Privacy Impact Assessment
for the

DHS Data Framework

DHS/ALL/PIA-046

November 6, 2013

Contact Point
Tom Bush and Michael Frias
Common Vetting Task Force
Department of Homeland Security
202-447-4750

Reviewing Official
Karen L. Neuman
Chief Privacy Officer
Department of Homeland Security
202-343-1717

Abstract

The Department of Homeland Security (DHS) is developing the DHS Data Framework, which is a scalable information technology program with built-in capabilities to support advanced data architecture and governance processes. This program will alleviate mission limitations associated with stove-piped IT systems that are currently deployed across multiple operational components in DHS. It will also enable more controlled, effective, efficient use and sharing of available homeland security-related information across the DHS enterprise and, as appropriate, the U.S. Government while protecting privacy. DHS is developing three systems to test different capabilities needed to implement the program: the Neptune Pilot, the Common Entity Index Prototype (CEI Prototype), and the Cerberus Pilot. Each of these has a separate Privacy Impact Assessment (PIA).

DHS is publishing this PIA because the DHS Data Framework will use personally identifiable information (PII) collected from members of the public using information technology. This PIA covers the overall approach and vision for the program. As DHS develops the DHS Data Framework, this PIA will be updated.

Introduction

Section 101 of the Homeland Security Act of 2002, Pub. Law No. 107-296 (Nov. 25, 2002), as amended, establishes the Department of Homeland Security (Department or DHS) as an executive department of the United States. The mission of the Department is, among other things, to prevent terrorist attacks within the United States, reduce the vulnerability of the United States to terrorism, minimize the damage and assist in the recovery from terrorist attacks that do occur within the United States, support the missions of its legacy components, monitor connections between illegal drug trafficking and terrorism, coordinate efforts to sever such connections, and otherwise contribute to efforts to interdict illegal drug trafficking. At the same time, the Department has the primary responsibility to ensure that the privacy, civil rights, and civil liberties of individuals are not diminished by efforts, activities, and programs aimed at securing the homeland. To enable the Department to carry out these missions, the Homeland Security Act eliminated information firewalls between government agencies by consolidating multiple agencies under DHS.

DHS is changing the way it structures its information architecture and data governance to further consolidate information in a manner that protects individuals' privacy, civil rights, and civil liberties. The existing architecture of DHS databases, however, is not conducive to effective implementation of the "One DHS" policy. [1] Existing information is subject to privacy, civil rights and civil liberties, and other legal and policy protections, and it is collected under different authorities and for various purposes. Since 2007, DHS has operated under the "One

[1] *DHS Policy for Internal Information Exchange and Sharing*, February 1, 2007.

DHS" policy, which was implemented to afford DHS personnel timely access to relevant and necessary homeland-security information they need to successfully perform their duties. Under this policy, DHS personnel requesting information maintained within another departmental component may access such information when the requestor (1) has an authorized purpose, mission, and need-to-know before accessing the information in performance of his or her duties; (2) possesses the requisite background or security clearance; and (3) assures adequate safeguarding and protection of the information. Currently, this access is cumbersome, time-intensive, and requires personnel to log on and query separate databases in order to determine what information DHS systems contain about a particular individual.

The DHS Secretary and Deputy Secretary developed the DHS Data Framework through the Common Vetting Task Force (CVTF)[2] and collaboration among the Office of the Chief Information Officer (OCIO), the Office of Policy (PLCY), the Office of Intelligence and Analysis (I&A); the "oversight offices" the Privacy Office (PRIV), the Office for Civil Rights and Civil Liberties (CRCL), and the Office of the General Counsel (OGC): and operational components. The goal of the DHS Data Framework is to provide a user the ability to search an amalgamation of data extracted from multiple DHS systems for a specific purpose and view the information in a clear and accessible format. The current process requires duplication of key DHS systems' data within a variety of other DHS systems to make it available to complete an organization's mission.

The DHS Data Framework will create a systematic repeatable process for providing controlled access to DHS data across the Department. The DHS Data Framework will enable the implementation of efficient and cost-effective search and analysis across DHS databases in both classified and unclassified domains. The searches will identify key DHS data associated with an individual or identifier. Adhering to the DHS Data Framework will ensure access to the most authoritative, timely, and accurate data available in DHS to support critical decision making and mission functions. Finally, the DHS Data Framework will enable controlled information sharing in both classified and unclassified domains in a manner that manages search parameters and access to the underlying data while maintaining the authoritative source of data at the source system.

