PRIVACY IMPACT ASSESSMENT (PIA) 3-YEAR REVIEW COVER PAGE

Component: *Science and Technology Directorate*

Name of Program/System: *Protected Repository for the Defense of Infrastructure Against Cyber Threats (PREDICT) DHS/S&T/PIA-006*

This system has undergone a PIA 3-Year Review on: *October 24, 2011*

The DHS Privacy Office works with DHS components to ensure that PIA reviews are conducted every three years.

DHS requires each component PIA to be reviewed in conjunction with the expiration of the accompanying PTA, in an effort to determine whether significant changes have been made to the system. This review ensures that each system continues to accurately relate to its stated mission.

Specifically, the PIA 3-Year Review Adjudication addresses each of the main areas of the PIA relating to: Legal Authorities; Characterization of the Information; Uses of the Information; Notice; Data Retention; Information Sharing; Redress; and Auditing and Accountability.

Since the publication of this PIA, the DHS Privacy Office has published the Portals PIA to cover the collection of registration information for portal system users, as well as address the privacy risks and mitigations. The registration/application information originally described in the PREDICT PIA is now covered under the Portals PIA.

S&T has conducted a separate privacy analysis on the actual dataset categories that are shared through the PREDICT program. Even though DHS will not receive or have access to any data, they still may raise privacy concerns. In this analysis, S&T examines the potential privacy risks associated with the sharing of such datasets and the privacy mitigations to address these risks. (Link to the privacy analysis)

The information technology certification and accreditation (C&A) approval has been extended to March 2, 2014.

# Privacy Analysis for PREDICT Databases

## 3-YEAR REVIEW

### DHS/S&T/PIA-006

**Privacy Risks**: Inaccurate or incorrect data is provided to the researchers, which may wrongfully implicate individuals of cyber crimes or other wrongdoing.

**Mitigation**: The datasets will only be used by the researchers for research purposes. Inaccurate or erroneous data may affect research results; however, it will have no impact on or implicate any individuals.

The Predict Coordinating Center (PCC) and DHS never gain access to the actual datasets. The datasets go directly from the data provider (through a third party data host) to the researchers.

The PREDICT program has established a detailed process for analyzing both data providers and the data they are offering to PREDICT. Each provider must undergo review for each dataset category that it is offered through PREDICT. During this review, the PCC conducts its own privacy review of the datasets. The review includes a comprehensive legal evaluation of how the data was collected by the provider, who collected it, and whether it may be disclosed to or used by third parties, such as researchers. This process closely analyzes issues against relevant privacy laws, including the Privacy Act of 1974 and the Electronic Communication Privacy Act, and the provider's privacy policies. Data providers must sign a Memorandum of Agreement (MOA) with the PCC prior to providing any data to researchers.

The quality of the PREDICT datasets will be maintained through authentication and authorization controls, regular backups, PREDICT policies and procedures (including MOAs regarding the use and handling of datasets) and audits of PREDICT operations regarding compliance with policies, procedures, and operational terms and conditions.

**Privacy Risks**: The data shared through PREDICT are not being use in ways consistent with their original collection.

**Mitigation**: The data providers will take the appropriate steps to sanitize, de-identify, anonymize, and otherwise clean the data of personally identifiable information (PII) prior to sharing it with researchers, unless consent for disclosure of any PII has been given or the data is already publicly available.

Additionally, thorough reviews are conducted by the PCC to ensure that the sharing of data aligns with the providers' privacy policies and relevant privacy laws including the Privacy Act of 1974 and the Electronic Communication Privacy Act.

**Privacy Risks**: Individuals are not provided with notice that the datasets may be shared with researchers.

**Mitigation**: The data providers own, collect, and share the data. The data providers establish and determine user notification policies. DHS does not obtain, use, collect, or share the data.

Prior to data providers sharing the datasets with researchers, the PCC conducts a thorough review of the datasets which includes examining how the data was collected by the data provider, who collected it, and whether or not it may be disclosed to third parties.

**Privacy Risks**: Data providers will disclose datasets to unauthorized users, or to individuals who intend to use the data for unauthorized purposes.

**Mitigation**: All researchers must complete an official PCC application to use any datasets. The PCC application serves as a MOA between the researcher and the PCC in which the researcher states their intended use of the datasets and provides information about their research team.

The data providers review and approve all researchers' applications and all uses of the datasets provided through PREDICT. Once the researchers are approved, the data host is responsible for the release of the data to the researcher, and is solely responsible for ensuring that any data it releases complies with its contractual agreement with the PCC.

Data providers also provide their terms and conditions for access to and use of the data, which researchers must agree to prior to accessing the datasets. These terms and conditions can include specific restrictions or permitted uses, minimum safeguards to protect the data, procedures for archiving the data, and restrictions on publishing or relating information about the data.

**Privacy Risks**: Data providers may supply too much information, including PII, to the researchers through PREDICT.

**Mitigation**: Prior to supplying data, data providers must enter into a MOA with the PCC. In the MOA, the data providers agree to sanitize, de-identify, anonymize, and otherwise clean any and all information in the data shared through PREDICT, as required. Data providers must also ensure that information provided is compliant or consistent with any policies, procedures, and understandings applicable to the data, whether explicit or implicit. Periodic audits are performed by data providers and third party data hosts to ensure each other's compliance with the PCC's requirements (such as anonymization) for the data.

Additionally, data providers shall not supply any data other than that which falls within the approved dataset categories.

Some of the datasets may include information like IP addresses. It is sometimes possible to trace IP addresses back to an organization, but usually difficult to track them to the user. The researchers, however, agree prior to receiving data, to make no attempts to re-identify any individual using the research data, unless specifically allowed by the provider.

The datasets may contain email content, but it is only content of unsolicited bulk email messages (or spam). This content is bulk-generated and sent to a large, targeted pool of IP addresses. It does not contain personal content or personal email messages.

**Privacy Risks**: Researchers may be able to re-identify individuals through data, and implicate individuals in potential cyber crimes.

**Mitigation**: Unless re-identification is approved by the data provider[1] in a MOA or terms of use,

[1] An example of when re-identification may be attempted / allowed is when conducting research on the effectiveness

the researcher must agree that he/she will not attempt to unlock, override, reverse engineer, or take steps to defeat any anonymization methods or tools that have been applied to the data. Any attempts will violate terms of use of the data.

Additionally, researchers will only use the data for researcher purposes, in accordance with the terms of the MOA and procedures in the researcher's original application for the data.

**Privacy Risks**: Researchers may share or disclose datasets or information within datasets with unauthorized users.

**Mitigation**: Pursuant to the Researchers' MOA with the PCC, Researchers are not allowed to disclose the datasets to any persons other than those identified in their original application as members of the research team. Upon termination of the MOA, researchers must dispose of all datasets and provide a certificate of disposal to the PCC. In some instances, the data provider will allow archival of the data to enable research results to be reproduced.

Researchers must immediately report to the PCC any unauthorized use or disclosure of the data and take all reasonable steps to mitigate the effects of such incidents.

**Privacy Risks**: Unauthorized users will gain access to the datasets provided through PREDICT.

**Mitigation**: Data will be provided to the researchers through secure means. Data security and access control requirements will be in place to prevent unauthorized access to the data, and ensure that only approved researchers can access the datasets.

Researchers will employ similar security and access control measures once they receive the datasets. Researchers will establish and maintain the appropriate administrative, technical, and physical safeguards to protect the confidentiality of the data and to prevent unauthorized use of and access to the data.

of anonymization methods.

