2007 Data Mining Report

DHS Privacy Office Response to House Report 109-699

July 6, 2007

Homeland Security

2007 Report to Congress on the Impact of Data Mining Technologies on Privacy and Civil Liberties

Respectfully submitted
Hugo Teufel III
Chief Privacy Officer
U.S. Department of Homeland Security
Washington, DC

July 6, 2007

TABLE OF CONTENTS

2007 Data Mining Report
DHS Privacy Office
July 6, 2007

I. Executive Summary

The Department of Homeland Security (DHS) Privacy Office is pleased to provide to the Congress its *2007 Data Mining Report: DHS Privacy Office Response to House Report 109-699.* This is the second report by the Privacy Office to Congress on data mining. This report describes data mining activities deployed or under development within the Department that meet the definition of data mining as mandated in House Report No. 109-699 – *Making Appropriations for the Department of Homeland Security for the Fiscal Year Ending September 30, 2007, and for Other Purposes*.[1] In addition, it provides an update on how the Privacy Office has begun implementation of the recommendations outlined in its first report on data mining entitled, *Data Mining Report: DHS Privacy Office Response to House Report 108-774*, issued July 6, 2006 ("July 2006 Report").[2] The Privacy Office recognizes the importance of these reports to Congress, since Congress mandated specifically data mining as one of the analytical tools that the Department should use in fulfilling its mission.[3]

In this report, the Privacy Office applied the new definition for data mining from House Report 109-699 in evaluating whether or not Departmental information processing activities were data mining activities.[4]Additionally, this report examines various definitions describing data mining and the impact those definitions have on the technology and analysis.

Since the definition of data mining activity used in this report is different from the one applied in the July 2006 Report, a number of the information processing activities

[1] Conference Report on HR 5441, DHS Appropriations Act, House Rept. No. 109-699, Sept. 28, 2006, H7784, at H7815.

[2] The July 2006 Report is posted at www.dhs.gov/privacy.

[3] The Homeland Security Act Section 201(d)(14) calls for the Department to "establish and utilize, in conjunction with the chief information officer of the Department, a secure communications and information technology infrastructure, including datamining and other advanced analytical tools, in order to access, receive, and analyze data and information in furtherance of the responsibilities under this section, and to disseminate information acquired and analyzed by the Department, as appropriate."

[4] The House Report requires this report to use the following definition for "data mining": "... a query or search or other analysis of 1 or more electronic databases, whereas – (A) at least 1 of the databases was obtained from or remains under the control of a non-Federal entity, or the information was acquired initially by another department or agency of the Federal Government for purposes other than intelligence or law enforcement; (B) a department or agency of the Federal Government or a non-Federal entity acting on behalf of the Federal Government is conducting the query or search or other analysis to find a predictive pattern indicating terrorist or criminal activity; and (C) the search does not use a specific individual's personal identifiers to acquire information concerning that individual."

discussed in the July 2006 Report are not discussed in this report, as they did not fall under the definition mandated in the House Report 109-699.

The activities in the July 2006 Report that do not meet the newer definition are: U.S. Customs and Border Protection's (CBP) Enterprise Data Warehouse (EDW); U.S. Immigration and Customs Enforcement's (ICE) Pattern Analysis and Information Collection System (ICEPIC); U.S. Citizenship and Immigration Services' (USCIS) Fraud Detection and National Security Data System (FDNS-DS); and USCIS' National Immigration Information Sharing Office (NIISO). This report summarizes these activities and why they do not meet the new definition.

In addition, a number of the information processing activities not reported on in the July 2006 Report now meet the definition mandated in the House Report 109-699 and are now included on the list of data mining activities. The new activities that have been added to this report that meet the definition are: Directorate for Science & Technology's (S&T Directorate) Analysis, Dissemination, Visualization, Insight, and Semantic Enhancement (ADVISE) Program Pilots;[5] CBP's Automated Targeting System: ATS-Inbound and ATS-Outbound (Cargo Analysis); and ICE's Data Analysis and Research for Trade Transparency System (DARTTS).

For each of the activities that meet the definition, the report provides a brief description of the activity, including highlights of the following: each activity's goals and plans; data sources, deployment status, and evidence of effectiveness; privacy documentation and privacy protections; data security and integrity measures.

As described below, the Privacy Office's compliance process requires programs using personally identifiable information to have completed federally-mandated privacy documentation, consisting of a Privacy Impact Assessment (PIA)[6] and a System of Records Notice (SORN).[7] The Privacy Office is working closely with four of the newly

[5] ADVISE may or may not fit the definition of "data mining" used in this report as defined by Congress depending on the data loaded into the system.

[6] Section 208 of the E-Government Act of 2002, Pub. L. 107-347 (December 17, 2002), specifically requires a federal agency when developing a new information technology to conduct an analysis of the impact that the new information technology will have on privacy. Further, Section 222 of the Homeland Security Act of 2002, Pub. L. 107-296, (November 25, 2002) enhances the Privacy Offices ability to conduct Privacy Impact Assessments on the use of technology at the Department through its duty to ensure "that the use of technologies sustain, and do not erode, privacy protections relating to the use, collection, and disclosure of personal information." 6 U.S.C. § 142(1).

[7] Privacy Act of 1974, 5 U.S.C. § 552a, as amended, requires federal agencies to place into practice the fair information principles when handling collections of personal information. If an agency maintains a system of records, 5 U.S.C. § 552a(a)(5), then that agency must provide notice regarding that collection

identified activities that have not fully completed the required privacy documentation: S&T Directorate's ADVISE Program Pilots, Office of Intelligence & Analysis's (I&A) I2F, and ICE's DARTTS and NETLEADS.

The S&T Directorate is working with the DHS Privacy Office to complete appropriate privacy-related documentation for the ADVISE Pilots. Pending these assessments, the Department has halted ADVISE. I&A is in the process of drafting a PIA for I2F. ICE is in the process of drafting PIAs and SORNs for DARTTS and NETLEADS.

In addition to monitoring and reporting on current data mining activities, the Privacy Office began to implement the recommendations outlined in the July 2006 Report. Section V of this report provides the status of the efforts made by the Privacy Office to move forward on these recommendations. The Privacy Office focused on educating components about the particular privacy issues raised by data mining activities. At this time, the Privacy Office is exploring the establishment of a coordination group to review data mining standards and harmonize development guidance across the Department.

The Privacy Office may hold a public workshop later this year as part of the process in educating the Department and the public about the data mining recommendations and to gather research on data integrity standards and validation models, as well as auditing and anonymizing technologies to help implement the recommendations. A workshop would educate both the Privacy Office and Departmental programs on privacy enhancing technologies such as anonymization and auditing tools and how best to move forward to implement privacy protections when conducting data mining activities. The Privacy Office is in communication with the relevant offices and components of the Department to determine whether there is sufficient need for a workshop, while providing direct assistance to the programs identified in this report.

II. Introduction

The Privacy Office operates under the direction of the Chief Privacy Officer, who is appointed by and reports directly to the Secretary of the Department. The Privacy Office serves to implement Section 222 of the Homeland Security Act of 2002,[8] and has programmatic responsibilities involving the Privacy Act of 1974, the Freedom of

that informs the public about the collection's purpose, who is affected by the collection, what types of information is contained in the collection, and how will the information be shared outside the agency, among other items. 5 U.S.C. § 552a(e)(4) ("System of Records Notice" or "SORN").

[8] Section 222 of the Homeland Security Act of 2002, as amended by the Section 8305 of the Intelligence Reform and Terrorism Prevention Act of 2004, Pub. L. 108-458 (December 17, 2004), 6 U.S.C. § 142.

Information Act ("FOIA"),[9] the privacy provisions of the E-Government Act of 2002, and DHS policies that protect individual privacy associated with the collection, use, and disclosure of personally identifiable information. Section 222 of the Homeland Security Act of 2002 calls on the Chief Privacy Officer to assume primary responsibility for privacy policy within the Department, as well as "assuring that the use of technologies sustain, and do not erode, privacy protections relating to the use, collection, and disclosure of personal information."[10]

The Privacy Office published its *Data Mining Report: DHS Privacy Office Response to House Report 108-774*, issued July 6, 2006 ("July 2006 Report"). That report was prepared pursuant to the requirements of House Report 108-774 – *Making Appropriations for the Department of Homeland Security for the Fiscal Year Ending September 30, 2005, and for Other Purposes.* The July 2006 Report provides a definition and description of the process of data mining and sets out the privacy and civil liberties concerns raised by the use of data mining technologies for homeland security. It identifies specific data mining activities and programs at DHS and provides information related to their purpose, data sources, and deployment dates. The report also describes the policies, procedures, and guidance that apply to each data mining activity identified. Looking forward, the report made a number of recommendations regarding DHS data mining activities aimed specifically at addressing the privacy concerns those activities may raise.

This second report on data mining is prepared by the Privacy Office pursuant to the requirements of House Report No. 109-699 – *Making Appropriations for the Department of Homeland Security for the Fiscal Year Ending September 30, 2007, and for Other Purposes*[11] ("The House Report"). The House Report states that DHS must furnish a report to Congress on data mining activities "consistent with the terms and conditions listed in Section 549 of the Senate bill." Section 549 of the Senate bill states:

> "*The head of each department or agency in the Department of Homeland Security that is engaged in any activity to use or develop data-mining technology shall each submit a report to Congress on all such activity of the agency under the jurisdiction of that official. The report shall be made available to the public.*

[9] Freedom of Information Act, 5 U.S.C. § 552.

[10] 6 U.S.C. §142.

[11] H. Conf. Rept. 109-699, Making Appropriations for the Department of Homeland Security for the Fiscal Year Ending September 30, 2007, and for Other Purposes, Sept. 28, 2006, H7784, at H7815 (Conference Report on HR 5441).