The DHS Data Framework defines four elements for controlling data:

(1) User attributes identify characteristics about the user requesting access such as organization, clearance, and training;
(2) Data tags label the data with the type of data involved, where the data originated, and when it was ingested;

[2] The CVTF is a Department-wide task force comprised of representatives from support and operational components dedicated to improving the efficiency of DHS's screening and vetting activities.

(3) Context combines what type of search and analysis can be conducted (function), with the purpose for which data can be used (authorized purpose); and

(4) Dynamic access control policies evaluate user attributes, data tags, and context to grant or deny access to DHS data in the repository based on legal authorities and appropriate policies of the Department.

DHS will log user activities to aid audit and oversight functions.

Initially, the data tags, context, and dynamic access will be tested. While the DHS Data Framework will incorporate a User Attribute Hub, which will maintain a listing of a system user's attributes for determining access control (e.g., component in which the individuals works, location, job series). This attribute hub is being developed through a different effort at DHS OCIO. The following capabilities will test the other three elements of the framework:

- ***Neptune Pilot:*** The Neptune Pilot, residing in the Sensitive but Unclassified/For Official Use Only (SBU/FOUO) domain, will ingest and tag data in a data repository known as "Neptune." This pilot will test the second element of the DHS Data Framework ("data tags"). Data in the Neptune Pilot will be shared with the CEI Prototype and the Cerberus Pilot, but will ***not*** be accessible for other purposes.
- ***CEI Prototype:*** The CEI Prototype, also residing on the SBU/FOUO domain, will receive a subset of the tagged data from the Neptune Pilot and correlate data from across component datasets. The CEI Prototype will test the utility of the Neptune-tagged data—specifically, the ability to ensure that only users with certain attributes are able to access data based on defined purposes using the dynamic access control process, which is described below in greater detail. This prototype will use "data tags" to test the third and fourth elements of the DHS Data Framework ("authorized purpose/use" and "dynamic access control," respectively).
- ***Cerberus Pilot:*** The Cerberus Pilot, residing in the Top Secret/Sensitive Compartmented Information (TS/SCI) domain, will receive all of the tagged data from the Neptune Pilot in a separate data repository known as "Cerberus." The Cerberus Pilot will test the ability to ensure that only users with certain attributes are able to access data based on defined purposes using the dynamic access control process. This pilot will also leverage the "data tags" to test the "authorized purpose/use" and "dynamic access control" elements of DHS Data Framework. The Cerberus Pilot will also test the ability to perform simple and complex searches across different component datasets using different analytical tools.

CVTF, operational component staff, OCIO staff, and oversight staff will closely review the testing of the capabilities. In its first phase, the Neptune Pilot will be limited to staging unclassified data contained in three DHS component databases:

(1) The U.S. Customs and Border Protection (CBP) Electronic System for Travel Authorization (ESTA);[3]

(2) The U.S. Immigration and Customs Enforcement (ICE) Student and Exchange Visitor Information System (SEVIS);[4] and

(3) The Transportation Security Administration (TSA) Alien Flight Student Program (AFSP).[5]

OCIO staff will test the processing of data in these databases to ensure (1) tagging of data is accurately occurring and supports controls necessary for protection of privacy, civil rights, and civil liberties, and (2) transferring of data to the CEI Prototype and the Cerberus Pilot.

The Cerberus Pilot will import the tagged data from the Neptune Pilot into its TS/SCI data repository and add additional classified tags as necessary. The CEI Prototype and Cerberus Pilot will demonstrate the feasibility of using automated policies to effectively control access to the tagged data in the CEI Prototype and Cerberus repositories.