Privacy Impact Assessment
for the

# Protected Repository for the Defense of Infrastructure Against Cyber Threats (PREDICT)

## DHS/S&T/PIA-006

February 25, 2008

Contact Point
Douglas Maughan
Cyber Security R&D
Science & Technology Directorate
(202) 254-6145

Reviewing Official
Hugo Teufel III
Chief Privacy Officer
Department of Homeland Security
(703) 235-0780

# Abstract

The Science & Technology Directorate's Protected Repository for the Defense of Infrastructure Against Cyber Threats (PREDICT) system is a repository of test datasets of Internet traffic data that is made available to approved researchers and managed by an outside contractor serving as the PREDICT Coordination Center (PCC). The goal of PREDICT is to create a national research and development (R&D) resource to bridge the gap between (a) the producers of security-relevant network operations data and (b) technology developers and evaluators who can use this data to accelerate the design, production, and evaluation of next-generation cyber security solutions, including commercial products. A key motivation of PREDICT is to make these data sources more widely available to technology developers and evaluators, who are currently forced to base the efficacy of their technical solutions on old, irrelevant traffic data, anecdotal evidence, or small-scale test experiments, rather than on more comprehensive, real-world data analysis.

# Overview

PREDICT is sponsored by the Department of Homeland Security (DHS) Science & Technology Directorate's Cyber Security R&D Group (S&T/CCI). PREDICT represents an important three-way partnership between government, critical information infrastructure providers, and the security development community (both academic and commercial), all of whom seek technical solutions to protect the public and private information infrastructure.

The purpose of PREDICT is to help accelerate the advancement of network-based cyber defense research, product development, and evaluation. Specifically, PREDICT provides researchers, developers, and evaluators with regularly updated network operations data sources relevant to cyber defense technology development. PREDICT datasets will provide researchers and developers timely and detailed insight into cyber attack phenomena occurring across the Internet.

PREDICT will facilitate the selection, indexing, and hosting of sets of data on Internet traffic from approved network operators and make those datasets available to approved researchers for studies on cyber security. Internet traffic data used in PREDICT does not include any content (e.g.: the text of email messages or webpages). Traffic data consists of datasets containing, for example, topology measurement data, blackhole address space data, packet traces and headers, firewall logs, border gateway protocol update messages and routing table dumps, and VOIP measurement data. DHS selected the data providers and approved the data that they would provide to the PREDICT project. DHS also contracted with data providers or other entities to store and host the datasets and provide access to them to approved researchers. DHS has contracted with a not-for-profit organization to serve as the PCC. The PCC will maintain the list of approved researchers, data providers, and data hosts, as well as, an index of the sets of data that are available. The PCC will not maintain any of the data sets, but rather will act as a library for researchers to review a catalog of metadata and determine which datasets will be of use to their particular research projects.

The concept of dataset repositories, created by research communities for the purpose of advancing the depth and quality of analysis techniques or to better understand the characteristics of operational systems, is not unique. Similar repositories have been developed in other types of technical fields for R&D activities. By creating a community of data providers and an operational structure for disseminating critical data sources to those seeking to field the next generation of cyber defense technologies, PREDICT has the potential to dramatically accelerate the R&D community's ability to develop effective and timely cyber defense technologies. It will also enable researchers to more closely monitor the emerging trends and patterns of attacks that propagate across the Internet.

The entire PREDICT operational structure is handled by contractors and a volunteer community of data providers. These datasets will be made available to approved researchers through the PCC, in accordance with established policies and procedures, for purposes of conducting cyber security research and development. Researchers will be bound by legal agreement regarding the use and security of the datasets. Researchers must establish and maintain the appropriate administrative, technical, and physical safeguards to protect access to the data to only authorized and approved persons.

In the operation of PREDICT, the PCC will establish and maintain administrative records on researchers, sponsoring institutions, data providers, data hosts, an advisory board, an application review board (ARB), and a publication review board (PRB). These records will contain personally identifiable information (PII) and will be for the internal administration of PREDICT. This data will not be accessed by anyone other than the approved and authorized administrators of the PCC and persons involved in PCC operations.

The administrative data collected by PCC is collected on behalf of DHS, and as such is covered by the DHS System of Records Notice 002 published December 6, 2004 69 Fed Reg 70460 DHS/ALL 002. As part of the user agreement individual users provide consent for the sharing that occurs between the data provider, the PCC, and the data hosts. As a rule, DHS does not have access to any of the administrative records or the data sets.

This PIA analyzes the privacy risks presented by (a) datasets used in PREDICT, and (b) the administrative records. It describes the steps that S&T/CCI, the contractors, and volunteer community have taken to mitigate those risks. While many of the test datasets contain no PII, some do contain Internet Protocol (IP) addresses. PREDICT minimizes privacy risks by anonymizing those elements of data in the test datasets that are not already available to the public and which may be linked to PII, notably IP addresses. The non-anonymized data does not include any direct PII, such as names, addresses, or telephone numbers. These datasets are described more fully in Section 1.0 of this document pertaining to "Characterization of the Information." The anonymization of the datasets is discussed in Section 5.0, "External Sharing and Disclosure."

In preparation for this project, S&T/CCI (1) analyzed the privacy risks associated with each type of dataset and the administrative records, (2) determined which test datasets are suitable for the PREDICT project, (3) worked closely with the PREDICT volunteer data provider community and data hosts to establish PREDICT operational rules and policies and procedures, and (4) determined privacy protections applicable for each test dataset and the administrative records. Intermediaries, such as DHS, the PCC, or advisory board do not have access to any of the datasets provided to PREDICT.

PREDICT seeks to facilitate new cyber defense solutions by making available traffic data sources that would otherwise be very difficult to obtain by cyber defense researchers and developers. The complexity of the Internet architecture and the sophistication of attacks require several different types of Internet data be available to researchers. For example, Domain Name Server (DNS) root server data, Internet topology measurement data, blackhole address space data, Border Gateway Protocol (BGP) update messages and BGP routing table dumps, packet headers, worm data, and firewall logs are some of the types of data required for effective cyber security R&D. Today, however, these datasets are not available to most of the research community.

At present, large-scale attack detection analyses are driven primarily through opportunistic data sources, such as firewall logs and intrusion detection alerts. While several studies have demonstrated the efficacy of firewall logs and intrusion detection data to detect some classes of self-replicating malicious code, these data sources also have inherent issues regarding their enormous overall size, data sensitivities, and threat coverage limitations. These data sources are used today, primarily, because they are available, not because they are the best data sources. The most widely used intrusion detection data source was created in 1998 by the Defense Advanced Research Projects Agency (DARPA). Given the age of this dataset, the value it offers current researchers is limited. Traffic data that is 10 years old cannot be used to effectively analyze today's attacks, viruses, malicious code, and traffic patterns. This necessarily limits the research community's ability to produce offensive and defensive privacy and security solutions and products.

Significant collections of diverse network operations data and statistics will allow researchers to discover alternative data sources that support the detection of large-scale attack phenomena with:

- Greater precision (i.e., less prone to perturbations that may lead to false positives);
- Greater efficiency than current data sources (more economical); and
- Lower sensitivity of available information (e.g., current IDS alerts and firewall logs have the potential to reveal the producing site's security posture).

For example, the types of issues that would be addressed by collecting a broad range of data from a diverse set of providers would allow researchers to:

- Map and predict how viruses and worms spread;
- Develop models of the velocity/acceleration of the spreading of viruses and worms;
- Predict the onset of attacks; and
- Develop models of network topologies and how they accelerate/decelerate an attack.

Research of this nature will significantly advance efforts to find technological solutions to protect privacy and defend against attacks and other forms of security breaches.

PREDICT OPERATIONAL STRUCTURE

DHS interviewed volunteer data providers who were willing to contribute data to the PREDICT project, selected datasets that were suitable, and selected the PCC. DHS also contracted with selected entities to serve as hosts for the datasets. DHS worked collaboratively with the data providers, hosts, PCC, and privacy experts to develop the PREDICT operational structure and associated policies and procedures.

