2016 IEEE International Conference on Big Data
Workshop: Computational Archival Science: digital records in the age of big data
December 8, 2016

Opening Up Dark Digital Archives Through The Use of Analytics To Identify Sensitive Content

Jason R. Baron[†]
Drinker Biddle & Reath LLP
Washington, D.C.
jason.baron@dbr.com

Bennett B. Borden[††]
Drinker Biddle & Reath LLP
Washington, D.C.
bennett.borden@dbr.com

***Abstract*—The Nation's history is going dark: without technological solutions, presidential and federal e-mail and other electronic records accessioned into the US National Archives will remain effectively inaccessible to the public due to sensitive content considerations, including most notably PII, for many decades. Analytics offers the means for achieving earlier public access to digital collections of public records while protecting the privacy of records creators and third-parties.**

Keywords—archives; sensitive data; PII; analytics; predictive coding; technology assisted review

I. INTRODUCTION

On January 20, 2017, at noon, the U.S. National Archives and Records Administration (NARA) will come into the legal custody of over 500 million e-mail messages and attachments constituting presidential and federal records of the Obama Administration.[1] As of that date, the cumulative number of presidential and federal records consisting of e-mail in NARA's legal custody, going back to the Reagan Administration, will be on the order of 750 million records – more than the combined holdings of all hard copy textual holdings in Presidential libraries beginning with the Hoover Administration.[2] As staggering as these numbers represent, they will soon be dwarfed by NARA's receipt of electronic records of all types (not just e-mail), from the entirety of the Executive branch; these permanently appraised electronic records will begin arriving in earnest after December 31, 2019 – the deadline set in the August 2012 Managing Government Records Directive (2012 Directive) requiring that future accessioning of all permanently appraised records into NARA be in approved electronic formats [1].

The modest aim of this paper is to attempt to frame a research agenda for using big data analytics to help solve the problems faced by NARA (and other public sector archives) associated with providing timely access to permanent archival holdings in digital form while respecting the privacy of document authors and third-parties. With respect to access considerations, the primary challenge NARA faces is how to open digital archives containing, in whole or in part, personally identifiable information (PII), or other sensitive, privileged, or exempt textual content. Absent greater use of existing technological solutions, the looming public policy issue archivists and historians increasingly face is their collective inability to meaningfully provide access to existing and soon to exist digital archives until the 21st century is almost over.

This paper is in three parts. First, we aim to set out the key legal and archival policy constraints preventing earlier access by the public to the government's records. Second, we describe the available analytical toolkit for identifying PII and other sensitive content. Third, we discuss a research agenda that would assist in moving towards a world of more accessible public records archives.

II. LEGAL AND ARCHIVAL BACKGROUND

A. *Legal Considerations*

1. Access to Presidential Records

The Presidential Records Act of 1978 (the PRA) provides for the U.S. Archivist, on behalf of NARA, taking permanent legal custody of presidential records at the end of each Administration.[3] The PRA prohibits general public access to presidential records for five years, and allows for a President to designate certain categories of records as exempt from public access up to 12 years.[4] (The PRA provides for special access to records during these time periods for designated

† J.D., Boston University School of Law. Mr. Baron is Of Counsel at Drinker Biddle. He previously served as Director of Litigation at the National Archives and Records Administration (2000-2013).

†† J.D., Georgetown University Law School; M.S. (Business Analytics), New York University. Mr. Borden is Chief Data Scientist, Partner, and Chair of the Information Governance and eDiscovery Group at Drinker Biddle.

[1] Based on conversations with NARA staff.

[2] *Id.*; *see also* National Archives and Records Administration, "Holdings of the Presidential Libraries" (stating that the thirteen existing Presidential libraries up through the George W. Bush Library hold a combined 400 million pages of textual records), http://www.archives.gov/presidential-libraries/about/holdings.html (accessed October 10, 2016).