Dynamic Access Control

DHS will change access control from the existing Role Based Access Control (RBAC) approach to one that includes a dynamic, granular access control mechanism and provides enhanced protection of privacy, civil rights, and civil liberties through definition and enforcement of who (User Attributes) is allowed access to individual data elements (Data Tags) for particular purposes (Context = Purpose + Function). This method requires that (1) users requesting access to information be described in a standard way (through user attributes), (2) source data be defined in a standard way (through data tagging), and (3) applications indicate for what purpose and how users will search (through context) the data. Each of these approaches will be determined through a systematic repeatable governance process led by a departmental governance body that includes both operational components and the oversight offices. With these three access control components in place, the governance body can create policies that allow the system to automatically evaluate access requests and automated policy-based decisions

[3] *See* DHS/CBP/PIA-007(c) - Electronic System for Travel Authorization (ESTA) Update, June 5, 2013; DHS/CBP/PIA-007(b) Electronic System for Travel Authorization (ESTA) - Internet Protocol Address and System of Records Notice Update, July 18, 2012; DHS/CBP/PIA-007(a) Electronic System for Travel Authorization (ESTA) Fee and Information Sharing Update, July 18, 2011; DHS/CBP/PIA-007 Electronic System for Travel Authorization, June 2, 2008, available at www.dhs.gov/privacy.

[4] *See* DHS/ICE/PIA-001(a) Student and Exchange Visitor System (SEVIS) Update national Counter Terrorism Center, June 23, 2011; DHS/ICE/PIA-001 – Student and Exchange Visitor Information System (SEVIS), February 5, 2005, available at www.dhs.gov/privacy.

[5] DHS/TSA/PIA-026 Alien Flight Student Program, December 4, 2009, available at www.dhs.gov/privacy.

to permit or deny access to information. For each classification level, an authorized user can then request DHS information (without the need for separate logins to each source system) with assurance that the information is obtained in accordance with DHS defined policy.

If successful, this multi-faceted approach will ensure that data from the original source systems will only be used by authorized personnel for authorized purposes in the CEI Prototype and the Cerberus Pilot; data in Neptune will not be accessible by any user because Neptune's intended use is data processing, not user access. In Cerberus, data tags will also enable data of various security levels to be stored in the same repository and will allow users with different degrees of access to perform searches. Data tags will ensure that each user only views the data which, by policy, role, and attribute, they have permission to access. Data tags will enable the CEI Prototype and the Neptune and Cerberus Pilots to preserve the lineage of the source data for its integrity and the purpose and intent for the data's collection when it is extracted from each system. Ultimately, the governance body (currently the CVTF) will manage the tagging of data, the assignment of user attributes, and the context of search. This will allow OCIO to develop policy rules that ensure that information is delivered only to authorized users in conformance with law, policy, and best-practice standards.

Element 1: User Attribute

User attributes, which will be stored in a User Attribute Hub, are characteristics of a user performing a query operation on a data element. User attributes will be used by automated policies in the decision to grant or deny access to specific data elements. The Identity Credentialing and Access Management group, a part of OCIO, is developing the User Attribute Hub. The ongoing development of the User Attribute Hub will be influenced by the lessons learned during the pilot projects. User attributes will likely include Department[6], Component, Job Series Code, Clearance, and Training completed.

Element 2: Data Tagging

The Neptune data repository will store DHS data and associated data tag values. Data will be tagged with metadata (i.e., data about data) such as lineage and source. Data will also be tagged for its data type, which will be used by automated policies to determine whether to grant or deny access to specific data elements residing in the CEI Prototype or the Cerberus Pilot. Metadata is the descriptive information characterizing the data, and, for the Neptune Pilot, will include the following:

- The name of the source system;

[6] For initial testing, DHS will only tag data collected and maintained by DHS. As the initiative develops, it is expected that this attribute could be used to denote users from other federal departments and agencies who have been provided appropriate access.

- A source system identifier, which will allow the specific data element to be traced back to the source system;
- The contact information for the component data provider; and
- The date the information was ingested in the Neptune Pilot.

As DHS tests the various aspects of the DHS Data Framework, additional metadata may be included, such as sensitivity of the data (e.g., law enforcement sensitive), applicable retention periods so that the data is automatically removed when no longer retained in the source system, United States person status, and status of individuals who are members of a protected class of data (e.g., information subject to special restrictions on intra-agency and external disclosure pursuant to 8 U.S.C. § 1367).

Data tags will identify the type of data involved, where the data originated, when it was ingested as authoritative mission data, and at least during the pilot phase, whether the data elements are designated as core, extended, or encounter.

