

Is NARA Ready Yet? The Newly Extended 2024 Start Date For Accessioning Records into the US National Archives Only in Electronic and Digital Formats, and What That Means

LUSTRE Workshop 2:
AI and Born-Digital Archives in the Government Sector
and Beyond: Challenges and Opportunities
CDDO, The White Chapel Building, London, 4 May
2023

Jason R. Baron
University of Maryland, College of Information Studies
College Park, Maryland





- How will NARA meet the challenge of accessioning especially born digital records when the June 2024 deadline does finally arrive?
- Are freedom of information laws up to the task of providing timely access in the future to those archival records?
- And how can AI tools assist in ensuring that public access to the government's archives remains real and not just aspirational?

A historical
journey....



Iran-Contra PROFS notes on a vintage 1987 desktop computer



From NSOLV --CPLA
To: NSJNP --CPLA

Date and time 12/04/85 02:02:55

*** Reply to note of 08/31/85 13:26

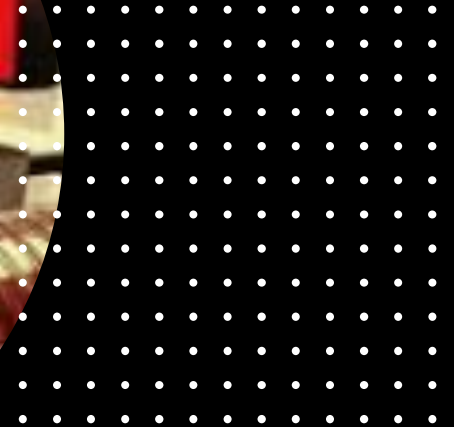
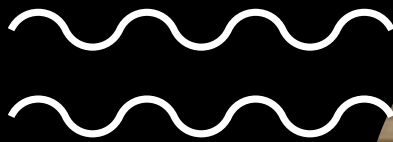
NOTE FROM: OLIVER NORTH

Subject: PRIVATE BLANK CHECK

Subject: Current Status of Operation Recovery: Following summary is based on discussions in Geneva among Kimche, Copp, Gorbenifahr and ~~IRANIAN FIRST CHANNEL B~~ which were concluded this weekend, subsequent discussions with Mendy Meron here in Washington which are continuing, and calls between Copp and Kimche's asset (Nimrod) who is "baby sitting" the Iranians in Paris (where they are are receiving guidance from Tehran).

The attempted transfer ~~through a~~ ~~European country~~ of 18 Hawk missiles went awry because the Iranians were in fact seeking a weapons system that would be capable of stopping Soviet reconnaissance flights along the Iranian/Soviet border and on the Iranian/Iraqi border. Gorba rptd that these flights occur regularly and as deep as 40mi inside Iranian airspace. Because Schwimmer and Ledeen were unfamiliar with the operational parameters of the HAWK, they agreed to ship 120 weapons that were totally inadequate to meet the reqmts established by the Iranians. This delivery has created an atmosphere of extraordinary distrust on the part of the Iranians: in Kimche's view, because the credibility of the Gorba/~~IRANIAN FIRST CHANNEL A AND B~~ mission has probably been seriously called into question.

PROFS Notes on NSC
open reel backup tapes

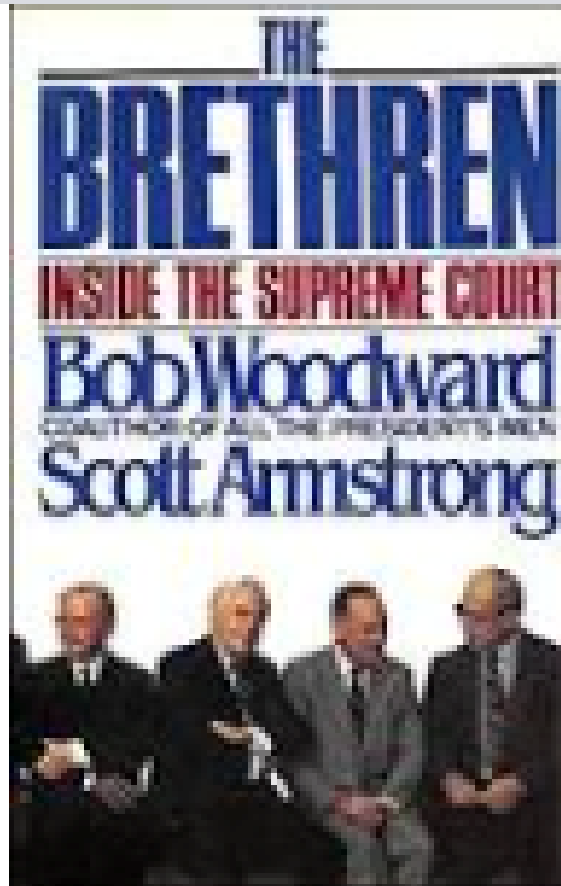


THE “PROFS CASE”

