Richard Purcell
Chair, DHS Data Privacy and Integrity Advisory Committee

January 31, 2012
Delivery by Hand

Hon. Janet Napolitano
Secretary, Department of Homeland Security
Washington, DC 20528

Mary Ellen Callahan
Chief Privacy Officer
Department of Homeland Security
Washington, DC 20528

Re: DHS Data Privacy and Integrity Advisory Committee Privacy Policy and Technology Recommendations for a Federated Information Sharing System

Dear Secretary Napolitano and Ms. Callahan:

I have the honor to convey to you the enclosed Report, which sets forth privacy considerations and recommendations for designing and building policy and technical protections into federated information sharing systems. Our recommendations are grounded in the Department's Fair Information Practice Principles policy framework. We believe that implementing these recommendations would protect the information privacy of individuals while strengthening the Department's efforts to efficiently use personal information held in DHS systems and furthering the Department's mission.

If I may be of any assistance to you concerning these recommendations, please do not hesitate to contact me.

Sincerely,

Richard V Purcell

Richard V. Purcell

Enclosure
cc: Members, DHS Data Privacy and Integrity Advisory Committee (by email)

Report No. 2011-01

Privacy Policy and Technology Recommendations for a Federated Information-Sharing System

This white paper reflects the consensus recommendations provided by the Data Privacy and Integrity Advisory Committee (Committee) to the Secretary and the Chief Privacy Officer of the Department of Homeland Security (DHS or Department). The Committee's charter under the Federal Advisory Committee Act is to provide advice on programmatic, policy, operational, administrative and technological issues within DHS that relate to personally identifiable information (PII), as well as data integrity and other privacy-related issues.

The Committee deliberated on and adopted these recommendations during a public meeting on 06 December 2011, in Washington, DC.

Preamble

This white paper is intended to be responsive to a specific tasking from the Chief Privacy Officer of the Department of Homeland Security (DHS). This paper is intended to provide guidance based on the information understood to date by the Data Privacy and Integrity Advisory Committee (DPIAC) and is intended to evolve as the Department continues to refine the federated information-sharing construct. The document below is intended to be read concurrently as a single paper and to provide advice and analysis in a timely manner regarding a federated information-sharing program that has yet to be built. By requesting DPIAC to issue this guidance early, DHS and the Privacy Office demonstrate their commitment to Privacy by Design. As the federated information-sharing program evolves, so too may this guidance.

Privacy Policy Guidance

I. Introduction

The Department of Homeland Security's Data Privacy and Integrity Advisory Committee is pleased to present this white paper in response to a request by the Chief Privacy Officer in a Tasking Letter to the Committee Chairman, dated December 30, 2010. The Tasking Letter indicated that the Department is in the process of creating a policy framework and a technology architecture for enhancing DHS's information-sharing capabilities. Currently, the information sharing environment at DHS is composed of individual systems intended to support the unique missions of the various DHS components. The data are used within those systems, which essentially comprise a series of stovepipes, to support the unique functions of the distinct DHS components. The new information-sharing project aims to create a federated system to facilitate efficient and effective data sharing among the various DHS components. We understand that there are two possible approaches in designing such a system. One approach envisions a centralized database at the "hub," which would contain pointers to participating component databases, the queries of users searching for information in other component databases, and the results of those queries. The hub also would contain an audit log. An alternative approach would be to retain far less information at the hub, limited to pointers to participating component databases and an audit log. The audit log would retain the queries and information on the users.

As discussed in detail in this white paper, the Committee believes the latter approach, with minimal data stored at the hub, would be preferable (assuming little or no reduction in effectiveness of the proposed data-sharing project), reducing the possibility of adverse privacy impacts and database management challenges.

We also understand that there may be sharing of data through the federated system between DHS and other federal agencies. Such a possibility reinforces our recommendation against a centralized, fulsome database at the hub.

II. Potential Privacy Issues

Although the various DHS components currently share data including personally identifiable information (PII) among themselves from their distinct databases, the move to an automated, federated system for information sharing inevitably will exacerbate certain privacy risks. The new system, whose purpose is to facilitate information sharing, will, by its nature and ease of use, increase the volume of information shared, the volume of new data created as a result of combining data sets, and the volume of inferences and decisions made possible by the new data sets. This white paper is intended to highlight the potential privacy policy and technology issues raised by the enhanced information sharing made possible by a federated system.

