Legibility Machines

Thank you for taking time away from the news, and your Friday home offices. When I applied to the iSchool PhD program 6 years ago I never imagined that I would be defending my dissertation from my basement. But I've been surprised by a lot of things along the way in my studies and research, and I will be describing some of that today.

In this presentation I will be sharing the key findings from my dissertation project that focuses on answering the question of how archival appraisal is being enacted in web archives, or archives of web content. A good friend of mine once advised me to lead with the punch line in presentations, so if you take nothing else from my presentation today please let it be this: expressions of appraisal value in web archives are intrinsically tied to the ways in which those records are used. Records in web archives have no intrinsic value in and of themselves. However, *use* can take many forms, including misuse and disuse. Recognizing what feminist scholar Sara Ahmed calls the *uses of use* and the *strange temporalities of use* is critical for understanding archival appraisal in the context of the web.

(Overview)

My dissertation project began with an encounter with a paradox, or conundrum, concerning the web as a medium.

The web is ephemeral. Web resources are in a constant boil, appearing and disappearing, as the links we once clicked on turn into *404 Not Found* errors. Estimates about the prevalence of *linkrot* vary. The size of the web and the fact that it's constantly changing means that any measure of linkrot is a function of how and when we ask the question. Be that as it may some estimate that about a quarter of all links break every 7 years (Ceglowski, 2011).

But the web can also be highly durable. The web sits atop the Internet, which was, at least in principle if not in practice, designed to "route around failure". Content delivery networks (CDNs) permit synchronized copies of documents to be made instantly available in geographically dispersed caches. Sites like Wikileaks and SciHub have been designed to seamlessly replicate themselves into multiple locations and onto the *darkweb*, in order to resist attempts by governments and powerful institutions to remove them (Firmino et al., 2018; Greshake, 2017). Social media platforms allow users to screenshot and re-share deleted content, which can then circulate in a multitude of slightly altered forms to evade filtering by content moderation systems (Baumgartner et al., 2019). Even media that is designed to be ephemeral can find new life on the web where it is mirrored and mutated as it achieves transmission (**???**). The web allows once fragmented collections to be reassembled as part of virtual reuinfication projects (Punzalan, 2014). And for about 20 years or so web archiving projects have worked to collect regions of the web in order to re-present it back on the web in so called web archives. The ability to rapidly collect this amount of documentation is something that the web itself has made possible.

It's this strange property of ephemeral-durability that led me into my research project to ask a seemingly simple research question: how are archivists deciding what to collect from the web?

Archival studies researchers will immediately recognize this as a question about *appraisal*, which is a topic of perennial interest to the field, and is widely regarded as being "the first responsibility of the archivist" (Cox & Samuels, 1988). A significant chunk of my dissertation contains a review of the rich and varied literature of archival appraisal, but the Society of American Archivists Glossary provides a useful definition for us here.

My literature review suggested that within this broad question about appraisal of the web there were two linked sub-questions: 1) how do the tools that are being used to archive the web shape what is collected; 2) and how do the characteristics of what we want to collect from the web shape the tools we use to do it?

Information studies researchers will recognize these questions as two sides of a single *sociotechnical* question, where phenomena aren't reducible to purely social or technical explanations but must be understood in terms of how the social and the technical *coproduce* each other (Jasanoff, 2006).

Nearly three decades ago Margaret Hedstrom identified the need for studies of archival systems that analyze the design and evolution of computer and organizational *interfaces* in terms of sociotechnical factors such as automation, remediation, markets, organizational change (Hedstrom, 1991 , 2002). This need has proven evergreen, particularly for the study of web archives. Recent work in this area has focused on the role that *provenance* plays in web archive infrastructures (Maemura et al., 2018), as well as how web archives participate as part of a diverse set of claim making practices (Ogden, 2019). Web archives are legibility projects and *epistemic sites* that have real material consequences in terms of the types of knowledge claims that they permit (and don't permit) to be made about the world (Acker & Chaiet, 2020; Ben-David & Amram, 2018).

However there has been very little analysis of how *appraisal* of web content is being performed as a sociotechnical practice. So my dissertation asks the following research questions:

- How are archivists deciding what to collect from the web?
- How does our understanding of what constitutes a web archive relate to the practice of appraisal?

I chose qualitative methods to answer these questions because I wanted to gain insight into the *motivations* for appraisal decisions, rather than measure, control for and predict that behavior. My dissertation contains three studies.