The PCC serves as PREDICT's operational body. In this role, the PCC:

- Facilitates the data flow between PREDICT participants;
- Processes applications from researchers for access to datasets or approval to publish research results;
- Develops and maintains a catalog of metadata about the datasets; and

- Coordinates the development of protocols, such as anonymization of IP addresses and restrictions on data (which are subject to DHS approval), to protect the confidentiality and integrity of data and direct its proper usage.

The following nine types of organizations participate in PREDICT:

1. PREDICT Coordination Center (PCC)
2. Data Providers
3. Data Hosts
4. Researchers
5. Sponsoring Institutions
6. Application Review Board
7. Publication Review Board
8. Advisory Board
9. Department of Homeland Security

These entities work in a coordinated fashion to help ensure that the operational policies and procedures are followed, privacy protections are effective, researchers are legitimate and sponsored by worthy institutions, and the datasets are handled in a secure and responsible manner.

The ***PCC*** receives and catalogs metadata about the test datasets and makes the metadata catalog available to approved researchers, subject to the terms and conditions set forth by DHS, PCC, and the data providers and hosts in Memorandum of Agreements (MOAs) between the various parties. DHS and the PCC do not store, maintain, or have access to any of the datasets; only the data hosts and approved researchers have access to specific datasets. Upon reviewing the catalog of metadata, and submitting an application to use specific datasets, approved researchers will obtain access information and will be able to access the datasets directly from the data hosts.

***Data providers*** include the operators of the Internet backbone, large and small Internet service providers (ISPs), and large enterprise network operators. They provide datasets of Internet traffic data that they own, or have permission to release, to PREDICT, subject to terms and conditions that they stipulate and in accordance with privacy requirements, which are set forth in the MOA between the data provider and the PCC. These terms and conditions are included in the MOA between the researcher and the PCC. Data providers may also require researchers to sign a separate MOA with them regarding the use of their data. The combination of data provider security requirements and PCC privacy requirements that are set forth in MOAs between the researcher, the PCC, and the data provider, help ensure that any potentially personally identifiable information in the datasets is anonymized[2] and that usage of the data is appropriate. Data providers may select a data host to receive and host their datasets, or they may host their own data.

The MOA between the data provider and the PCC requires the provider to agree that:

- The data provider will provide the PCC with metadata on the data they agree to make available

[2] IP addresses that are not anonymized in datasets that are already available to the public through other sources are not required to be anonymized for PREDICT.

to data hosts for release to researchers approved by the PCC.

- The data provider will make the data available to data hosts, for release to approved researchers and no others, under the terms and conditions for access and use as specified by them and the PCC.

- The data provider may provide terms and conditions for access to and use of the data, including identification requirements for the researcher; permitted uses and specific restrictions; minimum safeguards to protect the data; procedures for receipt, handling, control, dissemination, and return of data; and restrictions on publishing or releasing information about the data (which is addressed below under Publication Review Board). The data provider's terms and conditions for access and use of their datasets will flow down to the researcher and become part of the researcher's application and Memorandum of Agreement to use the dataset.

- The data provider is responsible for the release of the data to PREDICT, and is solely responsible for ensuring that any data it releases complies with all applicable statutes and regulations of governing or regulating bodies and contractual agreements. The data provider must also ensure that the release of the data is consistent with its privacy, security, or other policies and procedures applicable to the data. The data provider must certify that the data provided for use in the PREDICT program is in compliance with the foregoing and that the data has been sanitized, de-identified, or cleaned of any and all information that would not be in compliance or consistent with its policies and procedures and privacy requirements. The data providers have been asked to ensure that their privacy policies are available to the researchers.

- Non compliance with these requirements may result in the data provider's expulsion from the PREDICT project, including removal of the provider's datasets from the PREDICT project.

***Data hosts*** provide computing infrastructure to store datasets received from the data providers and provide methods by which researchers can access datasets after the researcher's application for particular dataset(s) has been approved.

Data hosts must enter into a MOA with PCC in which they agree to specific terms regarding access to the datasets, including that:

- The data hosts will accept data from approved data providers, for release to approved researchers, subject to the terms and conditions set forth by the providers and hosts.

- The data hosts will provide terms and conditions for access to, transfer, storage, and use of the data as required by the data provider and PCC, as well as any other restrictions the host deems necessary to accomplish efficient and secure access to the data.

- The data hosts acknowledge that the data access approval given to a researcher in any application will permit access to the requested data by that researcher, regardless of approval or denial of access to that researcher in any other previous PREDICT dataset application.

- The data hosts are solely responsible for ensuring that any data they release complies with the host's separate agreement with the data provider.

***Researchers*** are members of the cyber defense R&D community who complete an official PCC application, which serves as a MOA between the researcher and the PCC, requesting datasets from PREDICT for use in their research and who are affiliated with a sponsoring institution.

Researchers must agree that:

- They will not use the data for purposes other than described in their application.
- The researchers will not disclose the data to any persons other than those identified in their application.
- The researchers will establish and maintain the appropriate administrative, technical, and physical safeguards to protect the confidentiality of the data and to prevent unauthorized use of or access to the data.
- The researchers will permit others to use the data only in accordance with the terms of the MOA and the procedures in the researcher's application.
- If the researcher moves to a different institution, he/she must notify PCC and the sponsoring institution in writing regarding the disposition of all copies of the data and follow PCC's directions and the sponsoring institution's guidelines. A researcher may apply for a new PREDICT account through his/her new institution and, once approved, submit a new application for continued usage of the datasets.
- No findings, analysis, or information derived from the data may be released if such findings contain any combination of data elements that might allow for identification or the deduction of a person's or institution's identity. The Publication Review Board and associated review of documents submitted for publication is intended to help ensure compliance with privacy requirements.
- Any findings, results of analysis, or manuscripts proposed for public release, publication, or any other type of disclosure to persons not listed and approved in this application (e.g., abstracts, presentations (oral or written), publications) must be submitted for review by a Publication Review Board managed by PCC prior to publication or public release to assure that data confidentiality is maintained, entities or individuals cannot be identified, and the terms and conditions attached to the use of the data have been followed.
- The researcher must immediately report to the PCC any use or disclosure of the Data other than as permitted and will take all reasonable steps to mitigate the effects of such improper use or disclosure, cooperating with all reasonable requests of PCC towards that end.
- In the event PCC determines or has a reasonable belief that researcher has violated any terms of the MOA, PCC may terminate the MOA and require the researcher to return the data and all derivative files. PCC may also seek injunctive relief against the researcher or the sponsoring institution to prevent any unauthorized disclosure of data. In addition, PCC will report any misuse or improper disclosure of the data to the data provider and host and to appropriate authorities as required by applicable Federal or state law.
- The researcher will destroy all copies of the data when the MOA expires or take such action as specified in the MOA and will certify such destruction or return by signing and providing to PCC a Certification of Data Return or Destruction.

***Sponsoring Institutions*** are organizations that are affiliated with or otherwise sponsor researchers and validate their research and need for PREDICT data. Sponsoring Institutions could be corporations, academic institutions, national laboratories, or government research laboratories.

The ***Application Review Board*** (ARB), in conjunction with the PCC and the data provider, reviews and approves or rejects the application to use PREDICT datasets and forwards approved applications to data hosts to enable the researcher to have access to and use the requested data. The Application Review Board is chaired by the PCC and includes representation from data providers, data hosts, and a subject matter expert in cyber security.

The ***Publication Review Board*** (PRB) reviews and approves or rejects applications from researchers or sponsoring institutions to publish or otherwise release any information, study results, or other summaries or descriptions on, about, or relating to data or metadata previously made available through the PREDICT project. Publication constraints are defined by the data provider during the process when the data is initially inserted into the PREDICT system. The Publication Review Board is comprised of PCC personnel, the original data provider, and the hosting site for the provider data (if the host is not the provider).

