Privacy Impact Assessment Update
for

Neptune

DHS/ALL/PIA-046-1(a)

August 29, 2014

Contact Point
Paul Reynolds
Data Framework Program Management Office
Department of Homeland Security
(202) 447-3000

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

Abstract

The DHS Data Framework ("Framework") is a scalable information technology program with built-in capabilities to support advanced data architecture and governance processes. The Framework is DHS's "big data" solution to build in privacy protections while enabling more controlled, effective, and efficient use of existing homeland security-related information across the DHS enterprise and with other U.S. Government partners, as appropriate. Currently, the Framework includes the Neptune and Cerberus systems, and the Common Entity Index. Between November 2013 and August 2014, DHS deployed a pilot/prototype to test different capabilities needed to implement the Framework. After the successful completion of the pilot/prototype phase, DHS now intends to mature the Framework by entering into the next phase—limited production capability. DHS is updating the original Framework Privacy Impact Assessments, including this Neptune PIA, to reflect this transition to limited production capability.

Introduction

In a Privacy Impact Assessment (PIA) published on November 6, 2013, the Department of Homeland Security (Department or DHS) previously described the Department's development of the Framework. The *DHS Data Framework Overview* in that PIA is summarized here for ease of reference, followed by a description of the *DHS Data Framework Pilot/Prototype Phase,* and an explanation of the next phase in the maturation of the project.

DHS Data Framework Overview

1. Background

The Department's primary mission 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 the Department's 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 individuals' privacy rights, civil rights and civil liberties are not diminished by efforts, activities, and programs aimed at securing the homeland. To enable the Department to carry out these complementary missions, the Homeland Security Act of 2002 sought to eliminate information firewalls between government agencies by consolidating multiple agencies under DHS.

Since 2007, DHS has operated under the "One DHS" policy,[1] which was implemented to afford DHS personnel timely access to the relevant and necessary homeland security information they need to successfully perform their duties. DHS personnel requesting this information must: (1) have an authorized purpose, mission, and need-to-know before accessing the information in performance of their duties; (2) possess the requisite background or security clearance; and (3) ensure adequate safeguarding and protection of the information. The Department's existing architecture for its IT systems and databases, however, is not conducive to effective implementation of the One DHS policy because information exists in multiple separate databases. The result is a technically cumbersome, time-intensive process to determine what information DHS has about a particular individual.

The Secretary of Homeland Security and the DHS Deputy Secretary directed the development of the Framework to automate execution of the One DHS policy through a collaborative effort among the Department's Common Vetting Task Force (CVTF),[2] the Office of the Chief Information Officer, the Office of Policy, the Office of Intelligence and Analysis, the "oversight offices," including the Privacy Office (PRIV), the Office for Civil Rights and Civil Liberties (CRCL), the Office of the General Counsel (OGC), and DHS's operational components.

2. Objective

The Framework will create a systematic repeatable process for providing controlled access to DHS data across the Department. The 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 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 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.

In order to achieve the Framework's goal, DHS created two central repositories for DHS data: Neptune and Cerberus. Neptune serves as the repository in the unclassified domain. Cerberus resides in the Top Secret/Sensitive Compartmented Information domain. Through these systems, DHS applies appropriate safeguards for access and use of DHS data and delivers search and analytic capabilities.[3]

[1] *DHS Policy for Internal Information Exchange and Sharing*, February 1, 2007.
[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] During limited production capability, the search and analytic capabilities will be limited to the three basic search

The 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 based on the type of data involved, the authoritative system from which the data originated, and when it was ingested into the Framework;
(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 and/or Components.

The Framework uses the dynamic access control policies to enable the automated enforcement of access requirements so that a user sees only the information that he or she would otherwise be entitled to view as a matter of law and policy. The Framework includes these elements and related processes to ensure: (1) accurate data tagging; (2) data integrity as data is copied and transferred from its original location; and (3) enforced access control policies. The Framework also enables the Department to log user activities to aid audit and oversight functions.

