Privacy Impact Assessment Update
for the

DHS Data Framework –

Unclassified Use

DHS/ALL/PIA-046(e)

October 6, 2017

<u>Contact Point</u>
Paul Reynolds
Data Framework Program Office
Department of Homeland Security
(202) 447-3000

<u>Reviewing Official</u>
Philip S. Kaplan
Chief Privacy Officer
Department of Homeland Security
(202) 343-1717

Abstract

The Department of Homeland Security (DHS or Department) Data Framework (Framework) is a scalable information technology (IT) program with built-in capabilities to support advanced data architecture and governance processes. The DHS Data Framework is the Department'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. Beginning in October 2017, DHS intends to mature the Data Framework by modifying how unclassified data that is stored within the Framework is accessed in order to facilitate the unclassified use of that data. DHS is updating the Data Framework Privacy Impact Assessment (PIA) to account for the unclassified use of data within the Data Framework.

Overview

In a PIA published on November 6, 2013, and subsequent PIA updates, DHS previously described the Department's development of the DHS Data Framework.[1] The Framework creates a systematic repeatable process for providing controlled access to DHS data across the Department. Currently, the Framework includes the Neptune[2] and Cerberus[3] systems, as well as the Common Entity Index.[4]

Reason for the PIA Update

Initially, the Framework ingested data into the Neptune unclassified "data lake" that DHS uses to receive, store, and tag data from unclassified DHS IT systems. Once tagged, the unclassified DHS data sets from Neptune were transferred to Cerberus, which is the classified data lake that DHS currently uses to perform classified searches of unclassified DHS data sets. The Common Entity Index is an unclassified correlation engine that will allow DHS to connect disparate DHS data sets to view all available information about an identified individual.

Now, the Data Framework will allow approved users the ability to access, search, and use the Framework's unclassified data within the Neptune system. Approved Framework users will now be able to perform searches on the unclassified data contained within the unclassified system, Neptune, without having to access the exact same data through the classified system, Cerberus. Approved users, based on need-to-know and position roles, will be able to perform person-or entity-based searches, characteristic-based searches, and pattern-based searches amongst other searches on the unclassified data. The Framework uses dynamic access control policies to enable the automated enforcement of access requirements so that a user sees only the information that he

[1] *See* DHS/ALL/PIA-046 DHS Data Framework and subsequent updates *available at* www.dhs.gov/privacy.
[2] *See* DHS/ALL/PIA-046-1 Neptune and subsequent updates, *available at* www.dhs.gov/privacy.
[3] *See* DHS/ALL/PIA-046-3 Cerberus and subsequent updates, *available at* www.dhs.gov/privacy.
[4] See DHS/ALL/PIA-046-2 Common Entity Index Prototype, available at www.dhs.gov/privacy.

or she would otherwise be entitled to view as a matter of law and policy. DHS may develop analytic tools that will allow statistical analysis, geospatial analysis, link or pattern analysis, and temporal analysis among other analytic tools for approved users' use in the future. The availability of the various searches will be limited based on need-to-know and position/roles of users with the full range of search methodologies not necessarily authorized for each user type and/or role.

Allowing the unclassified use of data contained within Neptune will further the core purpose of the Framework to enable more controlled, effective, and efficient use of existing homeland security-related data within DHS. At this time data will not be shared outside of DHS. Furthermore, the unclassified use of data in Neptune is consistent with the original purpose for which the unclassified data was collected and will be subject to all of the privacy safeguards of the source system. Many of the risks identified in the original PIAs continue to apply to DHS's unclassified use of the Data Framework.

Privacy Impact Analysis

Authorities and Other Requirements

The data used by the Framework continues to be covered by the source system's System of Records Notices (SORN). All approved data sets are listed in Appendix A of the Data Framework PIA,[5] including the applicable source system SORNs. Appendix A will be updated as new data sets are added to the Framework. Neptune will continue to identify the source component for all records, which enables validation in the source system.

Data will be maintained according to the retention, use, and handling provisions of the respective SORNs for those source systems. The Data Framework prefers to rely on the source IT systems to notify the Framework of changes, deletions, or corrections to data. Preferably, the Data Framework will not delete data until it receives a deletion notification from the source IT system. Deletions will be applied as defined by the source IT system. As part of the metadata tagging process, the Data Framework tags each data set with a retention period, and therefore the Data Framework can remind the underlying source IT system of the upcoming retention expiration date if the Framework has not already received a deletion notification. Additionally, and as described fully in the Data Framework – Retention PIA[6], the Framework can use this retention period tag to delete data in the Framework that is from IT systems that are unable to accommodate the sending of delete notifications due to a number of constraints (*e.g.*, resources, legacy systems, disruptions to operational support).

[5] *See* DHS/ALL/PIA-046(b) DHS Data Framework Appendix A – Approved Data Sets, *available at* www.dhs.gov/privacy.