> *Each report submitted . . . shall include, for each activity to use or develop data-mining technology that is required to be covered by the report, the following information: (A) a thorough description of the data-mining technology and the data that is being or will be used; (B) a thorough description of the goals and plans for the use or development of such technology and, where appropriate, the target dates for the deployment of the data-mining technology; (C) an assessment of the efficacy or likely efficacy of the data-mining technology in providing accurate information consistent with and valuable to the stated goals and plans for the use or development of the technology; (D) an assessment of the impact or likely impact of the implementation of the data-mining technology on the privacy and civil liberties of individuals; (E) a list and analysis of the laws and regulations that govern the information; (F) a thorough discussion of the policies, procedures and guidelines that are in place or that are to be developed and applied in the use of such technology for data-mining in order to protect the privacy and due process rights of individual, and ensure that only accurate information is collected, reviewed, gathered, analyzed or used."*[12]

In addition, this House Report requires that the Privacy Office include information on how DHS has implemented the recommendations set forth in the July 2006 Report to address privacy concerns.

In this report, the Privacy Office provides the Congress with updated information about DHS data mining activities. It identifies data mining activities that have been under development or deployed since the July 2006 Report. It also reports on the Privacy Office's initial efforts to promote adoption and implementation of the privacy recommendations outlined in the July 2006 Report. As the recommendations set out by the Privacy Office have been available to data mining programs for only a short time, this document will serve as a status report rather than a final assessment of the extent and success of their implementation.

III. Background

A. Process of the Inquiry

To respond to the requirements of the House Report, the Privacy Office asked DHS components to update the information on data mining activities reported in the July 2006

[12] H.R. 5441, Engrossed as Agreed by Senate, Homeland Security Appropriations Act for FY 2007. This engrossed amendment was agreed to by unanimous consent and was included in the Senate-passed version of H.R. 5441 as Section 549. According to the conference report; however, Section 549 was deleted from the final bill, so the conference report included a statement on data mining and directed the DHS Privacy Officer "to submit a report consistent with the terms and conditions listed in section 549 of the Senate bill." House Conf. Rept. 109-699 at H7815.

Report. Specifically, the Privacy Office sought responses to two key questions. First, the components were asked to consider whether they were undertaking, or planning to undertake, any new data mining activities as defined in the House Report since the release of the July 2006 Report, and to provide information about those activities. Second, the Privacy Office asked components whose data mining activities were reported in the July 2006 Report to update the information previously reported and furnish information on the extent to which they had implemented the recommendations of the July 2006 Report with respect to those activities.

To elicit the information requested by the Congress, the Privacy Office posed a series of questions to DHS components using the definition of data mining contained in the House Report. Components deploying newly identified activities were asked somewhat different questions than those components that had activities that the Privacy Office had reviewed in the July 2006 Report. The Privacy Office wishes to acknowledge the assistance it received from each of the components included in this report. The Privacy Office will continue to work with the components to update and provide more robust information regarding data mining programs in future reports.

With respect to each newly identified activity, the inquiry asked about the goals and plans for these activities, the source of the data used, how long the activity had been in development or operation, and how effective the tool or activity is believed to be in providing accurate information consistent with the tool or its goals. The inquiry further asked about measures taken to ensure the privacy, security, and integrity of the data, including staff training and audit practices, and whether the activity had drafted a Privacy Impact Assessment.

The Privacy Office asked components, whose activities were the subject of the July 2006 Report, specific questions about the extent to which that activity had implemented each of the Privacy Office recommendations. Consistent with the recommendations, the questions asked whether the data mining activity uses tools principally for investigative purposes; whether the activity explicitly considers using anonymized data; what steps the activity takes to assure that the data used in the data mining activity is of a quality adequate to the analysis; whether the activity has adopted or undertaken to adopt a standard for validation of the models or rules derived from the activity; whether policies and procedures are in place to ensure security and integrity of data; how databases are secured; and whether the databases and systems are subjected to regular audits.

B. Definition of Data Mining

This report relies upon the definition provided in House Report No. 109-699 for the term "data mining." The House Report directs the DHS Privacy Officer to submit a report consistent with the terms and conditions listed in Section 549 of the Senate bill. Section 549 defines the term “data mining” as:

”[A] query or search or other analysis of 1 or more electronic databases, whereas –

(A) at least 1 of the databases was obtained from or remains under the control of a non-Federal entity, or the information was acquired initially by another department or agency of the Federal Government for purposes other than intelligence or law enforcement;[13]

(B) a department or agency of the Federal Government or a non-Federal entity acting on behalf of the Federal Government is conducting the query or search or other analysis to find a predictive pattern indicating terrorist or criminal activity; and

(C) the search does not use a specific individual’s personal identifiers to acquire information concerning that individual.”[14]

It is important to note that no consensus exists on what constitutes “data mining.” In colloquial use, data mining generally refers to any predictive, pattern-based technology and is narrower than the definition used in the July 2006 Report. Other government reports have used broader definitions.

- The Government Accountability Office (GAO) defines data mining in its May 2004 report entitled Data Mining: Federal Efforts Cover a Wide Range of Uses, as “the application of database technology and techniques – such as statistical analysis and modeling – to uncover hidden patterns and subtle relationships in data and to infer rules that allow for the prediction of future results.”
- The Congressional Research Service (CRS) defines data mining in its January 27, 2006, report to Congress entitled, Data Mining and Homeland Security: An Overview, in more generic terms. It states that data mining “involves the uses of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data sets.” The report describes data mining as using a “discovery approach” in which algorithms examine data

[13] Note that to meet this definitional requirement, at least one of the databases used in the data mining activity must be non-federal (e.g., a commercial database) or that the information used be obtained by a department or agency other than DHS for purposes other than intelligence or law enforcement (e.g., benefit claims or employment data). This definition, therefore, excludes activities that only use data obtained by the Department itself rather than from outside sources.

[14] This language expressly excludes any activity that entails using an individual’s name or other personal identifier to search for information about that individual. The definition in the Senate bill further specifies that “the term ‘database’ does not include telephone directories, news reporting, information publicly available via the Internet or available by any other means to any member of the public without payment of a fee, or databases of judicial and administrative opinions.”

relationships to identify patterns. It distinguishes this method from analytical tools that use a "verification based approach," where the user develops a hypothesis and then uses data to test the hypothesis. (This was the definition used in the Privacy Office's July 2006 Report.)

- The DHS Office of the Inspector General (DHS OIG) defines data mining in its August 2006 Survey of DHS Data Mining Activities, simply as "the process of knowledge discovery, predictive modeling, and analytics." It stated that this has traditionally involved the discovery of patterns and relationships from structured data bases of historical occurrences.

The definitions of data mining articulated by GAO, CRS, and the DHS OIG, as well as the definition the Privacy Office used in the July 2006 Report, encompass a different range of activities than those that fall under the Senate engrossed bill definition referenced in the House Report, as these other definitions do not limit the types of data sources or exclude data mining activities involving searches using a specific individual's personal identifier.[15] The reader is urged to consider the differences in definitions of data mining when reading this report.

C. July 2006 Report Recommendations and the Request for Updated Information

As noted above, the House Report specifically requests that the Privacy Office report on the extent to which components and offices are implementing the recommendations of the July 2006 Report.[16] These recommendations address the specific privacy concerns raised by data mining and go beyond traditional privacy and security protections, such as privacy impact assessments, Memoranda of Understanding between agencies that own source data systems, privacy and security training, and role-based access. The July 2006 Report recommended that in addition to those standard protections, programs that employ data mining tools and technologies put in place the following protections:

[15] For example, the DHS Office of Inspector General report identifies and discusses two activities of the Transportation Security Administration (TSA) as data mining under its definition that are not included in this report. TSA's Crew Vetting System (CVS) and Tactical Information Sharing System (TISS) fall under OIG's broader characterization of data mining as "the process of knowledge discovery, predictive modeling, and analytics." These two activities, however, do not fall under the definition set out by Congress for purposes of this Report. CVS is purely a name matching system, seeking matches against terrorist, law enforcement, and immigration databases. It does not seek patterns to predict terrorist or criminal activity. TISS is a law enforcement system that collects and disseminates suspicious activity reports. It also seeks trends that may reveal the need for a heightened alert, but relies entirely on law enforcement data and does not predict terrorist or criminal activity, thus taking it out of the Senate bill definition.

[16] House Conf. Rept. No. 109-699 at H7815.

1. Steps to ensure that the agency has authority to undertake such data mining activity, including the collection or aggregation of data required to perform the project, and that its authority is consistent with the purposes of the data mining project or program;
2. Written policies stating that data mining be used principally for investigative purposes and that no decisions may be made automatically regarding individual rights or benefits solely on the basis of the results produced by patterns or rules derived from data mining;
3. Anonymization of data whenever possible;
4. Adoption by DHS of data quality standards;
5. Adoption by DHS of standards for the validation of models or rules derived from data mining;
6. Implementation of review and redress policies and procedures for individuals identified for additional investigation by data mining activities; and
7. Demonstration of accountability through audit capabilities to record access to data, data marts and data mining patterns and rules.[17]

The Privacy Office remains committed to urging the adoption of these recommendations within DHS; however, additional time will be needed in order to inform programs fully and develop appropriate implementations for the recommendations. The Privacy Office will provide leadership in this area. Given that the Privacy Office is at the earliest stage of addressing the July 2006 recommendations, this report is only intended to provide a status report on the progress of their adoption, rather than a report card on their implementation.

D. Data Mining and the Privacy Office Compliance Process

Three federal laws form the foundation for privacy protections for data mining activities at the Department – Section 222 of the Homeland Security Act of 2002,[18] the Privacy Act of 1974,[19] and the E-Government Act of 2002.[20] Section 222 of the Homeland Security Act of 2002 states that the DHS Chief Privacy Officer is responsible for "assuring that the use of technologies sustains, and do not erode, privacy protections relating to the use,

[17] For a more complete discussion of the Privacy Office recommendations, see July 2006 Report, pp. 29-30. The recommendations from the July 2008 Report are excerpted as Appendix I of this report.

