Search Among Secrets

Douglas W. Oard
iSchool and UMIACS
University of Maryland, College Park, USA
On Christmas Eve, 1910, future president Warren G. Harding got out a photograph of himself, and on the back wrote an impassioned love note to his mistress.

“My Darling,” he began. “There are no words, at my command, sufficient to say the full extent of my love for you — a mad, tender, devoted, ardent, eager, passion-wild, jealous ... hungry ... love ...”

“It flames like the fire and consumes,” Harding, 45, who was married and would be elected the 29th president a decade later, wrote. “It racks in the tortures of aching hunger, and glows in bliss inexpressible — bliss only you can...
Held at The Library of Congress under court-ordered seal for the last 50 years, the trove of the original, often-juicy letters is scheduled to be opened to the public via the Internet on July 29.

The library is hosting a program to discuss the collection July 22, featuring library and outside experts and a Harding descendant.

The original Harding letters were sealed July 29, 1964, the library said. The Harding family donated them to the library in 1972, with the stipulation that they stay sealed until July 29, 2014.

But forgotten microfilm copies were discovered in an Ohio historical repository a decade ago by an author researching the president, and he used many of them in a detailed but little-known book.
Search Among Secrets: Separating the Wheat from the Buzzsaw

Douglas W. Oard

iSchool and UMIACS

University of Maryland, College Park, USA
Sensitive Content

• Privacy-sensitive personal interactions
  – Phone, email, SMS, ...

• Legally protected information
  – Doctor-Patient, Attorney-Client, Trade secret, ...

• National security
Outline

➢ The need for search among secrets

• Batch search among secrets

• Interactive search among secrets
Open Government

“Information maintained by the Federal Government is a national asset. My Administration will take appropriate action, consistent with law and policy, to disclose information rapidly in forms that the public can readily find and use.”

Presidential Memorandum, January 21, 2009
Different Errors Have Different Costs

• Incorrect withholding results in foregone value

• Incorrect disclosure can be the greater disaster

• Result: we withhold collections *because* we lack sufficient confidence that we can prevent:
  – Potentially harmful *disclosure* of items, or of specific types of elements in those items
  – Potentially harmful *inferences* at the collection, item, or element level
The “Mosaic Theory”

**Iceland.** Iceland is another “non-nuclear” country whose nuclear history remains incomplete. In Appendix B, Iceland is clearly the first blacked out country listed after Hawaii and before Johnston Island. Non-nuclear components were stored at the American base at Keflavik for a decade, from February 1956 to June 1966, and complete nuclear bombs were deployed there from September 1956 to September-December 1959.

Norris et al. (1999), Where They Were, Bulletin of Atomic Scientists, 55(6), 25-35
<table>
<thead>
<tr>
<th>COUNTRY</th>
<th>WEAPON</th>
<th>INITIAL ENTRY</th>
<th>WITHDRAWN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Guam (cont.)</td>
<td>Talos</td>
<td>Jul 65</td>
<td>Jun 69</td>
</tr>
<tr>
<td></td>
<td>Astor</td>
<td>Nov 65</td>
<td>Mar 74</td>
</tr>
<tr>
<td></td>
<td>ASROC</td>
<td>Jan 66</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Terrier</td>
<td>Mar 66</td>
<td>Jan 67</td>
</tr>
<tr>
<td></td>
<td>155mm Howitzer</td>
<td>May 66</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Polaris</td>
<td>Jul 66</td>
<td>Aug 66</td>
</tr>
<tr>
<td></td>
<td>Nike Hercules</td>
<td>Jun 68</td>
<td>Jun 69</td>
</tr>
<tr>
<td>Hawaii</td>
<td>Bomb</td>
<td>Jul 54</td>
<td>Jun 69</td>
</tr>
<tr>
<td></td>
<td>Depth Bomb</td>
<td>Dec 55-Feb 56</td>
<td>Jan-Mar 65</td>
</tr>
<tr>
<td></td>
<td>Regulus</td>
<td>Mar-May 56</td>
<td>Apr-Jun 63</td>
</tr>
<tr>
<td></td>
<td>Boar</td>
<td>Sep-Nov 56</td>
<td>Jun 75</td>
</tr>
<tr>
<td></td>
<td>Honest John</td>
<td>Jun-Aug 57</td>
<td>Jun 72</td>
</tr>
<tr>
<td></td>
<td>8-inch Howitzer</td>
<td>Oct-Dec 58</td>
<td>Jun 75</td>
</tr>
<tr>
<td></td>
<td>ADM</td>
<td>Jan-Mar 59</td>
<td>Oct-Dec 64</td>
</tr>
<tr>
<td></td>
<td>Hotpoint</td>
<td>Jan-Mar 60</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Nike Hercules</td>
<td>Jul-Sep 60</td>
<td>Jun 73</td>
</tr>
<tr>
<td></td>
<td>Little John</td>
<td>Apr-Jun 62</td>
<td>Oct 68</td>
</tr>
<tr>
<td></td>
<td>Talos</td>
<td>Oct-Dec 63</td>
<td>Aug 68</td>
</tr>
<tr>
<td></td>
<td>Astor</td>
<td>Apr-Jun 64</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Davy Crockett</td>
<td>Oct-Dec 64</td>
<td>Jun 69</td>
</tr>
<tr>
<td></td>
<td>155mm Howitzer</td>
<td>Mar 65</td>
<td>Jun 75</td>
</tr>
<tr>
<td></td>
<td>Terrier</td>
<td>Aug 65</td>
<td>Sep 66</td>
</tr>
<tr>
<td></td>
<td>Subroc</td>
<td>May 66</td>
<td>Jun 67</td>
</tr>
<tr>
<td></td>
<td>Falcon</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Nonnuclear Bomb</td>
<td>Feb 56</td>
<td>Jun 66</td>
</tr>
<tr>
<td>Johnston Is.</td>
<td>Nonnuclear Bomb</td>
<td>Sep 56</td>
<td>Sep-Dec 59</td>
</tr>
<tr>
<td></td>
<td>Thor</td>
<td>Dec 54-Feb 55</td>
<td>Jun 65</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Jul-Sep 64</td>
<td>Jun 71</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Jul-Dec 63</td>
<td>Jul 66</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Jul 61</td>
<td>Jun 65</td>
</tr>
<tr>
<td>Midway</td>
<td>Depth Bomb</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Nonnuclear Bomb</td>
<td>Jul-Sep 53</td>
<td>Jun 65</td>
</tr>
<tr>
<td></td>
<td>Bomb</td>
<td>May 54</td>
<td>Sep 63</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sep-Nov 57</td>
<td>Mar 61</td>
</tr>
</tbody>
</table>
One example of USSR's determination not to interfere in internal affairs of other countries is Iran, an ally of the United States. The Soviet Union does not want a revolution there and does not do anything in that country to promote such a development. However, the people of that country are so poor that the country has become a volcano and changes are bound to occur sooner or later. The Shah will certainly be overthrown. By supporting the Shah, the United States generates adverse feelings toward the United States among the people of Iran and, conversely, favorable feelings toward the USSR. This, of course, is to the US's own disadvantage. The Soviet Union does not sympathize with dictators or tyranny. This is the crux of the matter. No agreement seems to be possible on this
Outline