[6] *See* DHS/ALL/PIA-046(d) DHS Data Framework – Retention, *available at* www.dhs.gov/privacy.

<u>Privacy Risk</u>: There is a risk that unclassified users of the data contained within Neptune may not have the same level of training as unclassified users using that same data in the source system.

<u>Mitigation</u>: To mitigate this risk additional training will be required in coordination with the holder of the source system and the Data Framework Working Group (DFWG), which includes the Office of Civil Rights and Civil Liberties, Privacy Office, and Office of General Counsel, to ensure that approved users of unclassified data within Neptune are trained in the legal and appropriate use of the data to the same standard as if the users were accessing the data in the source system. The Framework will submit these training materials to the DFWG and component source system owners for approval.

<u>Privacy Risk</u>: There is a risk that unclassified users of the data contained within Neptune may not have the same level of supervisor approval as classified users of the Framework.

<u>Mitigation</u>: An appropriate supervisor must approve each individual, and their stated need to know, who will be allowed access to the unclassified data in Neptune to the same standard as for users accessing the data in Cerberus.

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

Characterization of the Information

There is no change to the information the Framework collects, uses, disseminates, or maintains. The sources of the information and how the information is collected are described in previous Framework PIAs and in Appendix A.

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

<u>Mitigation</u>: As noted above, the Framework is designed first and foremost to support DHS users' mission needs while mitigating privacy risk. One of the hallmarks of the Framework is the ability to restrict access to types of data, including PII, within the Framework 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 DHS. The Department plans to expand the Framework's approved tagging scheme elements (i.e., common information fields used across the data sets) and refine data controls/tagging as the Department operationalizes the Framework in accordance with the governance onboarding process, as approved by the DFWG.

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

Mitigation: To partially mitigate this risk, DHS has developed a Data Quality Plan. In accordance with the Data Quality Plan, the Data Framework Program Management Office is establishing a feedback mechanism to report data quality issues to source systems. In addition, the Limited Production Capability phase introduced data quality processing that assures all data is received unaltered and correctly processed; that data anomalies are identified, logged, and reported to the source data owners for potential correction; and that the Framework is generating data quality metrics for performance and compliance reporting.

This risk will not be fully mitigated until DHS develops a near real-time refresh capability. The Data Framework Program Management Office also identified the timelines for manually refreshing each existing data set, and began implementation of limited manual data set refreshes. For each new data set, DHS will use the onboarding process, described above, to identify and implement manual data refresh timelines. To provide additional mitigation, Framework users will continue to be trained to understand the risk associated with data latency (due to limited refresh capabilities). Users will also be required to verify information in the source system before issuing any raw intelligence, (*e.g.*, intelligence information report), completing any final analysis, or using the information operationally.

Uses of the Information

Neptune currently ingests data as collected by unclassified DHS systems within a common schema that includes core biographic data, extended biographic data, and encounter data related to individuals. Core biographic data is basic biographic information that includes 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 and includes information such as address, phone number, email address, passport number, and visa number. Encounter data is information that derives from a DHS screening, vetting, law enforcement, or immigration-related event or process, and is collected in accordance with DHS authorities and regulations. Expansion of common schema elements and refinement of data controls and tagging may occur in accordance with the governance onboarding process as approved by the DFWG.

Unclassified data, which in its source system is already used on an unclassified basis by users of that source system, will now be available for unclassified use in the Framework through the unclassified Neptune system to only DHS users. For all data, the originating source systems remain the authoritative source for the data.

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

Mitigation: Until the DFWG approves new uses based on the onboarding process and its mission use case methodology, DHS users will only use the data for immigration, border security, and counterterrorism purposes. The SORNs for the existing data sets within the Framework specify that DHS collected the information for these purposes. Any changes to the data sets, users, and uses will trigger a review to determine whether the purpose remains compatible and whether this risk is impacted by the addition of new data sets, uses, or users.

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

Mitigation: To minimize this risk DHS will continue to carefully evaluate each data set to determine whether its use is directly relevant and necessary to accomplish the purposes authorized under the Framework. Appendix A will be updated continually to document information about current and pending data sets to be added to the Framework.

Privacy 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. The number of system data delivery methods will decrease as the Framework can provide a more controlled release and transfer of information within the Department. Eventually, some data aggregation systems may be decommissioned as their capabilities are replicated and centralized within the Framework. To implement this mitigation DHS must successfully replicate the capabilities of other systems and build user support.

Privacy Risk: There is a risk that DHS users will use the data for purposes other than those authorized.

Mitigation: Only the DFWG can approve new users and uses with input from the Office of Civil Rights and Civil Liberties, Privacy Office, and Office of General Counsel. Once a user or use is approved by the DFWG, technical controls ensure the user is only able to access data for the use or uses for which he or she has been approved, thus limiting that user's access within the Framework. Access to data is determined by a user's purpose and function, and the Framework's policy-based controls will ensure that a user is only able to access information that is permitted for a particular purpose and function.