In the first I conducted 28 semi-structured interviews with practitioners and used thematic analysis to derive a set of sociotechnical factors in *what* archivists talked about when they talked about appraisal. Some of these factors included: storage, network bandwidth, system traceability, and the significant improvisation that archivists did to work with, and work around, web archiving tools.

In the second I used critical discourse analysis to do a close reading of a subset of these interviews in order understand *how* archivists talked about appraisal. I found that the architecture of the web disrupts the trust relationship between records creators and archivists that is so central to appraisal, and can be mended by attention to the positionality of web archives

In the third study, which forms the bulk of my dissertation, I conducted a year long ethnographic field study with the National Software Reference Library to understand *how* archivists appraised the web.

These three studies were used as part of a between-method triangulation that was designed to build a "thick description" by which to understand my informants beliefs about their appraisal practices.

It wasn't until I analyzed the results of my ethnographic field study with the NSRL that my primary contribution came into focus. (Field notes, and coding.) I will spend my remaining time sharing a few vignettes from this case study, describe what I think their significance is, and what they suggest for future research.

But first I need to tell you what the NSRL is, because it does not fit the shape of what we normally talk about when we talk about web archives.

NSRL

The National Software Reference Library (NSRL) is a project based at the National Institute for Standards and Technology in Gaithersburg, Maryland. For the last 20 years the NSRL has gradually assembled one the largest known collections of computer software in the world (NIST, 2018).

The NSRL began by collecting "shrink-wrapped" software (physical floppy disks, CDs and DVDs) and storing them in a library on NIST's Gaithersburg campus. Digital copies of the media were created to serve as a backup and also as the input data for a set of digital forensics services that the NSRL provides.

Over the last ten years the NSRL has transitioned to collecting software from the web, since the web has became the predominant distribution mechanism for computer software.

Though housed at NIST the NSRL was created in 2000 with funding by a diverse set of federal government agencies that found common interest in creating something called the NSRL Reference Data Set (RDS).

The RDS is a collection of digital content signatures, or *hashes*, of all the files, from all the software that NIST has collected since work on the NSRL began.

These digital signatures are used to rapidly weed out or "DeNIST" known files when conducting forensics investigations.

(NIST:2002 Sherlock image)

As of June 2020, the publicly released NSRL RDS tracked over 250 million hashes for files that were part of 213,770 versions of 171,567 software packages, from approximately 70,000 publishers.

Hopefully this short sketch of the NSRL and the RDS is enough context for you to understand these three short stories from the field. I've selected them because they highlight how the NSRL enacted appraisal of web content, and their values about what to collect.

The Customer

It wasn't long after I started attending NSRL staff meetings that I heard "The Customer" being mentioned. In one meeting team members were discussing how many new hashes would be made available in the next RDS release, Lara, one of the NSRL's managers reminded the team that "two million new hashes isn't what matters; what matters is that the release is driven by what the customer needs". Melissa, who is responsible for purchasing software for the NSRL, followed up by saying how she "always tries to think like a customer" and how "I've done a good job with the customers". This made me wonder who the *The Customer* was, and what it meant to "think like a customer"?

The Customer

Team members generally agreed that "thinking like a customer" meant thinking like law enforcement using the RDS to "DeNIST" a set of computer files. In an interview Lara described how DeNISTing led the NSRL to focus on popular software:

> So many people are using it [the RDS] just to get rid of known software. That's our number one use case, to support the efficiency of these investigations, by getting rid of known content so you don't have to search it. For that, for everybody, the metric is popularity. Have big popular things.

However I learned during my observations that interpreting what "popular" meant was no simple process. Popularity measures change rapidly, sometimes minute by minute for some platforms, they are shrouded in mystery to prevent artificial inflation by software publishers, and the world of software is large, so popularity means different things given the category you are looking at.

These details around popularity seemed to unravel as I studied them until I happened to attend a memorable all hands staff meeting on the subject of an award that had been given to the NSRL from the Norfolk Constabulary in the UK. I learned there that the huge focus on collecting video games was something that was initiated by the UK foreign government in their construction of their Child Abuse Image Database (CAID).

When I asked Mike during an interview about how the NSRL had started collecting games he had told me how the focus on collecting games came directly from a collaboration between the UK Home Office, CAID, and game distributors like Steam, Blizzard and Epic.