Armstrong v. Reagan (later renamed Armstrong v. EOP)

On the afternoon of the last day of the Reagan Administration, January 19, 1989, author & investigator Scott Armstrong and other plaintiffs filed for a temporary restraining order against the National Security Council, to stop the destruction of backup tapes with Iran Contra related PROFS notes on them.

Over the objections of the Justice Department, the federal district court judge ordered that 362 backup tapes be temporarily preserved beyond the end of the Reagan Administration.





The late Hon.
Charles R. Richey &
Armstrong v EOP

THE NEW YORK TIMES **LAW** FRIDAY, JANUARY 8, 1993



Judge Charles R. Richey has ordered the National Archives to prevent the White House and the National Security Council from destroying computerized records until strict new guidelines can be written and enforced. Judge Richey worked yesterday in his chambers.

Preserving History, and Trivia, in Computer Files

By STEPHEN LABATON
Special to The New York Times

WASHINGTON, Jan. 7 — The office chatter of the Federal Government — at least the chatter conducted via computer — is about to be preserved for the ages.

A plan by the White House to erase at least some of the millions of electronic-mail messages that have accumulated on computer tapes raised at least the idea that there could be something embarrassing or scandalous on them. A ruling by a Federal

The Government's electronic mail as a time capsule.

security adviser, came from electronic mail records. One of the five counts on which Mr. Poindexter was convicted involved the illegal destruction of those records. The convictions of the two men were ultimately reversed by a Federal ap-

role of the White House in billions of dollars worth of loans to Iraq from the Atlanta branch of the Banca Nazionale del Lavoro.

"This is a landmark decision," said Scott Armstrong, the founder of the National Security Archive, who filed the lawsuit with Ralph Nader's Public Citizen group and the American Civil Liberties Union. "It preserves important records, and it brings the entire Federal Government into the 20th century."

Among the Federal offices that some lawyers said had insufficient guidelines were those within the Justice and Labor Departments and sev-

preserve some types of computer correspondence, but they would also like to be able to destroy sensitive information.

"Every president wants to preserve his prerogatives," Mr. Armstrong said.

Judge Richey, who was appointed by President Richard M. Nixon in 1971, said the law requires each agency to impose strict guidelines for preserving historical records, and he accused the archivists of violating his duty of preventing their destruction. The White House had argued that electronic mail sent through its Oasis

ally have to
lent." At Mr. Lan-
sons of
"sub-
diminished
e" and
ake them
conclude
his
ending de-
leemed

including
room" and
ugh, tradi-
were
em,"
nd "mitti-
to add a
take one

so, so does
ly, legal
west en-
ingly, from
ena of sen-
irma mak-
re "date
" had
e."

s contribu-
ler," "air
lity" (to
th penalty
ere are
e, like nar-
eople do.
To "Mir-
ict of his
ber the so-

city in
that isn't
York,"
to destroy
a concert-
back-
in term
e New Re-
page Tri-
l. "There's
of it that's
at Mr.
ted that
Justice

From NSOLV --CPLA

Date and time 12/04/85 02:02:55

To: NSJMP --CPLA

Re: Reply to note of 08/31/85 13:26

NOTE FROM: OLIVER NORTH

Subject: PRIVATE BLANK CHECK

Subject: Current Status of Operation Recovery: Following summary is based on discussions in Geneva among Kimche, Copp, Gorbenifahr and ~~IRANIAN FIRST CHANNEL B~~ which were concluded this weekend, subsequent discussions with Mendy Meron here in Washington which are continuing, and calls between Copp and Kimche's asset (Nimrodi) who is "baby sitting" the Iranians in Paris (where they are receiving guidance from Tehran).