- *Core biographic* data is basic biographic information, to include name, date of birth, gender, country of citizenship, and country of birth.
- *Extended biographic* data is additional biographic information about an individual that is not considered core biographic information, such as address, phone number, email address, passport number, and/or visa number. Extended biographic data pertains to the subject of the encounter rather than associated third parties. Extended biographic data does not include DHS derogatory data or detailed information about DHS encounter(s) or transactions with an individual or associated third parties.
- *Detailed encounter* data is information that derives from a DHS screening, vetting, law enforcement, or immigration-related event/process and is collected in accordance with DHS authorities and regulations. The term "encounter" is used to describe a face-to-face meeting, an electronic or paper-based transaction (such as an application for a DHS-administered benefit), or information provided to the United States by a domestic or foreign government agency, aircraft operator, or other private entity. Detailed encounter data may contain DHS derogatory information, screening/vetting results, or information pertaining to third parties who help administer government programs, such as points of contact for exchange visitor sponsors or schools participating in the Student and Exchange Visitor Program.

Data inconsistencies and errors discovered while conducting the initial processing will be rejected and placed in a holding area within Neptune for incompatible records to be manually reviewed[7] by OCIO staff. The process of Neptune initially ingesting the data from the source

[7] The review of data inconsistencies and errors is part of the technical testing to determine how well the system is loading data.

systems and the subsequent ingest by the CEI Prototype and the Cerberus Pilot will also generate auditable information as to the number of records processed and rejected. The CVTF, including representatives from both the oversight offices and operational components will review the audit information as part of the initial testing.

Element 3: Context ("Authorized Purpose/Use and Function")

Context combines the concept of function, which describes what type of search or analysis the user is performing, with purpose, which describes why the user is engaging in a particular activity. Function will drive what types of search and/or analytical tools the user can apply to the data in question. Purpose will determine which datasets and what type of data within those datasets are accessible to the user.

DHS has identified three general types of search (Person or Entity-Based, Characteristic-Based, and Pattern-Based Searches) and four types analytical capabilities (Statistical, Geospatial, Link, and Temporal) that would be part of the DHS Data Framework. During the pilot phase the Cerberus will test basic search capabilities and CEI Prototype will test correlation across data sets and basic search capabilities. DHS is developing a governance process to approve different types of analytical tools that will be part of the dynamic access control process. As additional analytical tools are added, new purposes will be described, evaluated, and added as updates to the PIA for Cerberus and Neptune Pilots.

For the CEI Prototype, DHS has created three broad categories of use: National Security, Law Enforcement, and Benefits. For the Cerberus Pilot, these categories will be broken out into the function and purpose for subsequent testing and will be further developed as the programs mature.

Element 4: Dynamic Access Control

Based on the policies developed through the CVTF discussed above, Dynamic Access Control will be developed to determine whether to grant or deny access to each data element. The enforced policy rules will be developed through close coordination between the operational components and oversight offices within the CVTF. They will then be audited and reviewed by the same group to ensure that users are accessing information in an appropriate manner. The policy rules will evaluate the data tags, the attributes of the user, and the context of the request to determine who gets access to what data and can use what types of tools. For the initial pilots, the enforcement will be verified through review by the CVTF.

Governance Structure

In order to ensure that the four elements of the DHS Data Framework are working consistently within the confines of law and policy, DHS will establish a governance structure that includes the appropriate policy and oversight offices, the DHS components providing data, the users of the relevant system(s), and the components supporting the operation and

maintenance of the systems within the framework, or use an existing governance structure that meets these requirements. This governance structure will operate subject to the guidance and oversight of both the CVTF and the Department's Information Sharing and Safeguarding Governance Board (ISSGB).

Pilot and Prototype Efforts

The gradual introduction of the DHS Data Framework, beginning with pilot efforts, will allow the Department to optimize the policy-control process, and will help minimize the impact on established privacy protections. This fine-tuning will be ongoing and continual if the program reaches an operational state.

A team of OCIO and I&A information technology support personnel have met with each source system's data owner to do the following:

1) Review the data and how it is used;
2) Document the existing access controls for each system;
3) Categorize the existing user population; and
4) Discuss the policy and decision points that are used to grant/deny access to each system.

The team has consulted and coordinated with the operational component and the DHS policy and oversight offices, including PRIV, CRCL, PLCY, and OGC, to ensure the rules appropriately safeguard the data in accordance with requirements.