Hearing Mike describe this relationship, and how it evolved, made me realize how significant this network of relations between users of the RDS such as CAID, NIST, and software manufacturers were to the NSRL's web appraisal practices.

CAID and Games Archiving

Misuse: ByLock

In another staff meeting James, one of the NSRL's technical analysts announced that he had attended a digital forensics conference the week before, and happened to remark "Oh, and apparently Turkey is using the RDS? Yeah, someone gave a lightning talk, and described how he worked for the Turkish government, which had used the RDS to inspect 1.6 million devices during the investigation into the 2016 coup attempt, which led to 75,000 people being put in jail?"

Lara replied, "Great…NSRL data being used to prop up military dictatorships." To which Mike responded "Yay us?", and then the meeting moved on after a pause.

James, Lara and Mike's sarcasm here underscored their critical perspective on this use of the NSRL. A few years earlier on July, 15 2016 a faction within the Turkish Armed Forces had attempted a coup to remove President Recep Tayyip Erdoğan and his government. The coup against Erdoğan failed, and ignited a series of purges that led to hundreds of thousands of government employees losing their jobs and others being imprisoned. The Turkish intelligence service alleged that individuals had used a mobile application called ByLock to coordinate the coup, and installation of the app on mobile devices was used as evidence to arrest individuals (Gokce, 2018).

Disuse: Cabrinety

In 2012 the NSRL entered into a relationship with Stanford University to process collection of approximaely 50,000 software titles known as the Cabrinety Collection. This collection of software

Even four years the after the software was acquired it still had not been added to the RDS. There was very little pressure to add them since at ingest time in 2013 80% of the software titles were 25 years or older. This wasn't popular software. In fact it was the exact opposite, it was *rare* software.

In an interview with a Stanford curator I heard the Cabrinety collection described as the "gift that keeps on giving" because of the various projects that it

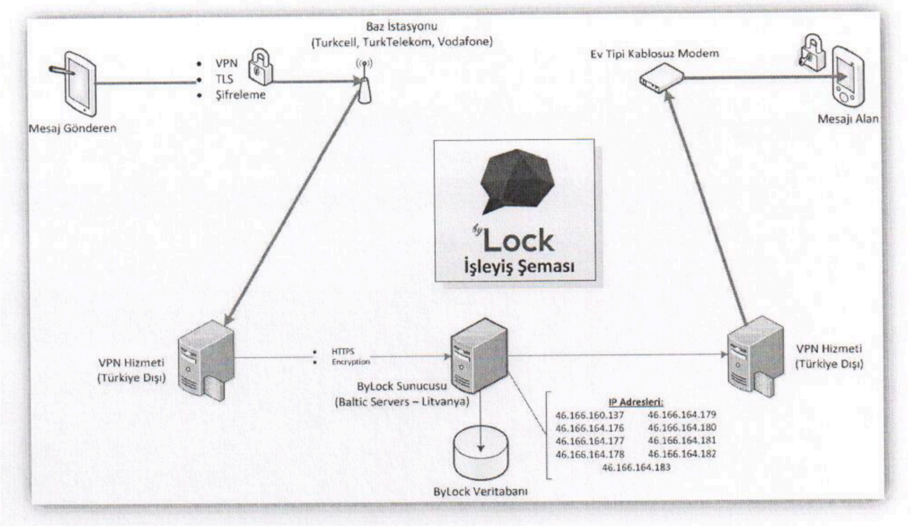

Figure 1: Turkish Intelligence Forensics Diagram from Fox-IT (2017)

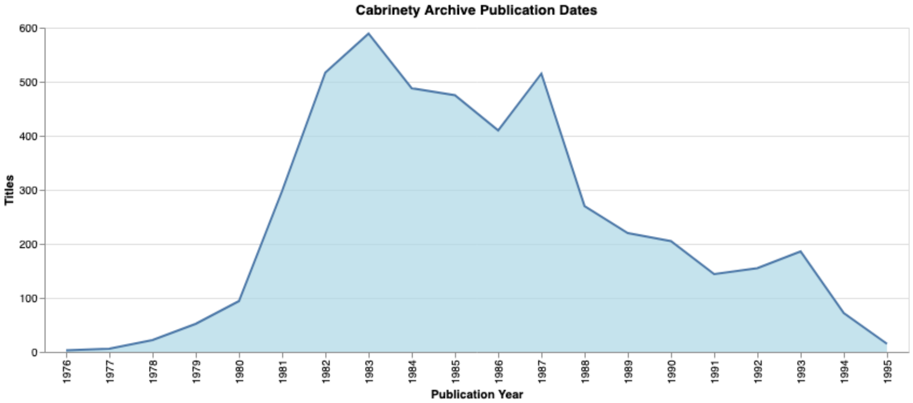

Figure 2: Age of Cabrinety Materials

has been at the center of over the years.

