level-nsa-analysts-have-powerful-and-invasive-search-tool/

or networks." This mandate is limited by "measures to protect the privacy of Canadians in the use and retention of intercepted information" (National Defence Act, S.273.64.2) as outlined in the criminal code.

Unlike other intelligence agencies, the Canadian CSE displays its public definition of metadata on its website:

> "Metadata is the context, but not the content of a communication. It is information used by computer systems to identify, describe, manage or route communications across networks. For example, metadata can refer to an internet protocol address or a phone number or the time of a transmission or the location of a device… While metadata reveals a certain amount of information about devices, users and transmissions, it is contextual and does not expose the content of emails, phone calls or text messages."[45]

The director of CSE recently reiterated this distinction between content and metadata as context in responding to a Toronto Star editorial calling for more oversight of the agency. "Context, not content."[46] But as we saw above, context in the eyes of these agencies is much different than how we conventionally understand the term and is deeply tied to communication content. As in the case of GCHQ, it is difficult to square this definition with the agency's continued use of XKEYSCORE.

A particular CSE surveillance program aptly reveals the power of metadata in surveillance activities. A document titled "IP Profiling Analytics & Mission Impacts" describes a trial program where CSE profiled the users of wifi networks at an international airport located on Canadian soil.[47] The CBC incorrectly reported this as involving the interception of wifi signals in the airport, but the actual practice is far more disturbing. Especially revealing is the statement by John Forster, then chief of CSE, called before the Senate hearing to explain the apparent violation of the prohibition on the tracking of Canadians:

> "This [CSE surveillance] exercise involved a snapshot of historic metadata collected from the global internet. There was no data collected through any monitoring of the operations of any airport. Just a part of our normal global collection."[48]

---

[45] https://www.cse-cst.gc.ca/en/inside-interieur/metadata-metadonnees

[46] Metadata is crucial, CSE insists | Toronto Star, Mar 3, 2016 https://www.thestar.com/opinion/letters_to_the_editors/2016/03/03/metadata-is-crucial-cse-insists.html

[47] http://www.cbc.ca/news/politics/csec-used-airport-wi-fi-to-track-canadian-travellers-edward-snowden-documents-1.2517881

[48] CBC, Spy agencies, prime minister's adviser defend Wi-Fi data collection http://www.cbc.ca/news/politics/spy-agencies-prime-minister-s-adviser-defend-wi-fi-data-collection-1.2521166

Through comprehensive capture, analysis and storage of internet communication, CSE spotted visitors to the airport based on the IP address of the airport's public wifi service. Analysts were then able to track individuals to other locations with identifiable IP addresses, both forwards and backwards in time, based on the UserIDs extracted from message content. Not only does this mis-reported case illustrate CSE's expansive interpretation of metadata, but the remarkably broad scope and fine detail of its domestic surveillance capabilities.

Findings

'Deep' versus 'shallow' metadata

This review of the meanings metadata in varied settings shows that notwithstanding a general recognition that metadata is ambiguous and difficult to distinguish from content, we can discern two distinct sets of meanings to the term. In the popular and legal discourses as well as in the official public statements by security agencies, we observe what can be termed narrow, conventional, 'shallow' metadata,[49] characterised by the various forms of data about a communication act or information object that can be read externally, without examining the actual content. In contrast to this is 'deep' metadata, which goes significantly beyond conventional metadata to include data that can be derived algorithmically from the content. It is this 'deep' metadata that we find defined in the archives field and operationalized in the big data surveillance activities of Five Eyes security agencies. One central conclusion is that attempts to maintain a clear distinction between communication content and metadata are untenable. Deep metadata practices are potentially even more revealing of peoples' lives than previously understood. They certainly meet the Supreme Court interpretation of the Criminal Code, which protects "any derivative of that communication that would convey its substance or meaning."[50] While the judiciary is showing signs of updating its understanding of metadata in light of contemporary practices, journalism and the law appear to be lagging.

Researching with the Snowden Archive

While we have been building and promoting the Snowden Archive for the past two years, this report reflects our first attempt to make use of the Archive ourselves for research purposes. Overall we have been generally pleased with the experience. Most obviously, the archive provides an opportunity to examine in a comprehensive and seamless way all the documents Snowden released and subsequently published. We did not face a priori divisions based on agency, national jurisdiction or publishing source. This enabled

[49] This terminology of 'shallow' vs 'deep' metadata is inspired in part by the similar distinction used in the *XKEYSCORE* document of Feb 25, 2008. See p 9 and 10. It also echoes the 'deep packet inspection' techniques employed by FVEY agencies in generating metadata from intercepted communication traffic.
[50] As quoted in Forcese, 2015, p. 137

us to draw a more holistic picture of the role of metadata in mass surveillance at the international level. We found document descriptions and links to media reports developed during the archival process particularly helpful to contextualize documents that are often heavily redacted and excerpted. We also became acutely aware of some of the imitations of the current implementation, notably the lack of a controlled vocabulary and technical limitations of the Greenstone software the archive is built on. But these were more of a nuisance than a major obstacle to our research.