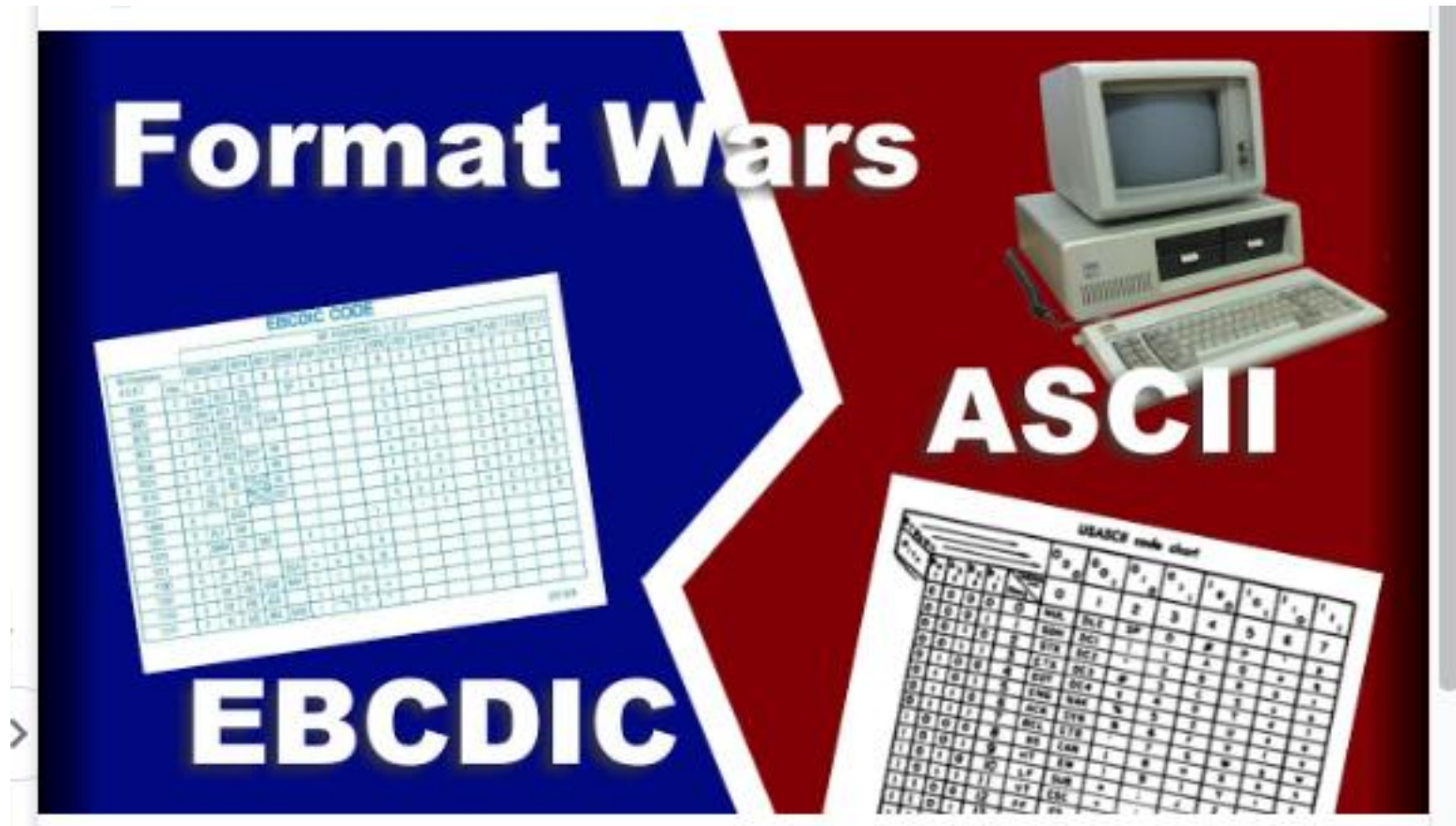
The attempted transfer ^{through a} ~~European country~~ of 18 Hawk missiles went awry because the Iranians were in fact seeking a weapons system that would be capable of stopping Soviet reconnaissance flights along the Iranian/Soviet border and on the Iranian/Iraqi border. Gorba rptd that these flights occur regularly and as deep as 40mi inside Iranian airspace. Because Schwimmer and Ledeen were unfamiliar with the operational parameters of the HAWK, they agreed to ship 120 weapons that were totally inadequate to meet the reqmts established by the Iranians. This delivery has created an atmosphere of extraordinary distrust on the part of the Iranians: in Kimche's view, because the credibility of the Gorba/~~IRANIAN FIRST CHANNEL A AND B~~ mission has probably been seriously called into question.