The ***Advisory Board*** advises and makes recommendations to the PCC on policy and issues relating to privacy issues and the general direction of the PREDICT project.

The ***Department of Homeland Security*** (DHS) selects the data providers and approves the data that they provide to the PREDICT project. DHS also contracts with data providers or other entities to store and host the datasets and provide access to them to approved researchers. DHS has contracted with a not-for-profit organization to serve as the PCC. DHS reviews the operational procedures of PREDICT and the security of the PCC computer system through its certification and accreditation process.

This PIA addresses both the administrative data collected by the PCC and the test datasets that are available to approved researchers for cyber security R&D. The administrative data is a system of records and requires a SORN and PIA. The administrative data, however, is not used by DHS, it is not accessed by DHS, and it does not support any DHS operations. It is not shared except as outlined in this PIA. The approved test datasets that are accepted into the PREDICT project and provided to approved researchers do not contain any data subject to the Privacy Act of 1974 and are not subject to a SORN. As noted above, they are included in this PIA as a voluntary measure by DHS to explain the privacy analysis that has been conducted with respect to the datasets.

# Section 1.0 Characterization of the Information

The following questions are intended to define the scope of the information requested and/or collected as well as reasons for its collection as part of the program, system, rule, or technology being developed.

## 1.1 What information is collected, used, disseminated, or maintained in the system?

### Contact and Information Related to PREDICT Datasets

The PCC collects administrative information from researchers, data providers, data hosts, sponsoring institutions, advisory board members, ARB members, and PRB members. The PCC maintains a catalog of metadata about the test datasets which is available to approved researchers, and it keeps an operational record of activities conducted within the PREDICT project. These operational records will be for the internal administration of the PREDICT project and this data will not be accessed by anyone other than the approved and authorized administrators of the PCC.

### 1.1.1 Administrative Information Collected from Researchers, Data Providers, and Data Hosts

The PCC collects the following information from researchers, data providers, and data hosts:

- First and last name
- Address (street, city, state, zip code)
- Telephone numbers (home, work, cell, and fax – only the work number is a required field)
- Email address
- Name of Sponsoring Institution
- Address of Sponsoring Institution (street, city, state, zip code)
- Name of authorized representative from Sponsoring Institution
- Telephone number of authorized representative
- Email address of authorized representative
- Approved researchers are assigned a username and password to access the PREDICT portal. The PCC also keeps a list of the datasets requested by Researcher and whether the request was approved or disapproved, including the reason for disapproval.
- The PCC also keeps record of requests from approved researchers to publish documents, articles, or materials pertaining to their research and whether the request was approved by the Publication Review Board or disapproved, including the reason for disapproval.
- The PCC keeps a record of the approved datasets that are provided from each data provider and which datasets the data hosts are hosting.

### 1.1.2 Application Review Board and Publication Review Board Members

The PCC collects the following information from ARB and PRB members:

- First and last name
- Address (street, city, state, zip code)
- Telephone number
- Email address
- Name of organization

### 1.1.3 Advisory Board Members

The PCC collects the following information from Advisory Board members:

- First and last name
- Address (street, city, state, zip code)
- Telephone number

- Email address
- Name of  organization

### 1.1.4 Sponsoring Institutions

The PCC collects the following information from Sponsoring Institutions:

- Name of  Organization
- First and last name of representative of Sponsoring Institution
- Title
- Address (street, city, state, zip code)
- Telephone number
- Email address

### 1.1.5 Dataset Information

The datasets used in PREDICT contain real traffic data provided by data providers in accordance with the terms of the MOA between the PCC and the data provider (see above section discussing Data Provider MOA obligations).  The data providers are not, however, collecting this information specifically for PREDICT. The data providers are already collecting the information to support their own cyber security responsibilities and are now choosing to share that same information with PREDICT. DHS will not have access to any of the administrative or traffic data provided for PREDICT.

TECHNICAL ASPECTS OF THE DATA

Content sent over the Internet is broken into packets.  The packets contain a header with the sender's and recipient's Internet Protocol (IP) addresses.  Generally, an IP address is randomly assigned to a computer at the time the user logs onto the Internet.  However, it is sometimes possible to trace IP addresses, even if they are not static, back to individual users.  Therefore, all IP addresses in PREDICT Phase I datasets that are not already available to the public will be anonymized by the data provider according to requirements in the MOA between the data provider and the PCC and any other data that could be used to trace the data back to an individual will be eliminated from the dataset.  The PIA will be modified to include information on different types of datasets that may be added to PREDICT in future phases.

The PCC, or another contractor, will conduct periodic compliance checks of the PREDICT project to ensure compliance with these privacy protections and other PREDICT operating procedures and contractual provisions.  The following types of datasets are being used in Phase I of the PREDICT project.  The described uses of these datasets are intended as examples.  As technology advances, traffic changes, or the innovative process advances, researchers may use the datasets differently than described below, however all uses of this data by researchers will always remain within the scope of the governing MOA.

Topology Measurement Data

This data is obtained from computers that the data provider puts on the network in order to map the network connections of the Internet connecting out from that point.  The computers send out probe packets with Time to Live (TTL).[3] The packet is owned by the data provider. It is sent out and comes back with information about routing, but no data is transmitted in the process. The data provider makes an Anonymous System (AS) core and ISP level map of Internet connectivity, based on the measurement of the TTL. The IP addresses in this dataset are not anonymized because this dataset contains measurement traffic[4] and not actual sender-receiver communication. The data provider requests that researchers not probe certain Internet Protocol (IP) addresses or disclose IP addresses to anyone else.

### Blackhole Address Space Data

The data provider owns a large number of IP addresses. Traffic to legitimate addresses owned by the provider is delivered, and the remainder goes back to the data provider because the traffic is targeting unassigned IP addresses. Since this traffic was trying to saturate a target population, it is generally hitting illegitimate IP addresses and is malicious traffic, such as scanners and worms. This traffic data can be useful for studying backscatter from distributed denial of service (DDOS) attacks, worm spread (growth rates, population size, and affected population), scanning and backdoor activity, and evaluating various honeypot responders. In addition, all IP addresses will be anonymized.

### Hi-Speed ISP Exchange Data from OC-48 operational network (packet traces)

This dataset is comprised of headers of traffic data that contain the source and destination IP addresses, but contains no content data. Pursuant to the terms of the MOA between the data provider and the PCC, the data provider must anonymize the IP addresses in headers. This dataset can be used for DDOS attack detection and characterization, malicious behavior detection (port scans and intrusion attempts), worm outbreak characterization, backscatter analysis, botnet distribution and characterization, analysis of impact of attack on normal traffic, spectral analysis of attacks and background traffic, and intrusion detection and firewall research and development.

### Full Packet Headers

Full packet headers contain IP addresses for the sender and recipient. The port numbers identify the application used (e.g., Internet browser or email). No packet content is included in this dataset, although a hash of the data associated with each packet header may be provided. Pursuant to the MOA between the data provider and the PCC, all IP addresses will be anonymized.

### Domain Name Server (DNS) Root Server Data

This is root server data from the host of major DNS root servers. The data will identify the user by IP address. It will contain the IP address for the requested site, but it will not indicate whether the person associated with the IP address actually connected to that site. Generally, requests are aggregated by multiple users, but some are not. All IP addresses will be anonymized. This data will be used for DNS root server traffic analysis and characterization and DNS root server attack analysis and characterization.

[3] The time to live (TTL) value can be thought of as an upper bound on the time that an IP datagram can exist in an internet system. The TTL field is set by the sender of the datagram, and reduced by every host on the route to its destination. The purpose of the TTL field is to avoid a situation in which an undeliverable datagram keeps circulating on an internet system, and such a system eventually becoming swamped by such immortal datagrams.

[4] Measurement traffic contains IP addresses of the machines that a packet touches as it travels along the network before it is sent back to the sender, allowing measurement of Internet topology and routing.

