Public Interest Declassification Board
Minutes of the Meeting
September 23, 2010

Summary of the Public Interest Declassification Board (PIDB) Meeting, Thursday, September 23, 2010 held in the Jefferson Room of the Conference Center at the National Archives Building in Washington, DC.

Members in attendance: Acting Chairman Martin Faga, Herbert Briick, Elizabeth Rindskof Parker, Jennifer Sims, David Skaggs, and Sanford Ungar. Also present: William J. Bosanko, serving as the Executive Secretary of the PIDB; John Powers, John Bell, Evan Coren and Neena Sachdeva, ISOO, serving as the PIDB staff. Also present were about 60 members of the public.

Guest Speaker: Jeff Jonas

Mr. Jonas presented his analysis of the challenge currently facing the National Declassification Center (NDC) as it seeks to review 410 million pages of records for public access by a Presidentially mandated deadline of December 31, 2013. Machine triage[1] is necessary. A system which accurately predicts declassification dispositions is also necessary. The best technological application for mass declassification would be a context accumulating system. Context means better understanding of surrounding conditions; that is, contextualizing relevance to the human reviewer. Context accumulation (through the ingestion of documents) increases predictability and helps sort the queue. To illustrate, Mr. Jonas used the metaphor of a jigsaw puzzle: the picture emerges by connecting the pieces. As more pieces are added, more realistic or predictable outcomes should occur. However, the pieces accumulated in context have many variables: incomplete, low quality, misinterpreted, misrepresented, duplicate, or missing information; no complete picture of the puzzle exists. Context accumulates observation. Unassociated, proximal, or connected assertions make it possible for the system, given new observations, to contradict or reverse earlier assertions. The system learns nuanced distinctions: more data would mean better, higher quality predictions.

Two policy questions emerged. What information exists in open sources and what damage or benefits could release incur? As the system learns through context accumulation and expert counting[2] it should better predict the disposition of a document (release or deny in full or in part). The second problem common within any organization is "enterprise amnesia" where one part of an organization has information that another

[1] Loosely defined as the "review of data by computers and selection of information for review by humans based on selection criteria," H. Bryan Cunningham testimony before US Senate Committee on the Judiciary, September 25, 2007. Available at: http://www.fas.org/irp/congress/2007_hr/092507cunningham.html

[2] See Mr. Jonas' article titled, *Smart Sensemaking Systems, First and Foremost, Must be Expert Counting Systems*, Proceedings of the International Risk Assessment and Horizon Scanning Symposium 2010 (IRAHSS). Expert counting recognizes when multiple references to the same entity are in fact the same entity; this is accomplished by counting discreet entities, including duplicate, inconsistent, and incorrect data.

part of the organization needs to know but does not know or cannot access. To nullify this problem, the context accumulation system ingests previously released and denied records to build its knowledge base.

Mr. Jonas presented a strawman architecture, an automated and human workflow system to manage declassification activity. The system would ingest data points consisting of 410 million documents, historical dispositions, previously released or denied declassification decisions, dirty words, open source material, etc. Once ingested, these data points would be fed through the extraction and classification process which contextually accumulates these data points to produce the predictive outcome. The prediction component sorts the documents into respective bins: classified, declassified documents, and those requiring human decisions. This workflow method assists the human reviewer in determining the disposition of the document. Disposition loops back to the front end (extraction and classification process) providing context for better predictions. To further improve the system and achieve effective workflow, "crowdsourcing"[3] may be employed; for example, a mix of subject matter experts (SMEs) from various Government agencies could collaborate to minimize poor classification decisions. The idea of an internal wiki used by SMEs to adjudicate declassification problems was also suggested. Mr. Jonas proposed using the declassification platform to assist in the classification process (the front end). The classification methodology would "pre-tag" documents to assist reviewers and the system in future classification and declassification decisions.

Mr. Jonas remarked that in relative terms, the 410 million pages was not a significant amount of data for a system to ingest. Further, building a knowledge base of the denied and released documents would not require any elaborate hardware.

Mr. Jonas concluded with four solutions to surmount the mass declassification challenge that the NDC faces: 1) context must provide sound predictions; 2) human action alone cannot overcome the volume of materials to be processed; 3) "human directing systems" are imperative to the daunting task of mass declassification; and 4) "data must find data." In other words, reviewers do not have be the repository of all classification or declassification knowledge, rather the system directs them to the relevant, historical information ("relevance must find the user").

Chairman Faga opened the session to questions and comments from the Board and audience. The Board engaged with Mr. Jonas regarding the theoretical basis of the proposed system and its validity in the information security work environment. A few themes emerged:

- creating a centralized mosaic system could involve risk to national security if transnational adversaries were to use US declassified information to plan attacks;
- there is a real threat of espionage vulnerabilities to any integrated system;

[3] Crowdsourcing a neologism combining crowd and outsource was coined by Jeff Howe, a contributing editor at Wired Magazine. The term is defined as the act of taking a job traditionally performed by a designated agent (usually an employee) and outsourcing it to an undefined, generally large group of people in the form of an open call. See: http://crowdsourcing.typepad.com/cs/2006/06/crowdsourcing_a.html

- to account for the Federal Government contracting cycle, the system could be developed incrementally;
- ingesting classification and declassification guides would be required to initialize the process;
- there may be difficulties in coordinating disparate systems across agencies.