As currently envisioned, the federated data sharing environment would be comprised of pointers, and possibly also of new data comprised of the queries. Although privacy protections already are in place for data sharing among the DHS components, a federated system might in fact be able to enhance existing privacy protections by allowing for the development and implementation of uniform, systemic safeguards that would apply to the data-sharing system as a whole rather than the individualized safeguards that currently apply to the siloed databases.[1]

We recommend that DHS fulfill its obligations under the Privacy Act with respect to the proposed federated system. We specifically urge DHS to be judicious in the exemptions it claims from the provisions of the Act. DHS should allow individuals to claim the maximum privacy rights created by the Act that are consistent with the government's legitimate national security and law enforcement concerns. Doing so upholds and complements the DHS Fair Information Privacy Principles and is an important aspect of an open and transparent government.

The Committee believes the key privacy policy issues associated with such a federated system fall within the following broad categories: (a) controlling access and use, (b) applicable privacy policies, (c) data integrity and quality assurance, (d) accountability and audit, (e) data security and data retention, and (f) redress. Each of these issues is addressed in the policy section of this white paper.

III. Privacy Policy Guidance

A. Controlling Access and Use in an Information-Sharing System

A key privacy consideration in sharing data is maintaining appropriate control over access to the data. Access includes both *who* can query the database and for what *purposes* such queries are permissible. If access controls are inadequate, sensitive information may end up in the wrong hands and be used for purposes for which it is not appropriate, either because of the inherent limitations of the database itself or because the use violates restrictions on how the data may be used. In a federated data-sharing system, controlling access is a principal means of ensuring that legal and policy limits on the uses of participating databases are respected.

There are at least two basic models for controlling access to shared databases: a centralized model, in which a headquarters organization within the Department determines who may access a particular database (and for what purposes), and a decentralized model, in which the

[1] We note that DHS has in place a process for conducting Privacy Impact Assessments ("PIAs"), which process is overseen by the DHS Privacy Office. The Committee recommends considering the issues raised in this white paper as part of the PIA process. The PIA template documents are available on the DHS Privacy Office's website at www.dhs.gov/privacy.

organization that creates and maintains the database (the "owner") controls access. As described below, each approach has advantages and disadvantages. We consider first the decentralized model because a primary advantage of the centralized model is that it reduces some of the disadvantages of the decentralized, owner-controlled, approach. We also briefly, but by no means exhaustively, consider hybrid systems, combining elements of centralized and decentralized systems.

Throughout this discussion, we are assuming a system that grants access to specific individuals (or perhaps individuals in specific positions) for specific purposes. More general access would raise a broader set of privacy concerns. We also assume that the system would permit queries based only on specific PII, such as a name, an address, or a phone number.[2] Given this assumption, there is little risk of users searching for potential patterns that conceivably could identify potential persons of interest. A system that would allow such pattern searches raises a far more significant set of privacy issues. Should the proposed system be altered to allow for pattern-based searches, this analysis would need to be revisited.

1 Decentralized Access Control

Under a system of decentralized control, the owner of a database would determine who is allowed to access the data. We understand that this is the current model within DHS: an individual or organization that seeks to obtain access to a particular database must contact the owner, provide information about the intended use of the data, and obtain the permission for access.

(a) Advantages

The organization that owns a particular database is likely to have the greatest knowledge of its strengths and limitations. The owner is aware of the sources of the data, the types of inaccuracies that may be present, and any restrictions that apply to secondary uses. If the owner has a clear understanding of the proposed use of the data for another user or purpose, the owner likely is in the best position to determine whether the data are appropriate for that use or purpose. Thus, decentralized control is the most effective way to minimize the risk of inappropriate uses of a database (including uses that violate restrictions on subsequent uses of the data).