NARA Issues Email Regulations First Issued in 1995

- Allow print-to-paper preservation policies
- Print-to-paper continued as the dominant method for preservation until 2016



Until expanding transfer formats circa 2003, NARA restricted electronic records transfer formats to EBCDIC and ASCII



Fast Forward





Boxes of presidential records at Mar A Lago: a retro-scandal involving paper records





Here are 99%+ of White House
presidential records

National Archives Wants to Use AI to Improve 'Unsophisticated Search' and Create 'Self-Describing Records'



<https://www.nextgov.com/analytics-data/2021/04/national-archives-wants-use-ai-improve-unsophisticated-search-and-create-self-describing-records/173417/>

Both of these efforts focus on answering three key questions:

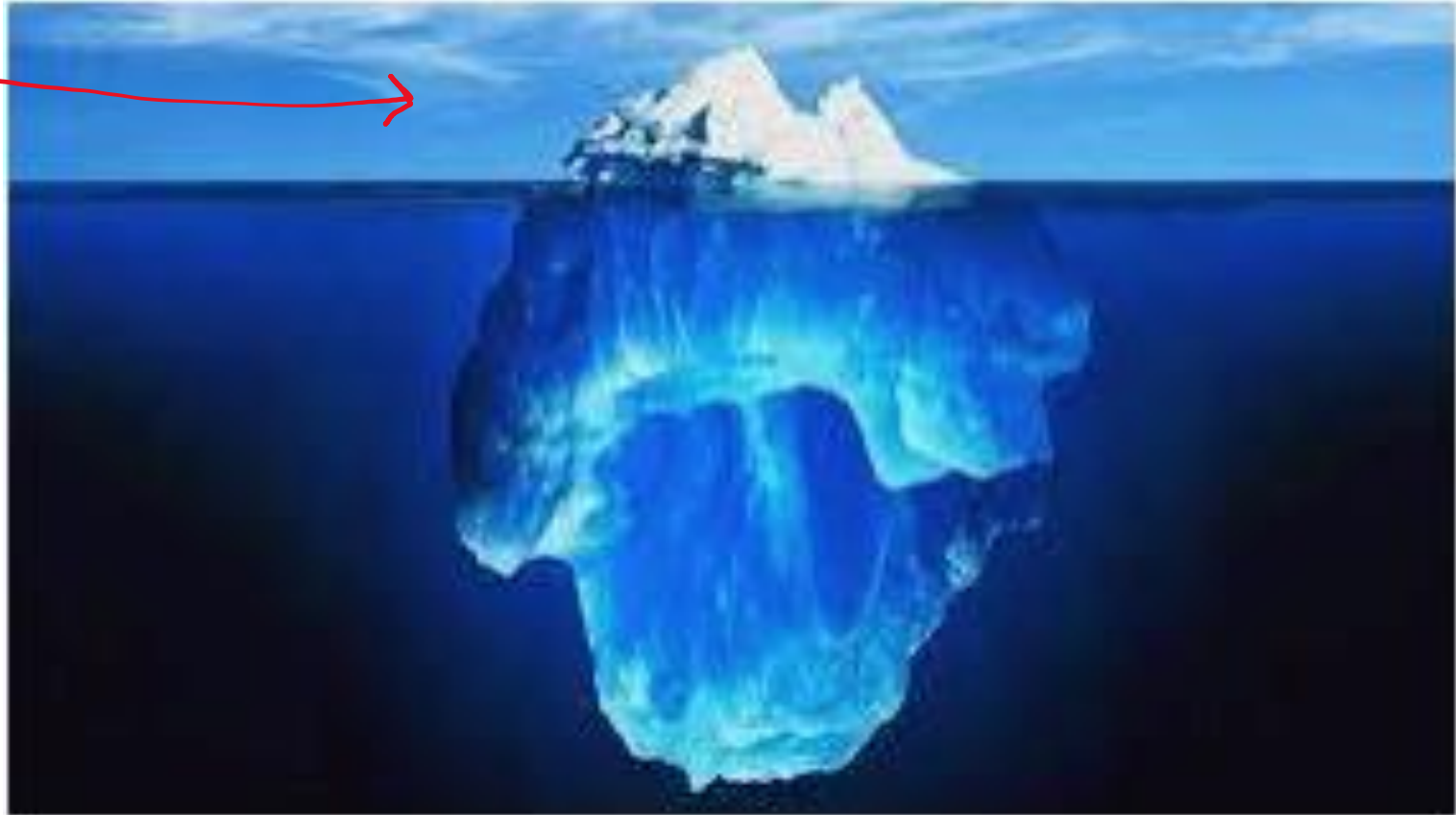
- How can NARA make records easier to find?
- How can NARA make records available more efficiently and quickly?
- How can NARA ensure its data's integrity?