Mr. Jonas addressed both the threat of espionage and risks of exposing national security information as a mosaic. These concerns would be diminished by "pre-tagging" documents with the appropriate classification controls. He used the analogy of a library card to limit or permit access to the human reviewer. In addressing the problem of Agency's lack of cooperation, a technique he developed would "anonymize" (make anonymous) data – the reviewer would not have access to the actual data but would be able to know who was permitted to see the data.

Guest Speaker: Tom Lee

Mr. Lee presented his essential approach the challenges facing the NDC as it seeks to review 410 million pages for public access. The technologies available for the past two decades could be employed to address the challenges posed by the large volume of records, developing a prioritization scheme, and actually conducting a declassification review.

The initial step would be to digitize the paper records by scanning; then technology would be applied, after which the images would be stored at high resolution and in a lossless[4] format. To make informed decisions about the documents, metadata, image data, and information about provenance need to be captured. At this stage, Optical Character Recognition (OCR) would be used to translate pixels into text and provide semantic sense. Mr. Lee noted that OCR engines vary in performance and the textual product may be imperfect.

Mr. Lee described a Sunlight Labs project to illustrate their experience using OCR and issues that arose out of the project. After the President nominated Elena Kagan to be an Associate Justice on the Supreme Court, the William J. Clinton Presidential Library released Ms. Kagan's archived e-mails to the public. The e-mails were released in "PDF" format; The Sunlight Foundation had an interest in showcasing and disseminating the e-mails to the public by posting them on their website and titled their project "Elena's Inbox." Mr. Lee described the process of how the Sunlight Foundation went about posting the e-mails on their site. First, he processed and gave them a familiar "Gmail" interface.[5] He explained that, during the Clinton Administration, these e-mails were captured automatically archived in the Automated Records Management System (ARMS) by the Office of Administration in the Executive Office of the President. Subsequently, at the end of the Clinton Administration, the archived e-mails were printed out before being sent to the Clinton Presidential Library. As Ms. Kagan's Senate confirmation

[4] A data compression algorithm which retains all the information in the data, allowing it to be recovered perfectly by decompression. Available at: http://dictionary.reference.com/browse/lossless

[5] Available at: http://elenasinbox.com/

hearings began, the e-mails were scanned using an OCR engine by the Clinton Library archivists. The Sunlight Foundation used these scanned images and posted them on the “Elena’s Inbox” site. Mr. Lee explained that this process was not ideal. Aside from the inefficiencies of the redundant and time consuming transfer process, in some instances, the transfer from digital to paper and then back to digital led to some data loss. For example, “1995” became “199S” in some cases.

When records are translated from digital to paper then re-digitized, OCR proved to be an imperfect, flawed process. Returning to the issue at hand on how to process 410 million pages for public access, Mr. Lee noted another limitation of the process is that classified information may not always be expressed textually; that is, an OCR engine cannot identify an image, such as a photograph or a map as potentially sensitive.

To counter this, Mr. Lee explained the notions of failsafe and fail secure lock systems as two alternative approaches: a failsafe lock opens in an emergency and a fail secure lock shuts and closes in an emergency. Declassification must be conducted in a fail secure manner. Classified documents must be identified first and declassified documents must be assigned a sensitivity score by ease or difficulty of release. Scoring would highlight sensitive and problematic records to eliminate these from the work queue. With the work queue reduced, workflow becomes more efficient, and most importantly, threats of accidental declassification or release are also reduced.

The next step in the process would be to determine the character of documents and assign them a declassification score. There are many methods which may be used to assess documents - from simpler methods as dirty words or pattern matching (such as social security numbers) to more sophisticated methods as natural language processing. In determining the relative frequencies of phrases or words, one may begin to define the classification of the documents.

Mr. Lee described two machine learning techniques, the Bayesian classifier model and the neural networks model. Both models rely on samples characterized by humans. The training corpus would require taking representative samples (which are statistically sound and have some element of randomization and selection) and then organizing sample data into buckets. By taking this training corpus and applying it to one of these models, it will learn the characteristics of those different categories of documents. Once the training phase has been completed the machine will have learned the rules and be able to comparatively apply them. Next the unobserved and un-reviewed documents could be ingested. This is known as the Bayesian classifier.