> … we were able to get thousands of titles from the collection off the original floppies, and data tapes into formats, disk images that we could put into the Stanford Digital Repository, into portable preservable objects. That was a big thing for us. That opened up the possibility of access to the collection in different ways.

The disuse of the Cabrinety items in terms of the RDS hinted at another use, which was the circulation of digital forensics as a practice, where disciplinary practices of defense, intelligence gathering, and law enforcement were articulated with digital preservation, archival science and historical inquiry. In fact this disciplinary work aligns with NIST's core mission:

> To promote U.S. innovation and industrial competitiveness by advancing measurement science, standards, and technology in ways that enhance economic security and improve our quality of life.

Discussion

These three vignettes have been drawn from a series of findings that I discuss in my dissertation because of the way they specifically speak to the role that use played in how software was selected from the web for the NSRL.

NSRL staff expressed an explicit value in collecting popular items, even as the metrics for popularity were contested, and escaped measurement due to the opacity of software distribution platforms. As long as there was general consensus that what was being collected was popular the work could proceed.

However the story of use in the NSRL has more value dimensions than this pragmatic argument about popularity and forensics initially suggests.

The example of the Turkish intelligence service using the RDS to identify "terrorists" by locating the ByLock app on suspect's devices is an example of *misuse*. NSRL team members recognized that their web collecting activities could be used by actors that did not share their values. The openness of the RDS dataset on the web confers the desired quality of neutrality that is so important for NIST as a standards body. But that same openness can lead to the RDS being used by a variety of actors whose interests do not necessarily align with the values of the NSRL and NIST.

And in the case of the Cabrinety Archive, which seemed to offer so little value in terms of the NSRL's forensic mission to collect popular software for the RDS. But which built technical competencies inside NIST and at Stanford, and thereby created new networks of knowledge/power between defense, law enforcement, academia and cultural heritage.

My analysis drew on the work of feminist scholar Sara Ahmed whose book *What's the Use* explores the "uses of use" and the "strange temporalities of use" (Ahmed, 2019). Ahmed provides a critical lens to the familiar utilitarian explanations of use which see *disuse* and *misuse* in purely negative terms–as things to be avoided and optimized away. She highlights how *queer uses* can function as a form of resistance, in opposition to dominant uses which so often are valorized.

The consideration of the use of archival records has been a bit of a third rail for appraisal theory. For example Richard Cox in his 12 Principles of Appraisal says:

> The archival appraisal selection criteria should rest not on unpredictable future research practices and trends but upon the more predictable sense of determining what are the salient and important features of contemporary institutions and society. (Cox, 1994, p. 24)

Here the archivist is forbidden from considering the use of records, because it is difficult to predict the future. Instead the archivist is asked to document society, as if this is somehow easier. This positions the archivist as an analyst/judge one step removed from the site of record creation. Science and Technology Studies scholars recognize this as an example of the God Trick which Donna Haraway memorably described:

> I would like to insist on the embodied nature of all vision and so reclaim the sensory system that has been used to signify a leap out of the marked body and into a conquering gaze from nowhere. This is the gaze that mythically inscribes all the marked bodies, that makes the unmarked category claim the power to see and not be seen, to represent while escaping representation. (Haraway, 1988, p. 581).

Failing to attend to the use of records lets us believe in the fantasy of their singular use, as if they are always and only the evidence of a particular thing. But allowing for use allows appraisers to appreciate the full set of contingent relations and value propositions that web archives participate in.

Archivists who divorce the collection of records from the use of those records will be forever chasing their own tail when trying to understand the value of records. The web and web archives are new enough that their architectures of use are noticeable; but I want to suggest that the architectures of use that we find in our traditional brick and mortar archives are also expressions of value of the use of use.

What's the Use?