The RFI seeks to answer those questions while investigating the best technical solution and acquisition strategy, including:

- Identifying and addressing data quality issues such as bias.
- Anonymization versus personalization of user-friendly search.
- Algorithms, frameworks and tools for creating AI solutions for the use cases, indicating outcome, addressing the strengths and limitations.
- Comparison of capabilities provided by different cloud providers such as AWS, Azure, IBM, Google, etc.
- Commercial off-the-shelf or cloud-based versus non-cloud based tools or frameworks.
- Related licensing costs, operating capabilities and required support.
- Storage and indexing capabilities agnostic to data format and type.
- Design of pipeline for development and delivery of AI and ML solution to production.
- Post-production activity needed such as infrastructure monitoring, debugging, job orchestration, etc.
- Areas for cost considerations.

A great percentage of unstructured data in public electronic repositories & archives are not from online sources

Last Slide



The
Looming
Problem

TECHNOLOGY

White House extends e-records deadline to June 2024



Justin Doubleday | @jdoubledayWFED

December 23, 2022 3:22 pm ⌚ 4 min read



A longtime goal to transition to a paperless government is once again a moving target, with the White House now giving agencies until the end of June 2024 to fully transition to electronic recordkeeping.

A 2019 directive had set a deadline of Dec. 31, 2022, for when the National Archives and Records Administration would stop accepting paper records from agencies.

Past and Future Deadlines in Transitioning to Electronic Recordkeeping



NARA will only accept born digital or digitized records after
December 31, 2019 (2012 Managing Govt Records Directive)
December 31, 2022 (2019 Transition to Electronic Records Memo)
June 30, 2024 (2022 Update to Transition to Electronic Records Memo)

Agencies since January 1, 2017 have been required to manage their **email** electronically (all email, both temporary and permanent in nature)



THE DIRECTOR

EXECUTIVE OFFICE OF THE PRESIDENT
OFFICE OF MANAGEMENT AND BUDGET
WASHINGTON, D.C. 20503



THE ARCHIVIST

December 23, 2022

M-23-07

MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES

FROM:

Shalanda D. Young

Director

Office of Management and Budget

Debra Steidel Wall

Acting Archivist of the United States

National Archives and Records Administration

SUBJECT:

Update to Transition to Electronic Records

Strong records management is necessary for transparency and accountability and underpins our democracy. Transitioning Federal agencies¹ to an electronic — or “paperless” — environment is a priority to enable and increase the ability of the public to engage with Government in new and more efficient and effective ways. It is critical that Federal agencies move beyond paper-based processes and embrace the opportunities afforded to improve Government by transitioning fully to an electronic environment.

Memorandum M-19-21, *Transition to Electronic Records*, issued jointly by the Office of Management and Budget (OMB) and the National Archives and Records Administration (NARA), established several requirements to move Federal agencies to a fully electronic environment, where appropriate. Agencies may have been delayed in their electronic records transition process during the COVID-19 pandemic. This memorandum reinforces the requirements established in M-19-21, reaffirms the underlying goal of the transition to electronic records, and updates the previous target dates described in M-19-21. This memorandum also describes NARA’s plans to issue updated guidance, as appropriate.



THE DIRECTOR

EXECUTIVE OFFICE OF THE PRESIDENT
OFFICE OF MANAGEMENT AND BUDGET
WASHINGTON, D.C. 20503



THE ARCHIVIST

December 23, 2022

M-23-07

MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES

FROM:

Shalanda D. Young

Director

Office of Management and Budget

Debra Steidel Wall

Acting Archivist of the United States

National Archives and Records Administration

SUBJECT:

Update to Transition to Electronic Records

Strong records management is necessary for transparency and accountability and underpins our democracy. Transitioning Federal agencies¹ to an electronic — or “paperless” — environment is a priority to enable and increase the ability of the public to engage with Government in new and more efficient and effective ways. It is critical that Federal agencies move beyond paper-based processes and embrace the opportunities afforded to improve Government by transitioning fully to an electronic environment.