Phase I: Framework Pilot/Prototype

Earlier this year, the Department successfully completed testing the initial Framework capabilities through the Neptune Pilot, Cerberus Pilot, and Common Entity Index (CEI) Prototype.[4] The Department used three data sets in the pilot/prototype phase: the U.S. Customs and Border Protection's (CBP) Electronic System for Travel Authorization (ESTA), the U.S. Immigration and Customs Enforcement's (ICE) Student and Exchange Visitor Information System (SEVIS), and the Transportation Security Administration's (TSA) Alien Flight Student Program (AFSP). The data sets were copied from the relevant component IT system, transferred into the Neptune platform and tagged, and then the tagged data elements were pushed to the CEI and Cerberus platforms. The pilot/prototype phase successfully demonstrated important foundational elements of the Framework, including, but not limited to those capabilities described below.

functions deployed in the pilot/prototype phase: person search, characteristic search, and trend search.
[4] The PIAs for these pilots and prototype are published at http://www.dhs.gov/privacy-documents-department-wide-programs.

- **Data Management and Transfer –** The pilot/prototype phase demonstrated that Neptune could ingest the data from the three data sets, apply access control tags and relevant metadata, and transfer the tagged data to the Common Entity Index and Cerberus. The data tags identified the type of data involved, when the data originated, when it was ingested as authoritative mission data, and whether the data elements are designated as core, extended, or encounter.[5]

- **User Authentication and Attributed-Based Access –** The pilot/prototype phase demonstrated users could be authenticated with appropriate certificates and that their attributes were properly set with predetermined functions and purposes. Upon login, a user's attributes were retrieved from an attribute authority.[6] Where a user had more than one function and purpose (i.e., the user needed access to data while acting in different capacities), the user was able to select the appropriate functions and purposes for accessing data. Once a user was positively authenticated, the user would have access to the Cerberus system and could request data.

- **Policy-Based Access Control –** The pilot/prototype phase demonstrated that DHS could apply policy-based access controls to determine the type of basic search tools[7] the user could use and what data the user could access. Given a particular user's attributes, an Access Control Server asked what function and purpose the user performs to then determine what privileges the user had. The user's function controlled the basic search tools (i.e., the type of query that could be performed) that the user could use. The user's purpose determined which data sets and which type of data (i.e., core, extended, or encounter) the user could access. The Department tested a variety of purpose and function combinations to test whether the Access Control Server gave the user access to the correct tools and data. In each instance, the

[5] 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. 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. For more information on these concepts, *see* DHS/ALL/PIA-046-1(a) Neptune PIA Update and DHS/ALL/PIA-046-3(a) Cerberus PIA Update, published concurrently with this PIA Update.

[6] During the pilot/prototype phase, attributes were self-asserted based on pre-defined choices. In a related effort, the Department is developing an authoritative user attribute hub that will include DHS and Intelligence Community user attributes. Once developed, the Framework will employ this authoritative attribute hub, called the Trusted Identity Exchange. The Trusted Identity Exchange will eliminate the need for users to self-assert their attributes. For the pilot/prototype phase, the Department used a Lightweight Directory Access Protocol server as a stand-in for the DHS Trusted Identity Exchange.

[7] Query tools included (1) [Specific] Person or Entity-Based Search; (2) Characteristic-Based Search; and (3) Pattern-Based Search. These queries are described in greater detail in the Fair Information Practice Principles analysis later in the document.

demonstrations showed that the policy-based controls were appropriately applied and that users only had access to the search tools, data sets, and types of data that they were permitted to access under DHS policy.

- **Audit Logging –** The pilot/prototype phase also demonstrated that DHS could log the application of policy-based controls as it was occurring. The policy decision log showed the policy enforcement when a user requested access and evaluating the policy rules to determine the user's privileges to data or tools. The audit log reader also captured the queries a user made and statistics regarding query results, aiding in audit and oversight, including verification of compliant data usage.

Lessons Learned

The Department intends to mature the Framework in an incremental manner and learned a number of important lessons as a result of the Framework's pilot/prototype phase, including those listed below:

1. Developing a scalable big data architecture means DHS needs to **establish a governance process** to evaluate the integration of new data, new missions, new users, and new analytical tools.

2. **Incremental development** of the Framework allows the Department to deploy new capabilities and then verify that those capabilities comply with legal and policy requirements. This approach allows DHS to ensure that it delivers new capabilities that support DHS's operational mission while protecting privacy, civil rights, and civil liberties. By incorporating these protections from the beginning, the Department is building a sustainable and scalable Department-wide big data architecture.