[3] *See* 44 U.S.C. 2201 *et seq.*

[4] *See* 44 U.S.C. 2204.

representatives of the President and Vice President, for designated Congress members, and upon court order.)[5]

Under the PRA, public access is provided in two ways: first, the U.S. Archivist may open records after providing a former President with the right to object as a matter of Executive Privilege.[6] Second, the PRA provides that after 5 years, the public may access presidential records under the Freedom of Information Act, 5 U.S.C. 552, subject to the proviso that NARA may not use Exemption 5 of the FOIA as a basis for withholding.[7] That is, NARA may not assert that some or all of a presidential record is exempt due to its having been part of a "deliberative process," or is otherwise considered "privileged" (e.g., under the attorney-client privilege).

In addition, the PRA states that Presidents, "to the extent practicable," should categorize records as either "presidential" (relating to official duties of the office), or "personal," to be filed separately.[8] As a practical matter, White House e-mail collections do contain personal records intermixed with presidential records.

Finally, the PRA provides that "the Archivist shall have an affirmative duty to make [presidential] records available to the public as rapidly and completely as possible" under governing laws.[9]

2. Access to Federal Records

The Federal Records Act (the FRA) provides for NARA taking legal custody of records determined to have sufficient historical value to warrant permanent preservation.[10] The statute provides for NARA obtaining such records "as soon as practicable . . . not later than thirty years after such records were created or received by the transferring agency.[11] As recently amended in 2014, the FRA allows for agencies making earlier transfers under "any shorter period" mutually agreed to.[12] "Federal agencies" include the Executive branch, as well as portions of the Legislative and Judicial branches.[13]

The 2012 Directive requires that after December 31, 2019, all Executive branch agencies accession permanent records created after that date to NARA in a permanent electronic format. The time for accessioning is governed by existing agreements co-signed by the agency and the US Archivist. NARA has issued guidance with respect to what it considers acceptable electronic formats for transfer and accessioning [2]. NARA Bulletin 2014-14 states that:

> "Before transferring permanent electronic records," agencies should "Identify sensitive information within records that may require screening before the records are available to the public: e.g., personally identifiable information (PII), such as Social Security Numbers of living individuals, law enforcement information, and proprietary information…."

The Bulletin goes on to state that "agencies are strongly encouraged to provide to NARA as much information as possible about the type and location of the sensitive information when transferring the records, in order to make NARA's screening more efficient and effective. For example, if an agency transfers a database with individuals' SSNs and dates of birth, the agency should indicate that these fields may be restricted."

The public's access to NARA's federal record holdings is accomplished in two ways. General public access to NARA's archival collections is governed under agency regulations, which recognize that a right to privacy exists which the agency respects in withholding some records or portions thereof.[14] NARA regulations permit withholding records up to 75 years if they contain sensitive content that would invade the privacy of a living individual, consistent with the Freedom of Information Act (FOIA).[15]

Individuals may also access federal records by filing a FOIA request, subject to exemptions.[16] For example, under Exemption 6 of the FOIA, certain records that "would constitute a clearly unwarranted invasion of personal privacy" may be withheld in whole or in part. Exemption 4 of the FOIA covers records containing trade secrets and commercial or financial information obtained from a person that is privileged or confidential. Exemption 7 of the FOIA covers records compiled for law enforcement purposes, including those which "could reasonably be expected to constitute an unwarranted invasion of personal privacy." Where material is exempted from disclosure, many agencies possess the technical means to partially mask records through automated redactions, including with some form of explanatory marking In general, records that are covered by one of nine FOIA exemption contain "sensitivities" that need to be flagged; for our purposes here, we focus in the main on a more limited subset of sensitive records containing PII, as defined below.

3. Definition of "PII"

There is no universally agreed upon definition of what constitutes PII in government records. The National Institute of Science and Technology (NIST) used the following definition of "PII" in its "Guide to Protecting the Confidentiality of PII" [3]:

[5] *See* 44 U.S.C. 2205.
[6] *See* 44 U.S.C. 2208.
[7] *See* 44 U.S.C. 2204(c)(1).
[8] *See* 44 U.S.C. 2203(b).
[9] *See* 44 U.S.C. 2203(g)(1).
[10] *See* 44 U.S.C. 2107
[11] *See* 44 U.S.C. 2107(a)(2).
[12] *See* 44 U.S.C. 2107(b).
[13] *See* 44 U.S.C. 2901(14).
[14] *See* 36 C.F.R 1256.28(a)(1) & (2) (access to records containing privacy-restricted information is available to qualified researchers or to all researchers if NARA "is able to make a copy of such records with all personal identifiers masked or deleted").
[15] *See* 36 C.F.R 1256.56.
[16] *See* 5 U.S.C. 552(b)(1)–(9).