Memorandum M-19-21, *Transition to Electronic Records*, issued jointly by the Office of Management and Budget (OMB) and the National Archives and Records Administration (NARA), established several requirements to move Federal agencies to a fully electronic environment, where appropriate. Agencies may have been delayed in their electronic records transition process during the COVID-19 pandemic. This memorandum reinforces the requirements established in M-19-21, reaffirms the underlying goal of the transition to electronic records, and updates the previous target dates described in M-19-21. This memorandum also describes NARA’s plans to issue updated guidance, as appropriate.



FEDERAL NEWS NETWORK

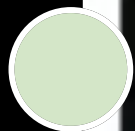
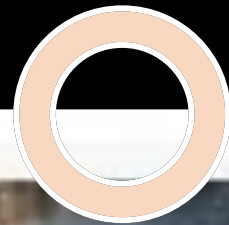
- The delay to the electronic record keeping deadline isn't surprising, according to Jason R. Baron, professor at the University of Maryland and former director of litigation at NARA.
- "It has been doubtful for some time whether the entire federal government could transition to electronic recordkeeping by the last day of 2022," Baron told Federal News Network. "Although the memorandum cites COVID-19 as a possible reason for agencies needing more time, I suspect that the real underlying reasons for delays involve the limited resources agencies face, and the complexity of the issues involved. These systemic issues will not be going away any time soon, and it remains to be seen whether agencies will be able to meet even the extended deadlines."



2022 Memo

- Real reasons for missing the target:
 - Failure on part of the federal government to be “ready” due to resources & bureaucratic inertia
 - Complexities of electronic recordkeeping
- Will the June 2024 deadline be complied with?
- What does the future look like in 2030 and beyond?





From Reagan to Trump
Over 600 million individual emails with
attachments
(~2 billion pages)

the 1990s, the number of people in the world who are under 15 years of age is expected to increase from 1.1 billion to 1.5 billion. The number of people aged 65 and over is expected to increase from 200 million to 400 million. The number of people aged 15 and over is expected to increase from 3.5 billion to 4.5 billion. The number of people aged 15 and over is expected to increase from 3.5 billion to 4.5 billion. The number of people aged 15 and over is expected to increase from 3.5 billion to 4.5 billion.

WH email in NARA's legal custody that are open and available for public access*

• Reagan era Iran-Contra emails	5,000
• Supreme Court Justice-related emails	
• John Roberts-related emails	60,000
• Elana Kagan-related emails	75,000
• Brett Kavanaugh-related emails	170,000
• Other emails released in litigation or FOIA	~100,000
TOTAL RELEASED	~410,000
TOTAL ACCUMULATED WH EMAILS	~600,000,000+
PERCENTAGE OPEN:	0.08%

* Note: figures represent distinct emails, not total pages. Page equivalents of emails & attachments > 2 billion (est.).

WH email in NARA's legal custody that are open and available for public access*

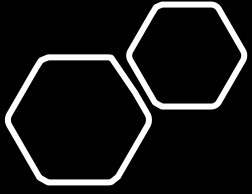
• Reagan era Iran-Contra emails	5,000
• Supreme Court Justice-related emails	
• John Roberts-related emails	60,000
• Elana Kagan-related emails	75,000
• Brett Kavanaugh-related emails	170,000
• Other emails released in litigation or FOIA	~100,000
TOTAL RELEASED	~410,000
TOTAL ACCUMULATED WH EMAILS	~600,000,000+
PERCENTAGE OPEN:	0.08%

* Note: figures represent distinct emails, not total pages. Page equivalents of emails & attachments > 2 billion (est.).

Capstone Email Policy



- Voluntary policy for US agencies to meet NARA 2016 mandate for management of email in electronic form
- Senior officials' emails deemed "permanent"; all other employees' emails retained for 7 years
- 250+ components of the federal government have elected to follow policy



Capstone email repositories

- Depending on the agency, senior official emails may already number in the hundreds of thousands or millions
- All other employees' emails may amount to millions, tens or hundreds of millions
- By 2030, billions of emails accumulated across entire Executive branch
- Capstone repositories soon will be broadened to include other forms of electronic messaging,



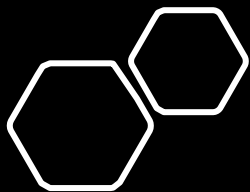
The Secret Service's deleted text message scandal, explained

The mysterious — and maybe criminal — disappearance of the Secret Service's January 6 text messages.