This PIA covers the overall framework for this program, which includes both policy-based decisions and technical implementation for how DHS will maintain and control its operational data through the pilots and going forward. The CEI Prototype, Neptune Pilot, and Cerberus Pilot are discussed in separate PIAs. The CEI Prototype and the Cerberus Pilot will act as use cases for demonstrating control over access to data. Through the governance bodies for these projects, DHS has already matured the access control models and will continue to mature, review, and then approve them for use. Testing will be performed to verify the accuracy of the policy rules on an individual basis and within mission use case contexts when several rules may be applicable and enforced at the same time. The DHS Data Framework is a DHS-wide program administered by the CVTF with the support of OCIO and I&A on behalf of the Department.

Fair Information Practice Principles (FIPPs)

The Privacy Act of 1974 articulates concepts of how the federal government should treat individuals and their information and imposes duties upon federal departments and agencies regarding the collection, use, dissemination, and maintenance of PII. Section 222(a)(2) of the Homeland Security Act of 2002 (codified at 6 U.S.C. § 142(a)(2)) states that the Chief Privacy Officer shall assure that information is handled in full compliance with the fair information

practices as set out in the Privacy Act of 1974. In response to this obligation, the DHS Privacy Office has developed a set of Fair Information Practice Principles (FIPPs) from the underlying concepts of the Privacy Act to encompass the full breadth and diversity of the information and interactions of DHS. The FIPPs account for the nature and purpose of the information being collected in relation to DHS's mission to preserve, protect, and secure.

DHS conducts PIAs on both programs and information technology systems pursuant to Section 208 of the E-Government Act of 2002 and Section 222 of the Homeland Security Act of 2002. Given that the DHS Data Framework is a framework and not a particular information technology system, this PIA is conducted as it relates to the DHS construct of the FIPPs. This PIA examines the privacy impact of DHS Data Framework operation as it relates to the FIPPs.

1. Principle of Transparency

Principle: DHS should be transparent and provide notice to the individual regarding its collection, use, dissemination, and maintenance of PII. Technologies or systems using PII must be described in a SORN and PIA, as appropriate. There should be no system the existence of which is a secret.

DHS has undertaken this PIA as well as PIAs on the specific technologies being deployed in order to provide transparency to the public about how DHS handles the data it collects. As part of the pilot process, DHS will determine whether additional notice at the point of collection is needed given the contemplated uses of the system. DHS will also determine whether the PIAs and System of Records Notices (SORN) for the source systems need to be updated to provide additional transparency.

2. Principle of Individual Participation

Principle: DHS should involve the individual in the process of using PII. DHS should, to the extent practical, seek individual consent for the collection, use, dissemination, and maintenance of PII and should provide mechanisms for appropriate access, correction, and redress regarding DHS's use of PII.

The DHS Data Framework does not change the initial collection of the information that takes place either directly from the individual or at the request of the individual. All data ingested by Neptune has been obtained for purposes that are consistent with DHS authorities. The privacy of individuals about whom DHS collects, maintains, and uses PII depend on the quality of the information—its completeness, accuracy, timeliness, and relevance. The information contributed initially from the three source datasets is information that was collected directly from the subjects of the information or at the request of the individual by a third party, a factor that enhances accuracy.

The opportunities available for individuals to decline to provide information are the

same as for the component source systems, as follows:

- ESTA information must be provided pursuant to applicable statutes for nonimmigrant aliens seeking to travel to the United States under the Visa Waiver Program (VWP) by air or sea. An individual who declines to provide information necessary to complete an ESTA application will not be permitted to travel to the United States under the VWP and will need to apply for a visa at a United States consulate or embassy.
- The failure to provide requested SEVIS information would preclude a foreign student or exchange visitor from participating in the Student and Exchange Visitor Program.
- Those who decline to provide AFSP information will not be eligible for flight training in the United States.

Although individuals do not have the opportunity to consent to specific uses of the information integrated into the DHS Data Framework, they consent to the collection of information by the source systems by voluntarily completing the required applications.

The DHS Framework will not impact the individual's ability to access and correct his or her information consistent with the published SORN for the source systems. For example, DHS/CBP provides access to and opportunity to seek correction of ESTA data if requested by or on behalf of the subject of the record to the extent the data was provided by that individual (i.e., first party requests). DHS will need to create a process for how to handle requests for data that are maintained both in the source system and Neptune and Cerberus Pilots to ensure that DHS is providing access to the different copies of the data it is retaining for operational use. Should an operational system be developed, such a mechanism would be required.