Implications (for discussion)

Our findings have significant implications for various actors in the surveillance arena:

For surveillance researchers

For those attempting to understand the nature and implications of contemporary big data surveillance practices across the private and public sectors, our findings point to the need to avoid being restricted to the conventional, 'shallow' definition of metadata. Notwithstanding protestations to the contrary, 'deep' metadata analysis based on algorithmic processing is widely practiced by the state security agencies, and likely by commercial enterprises with access to large volumes of personal information. It is clear that this form of data mining involves large-scale, systematic algorithmic analysis of communication content, to draw out potentially anything that may be of interest to the surveillance analyst. In short, metadata is a principal means for analysing and accessing 'content' and effectively inseparable from it.

Our experience studying metadata using the Snowden Surveillance Archive also suggests that researchers interested in big data surveillance, particularly by the Five Eyes agencies, may find the Archive a valuable research tool.

For security intelligence agencies

State security intelligence agencies like CSE, NSA and GCHQ are legitimately in the business of maintaining secrecy and deceiving perceived opponents, but in healthy democracies they are also ultimately accountable to the citizens they are mandated to protect. While it may run against institutional culture, maintaining the necessary public confidence and trust requires significant transparency, honesty and demonstrable compliance with legal and constitutional norms. Especially at a time when there is growing and well-founded skepticism that these agencies are neither in compliance wit the law , nor being effective in their missions, for their official statements to consistently rely on a narrow interpretation of metadata at odds with their actual practices is hardly reassuring. Misleading the public about such an important term as metadata only heightens concerns. Coming clean about what they really do with our data may initially

provoke adverse reaction, but in the long run is far better than fueling the vicious cycle of deception and public skepticism.

For those calling for state surveillance accountability and reform

For civil liberties and democratic governance advocates, parliamentarians, privacy regulators, journalists, action researchers and citizens interested bringing greater accountability and reform to state surveillance agencies, our findings point to several implications:

- Revise their working definitions and metaphors of metadata to include 'deep' metadata[51]
- Treat statements about metadata, especially by government officials, with skepticism
- Avoid repeating the conventional definition of metadata, but inform one's audience of the much more comprehensive meaning of metadata, its practical indistinguishability from message content and the greater privacy risks that deep metadata can pose to personal privacyChallenge official statements when they simply reiterate the conventional definition of metadata to be more specific and accurate
- Call to account state security and law enforcement that make misleading public statements and go beyond their legal authorizations
- Press for metadata receiving privacy protection equivalent to content (e.g. in the Criminal Code) .
- Support efforts for stronger transparency and accountability of organizations with access to large volumes of personal information, especially state security and law enforcement agencies. In particular, ensure that state security and law enforcement agencies are more transparent, accountable and operate within the norms of democratic governance.

References

Ball, J. (September 30, 2013). NSA stores metadata of millions of web users for up to a year, secret files show. *The Guardian*. Retrieved from: http://www.theguardian.com/world/2013/sep/30/nsa-americans-metadata-year-documents

[51] e.g. The OPC's 2014 Metadata and Privacy statement could be expanded to make explicit the forms of deep metadata we highlight above. See: https://snowdenarchive.cjfe.org/greenstone/collect/snowden1/index/assoc/HASHc296.dir/doc.pdf

Burke, A. (February 23 2016). Difficult to determine' scope of privacy breach in Five Eyes data sharing. *CBC News*. Retrieved from: http://www.cbc.ca/news/politics/cse-metadata-five-eyes-sharing-1.3459717

Criminal Code of Canada. R.S.C., 1985, c. C-46.

Forcese, C. (2015). "Laws, Logarithms, Libertie: Legal Issues Arising from CSE's Metadata Collection Initiatives." In Michael Geist (Ed.) *Law, Privacy, and Surveillance in the Post-Snowden Era*. (pp. 127-160) Ottawa, Ontario: University of -Ottawa Press

Fisher, D., DeLine, R., Czerwinski, M., & Drucker, S. (2012). Interactions with Big Data Analytics. *Interactions*, 19(3), 50–59.

Greenwald, G. (2014). *No place to hide : Edward Snowden, the NSA and the surveillance state*. London: Hamish Hamilton.

Gilliland, A.J. (2008). Setting the stage. In Murtha Baca (Ed.), *Introduction to Metadata* (pp. 1-19). Los Angeles, CA: Getty Research Institute. Retreived from: http://www.getty.edu/research/publications/electronic_publications/intrometadata/setting.pdf

Laprise, J. (2016) "Exploring PRISMS Spectrum." In Francesca Musiani et. al (Eds.) *The Turn to Infrastructure in Internet Governance*, (pp. 203-216). New York, NY: Palgrave Macmillan.

Lyon, D. (2014). Surveillance, Snowden, and Big Data: Capacities, consequences, critique. *Big Data & Society,* 1(2), 1-10.

National Defence Act. R.S.C., 1985, c. N-5.

Office of the Privacy Commissioner. (2014). Metadata and Privacy: A Technical and Legal Overview." Retrieved from https://www.priv.gc.ca/information/research-recherche/2014/md_201410_e.pdf

Sagiroglu, S., & Sinanc, D. (2013). Big data: A review. In *2013 International Conference on Collaboration Technologies and Systems (CTS)* (pp. 42–47). http://doi.org/10.1109/CTS.2013.6567202

Regulation of Investigatory Powers Act 2000 c. 23

U.S. House. 114th Congress. *H.R.2048 - USA Freedom Act of 2015*. Washington, Government Printing Office, 2015.