3. Establishing long-term operational utility and protecting privacy, civil rights, and civil liberties depends on DHS's ability to **refresh and update data** and to **incorporate appropriate redress** mechanisms into the Framework.

4. Conducting **more stakeholder engagement**—with mission operators, system administrators, and data stewards—will facilitate widespread adoption this Department-wide big data solution.

5. **Promoting transparency** will help the public understand how DHS is using its data and support a robust public dialogue on the appropriate use of big data solutions within the U.S. Government. Throughout the pilot/prototype phase, DHS published multiple privacy impact assessments, gave public briefings at the DHS Data Privacy

and Integrity Advisory Committee (DPIAC) meetings, and asked the DPIAC for recommendations to further promote transparency.

The DHS Data Framework Privacy Impact Assessment Update (published concurrently with the Neptune PIA update) describes the specific measures DHS is using to apply the lessons learned from the Framework pilot/prototype phase.

Reason for the PIA Update

Phase II: Limited Production Capability

Based on the Framework's success to date, the Department is moving from the pilot/prototype phase to a limited production capability for both the Neptune and Cerberus systems. During limited production capability, DHS will test the ability to refresh data from the original DHS IT system to the Framework. DHS has publicly declared[8] its intention to develop these capabilities prior to the operational use of data in the Framework, and the limited production capability provides the next step in implementing refresh.

The limited production capability shares many of the pilot/prototype phase conditions except that limited production capability provides for limited evaluation in the operational environment. For example, the data elements the source systems transferred to Neptune for ingestion remain the same as in the Neptune Pilot. The data tags remain the same except that Neptune will be adding metadata tags to support future access rules related to the sensitivity of or ability to release the data (i.e., data about persons who receive additional protections, such as U.S. Person minimization pursuant to Executive Order 12333 or the non-disclosure provisions of 8 U.S.C. § 1367 for certain special protected classes of aliens. In addition, limited production capability will introduce data quality processing that will validate ingest tagging, proper matching of source fields to the framework search and storage fields, and will generate data quality metrics for performance and compliance reporting.

As with the Neptune Pilot, the tagged data will be stored on the DHS-owned Neptune data storage system and accessed by a limited number of DHS staff in order to confirm that the Neptune tagging is successfully working. Incompatible data that cannot be processed during ingest (e.g., with improper record formats) will be placed in a holding area to be manually reviewed. The ingest process will also generate auditable information as to the number of

[8] *See* public briefing on the Framework presented during the DPIAC meeting on September 12, 2013 and January 30, 2014. Available on the DHS Privacy website at: http://www.dhs.gov/dhs-data-privacy-and-integrity-advisory-committee-meeting-information.

records processed and rejected. The overall approach to tagging will also allow the Department to consistently categorize information across the data sets.

The data tags necessary to support privacy, civil rights and civil liberties, and ultimately policies for access, use, and sharing of information remain subject to additional development through the direction of an executive steering committee and implementation by the Data Framework Program Management Office (PMO).

Neptune will ingest and tag data elements from three unclassified databases collected by DHS components.[9] These data elements will be extracted from the source systems and refreshed on a regular interval. One of the main goals of the limited production capability is to identify the timelines for refreshing each data set, test DHS's ability to refresh each data set, and begin implementation of limited data set refreshes. These refresh timelines will be based on operational need, available resources, and technical capabilities. Limited production capability will start with an initial bulk data ingest of the source systems (i.e., a 'snapshot' in time) into Neptune, followed by increased data updates as the limited production capability progresses and DHS tests its ability to refresh each data set. The goal is to have regular refreshes of data by the end of the calendar year, according to the refresh timelines established for each data set.

During the ingest process, the data will be tagged with metadata to enable further testing of access controls and information management policies based on data owner and privacy, civil rights and civil liberties, and safeguarding considerations. In addition, an audit log of loaded data will be created.

Privacy Impact Analysis

Authorities and Other Requirements

Neptune will continue to be a copy of data maintained in a source system of records and not result in the creation of any new system of records. The data used in Neptune continues to be covered by the source system System of Records Notices (SORN). Neptune will continue to tag the individual records to permit the component data providers to maintain control over the contents of the records.

[9] The Department will use the same three data sets from the pilot/prototype phase: the U.S. Customs and Border Protection's Electronic System for Travel Authorization (ESTA), the U.S. Immigration and Customs Enforcement's Student and Exchange Visitor Information System (SEVIS), and the Transportation Security Administration's Alien Flight Student Program (AFSP).