Notice

This PIA update provides additional notice to describe the unclassified access, search, and use of unclassified data through the Neptune system prior to the transfer of that unclassified data to the classified Cerberus system.

Privacy Risk: There is a risk that individuals may not be aware their PII is being compared against other DHS information in this DHS-wide big data project.

Mitigation: The existing SORNs for data sets incorporated into the Framework provide notice to the public that the information may be compared against other data sets and be subject to analysis for varying DHS missions. As DHS adds new data sets to the Framework, the Privacy Office will review the applicable SORNs to see if they provide appropriate notice. DHS will continue to review its PIAs when it adds new data sets, user bases, or capabilities to the Data Framework.

Data Retention by the Project

As described in the March 2017 update to the Data Framework PIA, the DHS Data Framework has developed data management capabilities to internally manage and comply with the source IT system's data retention rules. These data management capabilities provide significant enhancements to the DHS Data Framework's core offerings; however, it is still the preference of the DHS Data Framework Program Office to receive automatic notifications of changes, deletions, or corrections to business rules from the source IT system whenever possible. As described fully in the Data Framework – Retention PIA[7], the Framework can use the retention period tag to delete data in the Framework that is from IT systems that are unable to accommodate the sending of delete notifications due to a number of constraints (*e.g.*, resources, legacy systems, disruptions to operational support).

Privacy 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 apply when that information is ingested into the Framework. The Data Framework will delete data on receipt of a deletion notification from the source IT system, or will replicate the source IT system rules to ensure compliance when the source IT system cannot send delete notifications. As part of the metadata tagging process DHS tags each data set with a retention period, and, therefore, DHS can remind the underlying source IT system of the upcoming retention expiration date if the Framework has not received a deletion notification, or can enforce a deletion if the source IT system cannot send delete notifications.

[7] *See* DHS/ALL/PIA-046(d) DHS Data Framework – Retention, *available at* www.dhs.gov/privacy.

Privacy Risk: For data sets that cannot send delete notifications, there is a risk that the DHS Data Framework will not capture the retention rules accurately and maintain information longer than the National Archives and Records Administration (NARA)-approved retention period.

Mitigation: To mitigate this risk, the Department employed a formal governance process to address DHS Data Framework management of source IT system retention rules to ensure consistency and compliance with the retention period for the source system. As part of the governance process, the DHS Data Framework Program Office, in collaboration with associated stakeholders, prepared a request and documented the analysis of the retention and other business rules and compliance impacts associated with the internal retention management, including an assessment of potential risks and proposed mitigations, as well as consistency with DHS Data Framework compliance protections. The DHS Data Framework Program Office submitted these materials to the Oversight Offices and the DHS Data Framework governance structure for approval.

Information Sharing

This update does not impact internal or external sharing and disclosure, which is described in previous Framework PIAs.

Redress

This update does not impact how access, redress, and correction may be sought.

Privacy 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.

Mitigation: This risk remains during the unclassified use of data within the Framework and will remain until: (1) DHS has near real-time refresh and (2) DHS can 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. To partially mitigate this risk, users of the Framework must verify the accuracy of the data in the source IT system before using the data operationally.

Privacy Risk: There is a risk 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: This risk remains during the unclassified use of data within the Framework and will remain until DHS has near real-time refresh. To partially mitigate this risk, users of the Framework must verify the accuracy of the data in the source IT system before using the data operationally.

Auditing and Accountability

The DHS Data Framework Program Office will provide a report regularly to each respective Component source system owner and the Framework governance structure, in order to validate that data in the source IT system and the Framework are in sync.

Users will continue to be required to take general privacy training as well as training specific to data sets within Framework. Users are trained to verify information at the source system before completing any final analysis or using the information operationally. Additionally, users are required to refresh their training on the appropriate schedule as determined for all of DHS for privacy training and on the appropriate schedule as determined by the source Component for data set specific training in order to ensure that all training is completed regardless of whether the data is being accessed through the Component source system or the Data Framework.

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.

Furthermore, the Data Framework will require appropriate supervisor approval for use and scope of use. The supervisor of each user type must provide reasoning for the user to access and the purpose that a user is authorized to perform on the Data Framework and that user will be limited to that scope of search functionality. The approval must be registered by the Data Framework Working Group prior to the user being able to access the Data Framework.

Privacy 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: DHS has demonstrated that the Framework's audit capabilities were adequate to support an audit of whether PII 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. 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.

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

Mitigation: The Data Framework Program Management Office will pull a random selection of queries from Framework systems and manually review them to determine compliance with the Framework policies. The Program Management Office will present its findings to the DFWG.

Responsible Official

Donna Roy
Office of Chief Information Officer
Department of Homeland Security

David Bottom
Office of Intelligence and Analysis
Department of Homeland Security

Approval Signature

Original, signed copy on file with the DHS Privacy Office.

____

Philip S. Kaplan
Chief Privacy Officer
Department of Homeland Security