(b) Disadvantages

A potential advantage of a system of federated queries is that users may obtain access to data they did not know was available until they ran a query. In a decentralized access system, this advantage is sacrificed because a user must already know about the potential utility of the

[2] This assumption is based on the Tasking Letter dated December 30, 2010, which indicates that the federated database "would consist of a searchable index of biographic data" culled from certain DHS databases.

database and seek permission from its owner. Alternatively, a federated query might reveal to all users that there is relevant data in a particular database, but might still require the user to obtain access to those data from the owner. Depending on the database, however, the mere fact that information exists may itself be sensitive information. Moreover, users who know the information exists but do not have permission to access it may draw inappropriate inferences (e.g., there is a record for this person in a database that tracks deportation proceedings). Providing less information to someone who does not have permission to access the database reduces, but does not eliminate, the risk of inappropriate inferences. Knowing that there are data in a database of suspicious transactions is less sensitive information than knowing, for example, that there are 20 records in the last six months. It might also be possible to provide only the contact information for the owner of the database, without revealing precisely what database is involved.

Depending on the culture of a particular organization, owners of data may have an inappropriate "proprietary" interest in maintaining control over access. This can lead to a tendency to deny access to data in a database even when the requestor is seeking data that in fact would be useful and an appropriate use for the intended purpose. In short, owners of data may deny access when they should instead grant access.

Just as the owner of the data has the best information about their limitations and restrictions, the potential user has the best information about the intended use and why access potentially is important. Owners who do not fully understand the nature and significance of the intended use of the data inappropriately may deny access. If organizations are inappropriately possessive about access to their data, this problem is compounded.

A decentralized system presumes that the potential user is aware of the existence of a particular database and its relevance to a particular query. Potential users who are unaware of the possible utility of a database will not know to request access rights. It may not be reasonable, however, to presume that potential users (especially occasional users) are aware of the vast array of data included in various DHS systems, even though the data may have great value in particular cases. A potential advantage of federated queries of all DHS data is that relevant information may exist in databases that are unknown to the user; the decentralized system inevitably sacrifices this potential advantage. Although potential users likely are aware of databases that are relevant to routine queries, they may be less aware of those that are relevant to an idiosyncratic need for information. Even if they are aware of the potential relevance for an idiosyncratic query, the user's permissible uses may not include the unusual situation. Query-specific permission for access in such cases likely is too slow and cumbersome to support the Department's mission effectively.

2 Centralized Access Control

Under a system of centralized access control, a DHS-wide organization would assume responsibility for determining who may have access to different databases and for what purposes. This organization would need Department-wide authority to assure cooperation with the inevitable requests for information and compliance with its decisions. The Privacy Office, or some other Secretary-level unit of the Department (e.g., the Chief Information Officer, the Office of Policy or the Office of the General Counsel), may be the appropriate place for this function.

(a) Advantages

A centralized system is more likely to facilitate information sharing. If data owners are imposing inappropriate restrictions on sharing, the only effective solution is likely to be a centralized function that grants access to particular users for particular purposes. A centralized system also could reduce the problem of users being unaware of the potential utility of certain databases to serve their legitimate purposes because the centralized resource could assess the potential value of a given database and grant access accordingly. A centralized system could, for example, potentially decide that certain uses (e.g., finding possible locations for a suspected terrorist) are appropriate for any DHS database.

(b) Disadvantages

A centralized organization would have less information about the limitations of a particular database than the owner, and less information about the intended uses than the user. As an intermediary, it would depend on cooperation from both the owner and potential user to obtain the necessary information about whether access should be granted and the restrictions that should accompany the permission.

A centralized organization inevitably would develop its own sense of its purpose. If the organization views its primary purpose as facilitating information sharing, it may grant access in circumstances in which access is inappropriate. If it sees its primary purpose as preventing inappropriate access, it may deny access when access is both appropriate and potentially valuable.

3 Hybrid Systems

It is possible to imagine various combinations of the centralized and decentralized systems. These combinations could encompass virtually any mix of the two systems, and we do not attempt to explore them systematically. We note, however, a few possibilities that may take advantage of the strengths of both the centralized and decentralized approaches discussed above.

One possibility would be a basically decentralized system, but with a central organization that could override the decisions of the database owner. This organization might be an appeals organization, whose primary purpose is to resolve disputes between users seeking access and owners seeking to deny it, or it might offer access overrides, giving access regardless of the views of the database owner. The override approach in particular could avoid the problem of potential users who do not appreciate the utility of a particular database, if other parts of DHS are aware of its utility.

At the other extreme might be a basically centralized system, but with a right of database owners to override grants of access to particular users or for particular purposes. Like a centralized system, this approach could facilitate data sharing; like a decentralized system, it could assure that peculiarities of a database that render it inappropriate for particular uses would be reflected in access decisions.