## Internet Topology Data

Internet topology data is created by a program that tries to map the Internet. The program is able to determine which routers are capable of talking to other routers. Internet topology data only shows router connectivity within the Internet core and to external enterprise borders; it does not contain any identifiable information or internal enterprise topology information. This dataset can be used for worm outbreak modeling and simulation, worm containment and countermeasures, zombie distribution for DDOS attacks, vulnerability assessments, longitudinal studies of the evolution of Internet topology and address distribution, Internet topology and address map inference.

## Address Space Allocation Data

This data is used to determine which parts of the IP address space are used. This information will not identify persons or organizations. This dataset can be used for worm outbreak modeling and simulation, worm containment and countermeasures, zombie distribution for DDOS attacks, vulnerability assessments, longitudinal studies of the evolution of Internet topology and address distribution, Internet topology and address map inference.

## Enterprise Data

This data is internal traffic data from a large enterprise. It will consist of only headers with anonymized IP addresses. No content will be included. Nevertheless, the operating procedures of PREDICT and the PCC involves internal control points to ensure that the data provided has undergone an internal review by the provider, such as reviews by its legal counsel, regarding whether anonymization or notification to users is required. The principle security-oriented use of the enterprise datasets will be to serve as a representative example of real network traffic. By providing a large amount of real network traffic (versus developed test datasets), PREDICT's goal is to provide a resource for researchers to use in assessing the false positive rates and/or collateral damage of deploying proposed detection algorithms.

Prior to this effort, there have been no such enterprise background traces available, to the significant detriment of researchers attempting to devise enterprise-level network security mechanisms that will actually work soundly in practice. Such data is crucial in assessing security counter-measures since networks are massively heterogeneous and are filled with large amounts of benign traffic that is not malicious, but which could confuse attack detectors and falsely trigger defenses. More generally, there have been virtually no studies of what enterprise traffic looks like for nearly fifteen years, and, thus, these traces will provide a key first look at a hitherto unknown landscape. Since security breaches can be from internal or external sources, a clearer understanding of internal traffic data will help advance the development of cyber security technologies to counter insider attacks.

## BGP Routing Table Data

This dataset captures "snapshots" of the topological state of the Internet by archiving Border Gateway Protocol (BGP) routing tables from Internet routers in many locations around the world (these are called Internet Exchange Points). Each routing table expresses the "view" of the Internet from that router's point in the overall topology and, taken together, all of these views provide a relatively complete roadmap of the connectivity within the Internet Service Provider core of the Internet. This dataset contains only backbone topology information; it does not contain any packet header information or information which relates to individuals.

BGP Routing Table Data is used by researchers who study the overall growth patterns of the Internet over time, as well as those who are looking specifically at individual carriers, regions, or resources. It shows historical trends in the utilization of the two principal Internet resources, IP addresses and Autonomous System Numbers (ASN), and this presents the basic backdrop against which many other trends are tracked. Several organizations provide complementary, partly-overlapping, and slightly methodologically different partial views of a large and diverse environment.

### BGP Update Messages & BGP Routing Table Dumps

The BGP Routing protocol is used to exchange routing information between different autonomous systems in the Internet. Each BGP router sends information describing its own part of the Internet in the form of BGP Update messages, which are sent to every BGP device. ISPs pay each other to carry packets. Peering points are routing exchange points. Routing data is collected from several places on Internet. Routing sensors are responsible for providing two types of information: an accurate picture of the local routing topology and alerts on instability as viewed by the local network, indicating a potential resource-impacting event. The routing sensors receive this info by peering with critical routers near blackhole sensors. This data will contain raw BGP updates and routing table dumps (which are the same as the BGP Routing Table Data described above). When various networks are added (or if networks lose connectivity), the BGP protocol is responsible for propagating this information throughout the Internet. When a network taken down, it watches to see how it heals itself. This information is already being made publicly available by some institutions, such as http://www.routeviews.org. This dataset can be used to develop tools and techniques such as beacons, fault injection, other measurement methodologies for reachability and stability, and new routing protocols. It can also be used in the development of applications that will help determine the stability of the topology (convergence, flapping), route hijacking, reachability, and interconnectedness.

### VOIP Statistical Data (End-to-End Quality)

There are 2200 ISPs registered with data provider that collect data about the quality of VOIP connections. This dataset contains only statistical information about VOIP calls. It does not contain phone numbers or content of calls. It consists only of statistical information and a researcher would not be able to identify who was talking, with whom, or about what subject. The dataset is comprised of end-to-end data which characterizes the *quality* of the paths which Voice Over Internet Protocol (VOIP) telephone calls take across the global Internet and contains anonymized Session Initiation Protocol (SIP) teardown messages collected from both ends of the conversation. There are three primary pieces of information:

1. Prefix representing logical grouping
2. Autonomous system number of the ISPs at both ends of the connection.
3. Country code of Autonomous System Numbers (ASN).

All three pieces of information are currently publicly available.

This data is very useful to study the quality and security of VOIP calls. It is anticipated that the VOIP statistical data will be used by researchers who wish to compare differential quality of service in similar and dissimilar regions of the Internet, such as across different backbone carriers which utilize different technology or capacity-planning methodologies. These comparisons of quality of service including statistical analysis of the VOIP data (e.g., call duration, call frequency, termination endpoints) could indicate potential security weaknesses within the VOIP network. The data can also be analyzed to determine the availability of the VOIP network in the event of an attack to see if emergency communications can be accomplished over the Internet using the VOIP network.

### Firewall Logs

Firewalls detect distributed denial of service (DDOS) attacks and other malicious activity. Firewall logs contain detailed information regarding the end point that is directing harmful activity towards the network they are protecting. They contain the number of packets, origin, and where it went. All IP addresses will be anonymized.

### Traffic Flows via Netflow

This data consists of statistics regarding data collected from routers. It identifies two end points. Raw NetFlow data will indicate not only traffic totals, but the application breakdowns at each peering point. The individual flows will be stored in a method compatible with free analysis tools. All IP addresses will be anonymized. This dataset can be used to develop tools and techniques that will lead to a better understanding of the tradeoffs in netflow versus other techniques and in improving netflow and new netflow versions. It can also be used in the development of anomaly detection and intrusion detection applications and for traffic characterization applications, such as self similarity and bottleneck bandwidth estimation applications.

### Network Management Data

Network management data is generally represented in the form of messages that are transmitted using Simple Network Management Protocol (SNMP). The messages convey gross aggregated statistical information from various network devices. It is possible to study Internet behavior by analyzing these messages. SNMP messages are also used to transmit information conveying the physical well-being of these devices. They often reveal important causal information that is subsequently observed in other types of data logs. They do not pose an individual privacy risk, but can reveal information regarding the operation of the network. Anonymization will protect critical network operator information. The data provider currently makes this data publicly available.

### Intrusion Detection System (IDS) Logs

An intrusion detection system scans traffic to detect unauthorized or malicious activity. When it detects an attack, it can trigger protective actions. It is essentially a sensor that is watching for malicious activity. It is important to study IDS traffic in order to understand the evolution, rise, and decay of such traffic. It is possible to identify the end point responsible for originating the suspicious activity.

### Anonymized Internet Witty Worm Data

The data provider owns a large number of IP addresses. Due to its size, this network receives a large volume of unsolicited traffic trying to spread random-scan Internet worms. Packet header traces and tables summarizing infected host activity contain detailed information regarding the end point that is directing harmful activity towards the network. They contain the truncated packets (copies of the worm, no other data) and summarized information on the number of packets, country-level origin, and type of bandwidth. All IP addresses will be anonymized.