[18] Homeland Security Act of 2002, Pub. L. 107-296, Section 222 (1-2) (codified 6 U.S.C. § 142).

[19] Privacy Act of 1974, Pub. L. 93-579, 5 U.S.C. § 552a.

[20] E-Government Act of 2002, Pub. L. 107-347.

collection, and disclosure of personal information."[21] Further, the Chief Privacy Officer is also "responsible for assuring that personal information contained in Privacy Act systems of records is handled in full compliance with fair information practices as set out in the Privacy Act of 1974."[22]

In addition, the Privacy Act, among other requirements, mandates that agencies publish a notice when personally identifiable information is maintained in a system of records.[23] The System of Records Notice (SORN) is published in the Federal Register and identifies the purpose for the system of records, the categories of individuals in the system, what categories of information are maintained about the individuals, and how the agency discloses the information to other agencies (routine uses).[24] The SORN also provides the public notice regarding the available mechanisms to exercise the rights granted through the Privacy Act to access and correct the personally identifiable information an agency maintains.[25]

Section 208 of the E-Government Act of 2002 requires all federal government agencies to conduct Privacy Impact Assessments (PIAs) for all new technology that collects, maintains, or disseminates personally identifiable information.[26] OMB guidance for implementing the privacy provisions of the E-Government Act extends this requirement to substantially changed systems as well.[27] Although the E-Government Act does not require PIAs on national security systems or systems with information about federal employees and contractors, as a policy matter, the Privacy Office requires all information

[21] 6 U.S.C. § 142(1).

[22] 6 U.S.C. § 142(2).

[23] 5 U.S.C. § 552a(e)(4) ("[S]ubject to the provisions of paragraph (11) of this subsection, publish in the Federal Register upon establishment or revision a notice of the existence and character of the system of records..."); 5 U.S.C. § 552a(a)(5), the term "system of record" means a group of any records under the control of any agency from which information is retrieved by the name of the individual or by some identifying number, symbol, or other identifying particular assigned to the individual.

[24] *Id*.

[25] *Id*.

[26] E-Government Act of 2002, Pub. L. 107-347, 116 Stat. 2899, § 208(b) (44 U.S.C. § 3501 note).

[27] OMB Guidance for Implementing the Privacy Provisions of the E-Government Act of 2002, M-03-22, Sept. 26, 2003.

technology systems to conducts PIAs.[28] Nonetheless, classified systems may be exempted from the requirement to publish the PIA.[29]

As part of its overall compliance program, the Privacy Office seeks to identify programs that intend to engage in data mining through several different processes. First, the Privacy Office reviews all Office of Management and Budget (OMB)-300 budget submissions to learn of programs that employ personally identifiable information and determine whether the program or system addresses privacy appropriately. Second, the Privacy Office created the Privacy Threshold Analysis (PTA) document, a short form used for all information technology systems that are going through certification and accreditation (C&A) process required under the Federal Information Security Management Act (FISMA)[30] to determine whether the system maintains personally identifiable information. The PTA assists the Privacy Office in identifying those programs that use personally identifiable information and, of those, which need privacy impact assessments (PIA). Then, the Privacy Office identifies those systems that need a new or updated SORN as required by the Privacy Act.

In addition, the Privacy Office reviews the proposals for technology investment proposals that the DHS Enterprise Architecture Center of Excellence (EACOE) and the Integrated Project Review Team (IPRT) process to ensure that DHS investments in technology include a specific review for compliance with privacy protection requirements. Through these activities, the Privacy Office compliance process provides a number of opportunities to learn about proposed data mining activities and to engage program managers in discussions about potential privacy issues.

Moreover, the Privacy Office is developing now the Privacy Technology Implementation Guide (PTIG), designed specifically to augment the existing PIA process to identify privacy issues at the very earliest stages of technology development. The Privacy Office believes this new guide will ensure that privacy is considered from the very beginning of technology development. The PTIG will be particularly useful in helping developers of data mining tools to evaluate privacy implications and identify privacy protections that can be built into the design. Unlike the PIA, which focuses on a specific implementation of technology involving PII, the PTIG will enable researchers and technology designers

[28] *See* 6 U.S.C. § 142(1).

[29] The Privacy Office requires a PIA for federal employee and contractor systems that are deployed throughout DHS.

[30] The Federal Information Security Management Act is included under Title III of the *E-Government Act* (Public Law 107-347).

to consider privacy at the very earliest stages of technology development, even before the tool is deployed.

Whether conducting a PIA or implementing the new PTIG, all of the Privacy Office's compliance tools apply the Fair Information Principles (FIPs), the fundamental framework for privacy implementation at the Department. These principles are the principles of transparency, individual participation, purpose specification, minimization, use limitation, data quality and integrity, security, and accountability and auditing. In addition, these principles are reflected in agency SORNs, which are issued pursuant to the Privacy Act of 1974 and published in the Federal Register.

The Privacy Office is working with the various components to complete the privacy compliance documentation, PIAs and SORNs, required for each of the activities identified in this Report. Some of the activities identified in this report are covered by a DHS issued SORN and others are operating under what is referred to as a "legacy" SORN, meaning a SORN issued by predecessor agencies of DHS and carried forward into DHS when DHS was created. In addition to the SORNs specific to the programs discussed in this report, the Privacy Office is working closely with DHS components to update existing legacy SORNs as DHS updates the programs and technologies.[31]

Presently, the Department published 51 SORNs (either reissued and updated legacy SORNs or new system SORNs) and has another 207 legacy SORNs that need to be reviewed and either re-published or retired. In addition, Office of Management and Budget (OMB) requires that the Senior Agency Official for Privacy, which is the Chief Privacy Officer at DHS, review SORNs bi-annually to ensure that they are up-to-date and consistent with Privacy Act requirements. Given this challenge among other compliance responsibilities, the Privacy Office requested additional staffing for Fiscal Year (FY) 2008 to support the privacy compliance operation, which handles all SORNs and PIAs for DHS.

Additionally, a number of the components are in the process of completing PIAs for one or more of the activities discussed in this Report. Activities using technology that pre-dates the E-Government Act do not require a PIA unless the system has undergone a significant change. Generally, PIAs are required before a program is operational; however, in a few instances, data mining activities may have progressed without adequate attention to privacy documentation. This shortcoming is being addressed so that all of the activities noted in this Report will shortly have documentation.

[31] Pursuant to the savings clause in the Homeland Security Act of 2002, Public Law 107-296, section 1512, 116 Stat. 2310 (Nov. 25, 2002), the Department of Homeland Security (DHS), and its components and offices, have relied on preexisting Privacy Act systems of records notices.

IV. Reporting

A. Data mining Activities Identified Since The July 2006 Report

1. Analysis, Dissemination, Visualization, Insight, and Semantic Enhancement

The DHS Directorate for Science & Technology (S&T Directorate) maintains Analysis, Dissemination, Visualization, Insight, and Semantic Enhancement (ADVISE), a technology framework that can be used to analyze and visually represent relationships between different types of data. Development of the ADVISE Technology Framework began in 2003. The S&T Directorate developed the ADVISE Technology Framework to help DHS analysts quickly identify and retrieve the most relevant information for their research and reporting needs from the vast array of document sources already received. At the time of this writing, all work associated with ADVISE, testing or otherwise, has been halted pending review and approval of all related privacy assessments by the Privacy Office.

The ADVISE Technology Framework is not itself an operational system; it cannot collect, maintain, disseminate or otherwise use personally identifiable information; however, the ADVISE Technology Framework raises potential privacy protection issues, because the Framework supports the ability to combine multiple types of data from multiple data sources and represent relationships between individual pieces of information. Combining data raises potential privacy protection concerns, including potential errors as a result of the process, as well as the challenge of adhering to privacy compliance requirements to maintain the purpose and use consistent with the original collection.

Until the framework is applied to a particular set of data, it is uncertain if the ADVISE Technology Framework meets the definition of data mining used in this Report. Nevertheless, because the tool could be used to build a system that could meet the definition, it is discussed below.

a) Goals and Plans for the Program

The purpose of the ADVISE Technology Framework is to enable analysts to rapidly sift through the materials they already find to be the most pertinent and useful information and to open the source documents all through a single user interface. The ADVISE Technology Framework could be used with information about any topic and in a variety of formats, including structured (databases) and unstructured (text files). If analysts identify particular pieces of information about people and establish relationships between those pieces of information and other pieces of information (other people or places and events), then the ADVISE Technology Framework would allow analysts to visualize those relationships across all of the data sources where those same particular pieces of information were identified.

The ADVISE Technology Framework consists of three sets of tools. The first set of tools is designed to prepare and load source data. The second set of tools supports the analysis of the data – enabling analysts to identify potential relationships between people, places, and events). The third set of tools provides a visual interface to view and interact with the results of the analysis.

When data is brought into an implementation of the ADVISE Technology Framework, analysts can use the data loading utilities and text processing utilities to identify objects (i.e., people, places, things) and to establish the relationship between those objects. The identified objects and relationships are used to construct a semantic graph.[32] Attributes about the objects are stored outside the main graph. A document management system could be included to access the original source documents.

The ADVISE Technology Framework cannot be used to automatically identify particular pieces of information nor can it be used to find new data. The ADVISE Technology Framework can only be used to represent visually the particular pieces of information and the relationships between those particular pieces that analysts already identified during the data loading and mapping process.