4 Concluding Thoughts on Data Access

Because the present system is essentially decentralized, a critical question is how well it is performing. If appropriate decisions are being made in the current system about who can access data and for what purposes, retaining that system would be attractive. Creating a new centralized system inevitably would create risks of the problems discussed above. Implicit in this view, however, is that the only reason for a federated query system is to reduce the costs of obtaining the data to which potential users outside of the component that owns the data already have access.

If, however, appropriate access is not being granted under the current system, either because database owners are overly restrictive in granting access or because potential users are unaware that potentially useful information exists, a more centralized approach (or a hybrid approach) may be more appropriate.

A. Applicable Privacy Policies

Although the components already have privacy policies for their databases that will comprise the federated system, we recommend that the Department develop a privacy policy that would govern the use of the federated database. The “federated” privacy policy should describe, at a minimum, who has access to the system, the purposes for which the system may be searched, and who is accountable for ensuring compliance with the system’s privacy policy. The policy should take into account the Department’s “Federal Information Sharing Environment Privacy and Civil Liberties Protection Policy”[3].

[3] Memorandum 2009-01 of May 2009, available at www.dhs.gov/xlibrary/assets/privacy/privacy_crcl_guidance_ise_2009-01.pdf. See also the Advisory Committee’s paper, “Final White Paper on Department of Homeland Security Information Sharing and Access Agreements, at www.dhs.gov/xlibrary/assets/privacy/privacy_dpiac_issa_final_recs_may2009.pdf.

This new policy should acknowledge existing privacy policies that govern data use by DHS as a whole and by each of its components, particularly regarding restrictions on the way by which source data may be used. Thus, if a component has a privacy policy in place that restricts the sharing of data or disallows the combining of data, DHS should consider whether that component's database should be excluded from the federated system. DHS also should ensure that other relevant restrictions are honored. Although DHS and its components may amend and revise their privacy rules, absent any such change, use of the federated database should not trump a component's policy that protects individual data or data sets. In short, data should retain its original protections when it is accessed through the federated system.

B. Data Integrity and Quality Assurance

The privacy rights of individuals about whom DHS collects, maintains and uses PII depend on the quality of the information – its completeness, accuracy, relevance and timeliness. When information is shared outside the component that originally collected it, there is increased potential for misunderstanding its relevance and context, introducing inaccuracies, and allowing it to become outdated. Furthermore, the quality of the new information product obtained by combining data on an individual from multiple components is dependent on that of its sources, with any shortcomings potentially multiplied and magnified by the act of combination.

1. Creation of New Data

As a threshold question, we consider what data should be the subject of privacy concerns and policies, including those of data integrity and quality assurance. In the approach we recommend in Part I, the audit log would contain new data: the identity of the query submitters, the dates of queries, other metadata about the users, and the contents of the queries. For these data, the security of the system and strong access controls for the system as a whole (and the audit log in particular) would be the means of assuring the integrity of the data. Integrity and reasonable retention policies for these data are essential to providing individuals with appropriate redress.

We assume that the data consisting of the results of queries will reside not in the hub, but instead in the databases of the users, who would bear the responsibility for quality assurance. We address such data in our discussion below on data quality.

2. Data Quality

To enable a user of the federated system to make judgments about the quality of the data returned in response to queries, we recommend the creation of a knowledge management tool, which would contain basic information about the participating databases and would perhaps be stored in the hub. The knowledge management tool would identify the person responsible for a particular database, the rules governing access to it, the purpose for which the data were collected, the permissible uses of the data, the frequency with which the data and database are updated, and other relevant information. The contents of the knowledge management tool would be provided by the persons responsible for each database comprising the federated system.

The user who submits a query should be responsible for ensuring the quality of the search results. Accuracy and reliability are essential elements of data quality. The user submitting a query likely will have a different specific purpose for using the information than the source component, and may require a higher degree of reliability than was necessary for the original purpose. It is critical that the query submitter be able to obtain some understanding of the reliability of the information received from other components. To achieve this, the user would need to know the source of the data, including the underlying conditions of its collection. The user also would need to assess the accuracy of the association of the new data with the individual on whom the user is seeking information.

Timeliness also is of vital importance to data quality. A query result should indicate, perhaps through the use of metadata, the date of the last update and, if relevant, the regularity with which the information is updated by the source.