### Code-Red Worm Data

The data provider owns a large number of IP addresses. Due to its size, this network receives a large volume of unsolicited traffic trying to spread random-scan Internet worms. Tables summarizing infected host activity contain detailed information regarding the end point that is directing harmful activity towards the network. They contain the number of packets, country-level origin, and type of bandwidth. All IP addresses will be anonymized.

## 1.2 What are the sources of the information in the system?

- Researchers
- Data providers
- Data hosts
- ARB/PRB members
- Advisory Board members
- Sponsoring Institutions
- Academic institutions and private sector entities provide traffic data. No PREDICT Phase I data is provided by communications providers to the public.

## 1.3 Why is the information being collected, used, disseminated, or maintained?

### 1.3.1 Researcher Information

The information collected about researchers is essential to maintaining operational records indicating who has used specific PREDICT datasets, for what purpose, when, and what organization sponsored them. Additional collected information pertains to whether the researcher's application was approved or disapproved and whether any publication requests were approved or disapproved. Reasons for disapproval of requests for access to certain datasets or publication are kept by the PCC. This information is very useful to the Application Review Board and Publication Review Board and helps them ensure fairness and even treatment. PREDICT operational procedures ensure that the denial of access or publication requests is not used against a researcher when making subsequent requests.

### 1.3.2 Data Provider Information

The information regarding data providers is important to the operations of the PCC. The data providers must be notified when a researcher is requesting their data or requests permission to publish regarding the research conducted using the datasets. Additionally, the data providers are active members of the Application Review Board and Publication Review Board and set the terms and conditions for use of their data. Thus, it is important that the PCC maintain accurate information for each data provider and which datasets they are providing to the PCC.

### 1.3.3 Data Host information

The information collected and maintained about data hosts is also critical to PCC operations. The data hosts must be notified when a researcher is approved to access data sets that they are hosting. The data host must also provide any special terms or conditions that they want the researcher to comply with in accessing the data. It is important that the PCC maintain accurate information for each data host and the datasets which they are hosting for the PREDICT project to enable the PCC to operate correctly and conduct audits of compliance with PCC policies and procedures.

### 1.3.4 Application Review Board and Publication Review Board Member Information

The PCC must maintain contact information of the people who serve as Application Review Board and Publication Review Board members so they can contact them for assistance and decisions regarding applications and requests to publish.

### 1.3.5 Advisory Board Member Information

The PCC collects contact information from the people who serve on the Advisory Board so they can contact them with information regarding upcoming meetings, minutes of past meetings, questions that need to be considered, and the like.

### 1.3. Sponsoring Institution Information

The PCC collects the contact information of the person representing the institution who is sponsoring the research so it can contact the institution to verify the researcher's application, if necessary, resolve any questions that may arise, or conduct an audit of the institution's compliance with PCC policies and procedures and the terms of the MOAs.

### 1.3.7 Dataset Information

The datasets in the PREDICT project are collected in the form of test datasets to enable researchers to use them in the research and development of cyber security technologies that will better protect the security of networks and applications and the privacy of information.

## 1.4 How is the information collected?

PREDICT administrative information is collected through the PREDICT portal and via email and facsimile. The PREDICT test datasets are either retained hosted by the provider or the provider sends its approved test datasets to an approved data host.

## 1.5 How will the information be checked for accuracy?

PREDICT administrative information is provided directly by the person submitting it (e.g., researcher, data provider, data host, sponsoring institution, or advisory, ARB, or PRB members). The PCC verifies the information that has been provided prior to it being entered into the system and, for researchers, prior to assigning them an account. Communications between the PCC and these persons helps ensure the information remains accurate.

## 1.6 What specific legal authorities, arrangements, and/or agreements defined the collection of information?

PREDICT provides a Privacy Statement to users to inform them about how their personal information will be used and the purpose for which it will be used. Appendix A is attached with a copy of the Privacy Statement. The MOAs between the PCC and researchers, data providers and data hosts and corresponding provisions are discussed in the Overview section above. S&T/CCI has established contracts with the PCC for the administration of PREDICT and with the data hosts for the hosting of the test datasets.

## 1.7 Privacy Impact Analysis: Given the amount and type of data collected, discuss the privacy risks identified and how they were mitigated.

There is a risk associated with the collection of personally identifiable information. To mitigate this risk, the PCC has integrated robust security into its risk management plans and routinely tests the security of PREDICT. Further, the PCC has implemented best practices, such as auditing, to defend against misuse of the data and to monitor those with access to the information. In addition, in accordance with the DHS Sensitive Systems Handbook, PREDICT ensures effective security controls and authentication. By enforcing system policies and settings and strong passwords, the PCC protects the privacy of data to promote or permit public access to the PREDICT project and to protect the integrity of the data itself. Also, any contact information collected by the PCC is not searchable. Contact information is only secondary data and is used by PCC to validate the user requesting an account and access to PREDICT data or for administrative purposes.

PREDICT test datasets have privacy protections built in through the review process and required anonymization of certain fields. These protections are described for each dataset in 1.1.5 above regarding "Dataset Information."

### 1.7.1 Data About Researchers

The privacy of the data about researchers will be maintained through authentication and authorization controls, regular backups, PREDICT policies and procedures (including change management), and audits of PREDICT operations regarding compliance with policies and procedures and operational terms and conditions.

### 1.7.2 Data About Data Providers

The privacy of the data about data providers will be maintained through authentication and authorization controls, regular backups, PREDICT policies and procedures (including change management), and audits of PREDICT operations regarding compliance with policies and procedures and operational terms and conditions.

### 1.7.3 Data About Data Hosts

The privacy of data about data hosts will be maintained through authentication and authorization controls, regular backups, PREDICT policies and procedures (including change management), and audits of PREDICT operations regarding compliance with policies and procedures and operational terms and conditions.

### 1.7.4 Data About Application Review Board Members

The privacy of the data about Application Review Board members will be maintained through authentication and authorization controls, regular backups, PREDICT policies and procedures (including change management), and audits of PREDICT operations regarding compliance with policies and procedures and operational terms and conditions.

### 1.7.5 Data About Publication Review Board Members

The privacy of the data about Publication Review Board members will be maintained through authentication and authorization controls, regular backups, PREDICT policies and procedures (including change management), and audits of PREDICT operations regarding compliance with policies and procedures and operational terms and conditions.

### 1.7.6 Data About Advisory Board Members:

The privacy of the data about Advisory Board members will be maintained through authentication and authorization controls, regular backups, PREDICT policies and procedures (including change management), and audits of PREDICT operations regarding compliance with policies and procedures and operational terms and conditions.

### 1.7.7 Data About Sponsoring Institutions:

The privacy of the data about Sponsoring Institutions will be maintained through authentication and authorization controls, regular backups, PREDICT policies and procedures (including change management), and audits of PREDICT operations regarding compliance with policies and procedures and operational terms and conditions.

### 1.7.8 Dataset Information:

The PREDICT datasets will be secured by their respective PREDICT data hosts and will only be accessed by approved researchers who have been provided with terms and conditions for  accessing the datasets, downloading them, storing them, and destroying or returning them.  The quality of the PREDICT datasets will be maintained through authentication and authorization controls, regular backups, PREDICT policies and procedures (including MOAs regarding the use and handling of the datasets), and audits of PREDICT operations regarding compliance with policies and procedures and operational terms and conditions.

The dataset will be downloaded to the sponsoring institution site where the researcher is working. The researcher is bound contractually to use the data only as specified and within the bounds of the research described in the application.  In addition, the data provider may require an additional written agreement between the provider and researcher setting forth specific rights and restrictions regarding the use of the data.

# Section 2.0 Uses of the Information

The following questions are intended to delineate clearly the use of information and the accuracy of the data being used.

## 2.1 Describe all the uses of information.

### 2.1.1 The Intended Use of Information Collected From Researchers, Data Providers, Data Hosts:

The specific intended use of the information collected from researchers, data providers, and data hosts is described above (see 1.3.1, 1.3.2, 1.3.3).  This information will be used to manage PREDICT operations and conduct periodic audits to ensure PREDICT policies and procedures and operational terms and conditions are being followed.

