Privacy Impact Assessment Update
for

Cerberus

DHS/ALL/PIA-046-3(a)

August 29, 2014

<u>Contact Point</u>
David Hong
Program Manager
Department of Homeland Security
(202) 282-9632

<u>Reviewing Official</u>
Karen 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 Assessment 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 *Framework Overview* in that PIA is summarized here for ease of reference, followed by a description of the *Framework Pilot/Prototype Phase*, and an explanation of the next phase in the maturation of the project.

Framework Overview

1. <u>Background</u>

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, the Office for Civil Rights and Civil Liberties, the Office of the General Counsel, 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 functions deployed in the pilot/prototype phase: person search, characteristic search, and trend 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.

<u>Phase I: Framework Pilot/Prototype</u>

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.

- **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

[4] The PIAs for these pilots and prototype are published at http://www.dhs.gov/privacy-documents-department-wide-programs.

relevant metadata, and transfer the tagged data to the Common Entity Index and Cerberus. The data tags identified the type of data involved, where 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] When 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 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.

[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, 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.

- **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 PIA Update (published concurrently with this Cerberus PIA Update) describes in detail 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, common schema mapping, and will generate data quality metrics for performance and compliance reporting.

DHS now intends to use Cerberus for analytical purposes during a limited production capability phase in a limited operational environment with no external sharing. During the limited production capability phase, a controlled set of users will perform classified and unclassified searches against the Cerberus data cloud using the same search tool set described

[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.

during the pilot phase (Specific Person or Entity-Based Search, Characteristic-Based Search and Pattern-Based Search). Cerberus users will perform these queries in support of their border security and counterterrorism missions. Cerberus will use information about the user and the mission context of the requested use to perform dynamic access control to ensure that only authorized users with certain attributes are able to access the data for approved purposes.

On a regular basis, Cerberus will receive data refreshes from the source systems via Neptune. 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.

Because privacy is integral to maintaining public trust in DHS programs and initiatives, the Department designed Cerberus to embed privacy protections at the data level and uses the Framework to dynamically enforce access control. During the Cerberus Pilot, these access control protections demonstrated that only those who should have access to the data do have access, and only for their authorized purpose. Such data-level privacy protections ensure that even though multiple data sets are aggregated, access to that information is strictly limited to users who have a need to access the data to perform their authorized responsibilities.

To accomplish this controlled use, Cerberus employs the Framework's dynamic access control to enable the automated enforcement of access requirements. More specifically, the information a user may access is based on characteristics of the user (user attributes), the purpose and function of the search (context), and the characteristics of the data within Cerberus (data tags). In other words, users may view only that information that they would otherwise be entitled to view in the authoritative system as a matter of law and policy.

Underlying all of this is a robust Framework governance structure for decision-making and oversight. Using the lessons learned from the Cerberus pilot, the new Framework governance structure, described in greater detail in the DHS Data Framework PIA Update,[9] is overseeing the development and maturation of consistent, repeatable, data ingestion processes in support of ingesting future data into the Framework. These processes include standardization and oversight to ensure: (1) accurate data tagging in support of privacy, civil rights, and civil

[9] *See* DHS/ALL/PIA-046(a) DHS Data Framework PIA Update, published concurrently with this PIA.

liberties controls and protections occurs; (2) data integrity is maintained as it is transferred; and (3) that the access control policies are enabled and enforced.

Privacy Impact Analysis

Authorities and Other Requirements

The data used in Cerberus continues to be covered by the source system Systems of Records Notices (SORN). Those SORNs are either currently consistent or are in the process of being updated to reflect the potential Intelligence operational use. To the extent a record from the source system of records relates to counterterrorism, the Enterprise Records System (ERS) SORN remains applicable. To the extent that CBP users (e.g., CBP's Office of Intelligence and Investigative Liaison), rely on Cerberus data for Intelligence operational use to create analytical products, the existing Automated Targeting System (ATS) SORN[10] remains applicable.