A Neptune System Security Plan has been completed and the system Authority to Operate (ATO) will be received by September 2014. The current Federal Information Processing Standards (FIPS) is High-High-Moderate.

Neptune continues to rely on the source systems to manage retention on their data and to include retention-based deletions and/or status changes in data deliveries to Neptune. During the ingest process, the data will be tagged with metadata that indicates record creation data which match data retention rules for the source system data retention requirements and allow for retention rule compliance to be verified.

There is no change to the Paperwork Reduction Act requirements, which remain non-applicable to this effort.

Characterization of the Information

Neptune ingests, tags, retains, and transfers information originally collected by the source systems noted above. The data elements are transferred to Neptune based on agreements with the source system owners/data stewards. If additional data sets are added to Neptune for tagging, similar agreements will be established consistent with the governance structure outlined in the Framework PIA Update.

Neptune does not alter the delivered source data, but instead maintains a complete copy of the source information in Neptune. Neptune associates metadata during the ingestion of the source data to support standardization of fields for storage and query operations and retention compliance. At a minimum, the following Framework metatags will be associated with the data:

- The name of the source system;
- The date the information was replicated in Neptune;
- The applicable retention rule with any original system creation date;
- A source system identifier, which will allow the specific data element to be traced back to the source system; and
- The contact information for the component data provider.

To support existing Framework access control rules, the data transferred to Neptune is grouped into core biographic data, extended biographic data, or detailed encounter data. During the LPC phase, Neptune will be adding additional metadata to support future Framework access rules related to the sensitivity of or ability to release data (i.e., data about persons who receive additional protections, such as U.S. Person minimization pursuant to Executive Order 12333 or the non-disclosure provisions of 8 U.S.C. § 1367 for certain special protected classes of aliens).

Neptune currently ingests, tags, retains, and transfers information originally collected by the previously identified component database systems. The specific information collected in the source systems of record is set forth in each program's respective SORN and PIA and its safeguarding controls are described in the Framework PIA. The data ingested into Neptune does not include information from public or commercial sources.

Neptune relies on the accuracy of source systems. The information contributed initially from the three underlying data sets is information that was collected directly from the subjects of the information or at the request of the individual by the school or sponsor, a factor that enhances accuracy. Delivered data will be logged, stored, and available to verify proper ingest. During that ingest, data accuracy and consistency are evaluated to determine whether it conforms to defined formats and schemas (e.g., "date" format of "MM/DD/YYYY") and with unprocessable records (e.g., invalidly formatted or physically corrupt) rejected. Data inconsistencies/errors discovered while mapping the source data elements to Neptune data schema are also rejected and placed in a holding area for incompatible records to be manually reviewed. Manual review is conducted by the development team to determine whether the data was rejected due to problems in the ingest process or because of an actual problem with the source data. During the LPC period, data quality checks will be developed to perform verification of source fields mapped to the common storage/query fields to ensure data retrieval and association accuracy.

Risk: There is a risk that PII transferred outside of the original IT system and into Neptune will not be accurate, relevant, timely, or complete.

Mitigation: To mitigate this risk in the long-term, DHS must create a process to refresh the data provided from the original DHS IT system to the Framework, so that updates or corrections are replicated from the original DHS IT system into the Framework. One of the main goals of the limited production capability is to identify the timelines for refreshing each data set, test DHS's ability to refresh each data set, and begin implementation of limited data set refreshes. More information on data refresh is provided in the "Principle of Individual Participation" and throughout the PIA.

To mitigate this risk during the limited production capability, Framework users will be trained to understand the risk associated with data latency (due to limited refresh capabilities). Users will also be required to verify information at the source system before completing any final analysis or using the information operationally.

Risk: There is a risk that inaccuracies may result from any incorrect mapping of the source data to the "common schema" or common information fields used across the data sets, (e.g., a last name is mapped to a first name).