By Ben Jacobs | Updated Aug 1, 2022, 5:29pm EDT

[f](#) [t](#) [SHARE](#)





Post-2024:

Billions of digital e-mail
records hiding in plain
sight



FOIA is
Broken



Largely Manual Workflows Leading to Huge Delays, Frustration

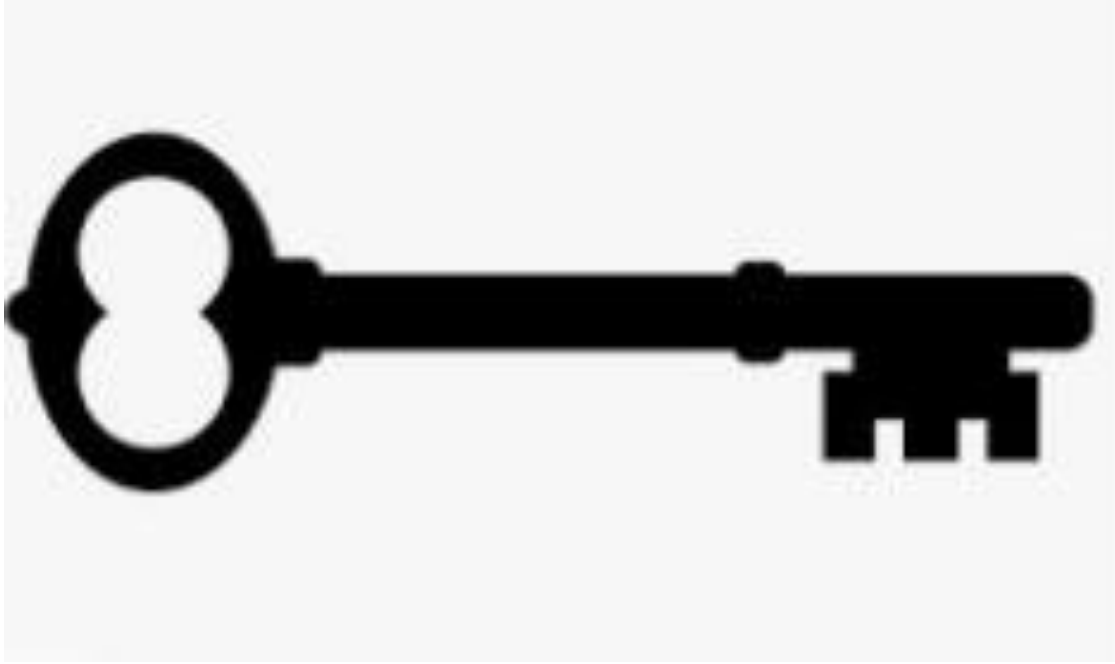


FBI Headquarters building as seen on August 23, 2016. | John Shinkle/POLITICO

Judge balks at FBI's 17-year timeline for FOIA request

By JOSH GERSTEIN | 07/29/2017 01:17 PM EDT

<https://www.politico.com/blogs/under-the-radar/2017/07/29/judge-balks-fbi-foia-timeline-17-years-241127>



Key to Unlocking Dark Archives:
AI searches + Filtering for Sensitivities

A need for archivists, records managers, and lawyers to filter sensitive content of all types from public records before access is granted



Categories of Personal Information (not exhaustive)

PII

- Names as metadata fields
- Social security numbers
- Telephone numbers
- Passport information
- Bank and financial information
- Credit card numbers
- Vehicle registrations
- Date of birth
- Height and Weight
- Asset information

Sensitive Personal Information

- Medical history
- Criminal history
- Sexual orientation
- Racial or ethnic origin
- Religious beliefs
- Political beliefs
- Mental health
- Genetic or biometric data

Dark Archives

- Sensitive content consisting of personal information closed between 100 and 110 years (less the age of the individual, if known). If the age of the individual is not known, for minors it is closed for the whole period, and for those deemed to be over the age of sixteen, 80 to 94 years.
- Source: Moss, Michael S. and Gollins, Tim J. (2017) "Our Digital Legacy: an Archival Perspective," *Journal of Contemporary Archival Studies*, vol.4, art. 3, <http://elischolar.library.yale.edu/jcas/vol4/iss2/3>
- Rule at the US National Archives is a presumption of closure for 75 years from date of creation of the document

“Technology Assisted Review” (TAR)

- A process for prioritizing or coding a collection of electronic documents using a computerized system that harnesses human judgments of one or more subject matter experts on a smaller set of documents and then extrapolates those judgments to the remaining document population.
- Also referred to as “supervised or active machine learning” or “computer-assisted review”
- Efficiently searches vast record collections in hours/days rather than weeks/months/years

Source: Adapted from Grossman-Cormack Glossary of Technology Assisted Review, v. 1.0 (Oct 2012)



Propositions:

- (1) The experience lawyers have with using machine learning goes a long way to demonstrating that algorithms used for relevance and the filtering of sensitive content can be successfully applied to a large archival repository
- (2) We have no alternative but to rely on algorithmic ways if we wish to carry out a central mission of allowing access to the world's accumulated knowledge.