> "PII is any information about an individual maintained by an agency, including (1) any individual's identity, such as name, social security number, date and place of birth, mother's maiden name, or biometric records; and (2) any other information that is linked or linkable to an individual, such as medical, educational, financial, and employment information."[17]

The NIST guidance identifies examples of information that may be categorized as PII:

- Name, such as full name, maiden name, mother's maiden name, or alias
- Personal identification number, such as social security number (SSN), passport number, driver's license number, taxpayer identification number, patient identification number, and financial account or credit card number. Address information, such as street address or email address
- Asset information, such as Internet Protocol (IP) or Media Access Control (MAC) address or other host-specific persistent static identifier that consistently links to a particular person or small, well- defined group of people
- Telephone numbers, including mobile, business, and personal numbers
- Personal characteristics, including photographic image (especially of face or other distinguishing characteristic), x-rays, fingerprints, or other biometric image or template data (e.g., retina scan, voice signature, facial geometry)
- Information identifying personally owned property, such as vehicle registration number or title number and related information
- Information about an individual that is linked or linkable to one of the above (e.g., date of birth, place of birth, race, religion, weight, activities, geographical indicators, employment information, medical information, education information, financial information). [3]

B. Archival Policy Considerations

"Archivists have more than a legal responsibility to provide access to records, they have a professional, ethical responsibility to do so" [4]. The Society of American Archivists' Code of Ethics states that "archivists actively promote open and equitable access to the records in their care . . ." The Code goes on to say that "[a]rchivists recognize that privacy is sanction by law. They establish procedures and policies to protect the interests of . . . [individuals] whose public and private lives and activities are recorded in their holdings." [5]

At Presidential libraries, textual holdings are subject to "systematic review" for purposes of opening to the public.[18] The systematic review process historically has been accomplished through manual, labor-intensive, document-level review. Additionally, staff in each PRA library have acted to open holdings in response to a large queue of FOIA requests, court orders, and Congressional demands. Efforts to provide access to presidential electronic records began in earnest with the Clinton Administration, "the first to conduct much of its official business on computers" [6]. NARA relies on staff performing keyword searching to find relevant e-mail records and associated documents in response to access requests directed to presidential libraries [7].

NARA's main headquarters and regional archives similarly deploy manual means for the processing of textual holdings as part of reference services. Public access to existing electronic holdings (mostly in structured databases maintained by the former Center for Electronic Records), is accomplished on an ad hoc basis by designated archivists charged with finding responsive records to individual reference and FOIA requests.

As a practical matter, with limited exceptions,[19] no direct, public access to NARA's vast presidential and federal record e-mail holdings presently exists.

III. Employing Analytics In The Archives To Isolate Sensitive Content

The volume of digital holdings at NARA, coupled with the need to isolate sensitive content prior to providing public access, fairly demands that more powerful analytical techniques be employed than continued reliance on manual review. The information tasks to be accomplished include (i) more efficient searching of large collections of records; (ii) using analytical methods to identify, filter, and redact sensitive content including but not limited PII.

In the past decade, the legal profession in particular has confronted the need to employ more efficient and accurate methods to search through and process for production vast quantities of "electronically stored information" that might be considered relevant evidence in legal proceedings. By means of what is termed "predictive coding" or "technology assisted

---

[17] NIST Publication 800-122 states that "[t]his definition is the GAO expression of an amalgam of the definitions of PII from OMB Memorandums 07-16 and 06-19. [and] GAO Report 08-536, Privacy: Alternatives Exist for Enhancing Protection of Personally Identifiable Information, May 2008, http://www.gao.gov/new.items/d08536.pdf.