Mitigation: This risk is mitigated through manual ingest validation and auditing of ingest processing. Data is only ingested if it meets the defined formats and valid values of the common schema. The Data Framework PMO will conduct a manual review of any rejected data to determine whether the data was rejected due to problems in the ingest process or because of an actual problem with the source data. Ingest and data quality processes generate auditable information that can be reviewed by the Data Framework PMO and/or the source data stewards to ensure proper mapping and to determine where improvements are needed. The ingestion and data quality processes generate auditable information as to the number of records processed and rejected. This information can then be passed back to the data owner of the source system for review and potential improvements to data quality at the source. The Data Framework PMO will establish and monitor a Framework Data Quality Mailbox to allow mission operators to flag data quality issues for redress initiated through the source systems of record. Data quality post-processing analytics will also perform verification of information mapped to the common schema to further ensure data retrieval and association accuracy.

Risk: There is a risk of errors occurring in the transfer of data from the source system.

Mitigation: This risk is mitigated by formalizing access and use authorization and technical exchange procedures between the source system owners and the Data Framework PMO developed during the pilot phase. All delivered extractions of the data will be logged, stored, and available to verify consistency and accuracy upon ingest. To secure the data and ensure its quality, each source system data delivery is encrypted while in transport and archived as negotiated with the source system owners. Incompatible data that cannot be processed according to the mapping of the source data elements to the standard data storage and query fields will be rejected and placed in a holding area to be manually reviewed.

Uses of the Information

Neptune ingests information from source systems of records into an integrated "big data" cloud tagged with metadata to support Policy Based Access Controlled data retrieval and analysis. Neptune does not have end users and does not support direct electronic searches, queries, or analysis nor will it be used to identify predictive patterns or anomalies. During the LPC period, Neptune will deliver the tagged information to Cerberus for limited operations that leverage applied tags and metadata to provide users access to the data commensurate with their roles and with rules that protect privacy, civil rights, and civil liberties, and enforce other legal protections. Use of the information in those Framework systems are described in their PIAs and controls on use such as training are described in the Framework PIA. Neptune will not have end-users and only delivers the tagged information to the Cerberus and CEI Framework systems.

Only DHS Office of the Chief Information Officer (OCIO) assigned Neptune system administrators will have controlled but direct access to Neptune to initiate data ingest/tagging,

export/delivery, and configuration of the Neptune system. There is an assigned DHS Headquarters Information System Security Manager that manages system administration user access to the deployed system. Neither source system data owners nor Oversight offices (including PRIV, CRCL, and OGC) have access to Neptune data through the system but will review the results of the ingest and tagging audits.

Risk: There is a risk that DHS will include data in Neptune for a purpose other than the purpose for which is was collected in the original DHS IT system.

Mitigation: During limited production capability, DHS users will only use the data for immigration, border security, and counterterrorism purposes. There are no non-DHS users of the Framework during the limited production capability. The ESTA, SEVIS, and AFSP System of Records Notices specify that DHS collected the information for these purposes. For example, the ESTA System of Records[10] states that "The purpose of this system is to collect and maintain a record of nonimmigrant aliens who want to travel to the United States under the [Visa Waiver Program (VWP)], and to determine whether applicants are eligible to travel to the United States under the VWP by vetting their information against various security and law enforcement databases and identifying high-risk applicants." The SEVIS System of Records Notice[11] notes that SEVIS allows DHS "to monitor the progress and status of lawfully admitted F/M/J nonimmigrants residing in the United States, to ensure they comply with the obligations of their U.S. admittance…" and that the information may be used "to support other homeland security and immigration activities…." The AFSP System of Records Notice[12] states that the purpose of the system includes the "[p]erformance of security threat assessments, employment investigations, and evaluations performed for security purposes that Federal statutes…" and "the retrieval of information from other terrorist-related, law enforcement, immigration and intelligence databases on the individuals covered by this system.

Risk: There is a risk that Framework users will access more PII than is necessary to accomplish their specified purpose.

Mitigation: One of the hallmarks of the Framework is the ability to restrict access to PII within a particular data set based on the user's specified purpose. To accomplish this, DHS has tagged elements from each data set as belonging to one of three categories—core biographic,

[10] *See* DHS/CBP-009 – Electronic System for Travel Authorization (ESTA), July 30, 2012, 77 FR 44642. Available at: http://www.gpo.gov/fdsys/pkg/FR-2012-07-30/html/2012-18552.htm.
[11] *See* DHS/ICE-001 – Student and Exchange Visitor Information System, January 5, 2010, 75 FR 412. Available at: http://www.gpo.gov/fdsys/pkg/FR-2010-01-05/html/E9-31268.htm.
[12] *See* DHS/TSA-002 – Transportation Security Threat Assessment System, May 19, 2010, 70 FR 33383. Available at: http://www.gpo.gov/fdsys/pkg/FR-2010-05-19/html/2010-11919.htm.

extended biographic, and encounter information—and users are only able to access the categories that are necessary to perform their function. This use of data tags allows DHS to minimize data access according to specified purpose, which is an improvement in the implementation of data minimization within the Department.