In short, the knowledge management tool should provide information sufficient to assist the users in making assessments of data quality.

As a result of combining query responses, the component that submitted a query might learn or suspect that some of the source information is inaccurate. The Department should consider whether the query submitter should report such potential inaccuracies to the source component, which would be responsible for assessing it and making any changes indicated.

C. Accountability and Audit

Although each DHS component currently is accountable for its respective databases, the transition to a federated system for purposes of sharing information raises new questions of accountability for compliance with the privacy policies governing the federated system, including accountability for the audit log of queries designed to facilitate information sharing. Accountability for the federated system implies responsibility for more than just controlling access to the system and the queries that are made. Accountability also requires attributing responsibility for actions taken and liability in the event of misuse.

Accountability for the new substantive PII that results from combining the results of queries, and the inferences and decisions made possible by the new data sets, rests with each user of the system.

Audit is an essential tool of accountability, and the ability to effectively audit system usage is critical to ensuring legal and policy compliance. Proper training, clear policies and procedures, and auditing can help mitigate potential abuses through prevention and early detection. Key questions are who should be responsible for audits and what types of audits should be conducted. Should audits of user access and permission rights be conducted at the component level or should auditing occur at the level of the federated system?

Audit requirements for the federated system must fit with the existing audits that we presume are part of current practice in both the component that owns the database and other components that have access to the data. We do not have sufficient familiarity with current audit practices to address the issue of fit directly. Nor do we address how to manage audits of classified queries. Finally, we are assuming that the federated system does not create a new database of queries at the hub, other than the audit log itself.

1. Audits by Components

(a) Advantages

Allowing the components to manage audits of use of the federated system by their own personnel is advantageous for three reasons. First, it may not be practical to reduce a component's complex responsibilities and workflows to a representation that an outside auditor could verify. Second, the components define many of their own policies, educate their own personnel about acceptable queries and appropriate access to substantive PII, and articulate consequences in light of their own understanding of the relevant mission and purpose of the organization and the applicable laws and policies. Third, the components might be able to more easily (1) establish a baseline evaluation of acceptable queries by appropriate individuals, (2) continuously monitor for changes in established controls, and (3) continuously monitor information flow [and achieve objectives such as asset safeguarding, data integrity, effectiveness, and efficiency.

(b) Disadvantages

Auditing at the component level has several disadvantages. Custom, culture and governance may inhibit irregularities, but they are not infallible deterrents. First, individuals may circumvent internal controls. Second, an individual responsible for exercising an internal control could abuse his or her responsibility. Three, there might be a strong bias to suppress information about identified breaches or violations that could reflect negatively on a component's reputation or image. Independent verification (whether through an independent party within DHS or an external party) and assurance of compliance with established data-sharing laws and policies provides the coordinating mechanism needed to foster effective information sharing. Independent third party audits may reduce the risk of inappropriate information sharing which would undermine public trust and confidence. In the absence of an independent audit function, it may be difficult for the public to trust that information will be shared in a manner consistent with the mission and purpose for which it originally was intended.

2. Audits by a Centralized Function

(a) Advantages

Audits at the level of the federated system, including an effective governing body, audit committee and external audit function, may constrain improper conduct. First, audits conducted at the level of the federated system would help identify problems with access rights and use of substantive PII at a level above that of the components. Second, audits of the log at the hub level might be most effective in identifying prototypical queries and acceptable users. This would help in curbing abuses. Third, an audit at the level of the hub might be most effective in identifying any changes to the underlying data. Fourth, such high-level audits might most effectively account for the laws and policies that apply to all of the components.

(b) Disadvantages

First, management-level auditors at the level of the federated database may not be able to comprehend fully the missions and objectives of each component database. Second, each component would better understand applicable security classifications of information, security clearances, relevant laws and policies, and reporting constraints. Third, the components necessarily have better knowledge about the practices, procedures, and techniques that provide for the authorization, completeness, and accuracy of application data. Fourth, the components might have a clearer sense of the audit environment, including an inventory of the infrastructure and security vulnerabilities.

B. Data Security and Data Retention

Protecting PII through appropriate safeguards, controls and training is essential to the Department's mission. Data security should protect against the unauthorized use, disclosure, access, destruction, modification and loss of PII. It also should safeguard against the unavailability of PII.