Home > ACM Journals > Journal on Computing and Cultural Heritage > Vol. 15, No. 1 > Providing More Efficient Access to Government Records: A Use Case Involving Application of Machine Learning to Improve FOIA Review for the Deliberative Process Privilege

RESEARCH-ARTICLE [PUBLIC ACCESS](#)

Providing More Efficient Access to Government Records: A Use Case Involving Application of Machine Learning to Improve FOIA Review for the Deliberative Process Privilege

Authors: [Jason R. Baron](#), [Mahmoud F. Sayed](#), [Douglas W. Oard](#) [Authors Info & Claims](#)



Test Collection:
Records of White House Assistant to the President for Domestic Policy Elena Kagan,
from the Clinton Presidential Library online database

Using a Classifier to Segregate Out Sensitive Content

- We show that when classifiers are trained and tested under consistent conditions it is possible to design classifiers that achieve F_1 measures between 70% and 83% (i.e., if tuned so that precision and recall were equal, we would expect that between 70% and 83% of the exempt material would be found, and that the same fraction of the content identified by the classifier as exempt would truly be exempt).
- We study the effects of differences between reviewers, between the materials held by different custodians, and within the topical content of the records being classified to identify which differences pose the greatest challenge for current text classifiers.
- We control for the effect of document type and recognizable characteristics of content items to study classifier effectiveness on the content and document types that human reviewers find most difficult.
- We suggest directions for future work, identifying a need to model contextual factors that require access to evidence beyond the boundaries of specific documents.
- We introduce a new freely distributable test collection that is annotated for the deliberative process privilege under exemption 5 of the FOIA.¹¹

J.R. Baron, M. Sayed & D. Oard, *Journal on Computing and Cultural Heritage*, 15:1, article 5: 1-19 (2022), <https://dl.acm.org/doi/full/10.1145/3481045>

Using a Classifier to Segregate Out Sensitive Content

- We show that when classifiers are trained and tested under consistent conditions it is possible to design classifiers that achieve F_1 measures between 70% and 83% (i.e., if tuned so that precision and recall were equal, we would expect that between 70% and 83% of the exempt material would be found, and that the same fraction of the content identified by the classifier as exempt would truly be exempt).
- We study the effects of differences between reviewers, between the materials held by different custodians, and within the topical content of the records being classified to identify which differences pose the greatest challenge for current text classifiers.
- We control for the effect of document type and recognizable characteristics of content items to study classifier effectiveness on the content and document types that human reviewers find most difficult.
- We suggest directions for future work, identifying a need to model contextual factors that require access to evidence beyond the boundaries of specific documents.
- We introduce a new freely distributable test collection that is annotated for the deliberative process privilege under exemption 5 of the FOIA.¹¹

J.R. Baron, M. Sayed & D. Oard, Journal on Computing and Cultural Heritage, 15:1, article 5: 1-19 (2022), <https://dl.acm.org/doi/abs/10.1145/3481045> (pre-print at <https://arxiv.org/abs/2011.07203>)



How well does ChatGPT 3.5 do in making FOIA determinations on withholding, and explaining its decisions?



“More Product, Less Process”

Mark A. Greene and Dennis Meissner,
“More Product Less Process:
Revamping Traditional Archival
Processing,” *American Archivist* 68:2
(2005)

- Waiting for every individual named in records to pass away before opening vast amounts of records is a strategy, but in my view not a very good one.
- We should continue to promote ML/TAR methods that efficiently find responsive records and accurately segregate personal content as best we can
- We should explore how generative AI can provide narratives explaining why documents or portions of documents have been withheld under freedom of information laws



Letting go of
a reluctance
to embrace
the AI world

Impenetrable



...but think of the black box as a “gift” to archivists to improve access to vast digital collections.





COLLEGE OF INFORMATION STUDIES

Jason R. Baron
University of Maryland
College of Information Studies
jrbaron@umd.edu