DHS has published a SORN for the CEI Prototype that allows individuals to request access to their information directly. For records ingested into the Neptune and Cerberus Pilots, the component SORN governing the data ingested in the Neptune data repository will also govern the ingested records in Neptune, as those records will remain under the control of the component data provider. New records generated in the Cerberus Pilot through the application of any search or analytic tools to ingested records by component analytic personnel will be stored in systems of records maintained by those components for the retention, use, and handling of such information and will therefore be subject to the provisions of the respective SORNs for those systems of records.. As part of the pilot, DHS will identify whether components need to update or create new SORNs to handle the information that is being created from the search and analytical tools available through the Cerberus Pilot.

3. Principle of Purpose Specification

Principle: DHS should specifically articulate the authority which permits the collection of PII and specifically articulate the purpose or purposes for which the PII is intended to be used.

The DHS Data Framework creates a structure that allows DHS to assess and decide how and when to give access to data in a consistent and repeatable process to ensure that DHS is using the information consistent with its authorities.

For the pilot stage of the DHS Data Framework, DHS has defined a basic set of authorized purposes for use of the data. These purposes will initially drive what types of searches and data the users can access. For the CEI Prototype, DHS has identified three broad authorized purposes for use of the data: National Security, Law Enforcement, and Benefits. For the Cerberus Pilot, these will be further developed to include both the function of the user and the purpose.

DHS has identified three general types of search tools and four types of analytical capabilities that will be part of DHS Data Framework during the pilot phase.

Types of Search:

- [Specific] Person or Entity-Based Search: A search to retrieve information about an identified individual or entity of interest using specific, detailed biographic information as the only search criteria. A person-based search is the narrowest type of search and therefore the type of search that will generate the highest percentage of responses that are truly responsive to the underlying query (e.g., Joseph A. Smith, DOB 3/29/70).
- Characteristic-Based Search: A search to retrieve information about an identified individual or entity of interest using limited and specific biographic information in combination with general attributes known or reasonably believed to apply to the subject of the search. A characteristic-based search is broader in scope than a person-based search, but is still designed to provide additional information concerning an individual based upon the limited identifying information available to the user at the time of the search (e.g., John or Joseph Smith, male, Citizen of X Country, arriving on an international flight at John F. Kennedy International Airport on xx/xx/2013).
- Pattern-Based Search: A general search to identify one or more individuals or entities of interest using specific criteria based upon credible available information that is reasonably indicative of a threat to homeland security. Unlike person-based or characteristic-based searches, which are designed to identify additional information pertaining to an individual already identified as a threat to homeland security, a pattern-based search is designed to identify previously unknown individuals who pose threats to

homeland security (e.g., Male, 18-35 years old, traveling from X City to Y City, between xx/xx/2013-xx/xx/2013).

Type of Analysis

The DHS Data Framework will use PII and various analytic tools to complete different types of analysis. DHS defines analytic tools as tools that assist in detecting or understanding trends, patterns, and emerging threats, and in identifying or understanding non-obvious relationships using the information maintained in a data repository. The tools will be available for use with data the user is authorized to receive based on the individual's user attributes and context of searches. During the pilot phase of the program these analytic tools will not be available to users. If the pilots are successful, the Cerberus and CEI Prototype PIAs pilot will be updated as these types of tools are introduced.

DHS may develop the following types of analytic tools:

- Statistical Analysis Tools: Modeling and statistical tools that can help analysts discover patterns or generalizations in the data. This analysis can produce models that can be used to identify similar patterns in other data or common characteristics among seemingly disparate data.
- Geospatial Analysis Tools: Tools that can display a set of events or activities on a map showing streets, buildings, geopolitical borders, or terrain. This analysis can help produce intelligence about the location or type of location that is conducive to a particular activity.
- Link Analysis Tools: Tools that can help analysts discover patterns of associations among various individuals and entities in the data repository. This analysis can produce a social network representation of the data.
- Temporal Analysis Tools: Tools that can display events or activities in a timeline to help an analyst identify patterns or associations in the data. This analysis can produce a time sequence of events that can be used to predict future activities or discover other similar types of activities.

There is a risk that elements of the data access and use control are insufficiently developed or incorrectly implemented. This risk is mitigated by the fact that DHS will test the user attributes, tags, and contexts through the Neptune Pilot before considering deployment in an operational system. Oversight offices will be extensively involved in the test development and testing results. If a decision is made to implement an operational system, the access elements will be developed and tested by a specially constituted working group that includes OCIO, I&A, the component data provider, PRIV, CRCL, PLCY, and OGC. Further mitigation is provided by the fact that the tags, and access controls (along with rules governing system operation) would be modified based on the pilot results and thereafter as needed.