Risk: There is a risk that DHS will include more data sets in the Framework than those which are necessary to fulfill the purposes authorized under the Framework.

Mitigation: To minimize this risk, DHS has carefully evaluated each data set to determine whether its use is directly relevant and necessary to accomplish the purposes authorized under the Framework. The pilot/prototype phase demonstrated that these three data sets were effectively used together to support DHS's immigration, border security, and counterterrorism missions. During the limited production capability, DHS will not be including any new data sets in the Framework.

Risk: There is a risk that the Framework will encourage DHS to replicate data sets across the Department, proliferating data across the Department.

Mitigation: An important goal of the Framework is to reduce the number of copies of data sets across the Department. By creating a Department-wide big data solution, DHS will actually reduce the number of copies of data sets across the Department in the long-term. Eventually, some data aggregation systems may be decommissioned as their capabilities are replicated and centralized within the Framework. To implement this mitigation, however, DHS must successfully replicate the capabilities of other systems and build operator support. The limited production capability is the next step in an iterative process toward these goals.

Risk: There is a risk that the elements of data access and control are insufficiently developed or incorrectly implemented and will fail to limit the use of the data to the purposes authorized for the limited production capability.

Mitigation: The pilot/prototype phase tested the user attributes, tags, and context to verify that the controls performed correctly. During limited production capability, DHS will continue to evaluate the application of these controls. Additionally, DHS provided demonstrations of these controls to subcommittees of the DHS DPIAC and requested recommendations from the DPIAC on what auditing and oversight capabilities DHS could develop to ensure that these controls are not circumvented.

Risk: There is a risk that data tags are insufficiently developed or incorrectly implemented to support mission uses and/or to sufficiently ensure data safeguarding.

Mitigation: The Framework pilots and prototype validated/proved the concept of tags and metadata supporting policy based access control rules. However, the process of evaluating and refining the mission use cases and their access requirements is ongoing. New tags and metadata controls are being developed based on input from data providers. The Data Framework PMO is working in conjunction with a specially constituted working group that includes the DHS OCIO, the component data providers, PRIV, CRCL, and OGC to support additional compliance polices.

Notice

Individuals do not have the opportunity to directly consent to the use of their data in Neptune. DHS provides public notice of the existence of Neptune, the data collected and maintained, and the routine uses associated with the information collected through its PIAs. In addition, DHS provides transparency through its SORNs and PIAs which are available on the DHS Privacy Office website, http://www.dhs.gov/privacy, and civil rights and civil liberties policies and procedures available on the DHS Office for Civil Rights and Civil Liberties website, http://www.dhs.gov/civilliberties. Ingested source systems are covered by their own PIAs and SORNs as described above, which are available on the DHS Privacy Office website, and each program provides a Privacy Act statement that provides specific notice about the collection and use of the relevant information to the individual.

Risk: There is a risk that individuals may not be aware their PII is being compared against other DHS information.

Mitigation: DHS has determined that the existing System of Records Notices for ESTA, SEVIS, and AFSP provide notice that the information may be compared against other data sets and be subject to analysis for DHS's counterterrorism and immigration missions. The ESTA System of Records[13] notes that DHS's purpose of collecting the information includes "…vetting [individuals'] information against various security and law enforcement databases and identifying high-risk applicants." The SEVIS System of Records Notice[14] notes one of its purposes is to support "…the analysis of information in the system for law enforcement, reporting, management, and other mission-related purposes." The AFSP System of Records

[13] *See* DHS/CBP-009 – Electronic System for Travel Authorization (ESTA), July 30, 2012, 77 FR 44642. Available at: http://www.gpo.gov/fdsys/pkg/FR-2012-07-30/html/2012-18552.htm.
[14] *See* DHS/ICE-001 – Student and Exchange Visitor Information System, January 5, 2010, 75 FR 412. Available at: http://www.gpo.gov/fdsys/pkg/FR-2010-01-05/html/E9-31268.htm.