### 2.1.2 The Intended Use of Information Collected From Application Review Board, Publication Review Board, and Advisory Board

### Members:

The specific intended use of the information collected from the Application Review Board, Publication Review Board, and Advisory Board members is described above (see 1.3.4 and 1.3.5). This information will be used to facilitate and manage PREDICT operations and may also be used from time to time in periodic audits to check compliance with PREDICT policies and procedures and operational terms and conditions.

### 2.1.3 The Intended Use of Information Collected from Sponsoring Institutions:

The specific intended use of the information collected from Sponsoring Institutions is described above (see 1.3.6). This information may also be used from time to time in periodic audits to check compliance with PREDICT policies and procedures and operational terms and conditions.

### 2.1.4 The Intended Use of the Datasets:

The intended use of the datasets will be specified by the researcher in their application. The application is reviewed by the Application Review Board, which consists of a PCC representative, a data provider representative, a data host representative, and an external cyber security expert. The researcher must use the data for the purposes stated in the application and in accordance with the restrictions specified by the data provider. The Application Review Board operates under a Non-Disclosure Agreement to protect the researchers' privacy and confidential/proprietary information contained in the researchers' applications.

## 2.2 What types of tools are used to analyze data and what type of data may be produced?

There are no tools, such as data mining tools, being used to analyze PREDICT information, and no data is produced.

## 2.3 If the system uses commercial or publicly available data please explain why and how it is used.

As noted in 1.1.5, "Dataset Information," some PREDICT datasets are already publicly available and some are provided by commercial entities, non-profit organizations, or academic institutions.

## 2.4 Privacy Impact Analysis: Describe any types of controls that may be in place to ensure that information is handled in accordance with the above described uses.

Information collected by the PCC is secured within PCC operations and will only be accessed by PCC personnel with a valid need-to-know for the information. The PCC is comprised of a small staff that has clear delineations in responsibility for managing the information collected from these sources. The PCC, in accordance with the DHS Sensitive Systems Handbook, ensures effective security controls and authentication mechanisms are implemented and working as intended.

#### 2.4.1 Administrative Data

As noted in 1.7 above, the PCC has implemented auditing checkpoints, tested controls, and deployed technical safeguards to protect administrative data. With respect to test datasets, PREDICT privacy requirements are embedded in the MOA between the data provider and the PCC. Likewise, terms for access, special handling and privacy/security that are set by the data providers are embedded in the MOA between the researcher and the PCC. These terms are also embedded in the MOA between the data host and the PCC.

#### 2.4.2 Dataset Information

The PREDICT project has analyzed each dataset and has directed data providers to anonymize any data that has the potential to be personally identifiable information, such as IP addresses. Anonymization of such data fields must comply with requirements set forth by the PCC and DHS. Researchers must agree to establish and maintain the appropriate administrative, technical, and physical safeguards to protect the privacy and confidentiality of the data and to prevent unauthorized use or access to the data. This specifically includes the use of locked storage facilities and strong passwords. Additional security provisions for access to the host system and the dataset may be included as terms of agreement in the researcher's application (which becomes the MOA between the researcher and the PCC). In addition, the data provider and data host may require a separate agreement with the researcher regarding security and use of the data. To ensure the long-term integrity of the datasets and efficacy of PREDICT policies and procedures, the PCC has established an advisory board that provides input on an ongoing basis regarding PREDICT operations, the security and integrity of the datasets, and privacy issues and concerns. In addition, periodic compliance checks of provider and researcher compliance will be conducted to ensure all parties are in compliance with PREDICT policies and procedures and the terms of the MOAs. As the PREDICT program expands, supplemental PIAs will be developed to accommodate new datasets and to maintain transparency of operations. The Certification and Accreditation (C&A) has been completed by the S&T CIO effective September 30, 2007.

# Section 3.0 Retention

The following questions are intended to outline how long information will be retained after the initial collection.

## 3.1 How long is information retained?

### 3.1.1 Data Collected from Researchers, Data Providers, Data Hosts, ARB and PRB Members, Advisory Board Members, and Sponsoring Institutions:

The PCC maintains all data collected from researchers, data providers, data hosts, ARB and PRB members, Advisory Board members, and sponsoring institutions and this information is retained for six years from the time they are no longer associated with the PREDICT project. This data retention is done offline as is specified in the System Security Plan (SSP).

### 3.1.2 Datasets

Upon termination of the right to use the data (when the researcher completes his/her research; when the researcher terminates his/her right to use the data; or when the PCC, data provider or data host terminates the researcher's right to use the data), the researcher must destroy all copies of the data or, where directed by PCC, return the data to the data host or destroy it per its instructions. The researcher must certify such destruction or return by signing and providing to PCC a Certification of Data Return or Destruction document.

The PREDICT data hosts maintain the datasets they host until such time that the data provider, data host, or the PCC removes them from the PREDICT project.

### 3.2 Has the retention schedule been approved by the component records officer and the National Archives and Records Administration (NARA)?

The retention schedule for the PREDICT administrative information has not been approved by the National Archives and Records Administration, but S&T is working with NARA to obtain a retention schedule.

### 3.3 Privacy Impact Analysis: Please discuss the risks associated with the length of time data is retained and how those risks are mitigated.

Retention of administrative records is per DHS policies and procedures as laid out in FISMA Certification and Accreditation.

## Section 4.0 Internal Sharing and Disclosure

The following questions are intended to define the scope of sharing within the Department of Homeland Security.

### 4.1 With which internal organization(s) is the information shared, what information is shared and for what purpose?

Neither PREDICT administrative information nor the test datasets are within the Department of Homeland Security. DHS does not have access to this data.

### 4.2 How is the information transmitted or disclosed?

Not applicable as DHS does not have access to PREDICT data.

### 4.3 Privacy Impact Analysis: Considering the extent of internal information sharing, discuss the privacy risks associated with the sharing and how they were mitigated.

Not applicable as DHS does not have access to PREDICT data.

## Section 5.0 External Sharing and Disclosure

The following questions are intended to define the content, scope, and authority for information sharing external to DHS which includes Federal, state and local government, and the private sector.

## 5.1 With which external organization(s) is the information shared, what information is shared, and for what purpose?

### 5.1.1 Information About Researchers:

The information collected about researchers will be shared with PCC personnel. Portions of it may be shared on an as-needed basis with Application Review Board and Publication Review Board members; data providers of the requested datasets; data hosts that host the requested datasets; personnel outside of PCC who may be conducting audits of PCC operations and compliance with PCC policies and procedures; and contractors working on PREDICT in a consulting capacity regarding PCC operations. In rare circumstances, the information about a particular researcher may be shared with the PREDICT project manager in the Department of Homeland Security.

### 5.1.2 Information About Data Providers:

Information collected from data providers will be shared on an as-needed basis with the PCC personnel; Application Review Board and the Publication Review Board members; data hosts who will be hosting the data provider's data; other data providers; personnel outside of PCC who may be conducting audits of PCC operations and compliance with PCC policies and procedures; contractors working on PREDICT in a consulting capacity regarding PCC operations; and the PREDICT project manager in the Department of Homeland Security.

### 5.1.3 Information About Data Hosts:

Information collected from data hosts will be shared on an as-needed basis with the PCC personnel; Application Review Board; data providers who will be using the data host to host their data; other data hosts; personnel outside of PCC who may be conducting audits of PCC operations and compliance with PCC policies and procedures; contractors working on PREDICT in a consulting capacity regarding PCC operations; and the PREDICT project manager in the Department of Homeland Security.