There is a risk that the IT developers are not building pursuant to the defined requirements for the system, due to lack of adequately developed requirements or failure of coordination internal to DHS. This risk would create a system that may not meet the requirements or include the controls specified in this PIA. This risk may also lead to information being used for an unspecified purpose. This risk will be mitigated through close coordination between the OCIO and CVTF, and through robust testing of the actual functions and performance of the system against the defined requirements, which includes the specific privacy controls detailed in this and the Neptune, CEI Pilot, and Cerberus PIAs.

There is a risk that the purpose and function that define which users have access to what data are not defined with sufficient detail so that there is no meaningful distinction of data access levels among users. This risk is being mitigated through the strong governance process being developed. The governance process will be tested as part of the pilot process.

4. Principle of Data Minimization

Principle: DHS should only collect PII that is directly relevant and necessary to accomplish the specified purpose(s) and only retain PII for as long as is necessary to fulfill the specified purpose(s). PII should be disposed of in accordance with DHS records disposition schedules as approved by the National Archives and Records Administration (NARA).

Currently, DHS copies entire datasets from one application to the next because of the stovepipe nature of DHS IT systems. This leads to the proliferation of the same data in different analytical systems, which increases the risk of storing inaccurate data, and also increases the complexity and costs of securing and overseeing the proper use of that data. Maintaining multiple copies of entire datasets also increases the risk of DHS granting unlimited access to the data. One of the goals of the DHS Data Framework is to reduce the number of copies of operational data made for analytical purposes and to tag the data so that a user may access only the minimum necessary to perform his or her job. This will also assist in the identification of storage hardware and software application modules that can be retired, thereby reducing the risk of a privacy incident from unnecessary copies of the data.

One risk with the DHS Data Framework approach is that the repositories might retain data for longer than the source systems. DHS is testing and evaluating the Neptune and Cerberus Pilots and the CEI Prototype to ensure the systems have the ability to retain data elements based on the retention guidelines of the source system. DHS will work with the National Archives and Records Administration (NARA) to establish appropriate retention schedules for data retained in the Neptune and Cerberus Pilots and the CEI Prototype. If the Pilots and Prototype are not approved for further development, all records will be deleted at the conclusion of the test and the component data providers will be notified of that deletion. In any case, data from the Pilots and Prototype will not be used for operational purposes.

5. Principle of Use Limitation

Principle: DHS should use PII solely for the purpose(s) specified in the notice. Sharing PII outside the Department should be for a purpose compatible with the purpose for which the PII was collected.

As the DHS Data Framework matures, DHS will address whether and how it gives users outside of DHS access rights to the data maintained in the repositories. Access would only be provided to users external to DHS pursuant to the applicable, published SORN governing the data provided and consistent with applicable information sharing access agreements (ISAA) identifying who can access what data and for what purpose and consistent with applicable laws, regulations, and policies. The DHS governance process will develop roles for external users, as appropriate.

6. Principle of Data Quality and Integrity

Principle: DHS should, to the extent practical, ensure that PII is accurate, relevant, timely, and complete within the context of each use of the PII.

Data Quality and Integrity is addressed at different points in the process. There is the initial process of tagging each data element to ensure the data can be linked back to the source system, thereby protecting the quality of the data. The governance process will also work to ensure that data tags are implemented correctly in the pilot and that the tags effectuate differences in access control as intended. This will be done through testing conducted by OCIO in collaboration with the source data owners. Over the long term, the governance process will need to remain engaged and robust to ensure data quality and integrity is sustained.

In addition to the decision of how and what to tag the data, there is the question of the method and frequency source systems should transfer data to the Neptune and Cerberus Pilot repositories so that the data remains accurate, relevant, timely, and complete. DHS will build on the process for securely moving data from the operational systems, to include handling updates, corrections, and deletions to the data. Testing this in future phases of the project will be integral to moving the system forward. In the interim, during the pilot/testing phase, a system notice will advise users that the data may not be used for operational purposes, which will mitigate any risk of an adverse impact on the individual due to data quality or integrity problems.

7. Principle of Security

Principle: DHS should protect PII (in all forms) through appropriate security safeguards against risks such as loss, unauthorized access or use, destruction, modification, or unintended or inappropriate disclosure.