I'd like to conclude with a few thoughts about the use of these results and specifically future areas of research:

- Can we develop appraisal frameworks that factor in the varieties of use, disuse and misues for records as they are being collected? Is it useful to

do this for archives more generally?
- What does the use of existing public web archives look like? Projects like Ian Milligan and Jimmy Lin's Archives Unleashed project have been approaching this from the perspective of academic research infrastructure, but we understand very little about the use of web archives in practices like journalism, law and political activism, to name just a few.
- My dissertation has really been an extended meditation on how our conception of web archives should not be limited to a particular architectural shape. What does the larger, hidden landscape of web archives look like, when don't truncate our understanding of web archives to be only about Wayback Machines and WARC data??
- As a field we are accustomed (especially here at UMD) to thinking of sociotechnical aspects of information in terms of interfaces and Human-Computer Interaction, and archives in terms of preservation and sustainability. But what if we center instead the varied genealogies of use that web archives participate in over time? For example the use of the CommonCrawl web archive, which finds its way into machine learning models like GPT-3 and tools like spaCy? Some call this Human Data Interaction, but another way of conceiving it is terms of tracing the history of uses that data participate in, or a Long Duree of Data.

Acker, A., & Chaiet, M. (2020). The weaponization of web archives: Data craft and covid-19 publics. *The Harvard-Kennedy School Misinformation Review*, *1*. https://misinforeview.hks.harvard.edu/wp-content/uploads/2020/09/acker_weaponization_web_archives_Covid19_20200927-1.pdf

Ahmed, S. (2019). *What's the use: On the uses of use*. Duke University Press.

Baumgartner, J., Bermejo, F., Ndulue, E., Zuckerman, E., & Donovan, J. (2019, March 26). *What we learned from analyzing thousands of stories on the christchurch shooting*. Columbia Journalism Review. https://www.cjr.org/analysis/christchurch-shooting-media-coverage.php

Ben-David, A., & Amram, A. (2018). The Internet Archive and the sociotechnical construction of historical facts. *Internet Histories*, *2*(1-2), 179–201.

Ceglowski, M. (2011). *Remembrance of links past*. https://blog.pinboard.in/2011/05/remembrance//_of//_links//_past/

Cox, R. J. (1994). The Documentation Strategy and archival appraisal principles: A different perspective. *Archivaria*, *38*(January), 11–36.

Cox, R., & Samuels, H. (1988). The archivist's first responsibility: A research agenda to improve the identification and retention of records of enduring value. *The American Archivist*, *51*(1-2), 28–42.

Firmino, R., Melgaço, L., & Kloza, D. (2018). The spatial bonds of Wikileaks. *Government Information Quarterly*.

Fox-IT. (2017). *Fox-it debunks report on bylock app that landed 75,000 people* *in jail in turkey.* https://blog.fox-it.com/2017/09/13/fox-it-debunks-report-on-bylock-app-that-landed-75000-people-in-jail-in-turkey/

Gokce, Y. (2018). The bylock fallacy: An in-depth analysis of the bylock investigations in turkey. *Digital Investigation*, *26*, 81–91.

Greshake, B. (2017). Looking into Pandora's Box: The content of Sci-Hub and its usage. *F1000Res*, *6*(541).

Haraway, D. (1988). Situated knowledges: The science question in feminism and the privilege of partial perspective. *Feminist Studies*, *14*(3), 575–599.

Hedstrom, M. (1991). Understanding electronic incunabula: A framework for research on electronic records. *The American Archivist*, *54*(3), 334–354.

Hedstrom, M. (2002). Archives, memory, and interfaces with the past. *Archival Science*, *2*(1-2), 21–43.

Jasanoff, S. (2006). *States of knowledge: The co-production of science and the social order*. Routledge.

Maemura, E., Worby, N., Milligan, I., & Becker, C. (2018). If these crawls could talk: Studying and documenting web archives provenance. *Journal of the Association for Information Science and Technology*, *69*(10), 1223–1233. https://tspace.library.utoronto.ca/handle/1807/82840

NIST. (2018). *Database of software "fingerprints" expands to include computer games.* https://www.nist.gov/news-events/news/2018/09/database-software-fingerprints-expands-include-computer-games

Ogden, J. (2019). *Saving the web: Facets of web archiving in everyday practice* [PhD thesis]. University of Southampton.

Punzalan, R. (2014). Understanding virtual reunification. *The Library Quarterly: Information, Community, Policy*, *84*(3), 294–323.