Notice[15] lists one of its purposes as "To permit the retrieval of the results of security threat assessments, employment investigations, and evaluations performed for security purposes; including criminal history records checks and searches in other governmental, commercial, and private data systems, performed on the individuals covered by this system." Additionally, DHS is updating this PIA and the PIAs for the DHS Data Framework and Cerberus to reflect deployment of the limited production capability.

Risk: There is a risk that individuals may not be aware that their PII is being used in the DHS-wide big data project, or that they may not understand the implications of the use of their PII in the big data context.

Mitigation: While the existing privacy documentation may permit the use of individuals' PII in this context, DHS is pursuing ways to provide transparency outside of the traditional privacy documentation process because of the privacy sensitivities surrounding big data technology and use. DHS promoted the Framework as part of the White House Big Data Review,[16] and the Framework is described in the White House's final big data report.[17] DHS has provided two public briefings on the Framework during meetings of its Federal Advisory Committee, the DHS DPIAC.[18] DHS plans to continue its public briefings at DPIAC meetings as the Framework progresses. Finally, DHS has tasked the DPIAC with developing recommendations regarding how DHS can further provide transparency into the Framework.

Data Retention by the project

Risk: There is a risk that data will be retained in the Framework for longer than is allowed in the original DHS IT system.

Mitigation: DHS has determined that the retention period for the original DHS IT system will also apply when that information is ingested into the Framework. Neptune relies on the source system to manage retention on its data and to include retention-based deletions in data deliveries to Neptune. During the limited production capability, Neptune will receive complete refreshes of data from the source system at least monthly to reflect correct retention of data.

Risk: There is a risk that original DHS IT system will not inform Neptune that

[15] *See* DHS/TSA-002 – Transportation Security Threat Assessment System, May 19, 2010, 70 FR 33383. Available at: http://www.gpo.gov/fdsys/pkg/FR-2010-05-19/html/2010-11919.htm.
[16] *See* the White House 90-Day Review for Big Data website for more information. Available at: http://www.whitehouse.gov/issues/technology/big-data-review.
[17] *See* the White House report "Big Data: Seizing Opportunities, Preserving Values," May 2014. Available at: http://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_5.1.14_final_print.pdf.
[18] *See* the DHS Privacy website for archived meeting materials. Available at: http://www.dhs.gov/dhs-data-privacy-and-integrity-advisory-committee-meeting-information.

information needs to be deleted to comply with retention rules.

Mitigation: During the limited production capability, this risk is mitigated by the agreements with the source system owners/data stewards that stipulate that the original DHS IT system must provide complete refreshes of the data on a monthly basis. In the next phases, data refreshes will be automated and will handle any required data deletions to reflect source system changes, including retention-based deletions. Even with this timelier syncing of data from the source systems, users will be trained to verify information accuracy at the source system of record when any use results in an action or decision based on that data.

Information Sharing

Neptune is being deployed to support CEI and Cerberus internally for DHS. There will be no information sharing outside of DHS via Neptune during limited production capability.

Risk: There is a risk that DHS will share PII outside of the Department for a purpose that is not compatible with the purpose for which the PII was collected.

Mitigation: DHS is not sharing information outside of the Department during the limited production capability.

Redress

The source system procedures for individuals to address possibly inaccurate or erroneous information are described in the respective Systems of Records Notices for the source systems. Neptune relies on the accuracy of the underlying component systems that supply the information. To the extent the current source systems collect information directly from the individual involved, the opportunity is provided for the individual to ensure the accuracy of the data submitted. An additional opportunity exists for individuals to request access to and/or correction of their record(s) in the underlying component systems, as permitted by law, DHS policy, and described in the applicable Systems of Records Notices.

The Data Framework PMO will establish procedures to communicate redress requests from the source system to other Framework component systems and ultimately to any mission users that may have consumed/leveraged the information in question. To support this, Neptune will make available a report capability that performs a person-specific retrieval of data with the use of personal identifiers. The execution of this report will be audited and limited to the same administrators that perform the data delivery export functionality.

Risk: There is a risk that an individual will not be able to receive appropriate access, correction, and redress regarding DHS's use of PII or that changes made to PII in the underlying DHS IT system as a result of correction and redress will not be replicated into the Framework.