The S&T Directorate initiated previously a set of pilot tests of the ADVISE Technology Framework. One of these tests involved validating the technology itself, to verify the functionally of the tools and the user experience in loading and operating the software. The other pilots focused on determining whether the ADVISE Technology Framework actually provides benefit to analysts, and to identify the type of source materials that would better lend themselves to the kind of analysis the ADVISE Technology Framework provides. As such, no pilot, either previous or anticipated, involves operational decision making. The S&T Directorate is following all DHS privacy compliance requirements for each of the anticipated test pilots of the ADVISE Technology Frameworks.

b) Data Sources, Deployment Dates, and Effectiveness

The ADVISE Technology Framework is a set of individual capabilities (tools) and as such, there is no actual data within the ADVISE Technology Framework. Each deployment of the ADVISE Technology Framework is designed specifically to match the environment and purpose of the DHS component that will use the resulting system. Part of the design of a particular deployment of the ADVISE Technology Framework includes

[32] Graph models that represent words and their relationships.

a determination of what data to use in the implementation. The decision on what data to use is made by the implementing organization in accordance with that organization's policies and related statutory authority.

c) Data Privacy

The ADVISE Technology Framework provides capabilities to enforce the privacy and security policies of the implementing organization. Each deployment of the ADVISE Technology Framework will be specifically designed to use only data that is relevant and necessary to the support of the mission of that DHS component.

Based on the nature of the ADVISE Technology Framework, the Privacy Office determined that the most effective and efficient method of integrating privacy protections would be to develop a new type of privacy guidance, one that could be adapted to match the architecture of the Framework itself.

This new privacy guidance document, the Privacy Technology Implementation Guide (PTIG), will provide step-by-step guidance to integrate privacy compliance requirements into the deployments of the ADVISE Technology Framework and any future technology tools.

The Privacy Office is actively working on an overall PTIG to address all uses of technology across the Department. Upon the completion of this overall PTIG, the Privacy Office will coordinate with DHS S&T to develop a PTIG specific for ADVISE. The PTIG will better enable the identification of privacy issues associated with development of a technology than the current PIA template, because the PIA template is best applied to a program as it makes decisions about specific uses of personally identifiable information.

The combination of privacy guidance for technology development and compliance review of specific implementations of developed technologies should identify opportunities to use specific privacy enhancing techniques such as anonymization to mitigate privacy risks. This process to consider privacy during technology development will also address the concerns expressed in a recent Government Accounting Office review of the ADVISE Technology Framework, which urged the Privacy Office to conduct a PIA on the ADVISE Technology Framework as soon as possible.[33]

[33] Government Accounting Office, Early Attention to Privacy in Developing a Key DHS Program Could Reduce Risks, GAO-07-203 (Feb. 2007).

The Privacy Office is currently conducting a review of the ADVISE Technology Framework and any associated pilots to ensure the completion of all appropriate privacy-related documentation.

d) Data Security and Integrity

Once data is in a system built from the ADVISE Technology Framework, it is secured using a Public Key Infrastructure (PKI)-based security infrastructure that supports the auditing of all user actions.

The ADVISE Technology Framework uses PKI designed for accreditation with Director of Central Intelligence Directive (DCID) Manual 6/3 Protection Level 3. The security layer includes access control and authentication services to ensure that only individuals who have received approval can access the system and that their access credentials are authentic. The ADVISE Technology Framework provides capabilities to limit access to data to only those with a need to know and in accordance with the policies of the implementing organization. Further, the ADVISE Technology Framework provides the capability to restrict access to data based upon the role(s) assigned to each individual. The ADVISE Technology Framework also implements the concept of communities of interest, which assigns data to a specific community of interest; authorized users are granted permission to access each community of interest based upon the user's need to know.

Users cannot change the data in the ADVISE Technology Framework. Corrections and updates are made through feedback to the organization providing the source data. The ADVISE Technology Framework supports auditing of all users actions. Determinations regarding policies and procedures for review, redress, transparency, and accountability will be made part of the design and development process for individual deployments of the ADVISE Technology Framework.

2. Automated Targeting System: ATS-Inbound and ATS-Outbound (Cargo Analysis)

CBP maintains the Automated Targeting System (ATS), which uses a common approach for data management, analysis, rules-based risk management, and user interfaces to support all CBP mission areas and the data and rules specific to those areas. ATS consists of six modules that provide selectivity and targeting capability to support CBP inspection and enforcement activities. Only two of these – ATS-Inbound and ATS-

Outbound – engage in data mining to provide decision support analysis for targeting of cargo for suspicious activity.[34]

ATS-Inbound and ATS-Outbound look at the data related to cargo in real time and look for indications of suspicious activity. ATS-Outbound and ATS-Inbound support CBP functions mandated by Title VII of Public Law 104-208 (1996 Omnibus Consolidated Appropriations Act for FY 1997), which provides funding for counter-terrorism and drug law enforcement. ATS-Outbound also supports functions arising from the Anti-Terrorism Act of 1997 and the 1996 Clinger-Cohen Act. The risk assessments for cargo are also mandated under Section 203 of the Security and Accountability for Every Port Act of 2006, Pub. L. 109-347, October 11, 2006 (SAFE Port Act).

a) Goals and Plans for the Program

The cargo analysis provided by ATS is intended to add automated anomaly detection to CBP's existing targeting capabilities to enhance screening of cargo prior to its entry into the United States. The data used in the development and testing of this technology is taken from bills of lading and shipping manifest data provided by vendors to CBP as part of the existing cargo screening process.

ATS-Inbound is available to CBP officers at all major ports (air/land/sea/rail) throughout the United States, and also assists CBP personnel in the Container Security Initiative (CSI) decision-making process. ATS-Inbound provides CBP officers and Advance Targeting Units (ATU) with an efficient, accurate, and consistent method for targeting and selecting high-risk inbound cargo for intensive examinations. ATS-Inbound assists in identifying imported cargo shipments that pose a high risk of containing weapons of mass effect, narcotics, or other contraband. ATS-Inbound processes data pertaining to entries and manifests against a variety of rules to make a rapid, automated assessment of the risk of each import. Entry and manifest data is received from the Automated Manifest System (AMS) and the Automated Broker Interface (ABI), both components of the Automated Commercial System (ACS) and the Automated Commercial Environment (ACE).

ATS-Outbound is the outbound cargo targeting module of ATS that assists in identifying exports that pose a high risk of containing goods requiring specific export licenses,

[34] News reports have identified ATS-P, involving passengers, as data mining; however, it does not fit the definition guiding this report. As noted above in section III.B., searches triggered by a specific individual's identifier does not meet the definition for this report. ATS-P develops a risk assessment for each traveler based on defined rules determined by human intelligence that pertain to specific operational and tactical objectives or local enforcement efforts. Each physical search or examination of a passenger is conducted as a result of a determination made by a CBP Officer following review of the traveler's personally identifying travel documents.

narcotics, or other contraband. ATS-Outbound uses Shippers' Export Declaration (SED) data that exporters file electronically with CBP's Automated Export System (AES). The SED data extracted from AES is sorted and compared to a set of rules and evaluated in a comprehensive fashion. This information assists CBP officers with targeting and/or identifying exports with potential aviation safety and security risks, such as hazardous materials and Federal Aviation Administration (FAA) violations. In addition, ATS-Outbound identifies the risk of specific exported cargo for such export violations as smuggled currency, illegal narcotics, stolen vehicles, or other contraband.

b) Data Sources, Deployment Dates, and Effectiveness

ATS does not collect information directly from individuals. The information maintained in ATS is either collected from private entities providing data in accordance with U.S. legal requirements (e.g., sea, rail and air manifests) or is created by ATS as part of the risk assessment and associated rules.[35]

ATS-Inbound and ATS-Outbound collect information about importers and exporters, cargo, and conveyances used to facilitate the importation of cargo into and the exportation of cargo out of the United States. This information includes personally identifiable information (e.g., name, address, birth date, government-issued identifying records, where available and applicable) concerning individuals associated with imported cargo, such as brokers, carriers, shippers, buyers, sellers, and crew. Similarly, this information includes individuals associated with exported cargo, such as exporters, freight forwarders, shippers, and crew.

The ATS-Inbound and ATS-Outbound became operational in 1997.

[35] ATS-Inbound collects information about importers and cargo and conveyances used to import cargo to the United States from destinations outside its borders. Information regarding individuals, such as importers, is collected in connection with the following items including, but are not limited to: Sea/Rail Manifests from Automated Manifest System (AMS), Cargo Selectivity Entries from Automated Broker Interface (ABI); Entry Summary Entries from ABI; Air Manifest (bills of lading (AMS-Air); Express Consignment Services (bills of lading); CCRA Manifest (bills of lading) from Canada Customs and Revenue CCRA); CAFÉ, QP Manifest Inbound (bills of lading) from AMS; Truck Manifest from Automated Commercial Environment (ACE); Inbound Data (bills of lading) from AMS; entries subject to Food and Drug Administration (FDA) Prior Notice (PN) requirements from the Automated Commercial System (ACS); and Census Import Data from Department of Commerce. ATS-Outbound collects information about exporters and cargo and conveyances used to transport cargo from the United States to destinations outside its borders. This information includes Shippers Export Declarations from Automated Export System (AES); Export Manifest Data from AES; Export Air Way Bills of Lading; and Census Export Data from Department of Commerce.

c) Data Privacy

The Privacy Office has been working closely with CBP to ensure that ATS programs satisfy the privacy documentation required for operation. A Privacy Impact Assessment was completed by CBP in November 2006 and is available on the DHS website. The System of Record Notice was published at 71 Fed. Reg. 64543 (Nov. 2, 2006).

Authorized CBP officers and other government personnel located at seaports, airports, and land border ports around the world use ATS to support targeting, inspection, and enforcement related requirements. Nonetheless, ATS supports but does not replace the decision-making responsibility of CBP officers and analysts. The information accessed in ATS does not form the conclusion about whether or not to take action regarding an individual, but it merely serves to assist the CBP officer in either refining the officer's analysis or formulating the officer's queries to obtain additional information upon which a decision will be based.