DHS is following the requirements for information assurance and security for the Neptune and Cerberus Pilots and the CEI Prototype. The Pilot and Prototype systems will

undergo the DHS security review process to ensure DHS standards for security policy, guidance, and architecture requirements are met before any source data will be permitted into those systems. DHS will test and evaluate in physically secure and approved government facilities, using secure and approved government equipment, and will perform periodic evaluations to ensure effectiveness of the security procedures and safeguards implemented. The systems have a waiver from the Chief Information Security Officer stating that sufficient security controls have been put in place to protect the data during the testing phase.

A final user list will be created for the CEI Prototype and Cerberus Pilot that will include roles based on the access control requirements collected from the different data owners (ESTA, SEVIS, and AFSP). Individuals participating in the test and evaluation of the Prototype and Pilot will be allowed to see certain data elements based on the authorized purpose of their search, the type of search, the assigned data tags, and the policies defined for access control.

The DHS Data Framework uses audit logging so that user requests and the results returned to those requests will be logged and include date and timestamps of these transactions. Additionally, only read-only access will be provided to individuals participating in the test and evaluation of the CEI Prototype so that source data cannot be changed in the Prototype. The CEI Prototype and the Cerberus Pilot will be tested by staff from the operational components for testing purposes only. The Neptune Pilot will have no staff from the operational components testing any capabilities because its only purpose is to tag the data. OCIO staff and oversight staff will review the results of the tagging in Neptune. The Neptune data store will be secured against accidental or deliberate unauthorized access, use, alteration, or destruction of information. For more details on the specific security practices used in the Neptune Pilot, such as auditing, accountability, and oversight, please see the separate Neptune Pilot PIA.

8. Principle of Accountability and Auditing

Principle: DHS should be accountable for complying with these principles, providing training to all employees and contractors who use PII, and should audit the actual use of PII to demonstrate compliance with these principles and all applicable privacy protection requirements.

DHS has an organizational commitment to accountability for external legal and privacy policy requirements as well as internal DHS policies and procedures. The DHS Data Framework puts privacy policies into effect through governance, technology, training and education, and compliance oversight and verification. This commitment includes transparency and, whe appropriate, the means for remediation and external enforcement.

For oversight purposes, data logs will provide information on what data is being accessed. For robust, continuous monitoring of data access, the DHS Data Framework will employ tamper-resistant audit logs, which will also provide metrics for assessing the

performance of the program and its compliance with stated practices. System audit logging will capture all successful and unsuccessful attempts to log in, to access information, and other meaningful user and system actions. The audit logs will contain the user ID and the query performed, but not the responses provided back. As part of the pilot, DHS will determine whether this is sufficient information to support an audit of whether PII was accessed properly and to determine the effectiveness of the dynamic access controls that limit the data that is viewed to the users who are permitted to view it.

DHS provides mandatory privacy training to all employees and contractors who have access to or use PII, and all users are required to complete information security training that addresses privacy as well as the proper and secure use of DHS applications. In addition, PRIV offers role-based training for agency employees involved with information sharing. CRCL offers several training products through its Civil Liberties Institute.[8]

Additionally, the Privacy Office will provide the CEI Prototype team members specific privacy training on the CEI Prototype. If CEI is made operational, personnel will be required to take source system (AFSP, ESTA, SEVIS) training, as well as DHS privacy and security training. Before user access is granted or any information is provided, the CEI Prototype will also test the effectiveness of rules intended to enforce the proper training (e.g., privacy, security) requirements based on a user's role and the data that he or she is trying to access.

[8] Accessible at http://www.dhs.gov/civillibertiesinstitute.

Conclusion

DHS has developed the framework specifically to ensure that it is consistently using DHS data for the purposes for which it was collected. There are several privacy risks to the overall DHS Data Framework that are being mitigated by starting with several pilots to see what is successful and what is not. DHS has explicitly stated that the pilots will not transition to operational purposes until DHS has demonstrated the ability to effectuate meaningful dynamic access controls.

Responsible Officials

Tom Bush

Common Vetting Task Force

Donna Roy

Office of Chief Information Officer

Clark Smith

Intelligence and Analysis

Approval Signature Page

Original signed on file with the DHS Privacy Office.

Karen L. Neuman
Chief Privacy Officer
Department of Homeland Security