1. Data Security

Concern about data security is one of the reasons not to create a fulsome database of queries and results at the hub of the federated system. Such a database would become a prime target for hackers (whether state-sponsored or otherwise). One safeguard is a policy of not retaining query results at the hub but instead returning them to the user's system, where they either are securely destroyed if they are not found useful or they are retained pursuant to the component's security policies.

The data created as a result of use of the federated system must be appropriately secured. The pointers that identify the participating component databases should be secured at the hub and made accessible only to authorized users, according to the federated system's access policy. The queries and query results, along with the metadata, should be retained in an audit log with very limited access for auditors, as designated by the system's access policy.

2. Data Retention

The results of queries that are incorporated into the databases of users of the federated system should be subject to the data retention policies of those databases. The new data, both the pointers at the hub and the queries and results in the audit log, should be subject to a retention policy specified in the system's privacy policy. The retention period should be lengthy enough to allow for audits and appropriate redress for individuals[4].

D. Redress

In an earlier report, the Committee provided a general description of redress and an overview of the elements of effective programs for providing redress to individuals.[5] Redress is particularly critical for DHS because of the serious impact on individuals of the decisions the Department makes. Those decisions often are based on PII collected by the Department, making the quality of that information of vital importance to the privacy and liberty interests of U.S. residents and other individuals.

A threshold question for redress is what constitutes a "wrong" that requires redress. The decisions made and actions taken by DHS may impinge on an individual's privacy in various ways, and when such decisions or actions are based to any extent on PII collected or maintained by DHS, then the affected individuals should have a means of challenging the quality of the information upon which the Department relied.

[4] We recommend consultation regarding this issue with the National Archives and Records Administration (NARA) and the Office of the General Counsel.

[5] Report No. 2010-10, "The Elements of Effective Redress Programs" (March 2010), available at www.dhs.gov/xlibrary/assets/privacy/privacy_dpiac_report2010_01.pdf.

We have addressed the issue of redress in our report entitled “The Elements of Effective Redress Programs.” Providing for redress with respect to decisions that involve information provided by multiple entities is always challenging. The elements of particular relevance for the proposed federated system are accountability and an integrated infrastructure for redress.

Accountability for redress for actions based on combined query results must rest with the component that created the new data (i.e., the query submitter) and that made the decision or took the action that affected the individual. It is unreasonable to place the burden of determining the original source of challenged information on the individual seeking redress. Of course, for a component to be able to investigate and potentially make any corrections indicated, the query responses must be traceable to the source component.

Privacy Technology Guidance

Throughout this discussion, we are assuming a system that grants access to specific individuals (or perhaps individuals in specific positions) for specific purposes. More general access would raise a broader set of privacy concerns. We also assume that the system would permit queries based only on specific PII, such as a name, an address, or a phone number.[6] Given this assumption, there is little risk of users searching for potential patterns that conceivably could identify potential persons of interest. A system that would allow such pattern searches raises a far more significant set of privacy issues. Should the proposed system be altered to allow for pattern-based searches, this analysis would need to be revisited.

A. Controlling Access to a Shared Database

Program decisions on the degree to which access controls should be centralized will be critical points for the DHS Privacy Office to provide input and guidance to the creation of a federated information sharing system. Federated access control systems contain many of the same issues as federated identity management structures and considerable guidance can be derived from the work done in the past in that area, and work currently being done on the National Strategy for Trusted Identities in Cyberspace program www.nist.gov/**nstic**.

Each federated database that contributes data to a DHS federated information sharing system will likely determine the classification of its data and prescribe rules on entities or individuals who should not access and/or receive the data. A centralized access control system will be necessary, especially given the lack of a full understanding of the potential uses of the federated data or of the classes of entities who may gain access such as non-DHS federal agencies and state/local/tribal organizations. From a process standpoint, the access control rules will have to be specifically delineated and made fully operational in corresponding technology solutions. Moreover, different account types will need to be identified, the conditions for group membership established, and access to the federated information sharing system should be predicated on specific conditions. These conditions should include multiple, auditable access control mechanisms, incorporating a variety of attributes important to the organization (e.g., role, intended use, physical location, case assignment), appropriate to the requested data, and to the source systems.