ATS-Inbound and ATS-Outbound rely upon the source systems that provide data for analysis to ensure accuracy and completeness. When a CBP officer identifies any discrepancy regarding the data, the officer can take action to correct that information, when appropriate. Since ATS is not the main system of records for most of the source data, ATS monitors source systems for changes to the source system databases. Continuous source system updates occur in real time, or near real time, from the Treasury Enforcement Communications System (TECS), which includes data from the National Criminal Information Center (NCIC), the Automated Commercial Environment (ACE), the Automated Manifest System (AMS), the Automated Commercial System (ACS), and the Automated Export System (AES). When corrections are made to data in source systems, ATS updates this information immediately and uses only the latest data. In this way, ATS integrates all updated data (including accuracy updates) in as close to real time as possible.

Individuals may gain access to their own data from source systems that provide input to ATS through procedures set out in the SORN for each source system. In addition, the Freedom of Information Act (FOIA)[36] provides an additional means of access to personally identifiable information held in source systems.

[36] 5 U.S.C. § 552.

Procedures for individuals to access ATS information are outlined in its SORN.[37] FOIA requests for access to information for which ATS is the source system may be directed to CBP.

CBP has created a Customer Satisfaction Unit in its Office of Field Operations to provide redress with respect to inaccurate information collected or maintained by its electronic systems, which include ATS. This process is available even though ATS does not form the sole basis of identification of enforcement targets.

System access to the ATS system is periodically audited by the process owner to ensure that only appropriate individuals have access to the system. CBP's Office of Internal Affairs also conducts periodic reviews of the ATS system to ensure that the system is being accessed and used in accordance with documented DHS and CBP policies.

d) Data Security and Integrity

ATS underwent the Certification and Accreditation process in accordance with DHS and CBP policy, which complies with Federal statutes, policies, and guidelines, and was certified and accredited on June 16, 2005, for a three-year period. A Security Risk Assessment was completed on March 28, 2006, in compliance with FISMA, OMB policy, and National Institute of Standards and Technology (NIST) guidance.

Further, access to the data used in ATS-Inbound and ATS-Outbound is restricted to persons with a clearance approved by CBP, approved access to the separate local area network, and an approved password. All CBP process owners and all system users are required to complete bi-annual training in privacy awareness and must pass an examination. If an individual does not take training, that individual will lose access to all computer systems, which are integral to his or her duties as a CBP Officer. Lastly, CBP employees are required to have access to TECS as a condition precedent to obtaining access to ATS. All TECS users are required to pass a specific computer-based training program regarding individual privacy, safeguarding data, information security, and CBP policies relating to dissemination and control of the data maintained within TECS. Such training is required prior to obtaining access to TECS and then bi-annually thereafter to maintain access.

3. Data Analysis and Research for Trade Transparency System

ICE maintains the Data Analysis & Research for Trade Transparency System (DARTTS), which is the primary automated investigative tool of the ICE Trade Transparency Unit (TTU). The TTU examines U.S. and foreign import, export, and financial data to identify

[37] See 71 Fed. Reg. 64543 (November 2, 2006).

anomalies in patterns of trade. DARTTS was created from a theoretical approach to identify criminal activity indicative of trade-based money laundering or other trade-related crime. DARTTS was designed to answer common investigative questions about criminal activity. DARTTS analyzes import/export and financial data from the United States and its foreign partners to identify trade anomalies and financial irregularities. These anomalies and irregularities often serve as indicators of money laundering, customs fraud, contraband smuggling, and even tax evasion (e.g. value added tax, income tax, duty, tariff). DARTTS allows for data mining and analysis not available in other systems. For example, DARTTS allows the user to produce aggregate totals for importations of currency, and then sort out any number of variables, such as country of origin, party name, or total currency value.

Information imported into DARTTS from various sources is compared and analyzed by the TTU. ICE Special Agents and Intelligence Research Specialists with experience in conducting financial, money laundering, and trade fraud investigations use the system analysis to identify criminal activity, initiate criminal money laundering and customs fraud investigations, and provide support to ongoing field investigations. DARTTS allows the TTU to identify anomalies in patterns in trade,[38] analyze Bank Secrecy Act information,[39] analyze the movement of cargo,[40] and analyze foreign financial information.

a) Goals and Plans for the Program

DARTTS is a small-scale, stand-alone system that uses commercial off-the-shelf (COTS) software to aid in the analysis of structured data found in databases. System owners from the ICE Financial and Trade Investigation’s TTU are collaborating with the ICE Office of the Chief Information Officer and the Information Systems Security Manager to provide DARTTS to users in a web environment.

[38] This analysis requires the comparison of U.S. and foreign import and export data for the following factors: Harmonized Tariff Schedule, country of origin, manufacturer, importer, broker, unit price, commodity activity by time period and port of import/export.

[39] This information is analyzed to identify patterns of activity involving the importation and exportation of currency, deposits of currency in financial institutions, reports of suspicious financial activities, and the identity of parties to these transactions.

[40] This analysis involves use of manifests, bills of lading and commercial data sources that identify carriers, containers, and their movements.

b) Data Sources, Deployment Dates and Effectiveness

DARTTS currently contains information about U.S. imports from CBP's Automated Commercial System (ACS) (the system used to track, control, and process all commercial goods imported into the United States); U.S. exports from the Department of Commerce and the Bureau of Foreign Trade Statistics; commercial trade data from PIERS[41] and Tradebytes;[42] foreign trade data from foreign countries; and currency and monetary instruments reports, suspicious activity reports, IRS data and currency transaction reports from Financial Crimes Enforcement Network (FinCEN).

Information in DARTTS is regulated under the Privacy Act, the Trade Secrets Act, and the Bank Secrecy Act.

DARTTS has been deployed in its current configuration for approximately two years but was built upon an older analysis tool, formerly known as the Numerically Integrated Profiling System (NIPS) that was used for approximately 10 years by the previous U.S. Customs Service.

Information identified by using DARTTS is used to initiate as well as to support numerous international and national ICE-led criminal investigations.

c) Data Privacy

DARTTS is a legacy system brought to DHS from the Treasury Department. ICE is currently working to complete a PIA and SORN for DARTTS, but the privacy documentation is not yet complete. The accuracy of the data in DARTTS is dependant on the accuracy of the source data. Accuracy can always be verified by querying the underlying systems from which the source documents are obtained. Access to data is limited to staff appropriately cleared with ICE and who have a need-to-know for all information processed in the system. All users of DARTTS must complete the annual ICE Computer Security Awareness Training.

Due to the nature of the data and the way data interact with the analytical application, regular audits are conducted each time an entire data set is replaced and backed up during the data updating process.

[41] PIERS is a comprehensive private sector database of import and export information on the cargoes moving through ports in the U.S., Mexico, Latin America, and Asia. PIERS collects data from over 25,000 bills of lading everyday, then translates the raw data into information companies use to analyze trade trends and forecasts as well as their competition.

[42] TradeBytes is the product of a Canadian company that compiles U.S. Government documents showing, on a trade-by-trade basis, what American buyers are importing. TradeBytes market data includes the name and address of each buyer and seller, plus a detailed product description.

d) Data Security and Integrity

DARTTS operates on a closed LAN with no Internet connectivity. The system is physically secured within ICE headquarters in Washington, D.C. No environmental or technical factors raise special security concerns. DARTTS can only be accessed through specially designed clients, which are provided only to authorized users. All authorized users have been issued a DARTTS user name and password. Because DARTTS is a research and analysis tool, users are only allowed to view the available data made available through the custom-designed interface. General users are prevented from directly accessing the original source data and may not change data directly.

The DARTTS system has approximately 30 individual users, which include ICE Senior Special Agents and Intelligence Research Specialists. Federal employees and contract staff assigned to ICE Office of Investigations also support the system.

4. Freight Assessment System

a) Goals and Plans for the Program

The Transportation Administration Agency's (TSA) Freight Assessment System (FAS) is a cargo risk assessment tool used to identify cargo that may pose a heightened risk to passenger aircraft transporting that cargo. The identified cargo will be flagged and set aside for further inspection by air carriers. To reduce the current reliance on random inspections, FAS plans on using a human-developed risk rules engine. TSA's future plans include incorporating automated analysis employing machine-based rules and using additional data sources to identify and assess high-risk cargo. The FAS designers are trying to identify unique information elements for pattern recognition and TSA plans in the future to develop and incorporate predictive indicators.

FAS is now preparing for a pre-system test to start late fiscal year 2007, with five (5) industry participants. The first testing period is scheduled for 60 days of steady state operations, after which post analysis reports will be completed. The second testing period will commence before the end of calendar year 2007, with the same participants as the first period and including some additional participants. As currently planned, no personally identifiable information will be used as part of the FAS, so no privacy documentation is required.

b) Data Sources, Deployment Dates and Effectiveness

The data sources FAS uses include the following: the air carriers' house and master airway bills (no PII from airway bills will be included), the TSA's Performance and

Results Information System (PARIS) (company names only);[43] the TSA Indirect Air Carrier Management System (IACMS)[44] (company names); the TSA Known Shipper Management System (KSMS) (company names);[45] Dun & Bradstreet; and public crime data (statistical data only from a website hosted by the University of Michigan).

The pre-testing deployment is planned for late fiscal year 2007. The pre-system testing will provide opportunity to validate risk assumptions, measure and observe effects on operations, and gain feedback from industry. Findings will lead to the refinement of the tool.