Mitigation: To mitigate this risk in the long-term, DHS will develop (1) a process to provide an individual with the same access and redress opportunities in the Framework that he or she would have in the original DHS IT system[19] and (2) the ability to refresh the data that is ingested into the Framework.

With respect to access and redress, the Data Framework PMO will employ the formal Framework governance structure to create this permanent process moving forward. The absence of an access and redress process that extends from the original DHS IT system to the Framework is one of the reasons that DHS chose to deploy a limited production capability instead of pursuing full operational use of the Framework.

With respect to data correction, DHS must create a process to refresh the data provided from the original DHS IT system to the Framework. One of the main goals of the limited production capability is to identify the timelines for refreshing each data set, test DHS's ability to refresh each data set, and begin implementation of limited data set refreshes. These refresh timelines will be based on operational need, available resources, and technical capabilities. Limited production capability will start with an initial bulk data ingest of the source systems (i.e., a 'snapshot' in time) into Neptune, followed by increased data updates as the limited production capability progresses and DHS tests its ability to refresh each data set. The goal is to have regular refreshes of data by the end of the calendar year, according to the refresh timelines established for each data set.

To help mitigate this risk during the limited production capability, DHS requires users to go back to the original DHS IT system and verify that an individual's information has not been updated pursuant to redress or correction before completing any final analysis or using the information operationally. Requiring users to verify information in the original DHS IT system will ensure that any updates pursuant to redress or correction will be incorporated into any final analytical product or before the information is used operationally.

Auditing and Accountability

The Neptune data store is secured against accidental or deliberate unauthorized access, use, alteration, or destruction of information. Neptune ensures that the information is used in

[19] The Framework does not impact an individual's ability opportunity to receive appropriate access, correction, and redress in the original IT system.

accordance with the practices stated in this PIA through specific auditing, accountability, and oversight measures.

DHS provides mandatory privacy training to all employees and contractors who have access to or use PII, and all users are required to complete mandated information security training that addresses privacy as well as the proper and secure use of DHS applications. In addition, the DHS Privacy Office offers role-based training for agency employees involved with information sharing. The Office for Civil Rights and Civil Liberties offers several training products through its Civil Liberties Institute.[20] Privacy training for Cerberus users will be detailed in the Cerberus PIA.

The Data Framework PMO will manage the review of information sharing agreements, additional uses, policy rule development and the potential addition of new source system data, in conjunction with DHS and component compliance, governance, and mission stakeholders, and data stewards.

Neptune is only accessible by approved system administrators. Administrative access to Neptune will be managed by the system DHS HQ Information System
Security Manager. Neptune has established reporting to support data quality and redress activities without providing system access to external organizations.

Risk: There is a risk that the use of PII will not be auditable to demonstrate compliance with these principles and all applicable privacy protection requirements.

Mitigation: As part of the pilot/prototype phase, DHS determined that the Framework's audit capabilities were adequate to support an audit of whether personally identifiable information was accessed properly and that the dynamic access controls could sufficiently limit the data that is viewed to the users who are permitted to view it. During limited production capability, the Framework will continue to employ tamper-resistant audit logs, which will also provide metrics for assessing the capture of 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 name and the query performed, but not the responses provided back.

Additionally, DHS provided demonstrations of the audit log capabilities to subcommittees of the DHS DPIAC and requested recommendations from the DPIAC on what auditing and oversight capabilities DHS could develop to ensure that these controls are not circumvented.

[20] *See* http://www.dhs.gov/civillibertiesinstitute.

Risk: There is a risk that DHS will not perform reviews of the audit logs to determine compliance with the Framework policies.

Mitigation: During the limited production capability, the PMO will pull a random selection of queries from Framework systems and manually review them to determine compliance with the Framework policies. The PMO will present its finding to an executive steering committee, which includes PRIV, CRCL and OGC.

Additionally, to mitigate this risk in the long-term, DHS has tasked the DPIAC with developing recommendations for how DHS can use audit logs in a meaningful way to ensure robust oversight.

Responsible Official

Donna Roy
Executive Director, Information Sharing Environment Office
Office of the Chief Information Officer

Approval Signature

Original signed copy on file with DHS Privacy Office.

_____

Karen L. Neuman
Chief Privacy Officer
Department of Homeland Security