The Bayesian classifier, used by Google’s Gmail spam filter, measures the frequency of words and their position in incoming emails. It takes this information, appends it into the Bayesian model, weighs these observations, then tunes the training model. The training set is infinite, and the training model adjusts based on human observation (each time the user hits the report spam button the model adjusts to improve filtering capabilities). This model would be the same operation that professional reviewers would assemble for the current declassification project. The Bayesian model presents sufficient results, but does not allow for the interplay between different inputs: one term may be problematic in context to another, but the Bayesian classifier does not grasp contextual distinctions.

The neural network model recognizes contextual distinctions since it mimics biological systems. The model works simply: each neuron behaves by excitation or input signal then fires and excites other neurons. The strength of these connections varies and that variable strain is determined during the training phase. By feeding input of linguistic observations about source documents and upon the completion of the training phase, the model would be able to produce a score comparing documents to similarly scored decisions (classification or declassified status). Ultimately, an interface, similar to a bug tracking system, would be created to provide a work queue for multiple users and a prioritization queue with metadata concomitant with each task.

Mr. Lee concluded with a proposed methodology to build a system for the NDC to use as it seeks to review 410 million pages for declassification and public access. To create a workable solution he would execute the following steps: (1) digitize the data; (2) convert the data into characters and words (OCR); (3) determine the veracity of observations using statistical interrogation; (4) determine the relative dimensions of documents to distinguish them from one another; (5) employ machine learning models to assign scores to documents; and (6) use a work queue to triage prioritization for reviewers.

Guest Speaker: John Verdi

Mr. Verdi's experience is largely with issues concerning civil rights, civil liberties and government transparency at the Electronic Privacy Information Center (EPIC). In seeking access to classified documents for litigation and administrative proceedings, EPIC has utilized both the Freedom of Information Act (FOIA) and mandatory declassification request (MDR) processes. Mr. Verdi emphasized that technology solutions adopted in the Board's future recommendations would matter to the public and would have an effect on usefulness of information made available to the public.

The pinnacle for open government advocates has been to have the Federal Government construct a large, centralized robust publicly available database of documents from all agencies. This database would be of use to the public for both historical research reasons as well as for policy review purposes. From past experience, EPIC has observed that individuals and organizations requesting access to or release of national security information do not merely benefit the requestor, but the larger public as well. Documents requested by FOIA or MDR indicate expressed interest by requestors and at the same time benefit the public at large. This interest should be placed high on the prioritization queue.

In addressing the challenge of reviewing ever increasing volumes of the electronic records that the Federal Government creates, one important item to keep in mind is that these records should be easily searchable and manipulable in digital format. Therefore, there is no reason to print these documents to paper then re-scan them to the lossy[6] process of OCR. Mr. Verdi recounted EPIC's experience requesting classified documents from Agencies. As an example, EPIC requested documents held by an Agency in digital format. In response, the Agency printed the digital documents to paper, redacted the paper versions, and re-scanned the redacted documents. The OCR process devalued the search capability inherent in the "born digital" documents. For large volumes of data used in the legal production of documents, browsing is useless; search capability becomes critical to access this data.

Mr. Verdi concluded that, while it was commendable to declassify the information, the ultimate goal is to make this declassified information widely available and openly accessible to the public in a comprehensive, relevant, serviceable manner that is most useful to the user.

At the conclusion of Mr. Verdi's presentation Mr. Faga opened the session to questions and comments from the Board and audience. The Board's inquiries included: scanning technologies and practices; the tension between the prioritization as the public demanded and the technological drive to improve declassification efforts; the varied and diffuse wishes of the public for openness. Of particular significance was Chairman Faga's question to Mr. Verdi regarding the paradox concerning privacy and transparency. Mr. Verdi responded that transparency and privacy are complementary values. The individual has the right to privacy and Federal Government operations must be conducted transparently. Medical, financial and other personally identifiable information (social security numbers, credit card numbers) remain unconditionally private. Private information would be processed with similarly exacting standards used to review and declassify national security information and would be processed using the same methodologies recommended by Mr. Lee.

Ms. Merlyn Fowler asked the Board what would become of the speakers' recommendations and proposed solutions. Specifically, she wanted to learn if there would be implemented, considered, or would undergo experimentation. Mr. Faga answered the presentations and recommendations would be detailed as public record and accessible on the Public Interest Declassification Board's website. Additionally, a report of the Board's recommended findings would be submitted to the President though the National Security Advisor. Mr. Ungar clarified that the Board's function was advisory in nature.

Mr. Faga thanked the participants and audience and adjourned the session.

[6] A term describing a data compression algorithm which actually reduces the amount of information in the data, rather than just the number of bits used to represent that information. The lost information is usually removed because it is subjectively less important to the quality of the data (usually an image or sound) or because it can be recovered reasonably by interpolation from the remaining data. Available at: http://foldoc.org

Additional Documents:

Jeff Jonas

1. Mass Declassification
2. Strawman Architecture

Tom Lee

1. Sunlight
2. Bomb Image
3. Bayesian Classifier
4. Bug Tracker