Validity testing was completed in a proof of concept that scored live airway bills with a prototype model of FAS. TSA found results to be inline with expectations and inline with CBP's Automated Targeting System determinations. TSA was able to inspect cargo that the tools identified as elevated risk and found that physical inspection confirmed the higher risk determination. Standards around validating the data mining models are part of the COTS software solution selected to support the pre-system test. Within the software, TSA has an internal analyzer tool that provides reports against the data received and audit capability that TSA's risk analysts will use to view scores, make updates, and manual adjustments to the model.

c) Data Privacy

Because TSA specifically designed the FAS system not to hold any personally identifiable information in the FAS system, no privacy documentation beyond the PTA

[43] PARIS compiles the results of cargo inspections and the actions taken when violations are identified. The PARIS database provides TSA a web-based method for entering, storing, and retrieving performance activities and information on TSA-regulated entities, including air carriers and indirect air carriers. PARIS includes profiles for each entity, inspections conducted by TSA, incidents that occur throughout the nation, such as instances of bomb threats, and investigations that are prompted by incidents or inspection findings.

[44] IACMS is a management system used by TSA to approve and validate new and existing Indirect Air Carriers. This management system and application is intended for freight forwarders wishing to receive TSA approval to tender cargo utilizing an Indirect Air Carrier certification. The IACMS is not intended for individuals wanting to ship cargo. An Indirect Air Carrier means any person or entity within the United States not in possession of an Federal Aviation Administration air carrier operating certificate, that undertakes to engage indirectly in air transportation of property and uses, for all or any part of such transportation, the services of a passenger air carrier. See PIA on TSA's Air Cargo Security Requirements, published on the DHS Privacy Office web site on May 25, 2006, which provides additional information on the privacy impact of the IACMS and Known Shipper.

[45] KSMS uses commercial databases to verify the legitimacy of shippers. Known shippers are entities that have routine business dealings with freight forwarders or air carriers and are considered trusted shippers. In contrast, unknown shippers are entities that have conducted limited or no prior business with a freight forwarder or air carrier.

will be required. Dun & Bradstreet data is used to confirm that the Indirect Air Carrier certification issued to a business owner matches the name on the certificate issued by TSA's Indirect Air Carrier regional coordinators. The information regarding the business owner is not entered into FAS as a factor. It is only used to verify the name of the business owner.

d) Data Security and Integrity

The FAS program has completed a Privacy Threshold Analysis and is undergoing a full Certification and Accreditation (C&A) in accordance with DHS and TSA information technology security requirements. System access is limited to authorized users. Unauthorized access is controlled through system lockdowns, role based access control, and the use of authentication servers. All systems security services will maintain DHS and TSA Certification and Accreditation standards. The systems will be hardened according to DHS and TSA standards. HTTPS (TLS) will be the only accepted method of communication with the web server (no HTTP). Audit capability exists within FAS and will be used to review the reasons for system hits.

B. Update on Data mining Programs Identified in the July 2006 Report

1. Enterprise Data Warehouse

CBP created the Enterprise Data Warehouse (EDW) to compile workforce statistics to assist CBP in determining the most efficient staffing of operational personnel to meet its ongoing trade compliance and border enforcement missions. Specifically, the data sources EDW uses allow queries to help plan, for example, staffing of shift work related to cargo inspections and passenger processing. EDW is comprised of several data marts, which are groupings of information taken from CBP law enforcement systems including TECS (Treasury Enforcement Communications System) and SEACATS (Seized Asset and Case Tracking System). The information taken from both TECS and SEACATS consists of case management data and statistics such as the number and types of seizures of imported merchandise occurring at specific ports of entry or the number and types of positive examinations. CBP operating units use the data mart reports to improve performance of their operations.

Although the July 2006 Report included EDW as a data mining activity, the Privacy Office does not believe it meets the criteria set forth for this Report. All of the data sources EDW uses are owned and compiled by the federal government and were created for law enforcement purposes. Additionally, the queries of the EDW are not conducted to find a pattern of terrorist or criminal activity, but rather to allocate resources and prioritize CBP staffing.

2. Law Enforcement Analysis Data System

ICE uses the Law Enforcement Analysis Data System (NETLEADS) program to facilitate law enforcement activities and intelligence analysis by enhancing the efficiency of searches involving multiple data sources and patterns and trends analyses. NETLEADS was created in 1997 by the Department of Justice Immigration and Naturalization Service to share sensitive law enforcement data between immigration inspectors, criminal investigators, and Border Patrol. NETLEADS tools are now used within ICE and CBP for intelligence and investigative analysis.

a) Description and Purpose of the Program

The NETLEADS program is a tool suite designed to provide a means of performing more efficient searches on a combination of structured data, such as Oracle, Microsoft and mainframe databases, and unstructured data, such as textual reports, open source documentation, Web pages, reports of investigation narratives from ICE databases, and images such as PDF files. The program includes two data mining and visualization features – a link analysis, which permits analysis of connections between entities such as individuals and organizations, and a trend analysis, which permits the identification of trends across cases. In link analysis, a visual display can be created to demonstrate links between entities; the diagram can be stored, and then compared to similar diagrams generated earlier or later with different versions of the data through the Timeline Analysis feature. NETLEADS also permits its users to look for trends across cases. These tools allow disparate data elements to be examined within the analysis constructs of ICE investigations and intelligence operations.

NETLEADS searches are conducted using “fuzzy” logic, a form of algebra employing a range of values that is used in decision-making with imprecise data. The search results display weightings indicating how closely they match the requested data, allowing the analyst to judge its relevance. When an analyst selects a particular result, the detailed data is extracted directly from the source system. NETLEADS uses groupings of key words seeking pattern analysis associated with patterns and trends, which produce a graphical link analysis from data in the databases.

NETLEADS includes the ICE nationwide Significant Event Notification (SEN) application, providing real time notification to the ICE Operations Center incident and intelligence activities. The SEN also provides e-mail alerts notifying key ICE managers that new incident data is available. NETLEADS uses dynamic, ad hoc data queries allowing agents and analysts to search any topic, as well as a standard search query for routine searches.

b) Data Sources, Deployment Dates, and Policies and Procedures

NETLEADS performs simultaneous search and analysis of records, documents, and images from fifteen data sources. NETLEADS tools perform searches on authorized databases maintained by DHS via the protected DHS secure intranet. ICE is creating and executing Memoranda of Understanding and Service Level Agreements that authorize data sharing efforts with other federal and state government law enforcement and intelligence agencies, such as Department of State, Bureau of Prisons, California Department of Justice, and the Law Enforcement Information Exchange (LInX) Governance Board.

The data sources for NETLEADS include sensitive government law enforcement databases as well as public sources, such as news services. The government data sources provide the raw reporting from multiple agencies government-wide with data on topics such as terrorism, human smuggling, contraband, criminal operations, and investigations. These data sources are protected by the originating agency and access is gained via password protections once Memoranda of Understanding between ICE and the releasing agencies have been authorized. In addition to government databases, NETLEADS tools access data from commercial sources, such as the Associated Press, News Edge, and public internet search engines.

The NETLEADS project has been in operation since approximately 1997. It was established to share of sensitive law enforcement data between immigration inspectors, criminal investigators, and the Border Patrol. NETLEADS tools are now used within ICE and CBP. Over 10,000 ICE Special Agents, Border Patrol agents, the Border Patrol Field Intelligence Center, Sector Intelligence offices, and ICE Intelligence and CBP offices in ports of entry, sea ports, and airports worldwide have access to NETLEADS.

c) Data Privacy

NETLEADS is used for investigative and intelligence purposes only. No application within the tools suite is used to grant benefits or make unevaluated automated decisions. The underlying government databases searched by NETLEADS tools are covered by various SORNs. Nevertheless, ICE is drafting a DHS SORN for the project. In addition, although NETLEADS predates the E-Government Act of 2002, and therefore would not ordinarily require a PIA unless it had experienced significant modifications since the Act, the Privacy Office is now requiring NETLEADS to undergo a PIA concurrently with drafting a DHS SORN. As noted in the discussion of the Privacy Office compliance process above, the Privacy Office is working closely with ICE and other DHS components to update and re-issue the legacy SORNs as DHS SORNs, as necessary.

d) Data Security and Integrity

The ICE intelligence information technology group implements, tests, and evaluates all ICE systems against DHS-directed standards. NETLEADS' open systems architecture allows for the implementation of enhanced security as DHS policies are completed and implemented. All employees who use NETLEADS must comply with the annual requirement to complete information security procedures and policy training.

Information in NETLEADS is used only by authorized DHS employees who have a need to know and have been subject to appropriate security and background investigations. Users with access to NETLEADS tools have role-based access to underlying databases. Prior to being granted access to these databases, all users receive annual automated data processing security training. As an ICE enterprise application development project, NETLEADS is subject to review by the Chief Information Officer and ICE business counsel for operation and design concept and proposed implementation. The System Lifecycle Management process for ICE and DHS includes security and privacy validation prior to approval of new systems procurement or development.

Audit monitoring is conducted seven days a week, 24 hours a day, on all transactions for the NETLEADS system by the Internet protocol (IP) address of the user and user account. These are screened by systems administrator and are triggered by automated attack or penetration attempts. Entries into databases are audited by user account identification.

The NETLEADS database comprises intelligence and investigative data derived from other data systems. The quality of the data in NETLEADS is determined by the quality of the data in the system from the source systems. The data in NETLEADS is updated from those systems on either a daily or weekly basis as appropriate.

3. ICE Pattern Analysis and Information Collection System

ICE developed its Pattern Analysis and Information Collection System (ICEPIC) to assist investigators in disrupting and preventing terrorism activities. ICEPIC is a toolset that assists ICE law enforcement agents and analysts in identifying suspect identities and discovering possible non-obvious relationships among individuals and organizations that are indicative of violations of customs and immigration laws as well as possible terrorist threats and plots. All ICEPIC activity is associated with ongoing and valid law enforcement investigations.