• The need for search among secrets

• Batch search among secrets
  ➢ Transparency in government
    – Privilege review in e-discovery

• Interactive search among secrets
Present Practice

Years since creation

Reviewed Fraction

100%
ISSUE NO. 2: Prioritizing the Declassification Review of Historically Significant Information.

There is no satisfactory means at present of identifying historically significant information within the vast body of information that is being reviewed and declassified. Accordingly, no priority is given to the declassification and release to the public of such information.
“Value-Sensitive” Prioritization

Such a system might operate as follows: A board consisting of prominent historians, academicians, and former Government officials would be appointed by the Archivist to determine which events or activities of the U.S. Government should be considered historically significant from a national security and foreign policy standpoint for a particular year.

PIDB Improving Declassification report, December, 2007
Value Estimation

• Evidence for scholarly interest
  – Scholar-specified criteria
  – Scholarly literature
  – FOIA and Mandatory Declassification requests
    • Both the requests and the requested materials

• Evidence for gaps in the scholarly record
  – Passage-level novelty
  – Contradiction
Effect of *Early* Value Sensitive Review

![Graph showing the effect of early value sensitive review over years since creation.](image-url)
The “Mosaicing” Chain

• Information
  – Existing (public and private)
  – Newly declassified
  – Future creation or release

• Aggregation
  – Physical access (digitized?, networked?)
  – Discovery

• Inference
Releasability Estimation

• Declassification guidance
  – Content
  – Classification markings

• Evidence from relationships
  – Decisions on other versions of the same document
  – Decisions on quoted content
  – Decisions on similar documents
Open Source Mosaicing

• (Reasonably) Comprehensive Collection
  – Public documents
  – Declassified documents

• Link detection
  – “Wikification” on steroids

• Knowledge-Base contruction

• Inference
  – Automated inference
  – Automated support for human inference
Outline

• The need for search among secrets

• Batch search among secrets
  – Transparency in government
    ➢ Privilege review in e-discovery

• Interactive search among secrets
E-Discovery

Clinton White House
32 million emails

National Archives
hired 25 persons

search request

Tobacco Policy
80,000

200,000

for 6 months …
An E-Discovery Process Model

- Formulation
  - Production request
  - Acquisition
  - Collection
  - Review for Relevance
  - Responsive ESI
  - Review for Privilege
  - Production
  - Sense-making
  - Insight
Sampling Error: Downside Risk

Topic = GCRIM ; Frequency = 4.01%

Annotation Budget (80% training / 20 % test)

RCV-1v2, SVMperf optimized for F1
Measurement Error: The Larger Risk
Outline

• The need for search among secrets

• Batch search among secrets

➢ Interactive search among secrets
FOIA Guidance

“[identify the subject(s) or record(s) as clearly and specifically as possible -- for example, all previously released National Intelligence Estimates (NIEs) on the former Soviet Union's space program].”

Sudheendra Hangal, Providing Access to Historical Email Archives for Historical Research, *Personal Digital Archiving*, 2013
Date: March 2, 2001 11:24am
From: Katbarce@
To: creeley@
Subject:

... Robert,

..... Central Michigan (.....

..... )

.....

.....

..... Portland

..... $...

..... $...

.....!

.....

.....

Kathryn

Kathryn Barcos
The Steven Barclay Agency

.....

Petaluma, ......
Thinking More Generally

• Things people want:
  – Access to collections
  – Access to specific items in those collections
  – Access to specific elements of those items

• Ways we can provide that:
  – Give them the collection
  – Provide them with a service
  – Provide them with an experience
What Needs to be Withheld?

• Some specific items
  – e.g., privileged documents in e-discovery

• Some specific types of elements
  – e.g., PII in medical records

• Some types of inference
  – e.g., in declassification
Some Implications

• A challenge of increasing scale
  – Caused by intermixing
  – Natural byproduct of informal interaction
  – Nearly every word produced each day is interactive

• Technology has implications for policy
  – Without technology we withhold nearly everything
    • Paradoxically, ability to withhold makes more open
  – Requires accepting some risk
    • We need to be able to characterize that policy