Cerberus continues to rely on the source systems to manage retention of their data and to include retention-based deletions and/or status changes in data deliveries via Neptune. During the ingest process in Neptune, the data will be tagged with metadata that indicates record creation data and allows enforcement of applicable retention rules and verification of compliance. There is no change to the Paperwork Reduction Act requirements, which remains non-applicable to this effort.

Characterization of the Information

Cerberus will continue to receive information originally collected by DHS components, consistent with their respective missions, via a computer-readable export from Neptune. Neptune provides tagged information originally collected by DHS and stored in unclassified DHS database systems. During the limited production capability phase, automated cross domain transfer will be explored. In the meantime, manual data transfer of encrypted exports between Neptune and Cerberus will continue.

DHS will maintain PII in Cerberus to conduct investigations in support of the users' border security and counterterrorism missions, which is consistent with the authority and mission for which DHS originally collected the information. The SORNs of all DHS data sources used

[10] *See* DHS/CBP-006 - Automated Targeting System (May 22, 2012), 77 FR 30297, *available at* http://www.gpo.gov/fdsys/pkg/FR-2012-05-22/html/2012-12396.htm. "ATS contains…information created by CBP, including: Operational and analytical reports and/or projects developed that may include public source information and/or classified information obtained by users/analysts for reference or incorporation into the report or project."

in Cerberus provide notice to record subjects that their information may be used for such investigations. To maximize data quality and integrity, all the policy and legal controls of the source systems are maintained via the Framework access control mechanisms while the data is stored, used, or shared in or through Cerberus.

The unclassified data maintained in Cerberus remains unclassified. If a specific piece of unclassified information is connected with classified material, then that association or connection will be classified. The fact that the data will be maintained on the TS/SCI environment does not change the nature of the data. The classification of the data will follow the authoritative system of records.

Cerberus continues to rely on the accuracy of source systems and the data validation during ingest and data quality processing to ensure accurate tagging and index. The Data Framework Program Management Office (PMO) is establishing timelier refresh of information from the source systems with near real-time being the refresh goal. Any risk that the data will not accurately reflect ongoing data changes at the source system because of refresh latency is mitigated by not relying on the system data for any operational actions or decisions. The systems of record are the source data systems so users retrieving information via Cerberus will be trained to understand the risk associated with the latency of data and trained to verify that information at the source system.

Risk: There is a risk that PII transferred outside of the original IT system and into Cerberus 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.

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.

Uses of the Information

To support limited operations, a controlled set of users will perform classified and unclassified searches against the Cerberus data cloud using the same search tools set described during the pilot phase (Specific Person or Entity-Based Search, Characteristic-Based Search, and Pattern-Based Search). Cerberus will use information about the user and the mission context of the requested use to perform dynamic access control to ensure that only authorized users with certain attributes are able to access the data for approved purposes.

Cerberus users will perform these queries to identify individuals and to support border security and counterterrorism mission purposes. No new records will be generated in the Cerberus limited production capability through the application of any search or analytic tools to ingested records by component analytic personnel. If, during the use of those tools, information of operational value is found, Cerberus component users are required to validate the information in the source systems before using the information in any mission systems or products. The use of that data will be in the mission systems of records and will be maintained according to the retention, use, and handling provisions of the respective SORNs for those mission systems of records.

At this time, there are only access rules defined for the three Framework-approved search tools internal to Cerberus. These rules and resulting dynamic access control were proven during the pilot phase. During the limited production capability period, additional analytical tools may be introduced through the Data Framework PMO to access Cerberus in order to enhance DHS mission analytic value. After Framework process approval, the execution of these new mission tools against Cerberus data will be controlled with policy rules developed consistent with the Framework policy-based access controls. This process includes governance structure vetting of the rules, applicable users and uses, as well as privacy impacts.

Risk: There is a risk that DHS will include data in Cerberus 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. Non-DHS users will not have access to 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[11] 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[12] 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[13] 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, 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.

[11] *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.

[12] *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.