Although ICEPIC was included in the July 2006 Report, it does not meet the definition for this report in that all queries using the tool are targeted using a personal identifier in order to find information about the person or about persons related to the target of the investigation. ICEPIC reveals relationships to an identified target, but it is not used to reveal a predictive pattern. From the relationships identified by the tool, ICE agents will develop specific leads and intelligence for active and new investigations. A draft PIA and SORN for ICEPIC are currently being reviewed by the Privacy Office.

4. Intelligence and Information Fusion

The Office of Intelligence and Analysis (I&A) designed the Intelligence and Information Fusion (I2F) to provide intelligence analysts with an ability to view, query, and analyze multiple data sources. It is a national security system and therefore exempted from certain statutory privacy documentation; however, the Privacy Office requires such systems to conduct a PIA, but does not publish the reviewed PIAs.

a) Description and Purpose of the Program

I2F uses commercial off-the-shelf (COTS) tools, including tools for search, link analysis, entity identification and extraction, and place name recognition. The tools include the ability to discover patterns and relationships across multiple sources of data, which are sometimes from classified sources. I2F enables analysts to look at information within a data set and identify patterns and relationships using non-algorithmic search technology. The searches are based upon human analysis using analytical tools to help reveal patterns; however, at this time, I2F is not applying predictive models.

b) Data Sources, Deployment Dates, and Policies and Procedures

I2F only incorporates data from government sources. I2F uses electronic media in various text formats including ASCII, MS Word, and PDF. Data are expected to be validated for accuracy and internal consistency through comparison across multiple data sources.

Since the July 2006 Report, a new release of I2F entered security testing. The new releases contain additional features, including a new search engine, entity extraction tool, web-based visualization user interface, and geographic place name recognition tool.[46]

Data is collected and retained under the authority of Executive Order 12333 – United States Intelligence Activities.[47] Information contained in I2F is used within I&A in preparation of intelligence products or to prepare responses to intelligence requests for

[46] The tools included a Convera search engine, the InXight entity extraction tool, an I2 web-based visualization user interface, and the MetaCarta geographic place name recognition tool.

[47] 46 Fed. Reg. 59941; 3 CFR § 1981.

information. There are no other authorized uses for the I2F system. Information is not shared with any other organization outside of I&A.

c) Data Privacy

A PIA is being conducted for I2F, but will not be made public, since I2F is a national security system and the PIA contains sensitive information. As noted before, the E-Government Act exempts classified systems from the requirement to conduct a PIA; however, the Privacy Office requires a PIA, but does not publish the document. I2F is covered by the Homeland Security Operation Center SORN.[48] I&A is currently working with the Privacy Office on updating its System of Records Notice as part of various re-organizations that have affected the former Information Analysis and Infrastructure Protection Directorate.

I2F is used exclusively to develop intelligence products to provide indications and warnings of possible threats to homeland security. Data is validated for accuracy and internal consistency through comparison of multiple data sources. When analysts view the results of a data search, I2F highlights relevant people, places, and things described in the retrieved document. The analyst makes constructs by connecting the highlighted passages in the retrieved documents (i.e., "Person X traveled to Boston with the weapon.") The analyst may select these constructs for visualization in a link chart that appears in a window in the web browser. By combining several constructs, a graphical representation of the relationship between objects in the source material can be viewed.

Extensive, established written procedures are followed for the vetting and review of finished intelligence products. I2F data are not used to make automatic decisions about individuals. Intelligence analysts use the data for the purpose of determining whether an individual, organization, or set of events pose a threat to homeland security. That information is communicated in the form of an intelligence product, which may be read by an agency that would initiate appropriate action including the opening of an investigation.

When data include personally identifiable information, some data elements are redacted in finished products to comply with intelligence oversight regulations and Executive Order 12333.[49] According to I&A implementation guidelines pertaining to Executive

[48] Homeland Security Operations Center Database, 70 Fed. Reg. 20156 (Apr. 18, 2005).

[49] For agency members of the Intelligence Community, Executive Order 12333 defines the applicable standards for the collection, retention, or dissemination of information concerning U.S. persons reflecting a careful balance between the needs of the government for such intelligence and the protection of the rights of U.S. persons, consistent with the reasonableness standard of the Fourth Amendment, as determined by

Order 12333, data can be kept for up to 180 days before the evaluation is conducted. At that time, an analyst reviews the data to determine whether the data may be retained.[50] Intelligence analysts review retained data periodically for continued relevancy and personally identifiable information no longer relevant to the mission is deleted.

Data used by I2F consists of field reports of activities and incidents related to terrorism. There are occasions where the original source of the data intentionally creates false information[51] that is gathered during the normal course of field reporting. I2F assists analysts in revealing this false information by using many different data sources as possible to validate a particular intelligence hypothesis. The cross referencing and correlation process is used to evaluate the quality of the data received and the reporting source. Inaccurate information is retained and flagged as unreliable. This information must be kept to document the validity of the information source. At this time, the I2F program does not yield predictive models and, as such, the program has not adopted standards for validation of such models.

Given the sensitive nature of the data within I2F, I2F does not permit individual "data subjects" to review information it maintains, and it does not have redress procedures for individuals identified for additional investigations. Additionally, as a national security system containing intelligence data, I2F is exempt from the relevant and necessary standards under the Privacy Act of 1974.

d) Data Security and Integrity

I2F incorporates an auditing system, which is inspected and approved by an independent security certification during the security certification and accreditation process. The certification and accreditation findings are evaluated by the Designated Approval Authority before a system Approval to Operate is issued.

Audit trails are generated at multiple points of access – web portal, database, and operating system - to search for patterns of attack that may indicate that someone is

factual circumstance. Under this process, the identity of U.S. persons must be protected by the agency and may only be available under three specific situations: 1) when the U.S. person has consented to such us of his or her identity, 2) when the information with the identity of the U.S. person is publicly available, or 3) when the identity of the U.S. person is necessary to understand or assess foreign intelligence or other categories of information pursuant to the Executive Order. If one of the above exceptions does not apply, the identity of a U.S. person must be redacted and not collected.

[50] Data can be retained if it falls into one of the following categories: information obtained with consent; publicly available information; terrorism information; vulnerabilities information; international narcotics activity information; border security information; information related to threats to safety; administrative information.

[51] This would be information that is fake, misleading, or purposely inaccurate.

trying to crack the system and patterns of abuse that may indicate that someone is conducting a mass data extraction.

Audits are automatic and not random. The Information Systems Security Officer (ISSO) is required to review the automatic audit results weekly. Audit results are maintained for five years. On a retroactive basis, should some incident occur, logs can be reviewed after the fact and determine who accessed what information. Employees are notified that audits are conducted. Warning banners appear at log-on to inform users that all activity at every workstation is monitored.

5. Fraud Detection and National Security Data System

USCIS developed the Fraud Detection and National Security Data System (FDNS-DS) as a case management system to record, track, and manage immigration inquiries, investigative referrals, law enforcement requests, and case determinations involving benefit fraud, criminal activity, public safety, and national security concerns. The FDNS-DS system is an upgrade of the Fraud Tracking System (FTS).

FDNS-DS was included in the July 2006 Report; however, this program is not yet operational and does not meet the definition of data mining required for this Report. FDNS-DS does not conduct searches to find predictive patterns of terrorist or criminal activity. FDNS-DS is envisioned in the future to be able to identify new fraud schemes and new associations nationwide to provide additional leads for investigation; however, there are no existing plans to develop FDNS-DS to perform predictive data mining.

At present, FDNS-DS is a case management system used to track fraud leads, cases where a suspicion of fraud has been articulated, referrals to ICE, and requests for assistance from law enforcement agencies. FDNS-DS tracks immigration related fraud, public safety referrals to ICE, and national security concerns discovered during the background checks performed on persons applying for immigration-related benefits. FDNS consolidates status information and results of administrative investigations, background checks, adjudication processes, and benefit fraud assessments required for completion of immigration eligibility petitions and application determinations in order to identify possible fraud.

6. National Immigration Information Sharing Office

The National Immigration Information Sharing Office (NIISO) was included in the July 2006 Report; however, it does not meet the definition used for this Report. USCIS maintains NIISO, which is responsible for responding to requests for information regarding immigration files from other DHS components, external law enforcement, and intelligence agencies. NIISO only uses DHS data sources, specifically USCIS' Computer-Linked Application Information Management System (CLAIMS), which is a repository of alien benefit information. It will later use the FDNS-DS as it is developed. The DHS Office of Intelligence and Analysis is preparing to expand the NIISO program. This expanded NIISO will be a distinctly different program than the NIISO currently managed by USCIS, and is not yet in existence. All NIISO activity will be based on predicated search information and will not employ data mining techniques or programs. Nonetheless, the Privacy Office is working with I&A to address the impact the new version of NIISO may have on privacy through the privacy compliance process.

V. Progress on Recommendations for DHS Data Mining Activities

The Privacy Office continues to engage with the programs involved in data mining activities regarding the implementation of the recommendations outlined in that report. The following update reflects progress made to date on those recommendations. The full text of the recommendations in the earlier report is presented as an appendix to this report.

A. Centralized Oversight of DHS Data Mining Technologies

The Privacy Office began internal discussions with the S&T Directorate and the OCIO regarding the creation of a focused effort to advance the management of data mining activities across DHS. This effort will ultimately draw upon the various components and programs identified in this report, as well as the Office of the Chief Information Officer, the Office of the Inspector General, the Office of the General Counsel, and the Policy Office to establish standards, policies, and practices for data mining activities throughout the Department.