The Privacy Office will need to have dedicated resources to help both guide the creation of this mix of centralization and federation, and to provide oversight of the regular risk assessment as to whether the system is behaving appropriately. Additional access control systems and processes will be required to be put in place for the log data created to provide reasonable

[6] This assumption is based on the Tasking Letter dated December 30, 2010, which indicates that the federated database "would consist of a searchable index of biographic data" culled from certain DHS databases.

security and accountability. A determination must be made at the outset over whether the Privacy Office should operate these systems, establish a technical reporting system for their operation, provide general assessment and oversight, or some combination of these roles. An explicit determination of who will take on these roles, and why, should be determined early in the design phase. In any case, the Privacy Office should play a central role in the development, testing, deployment and oversight of the system's design, function, and operation.

B. Data Integrity and Quality Assurance

The issue of interoperability of the data structures of the federated databases requires immediate technical attention, with guidance from the DHS Privacy Office. It is unlikely the data in the source databases are currently stored in a manner that allows for easy and accurate data relationships among them. DHS may need to create a template middleware translation to allow for similar, but different, data fields, formats and values to be combined.

In accordance with well-known business warehouse architecture concepts, attention will need to be paid to the major data layers of the system - data acquisition, data storage, and data presentation. Each of these layers plays an important part in assuring integrity and quality. For example, the data acquisition layer may address data in the different source databases and either load them into a data warehouse or 'normalize' the data to prepare it for queries. A main challenge may likely be the use of different attribute values across the different source databases. Data cleansing rules will need to be created to recognize the relationships between different types of data and to ensure their accuracy. The storage system for these data cleansing rules will itself need appropriate access control management processes and audit structures.

Another significant challenge will be to the need to automatically identify and resolve data conflicts that flag quality issues (e.g., two systems reporting different dates of birth for the same social security number). These conflicts will need to be logged and then communicated to the systems of record for resolution, which will also have to be overseen. Much of this process may be manual and may have privacy implications for individuals (e.g., determining which birth date is correct). This process may also present significant cost implications for the government, so prior similar efforts, both in DHS and elsewhere, should be analyzed before undertaking this effort.

An additional requirement will be the development of machine and system readable metadata tags and rules that would enable the management of data utility against policy requirements. For example, stale data may not be reliable for certain applications or functions. Likewise, confidence in the quality of certain data may be an attribute relevant to certain uses. As systems are designed (and redesigned) for future integration with the federated information sharing system, attention needs to be paid to meta-data tagging associated with data types, data elements, data sources, data time-stamps, data retention periods and other factors that may be material to the reliability and quality of data for particular purposes. This is distinct from data accuracy, in that accuracy *per se* does not necessarily address relevance and fitness for specified uses. Policies need to be developed addressing the appropriateness of data. This is an area where the DHS Privacy Office can contribute by developing policies and review and approval processes to manage this aspect of data.

C. Redress

During the requirements definition phase for any resulting system, it is important to address the opportunity for automated redress within the context of the system of records. The DHS Privacy Office will need to work closely with the program management team to collaboratively develop requirements that ensure inclusion of redress goals. At a minimum, there should be one level of redress required for implications to the individual from the results of the centralized query to the federated information sharing system. However, to the degree the redress request requires an update to the system of record, there should be an automated way to process that request such that it reveals from which systems of record the original data came. Effectuating this redress mechanism will likely require the centralized database to understand and log from which systems the initial data came. The inclusion of this data in the centralized system will create additional access control and security requirements for that centralized log.

D. Secondary Uses and Onward Transfers

Because the system will provide responses to specific queries, users will draw inferences from the combined data. This is a stated goal of all federated systems and drives the need to have a mechanism to make certain the queries to the federated databases, and the use of the resulting inferences, do not violate privacy commitments made by the source system of records. It is unlikely such a mechanism can be manual, so a serious design effort for an automated system must be undertaken.   These commitments will include representations made in Privacy Impact Assessments and Systems of Record Notices for Federal Systems, but may also include policy commitments, state/local/tribal laws and published privacy policies. Setting up the centralized mechanism for transferring these requirements from the systems of records, matching them with the proposed new uses or transfers, and maintaining auditable logs to understand how the decisions will be made, will be a significant undertaking. This work will require substantial resourcing from the DHS Privacy Office. There are other efforts within the Federal Government actively addressing these kinds of requirements in software systems and those technology developments should be leveraged herein.   The DHS Privacy Office will also need to work with their component privacy professionals who have oversight responsibility for the underlying systems.