[18] *See, e.g.*, correspondence dated March 2, 2016, to Neil Eggleston, White House Counsel to President Obama, from B. John Laster, Director, Presidential Materials Division, NARA, https://www.archives.gov/foia/pra-notifications/pdf/bush41/rn-lpgb-2016-035.pdf.

[19] Examples include: a relatively modest number of e-mails associated with the nomination to the Supreme Court of Justice Elena Kagan, as well as additional e-mails searched in response to various FOIA lawsuits, have been put up online. *See, e.g.*, http://www.archives.gov/news/elena-kagan/

review," lawyers are able to search through millions of electronic documents to isolate relevant information by subject matter category [8,9,10,11]. Research based on datasets used in the NIST Text Retrieval Conference (TREC) Legal Track has shown rates of accuracy in finding relevant documents through advanced search methods as high as 80% recall [12]. As of 2012, courts have begun to address and ratify the use of advanced search technologies to satisfy a party's legal obligations [8,13].

Technology assisted review as presently used in e-discovery is based on several predictive analytics techniques that could be used to identify and isolate sensitive content in archives. Predictive analytics leverage various types of algorithms or models to group similar types of data together. In recent years, these predictive models have become extremely accurate when applied to unstructured (or more accurately, semi-structured) data, like e-mail.

Descriptive metadata contained in specific sub-collections may also be useful in isolating known sensitive content [14]. E-mail contains header information and metadata that is like structured data. This metadata includes information regarding who sent and received an e-mail and when, its priority and routing, etc. An e-mail also contains unstructured text that can be analyzed with text mining techniques to derive additional features from the e-mail. Metadata and textual features are used by predictive models to group similar types of e-mail together with a high degree of accuracy. With sufficient examples of sensitive data, accurate predictive models can be developed to automatically identify and isolate potentially sensitive information.

As a corollary to using advanced search techniques, the legal community, like many other communities, also has embraced the use of regular expressions to isolate specific forms of PII found in document sets. A "regular" expression is "a pattern that the regular expression engine attempts to match in input text."[20] Certain forms of PII within the NIST definition given above conform to expressions which software can easily isolate, the most prominent example of which are SSNs (always in the numeric form xxx-xx-xxxx). There are libraries of publically available regular expressions for many forms of PII, such as credit card numbers, phone numbers, and addresses. Moreover, the software that identifies these regular expressions can be programmed to identify any alphanumeric string. Thus, if an agency knows that it uses specific alphanumeric strings in its data (i.e., patient IDs, test IDs, customer IDs, invoice numbers, etc.), these can be easily programmed into the software. In fact, most of these search tools will automatically identify often-recurring regular expressions. The same is true for entity extraction, where the software will automatically identify the names of people and organizations and generate a report. This is especially helpful in identifying sensitive data where, for example, a person or organization is referred to in the body of an e-mail but is not one of the correspondents or domains in the e-mail's metadata.

[20] Microsoft, "Regular Expressions Language – Quick Reference" https://msdn.microsoft.com/en-us/library/az24scfc%28v=vs.110%29.aspx (accessed Oct. 10, 2016),

Three other analytical approaches are good candidates for identifying sensitive content in digital archives.

First, social network analysis could be used to identify potentially sensitive information, and/or identify e-mail that likely is *not* sensitive for further sampling [15]. Sensitive information often may come from and be sent to only certain designated persons within an organization component. Using social network analysis, e-mail records among or between certain persons or even organizations can be isolated for further analysis, including by other techniques.

Second, sentiment analysis may also prove to be a useful technique for identifying certain types of sensitive information. Through this method of inquiry, the text of a document is analyzed through determining the opinion or emotion of the author, among other attributes. Some of the opinions or emotions detectable by sentiment analysis can be quite subtle, including, for example, those regarding respect, rectitude, and morality[16]. It may be possible that certain sentiments expressed in a candid medium such as e-mail are correlated with categories of sensitive information generally.