Although, the July 2006 Report suggested using the DHS Data Integrity Board[52] to review data mining activities, given the complexity of the concepts presented by data mining programs and the expertise necessary to evaluate properly data mining technologies, the Privacy Office now recommends the creation of a coordinating group of all the data mining programs at the Department to harmonize policies and processes to

[52] The July 2006 Report referred to the DHS Data Integrity Board as the "DHS Privacy and Data Integrity Board." The DHS Data Integrity Board examines and approves data sharing agreements between DHS and other departments as required under the Privacy Act. The Computer Matching and Privacy Protection Act of 1988 (Pub. L. 100-503, October 18, 1988) amends the Privacy Act of 1974 to establish procedural safeguards affecting agencies' use of Privacy Act records in performing certain types of computerized matching programs. The Act requires agencies to conclude written agreements specifying the terms under which matches are to be done. It also provides due process rights for record subjects to prevent agencies from taking adverse actions unless they have independently verified the results of a match and given the subject 30 days advance notice. Oversight is accomplished in a variety of ways: by having agencies (a) publish matching agreements, (b) report matching programs to OMB and Congress; and (c) establish internal boards to approve their matching activity.

ensure appropriate privacy protections accompany the use of data mining at the Department.

B. Policies Against Automated Decision Making

DHS has not yet issued any department-wide policies addressing the use of data mining or the use of automated decision making; however, no program described in this Report makes automated decisions about individuals. The data mining tools used at DHS today solely support human decision making. In general, the programs implementing data mining technologies understand the newness of the technology and wish to incorporate it into existing analytical processes rather than replace human decision making. The Privacy Office plans to work with DHS leadership, and in particular the S&T Directorate and OCIO, to develop policies to protect against automated decisions about individuals as outlined in the July 2006 Report.

C. Anonymization

The Privacy Office began discussions with the S&T Directorate to explore the use of anonymizing techniques in DHS programs, including integration with data mining activities. One of the initial steps will be for the Privacy Office to examine the effectiveness of anonymization techniques in terms of specific metrics and possible standards.

The Privacy Office discussed concepts of deploying anonymizing technologies with software and technology developers and continues to monitor progress in this area. The Privacy Office communicated these concepts with system developers within DHS to understand how these technologies may be integrated into the development of these programs. The Privacy Office will use the results of the research into these technologies to determine how best to integrate anonymization within DHS systems.

This research will likely involve determining which systems contain personally identifiable information, assessing the effectiveness of anonymization techniques and technologies in the context of the pragmatic missions and environments of those systems, and, finally, evaluating the manner in which proven anonymization techniques or technologies could be incorporated into the DHS system development life cycle and operational management environment. The Privacy Office will also look to understand the levels of de-identification that can be achieved through different anonymization techniques. To assist the Privacy Office in this research, the Privacy Office is considering holding a public workshop later this year about data mining and to explore anonymizing tools and their application to data mining.

D. Data Quality

DHS operates an internal organization entitled the “Data Management Working Group,” which is part of the Enterprise Data Management Office (EDMO) within the Office of the Chief Information Officer. One of the responsibilities of this group will be to coordinate with DHS components in development of data related standards for use across the Department. The EDMO works with the Standards Executive in the S&T Directorate to develop data quality standards and then to promote adoption of these standards by DHS. The Privacy Office will continue to work with the EDMO on identifying standards for data quality in the context of the various sources of data used by DHS systems.

E. Standards for Validating Data Mining Models

The development of data mining systems within DHS often begins with initial research conducted by the S&T Directorate, so it is appropriate that S&T Directorate play a leading role in developing a framework for validating data mining models. The operation of individual components of a data mining system should be evaluated against specific research data and measured for performance in terms of accuracy, speed, and scalability. Accuracy is measured by evaluating a component’s ability to locate specific relevant data within a larger collection of irrelevant data. Speed is measured by time and scalability is measured by evaluating speed over large data sets. Further work needs to be done to define the standards for validating data mining models. Standards for validating models will be one of the topics at the workshop on the data mining recommendations the Privacy Office is considering holding later this year.

F. Review & Redress Policies, Procedures & Training

DHS has department-wide security protocols that implement FISMA requirements and NIST standards; however, DHS does not yet have a department-wide policy specifically on redress and training for data mining programs.[53] Specific redress programs have not been established for the programs described in this report, as they are either not yet operational or do not make decisions affecting individuals. Nonetheless, all underlying data used in data mining must reside in a system of records and may be accessed and corrected to the extent allowed within the SORN for that particular Privacy Act system. In addition, the Privacy Act and Freedom of Information Act enable individuals to

[53] DHS has launched the DHS Travel Redress Inquiry Program (DHS TRIP) to provide a single point of contact for individuals who have inquiries or seek resolution regarding difficulties they experienced during the travel screening process at U.S. transportation hubs, such as airports and train stations, or across U.S. land and sea borders. DHS TRIP is a redress program for traveler experiences related to security screening or processing. Current DHS data mining programs do not affect the general public’s traveling experience.

request access to information. With regard to training, all DHS employees and contractors undergo security training and will receive privacy awareness training later this year as the Privacy Office rolls out its new online training program.

G. Audits

The programs described in this Report all implement FISMA, including requirements for audit controls. This is largely the result of departmental policies that mandate all systems to be audited for Certification & Accreditation and against DHS Security 4300A audit control requirements. The Privacy Office intends to examine the various types of audit controls being employed and to work towards creating a unified approach to audits of data mining programs across the agency.

VI. Conclusion

The Privacy Office is pleased to provide Congress its second report on DHS data mining activities. Although there are many ways to analyze data, data mining may represent an approach that warrants particular attention. As part of its statutory obligation to ensure that DHS use of technology sustains and does not erode privacy, the Privacy Office requires all programs to conduct a Privacy Impact Assessment to describe their use of analytical tools and to consider their impact on privacy. Such tools can include risk assessments and link analysis, as well as data mining, which Congress mandated in the Homeland Security Act as one of the analytical tools the Department should use in fulfilling its mission.

The DHS Office recognizes its unique role in monitoring the use of technology and its impact on privacy. The recommendations of the July 2006 Report have been in place for only a short time, and programs will need more time to consider whether and how the Report recommendations apply. The Privacy Office will continue to work with S&T, the Department's research component, to develop departmental guidance for data mining activities. The Privacy Office will also work to educate the Department on how to apply privacy protections to data mining programs and to implement the recommendations of the July 2006 Report.

VII. Appendix I: Conclusions and Recommendations (Excerpted from the July 2006 Report)

Based on our analysis of DHS activities that involve current and projected future uses of data mining, we note that data mining is usually only one part of a larger set of analytic activities and tools. Such analytic activities include searches and traditional analyses.

Although DHS programs that employ data mining tools and technologies also employ traditional privacy and security protections, such as Privacy Impact Assessments, Memoranda of Understanding between agencies that own source data systems, privacy and security training, and role-based access, we recommend additional protections that are aimed specifically at addressing the privacy concerns raised by data mining and we will take steps to implement these recommendations within the Department.

1. Prior to the start of any data mining activity, the authority of the agency to undertake such activity should be determined to be consistent with the purposes of the data mining project or program. The authority to collect or aggregate data required to perform the data mining should also be ascertained, whether the project involves collection of new data or aggregation of existing data from various sources. While oversight functions exist in different components within DHS,[54] the Department as a whole could benefit from more centralized oversight with a broader view of DHS activities and data holdings. One such body, which could assist the function of overseeing DHS data mining program, is the DHS Privacy and Data Integrity Board, an internal privacy board that is charged under the Privacy Act to examine and approve data matching agreements between DHS and other departments and that considers Departmental privacy issues. The Board, which includes representatives from all DHS components, the Office of the Inspector General, the Chief Information Officer, and the Office of General Counsel, and is chaired by the DHS Chief Privacy Officer, could provide oversight and confirmation to ensure responsible application of data mining tools and technologies. The Board assisted with the collection of data for this report concerning current and planned data mining programs within DHS.
2. As discussed in the report, data mining searches for patterns, relationships, and rules in the data without basing this search on observations or a theoretical model. Because the existence of patterns in the data may not reflect cause and effect, data mining tools should be used principally for investigative purposes, DHS components that use data mining tools should have written policies stating that no decisions may be made automatically regarding individual rights or benefits solely on the basis of the results

[54] Including the DHS Office of Civil Rights and Civil Liberties.

produced by patterns or rules derived from data mining.

3. Because the patterns, relationships, and rules in the data may not be derived from specific personal identifiers, such as a name or Social Security Number, when a data set includes personally identifiable information, data mining projects should give explicit consideration to using anonymized data in data mining activities. A discussion of the extent to which anonymization was considered should be included in the Privacy Impact Assessment for such a data mining project.
4. Data quality plays an important role in the ability of data mining techniques to produce accurate results. DHS should adopt data quality standards for data used in data mining. Application of these standards, which should affect systems using both data from government and commercial sources, should be ensured prior to the deployment of data mining models or predictive rules for use in the field.
5. In order to ensure that data mining models produce useful and accurate results, DHS should adopt standards for the validation of models or rules derived from data mining. Evaluation of the model validation process and the ability to meet these standards should be reviewed and documented prior to the deployment of the model to the field.
6. Each DHS component that employs data mining should implement policies and procedures that provide an appropriate level of review and redress for individuals identified for additional investigation by patterns, relationships, and rules derived from data mining. Because data mining algorithms most times produce highly complex patterns, relationships, and rules that may not be fully understandable as to the particular reasons for identifying the individuals, a complete procedure should include a step that a person, acting independently from the data mining process, substantiates the particular identification of individuals prior to any determinative processes and procedures. To ensure a complete understanding of the capabilities and limitations of data mining in this regard, employees who use data mining processes should be required to complete training on these policies and procedures.
7. In order to provide demonstrable accountability, each component that employs data mining should include strong, automatic audit capabilities to record access to source data systems, data marts, and data mining patterns and rules. Programs should conduct random audits at regular intervals, and all employees should be given notice that their activities are subject to such audits. These actions help underline the importance of transparency.