E. Applicable Privacy Policies and Standards Development

The federated information sharing system should have a machine readable privacy policy to help manage secondary use and onward transfer and other privacy management requirements. There is considerable history in machine readable privacy policies and their technical implementation. DHS should consider mandating the use of machine readable privacy policies for the databases that will comprise the federated information sharing system.

Work is currently underway in the standards development community, including OASIS, ISO/IEC and other recognized standards bodies, to develop standards that can be used to build automated implementations of privacy management controls. The Privacy Office technology staff should explore engagement with this important work in standards to inform the process and help provide use cases applicable to DHS needs. Additionally, the Privacy Office should coordinate with other government agencies, such as NIST, while also recognizing the additional value that may come from direct engagement in the standards process. Ultimately, DHS should determine the appropriateness of adopting specific standards applicable to its systems. Contributing to the standards development process can also help drive technology innovation and the integration of the standards into commercial, off-the-shelf products.

F. Accountability

For the appropriate oversight personnel (over both the information sharing system and the federated source systems) to be accountable for the commitments described above, it will be necessary to provide the technical ability for them to perform periodic risk assessments of the system and to understand the results of those risk assessments (if they have oversight responsibility of one of the federated databases). Tools will need to be developed, or acquired, to provide oversight officials with the appropriate access and log data to assess the system.

Technologies that support governance, risk management, and compliance (often referenced collectively as GRC) should be integral components of the federated information sharing system. While distinct, these three GRC components are inter-related and their integration within the federated architecture, system design and operational reporting systems will enhance oversight and visibility into the overall system and its trust posture.

Governance tools are critical because they will support the development and management of organizational policy requirements and the chain of control needed to ensure oversight, management awareness and remedial action. In the federated information sharing environment, high-level management and oversight are critical to ensure privacy and public trust in the system. Similarly, risk management and compliance controls and supporting technologies are critical for meaningful, ongoing oversight of the system once operational. Technical and personnel components are in constant flux, and reside in an ever-changing threat landscape. Risk management and compliance technical tools will enable appropriate insight into risks, risk management adjustments, and compliance reporting.

It will be important for the DHS Privacy Office to have formal points to engage in the further development of this system. Further, there will be value in this Committee re-engaging at the point when formal requirements with traceability to specific governing policies and regulations are specified, and when major system development or acquisition decisions will be made. The Committee can also provide additional value at the points where major policy decisions will be made, and when the audit tools are being developed.

G. Audits of System Usage

Logs should be kept in a data warehouse that can be queried and used to generate reports. This will also allow searching for patterns through data mining to shed light on the who, what, when of a data access event, and also potentially how, when and with what other data it is being integrated. Automated tools are available and should be used to carry out these pattern searches and reporting on anomalies should be done on a close-to-near-real-time basis. Using automated tools will allow for fewer people to view the personal data, and may thereby be privacy enhancing. In addition, these automated tools should be designed to automatically detect privacy issues (deviations from obligations) by testing queries access of data in the source systems against defined rule sets.

These tools need to account for issues that may arise from classified queries. If the individuals who are performing audits of the source systems do not have authorization to access the classified query, then DHS needs to provide appropriate protection of this classified information while also preserving the integrity of the source system audit.

H. Data Retention

The data retention policies for the information sharing system should be predicated on the following two principles. 1) The actual queries (not the data retrieved therefrom) should be saved for the longest regulatory period, so that audit logs can be effective in understanding what people are querying and why; 2) The data inferred from those queries should be saved for the shortest regulatory period possible (essentially consistent with the reason why that query/data was assembled in the first place).

I. Data Security

Assuming that a DHS information sharing system will support long-term storage of aggregated data, this system creates an attractive centralized target for malicious actors. The appropriate level of baseline controls are those specified in NIST SP 800-53 for high-impact systems where the baseline is set to protect against threats from highly skilled, motivated, and well-resourced threat agents. Due to the aggregation of clearly sensitive data from multiple sources, the security controls implemented within this system must be high, and the program managers will need to work closely with the DHS information security staff, while continuing to seek input from the DHS Privacy Office to ensure that the selected security controls are appropriate. The Privacy Office should work with the appropriate DHS security offices to develop continuous monitoring policies. They should also review business process requirements and establish reporting instruments with respect to an effective continuous monitoring regime.