Third, visual analytics may prove to be a useful tool for extracting meaning from large digital collections and identifying documents with various sensitivities [17,18].

The most effective predictive models leverage a combination of these techniques to identify the target data, in this case sensitive data. These techniques are well known in the data science community and tools using them are easily available in the market. Applying these tools to archived government records would be a relatively straightforward exercise if supported as a matter of research and government policy.

IV. A Proposed Research Agenda

To provide more timely access to public sector digital holdings, the above discussion suggests that the following research agenda be pursued. Our focus is intended to be on NARA, but the discussion has broader applicability to all public record archives at the local and state levels as well.

First, the greater archival community should work with academics and private industry to develop a standard set of regular expressions, to be routinely applied as filters against newly accessioned presidential and federal records.

Second, archivists should work with academics and private industry to pilot one or more analytical methods identical to software used by the legal community in performing predictive coding or technology assisted review, for the purpose of isolating sensitive content in presidential and federal e-mail records, and other records, that cannot be identified using forms of regular expressions. In this regard, a preliminary research program of the U.K. National Archives in this area appears highly relevant [19].

Third, archivists should test software for how well it performs in categorizing or isolating documents within the scope of FOIA exemptions 4, 6, and 7 (and corresponding state law freedom of information exemptions).

In connection with these efforts, archivists and government officials should determine as a matter of policy what tolerance limits exist in allowing presidential and federal records to be placed online where there may be residual uncertainty over whether PII or sensitive content existing within a collection has been successfully masked. Precedent exists for placing collections online without federal agency imposition of a zero tolerance rule for PII being exposed: FERC published online the Enron email collection a decade ago, containing a large number of SSNs and other sensitive content. In the interim, the Enron email dataset has functioned as a test collection for information retrieval research, which have included efforts to cleanse it of PII for a stable, public-use test set [12,15,20].

Any effort to use analytical means for satisfying access demands will necessarily involve the "triaging" of e-mail and other electronic records holdings, for purposes of isolating sub-collections for earlier public access. Archivists have familiarity with the concept of triaging and should be comfortable with prioritizing their efforts in carrying out systematic review [21].

V. Conclusion

The flood of digital public records in the form of successive waves of White House email collections from the Reagan through the Obama Administrations is only the beginning for NARA. Soon, sometime after 2019, the full effect of the 2012 Directive will begin to be felt, transforming NARA from an agency with primarily traditional hard copy holdings in records centers and regional and headquarters archives open and available to public access, into an agency with 99% of its individual record holdings in digital form, constituting dark, inaccessible archives absent the ad hoc FOIA or litigation request. Other public sector institutions face a similar problem of vast accessions of digital records containing sensitive content. Technology-assisted solutions exist that can be employed by archivists, working together with experts in information retrieval and computer science, to find sensitive content and mask it for the purpose of providing online access to the public many decades earlier than waiting 75 years or more.

References

[1] Office of Management and Budget and National Archives and Records Administration, "Managing Government Records Directive, M-12-18" (August 24, 2012), https://www.whitehouse.gov/sites/default/files/omb/memoranda/2012/m-12-18.pdf

[2] National Archives and Records Administration, "Revised Format Guidance for the Transfer of Permanent Electronic Records" (Jan. 31, 2014), https://www.archives.gov/records-mgmt/bulletins/2014/2014-04.html.

[3] E. McCallister et al., "Guide to Protecting the Confidentiality of Personally Identifiable Information (PII) : Recommendations of the National Institute of Standards and Technology," Pub. 800-122 (Apr. 2010), http://csrc.nist.gov/publications/nistpubs/800-122/sp800-122.pdf.

[4] R. Pierce-Moses,"Caught in the Middle: Access to State Government Records in the United States," Japan-U.S. Archives Seminar (May 2007), http://www.archivists.org/publications/proceedings/accesstoarchives/09_Richard_Pearce-MOSES.pdf

[5] Society of American Archivists, SAA Code of Ethics (Jan. 2012), http://www2.archivists.org/statements/saa-core-values-statement-and-code-of-ethics.