[13] *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.

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.

Notice

Individuals do not have the opportunity to consent to the use of their data the Cerberus. DHS does not provide an opportunity to consent for data used by DHS for counterterrorism analytical purposes because this is the purpose for which the authoritative data was collected by the DHS component. Cerberus is a copy of the data from the authoritative system and follows the existing policies for use of the data. If as part of the limited production capability, DHS identifies records in all three data sets that relate to counterterrorism, the ERS SORN will apply to these records.

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

Mitigation: 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[14]

[14] *See* DHS/CBP-009 – Electronic System for Travel Authorization (ESTA), July 30, 2012, 77 FR 44642. Available

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[15] 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 Notice[16] 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 Neptune 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,[17] and the Framework is described in the White House's final big data report.[18] DHS has provided two public briefings on the Framework during meetings of its Federal Advisory Committee, the DHS DPIAC.[19] 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

Cerberus relies on the source system to manage retention of its data and to include retention-based deletions and/or status changes in data deliveries to Neptune. Information in the

at: http://www.gpo.gov/fdsys/pkg/FR-2012-07-30/html/2012-18552.htm.

[15] *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.

[16] *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.

[17] *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.

[18] *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.

[19] *See* the DHS Privacy website for archived meeting materials. Available at: http://www.dhs.gov/dhs-data-privacy-and-integrity-advisory-committee-meeting-information.

audit log is covered by General Records Schedules, GRS 20, approved by the National Archives and Records Administration (NARA).

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 and Cerberus therefore rely 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 or that Neptune will not information Cerberus that 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

The use of Cerberus during limited production capability will be limited to approved DHS users and tools determined by dynamic access control. If, during the use of those tools, information of operational value is found, Cerberus users are required to validate the information in the source systems before using the information in any mission systems or products. Data will be used, maintained, and shared consistent with the source system of records notice. The potential impact of any sharing of analytical work products based on the validated source system information are discussed in those system PIAs.

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

Individuals may seek access to their records based on the directions outlined in the authoritative system System of Records Notices. Data maintained in Cerberus remains tagged separately and is part of the authoritative system of records although stored in a different location. Component Freedom of Information Act (FOIA) offices will work with Data Framework PMO staff to access the unclassified information on Cerberus limited production capability in order to respond, as appropriate, to Privacy Act and FOIA Requests.

Because the authoritative systems that are ingested into the Cerberus limited production capability may contain classified and sensitive but unclassified information related to intelligence, counterterrorism, homeland security, and law enforcement programs, certain records will be exempted from notification, access, and amendment to the extent permitted by subsection (j) and (k) of the Privacy Act and as described in the Code of Federal Regulations.

A request for access to non-exempt records in this system may be made by writing to the FOIA Officer, Office of Intelligence and Analysis, Department of Homeland Security, Washington, DC 20528, in conformance with 6 C.F.R. Part 5, Subpart B, which provides the rules for requesting access to Privacy Act records maintained by DHS.

Procedures are being established within the Data Framework PMO for Framework systems 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 protocol, Cerberus 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[20] and (2) the ability to refresh the data that is ingested into the Framework.

With respect to access and redress, the 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 Cerberus cloud is secured against accidental or deliberate unauthorized access, use, alteration, or destruction of information. Cerberus ensures that the information is used in accordance with the Framework 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

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

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.[21]

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

The management of users accessing the Cerberus limited production capability system will be managed by the system DHS Headquarters Information System Security Manager. All users will have the requisite Framework attributes for the dynamic access controls to evaluate during data and tool use. DHS 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 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 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 the Executive

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

Steering Committee, which includes the Privacy Office, Office for Civil Rights and Civil Liberties, and Office of the General Counsel.

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

Clark Smith
Director, Knowledge Management Division and Chief Information Officer
Office of Intelligence and Analysis

Approval Signature

Original signed copy on file with DHS Privacy Office.

____

Karen L. Neuman
Chief Privacy Officer
Department of Homeland Security