[6] National Archives and Records Administration, "Report on Alternative Models for Presidential Libraries Issued in Reponse to the Requirements of PL 110-404 (Sept. 25, 2009), http://www.archives.gov/presidential-libraries/reports/report-for-congress.pdf

[7] G. Paul and J.R. Baron, "Information Inflation: Can the Legal System Adapt?," 13 RICHMOND JOURNAL OF LAW AND TECHNOLOGY 10 (2007), http://jolt.richmond.edu/v13i3/article10.pdf

[8] The Sedona Conference ®, The Sedona Conference Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery," 15 THE SEDONA CONFERENCE JOURNAL 217 (2014), http:/www.thesedonaconference.org/publications

[9] M.R. Grossman and G.V. Cormack, "Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review," 17 RICHMOND JOURNAL OF LAW AND TECHNOLOGY 3 (2011)

[10] D. Oard and W. Webber, "Information Retrieval and E-Discovery," FOUNDATIONS AND TRENDS IN INFORMATION RETRIEVAL, Vol. 7, Nos. 2–3: 99–237 (2013), https://terpconnect.umd.edu/~oard/pdf/fntir13.pdf

[11] J.R.Baron, R.C.Losey, and M.D.Berman, M.D. (eds.), PERSPECTIVES ON PREDICTIVE CODING AND OTHER ADVANCED SEARCH TECHNIQUES FOR THE LEGAL PRACTITIONER (American Bar Ass'n 2016) (in press)

[12] National Institute of Standards and Technology Text REtrieval Conference, Legal Track Annual Reports (2006-2011), https://trec-legal.umiacs.umd.edu/.

[13] B.B. Borden and J.R. Baron, "Finding The Signal in the Noise: Information Governance, Analytics, and The Future of Legal Practice," 20 RICHMOND JOURNAL OF LAW AND TECHNOLOGY 7 (2014), http://jolt.richmond.edu/v20i2/article7.pdf

[14] D.F. Gleich, et al., "Some computational tools for digital archive and metadata maintenance," BIT Numer Math 51: 127-154 (2011), http://stanford.edu/~farnaaz/files/cads.pdf

[15] A. McCallum, A. Corrada-Emmanuel, X. Wang, "Topic and Role Discovery in Social Networks with Experiments on Enron And Academic Email," JOURNAL OF ARTIFICIAL INTELLIGENCE RESEARCH 30:249-272 (2007), ihttp://maroo.cs.umass.edu/pdf/IR-406.pdf

[16] General Inquirer, "Descriptions of Inquirer Categories and Use of Inquirer Dictionaries." http://www.wjh.harvard.edu/~inquirer/homecat.htm (accessed Oct. 10, 2016)

[17] V.L. Lemieux and J.R.Baron, "Overcoming the Digital Tsunami in E-Discovery: Is Visual Analysis the Answer?" CANADIAN JOURNAL OF LAW AND TECHNOLOGY 9, no. 1: 33-48 (2011).

[18] J.R.Baron and S.J.Attfield,"Where Light in Darkness Lies: Preservation, Sensemaking, and Access Strategies for the Modern Digital Archive," UNESCO Memory of the World in the Digital Age Conference: Conference Proceedings, Vancouver, B.C.: 580-595 (2012), http://www.ciscra.org/docs/UNESCO_MOW2012_Proceedings_FINAL_ENG_Compressed.pdf.

[19] UK National Archives, "The application of technology-assisted review to born-digital records transfer, Inquiries and beyond," Research Report (2016), http://www.nationalarchives.gov.uk/documents/technology-assisted-review-to-born-digital-records-transfer.pdf

[20] Electronic Discovery Reference Model (EDRM), "New ENRON Email Data Set," http://www.edrm.net/resources/data-sets/edrm-enron-email-data-set (accessed Oct. 10, 2016).

[21] P. McCarthy, "The Management of Archives: A Research Agenda," AMERICAN ARCHIVIST 51:52 (1988), http://americanarchivist.org/doi/abs/10.17723/aarc.51.1-2.23059213u2350248