### 5.1.4 Information About Application Review Board Members:

Information collected from Application Review Board members will be shared on an as-needed basis with the PCC personnel; other Application Review Board members; data providers who need the information to interact with the ARB; personnel outside of PCC who may be conducting audits of PCC operations and compliance with PCC policies and procedures; contractors working on PREDICT in a consulting capacity regarding PCC operations; and the PREDICT project manager in the Department of Homeland Security.

### 5.1.5 Information About Publication Review Board Members:

Information collected from Publication Review Board members will be shared on an as-needed basis with the PCC personnel; other Publication Review Board members; data providers who need the information to interact with the PRB; personnel outside of PCC who may be conducting audits of PCC operations and compliance with PCC policies and procedures; contractors working on PREDICT in a consulting capacity regarding PCC operations; and the PREDICT project manager in the Department of Homeland Security.

### 5.1.6 Information About Advisory Board Members:

Information collected from Advisory Board members will be shared on an as-needed basis with the PCC personnel; other Advisory Board members; personnel outside of PCC who may be conducting audits of PCC operations and compliance with PCC policies and procedures; contractors working on PREDICT in a consulting capacity regarding PCC operations; and the PREDICT project manager in the Department of Homeland Security.

### 5.1.7 Information About Sponsoring Institutions:

Information collected from Sponsoring Institutions will be shared on an as-needed basis with the PCC personnel; Application Review Board and Publication Review Board members; data providers who need the information to interact with the ARB or PRB; personnel outside of PCC who may be conducting audits of PCC operations and compliance with PCC policies and procedures; contractors working on PREDICT in a consulting capacity regarding PCC operations; and the PREDICT project manager in the Department of Homeland Security.

### 5.1.8 Dataset Information:

The PREDICT datasets will be shared with approved researchers who are affiliated with a sponsoring institution and who have completed and signed an application (which also serves as an MOA) setting forth the research they will be conducting and agreeing to the terms of use of the dataset as set forth by the data provider and data host and the terms of PREDICT. If a researcher leaves their sponsoring institution, the rights to use the dataset are terminated.[5] The researcher may, at that time, submit a new application through the new sponsoring institution to continue using the data. Once the new application is approved, the researcher will be allowed to continue using the data. At no time will DHS or PCC have access to the data.

## 5.2 Is the sharing of personally identifiable information outside the Department compatible with the original collection? If so, is it covered by an appropriate routine use in a SORN? If so, please describe. If not, please describe under what legal mechanism the program or system is allowed to share the personally identifiable information outside of DHS.

The PREDICT administrative information is never inside DHS. It is collected and maintained solely by the PCC. As noted in the Overview, this data collection is covered by the DHS System of Records Notice 002 published December 6, 2004 69 Fed Reg 70460 DHS/ALL 002.

[5] If the researcher is also the sponsoring institution (e.g., a corporation who is acting as both sponsoring institution and researcher) and a researcher leaves the organization, that organization continues to have the right to use the data and may submit the name of another researcher.

### 5.3 How is the information shared outside the Department and what security measures safeguard its transmission?

How PCC administrative information is shared outside the PCC is set forth in 5.1.1 through 5.1.7. Only the minimum amount of information is shared to facilitate PREDICT operations. The ARB, PRB, and Advisory board members sign non-disclosure agreements with the PCC. Internal and external audits also test control points to monitor the effectiveness of the controls and the security of shared information.

### 5.4 Privacy Impact Analysis: Given the external sharing, explain the privacy risks identified and describe how they were mitigated.

The risks to PREDICT PII were identified through a review and analysis of process flow and controls conducted by PCC personnel, audit personnel, PREDICT contractors, and the DHS PREDICT project manager. In addition, in accordance with the DHS Sensitive Systems Handbook, PREDICT ensures effective security controls and authentication. By enforcing system policies, settings, and strong passwords, PCC mitigates risks of disclosure of PII when shared.

## Section 6.0 Notice

The following questions are directed at notice to the individual of the scope of information collected, the right to consent to uses of said information, and the right to decline to provide information.

### 6.1 Was notice provided to the individual prior to collection of information?

General notice of this collection is covered by the DHS System of Records Notice 002 published December 6, 2004 69 Fed Reg 70460 DHS/ALL 002. Additionally, The PREDICT portal provides a Terms of Use and Privacy Statements for general users to review prior to registration for access. This user agreement allows for the individual data users to consent to the sharing of their information with the data providers. The PREDICT portal also contains the PREDICT Operations Manual, which details PREDICT operations and how information is used.

### 6.2 Do individuals have the opportunity and/or right to decline to provide information?

All individuals requesting access to PREDICT information have the opportunity and right to decline to provide personally identifiable information, however with their declination they will not have access to PREDICT and its information.

#### 6.2.1 Researchers, Data Providers, and Data Hosts

All researchers, data providers, and data hosts must have a PREDICT account, which requires them to complete all the requested information listed in 1.1, except the home, cell, and fax telephone numbers are optional. As part of the user agreement, the researcher consents to their information being shared with the data providers and the data hosts so as to facilitate the information flow. If all of the required fields are not provided, they may not participate in the PREDICT project. The information collected is used only for PREDICT operations.

### 6.2.2 Application Review Board, Publication Review Board, and Advisory Board Members

All Application Review Board, Publication Review Board, and Advisory Board members must provide the information set forth above in 1.1 or they may not participate in the PREDICT project. There is no procedure to enable any person associated with one of these Boards to decline to provide information or consent to only particular uses of the information.

### 6.2.3 Sponsoring Institutions

All Sponsoring Institutions must provide the requested information set forth above in 1.1 or their researchers may not participate in PREDICT. There is no procedure for a sponsoring institution to decline to provide the required information or to consent to only particular uses of the information.

### 6.2.4 Datasets

The data provider must consent to its data being used in PREDICT. DHS must also consent to the inclusion of any dataset in the PREDICT project. Prior to DHS granting such consent, it undertakes a review of data provider privacy policies and privacy risks associated with the dataset. Once a dataset has been accepted by DHS, no consent beyond that of the Application Review Board is required in order for researchers to use the datasets. Each data provider is responsible for the release of their data to the PREDICT project and is solely responsible for ensuring that any data it releases complies with all applicable statutes and regulations of applicable governing or regulating bodies and contractual agreements, and is consistent with the data provider's privacy, security, or other applicable policies and procedures. The only data being provided is non-content data and all personally identifiable forms of data are being anonymized, except for topology and measurement data that does not encompass sender-receiver communication. The data provider also agrees through the MOA that the data provided for use in PREDICT has been sanitized, de-identified, or cleaned of any and all information that would not be in compliance or consistent with any policies, procedures, and understandings applicable to the data, whether explicit or implicit.

## 6.3 Do individuals have the right to consent to particular uses of the information? If so, how does the individual exercise the right?

By submitting their information, the individual agrees that their information can be used in accordance with the PREDICT Privacy Statement, Operations Manual, and policies and procedures. All persons participating in the PREDICT initiative must abide by all of the Terms of Use and Privacy statement as well as PREDICT policies and procedures.

### 6.4 Privacy Impact Analysis: Describe how notice is provided to individuals, and how the risks associated with individuals being unaware of the collection are mitigated.

The PREDICT portal provides a Terms of Use and Privacy Statement for general users to review prior to registration for access and on every page after entry to the portal. The end user must also agree to the following every time they access the PREDICT portal:

> **WARNING** This system is for the use of authorized users only. Individuals using this computer system without authority, or in excess of their authority, are subject to having all of their activities on this system monitored and recorded by system personnel. In the course of monitoring individuals improperly using this system, or in the course of system maintenance, the activities of authorized users may also be monitored. Anyone using this system expressly consents to such monitoring and is advised that if such monitoring reveals possible evidence of criminal activity, system personnel may provide the evidence of such monitoring to law enforcement officials.

Access to PREDICT is voluntary. Users must agree to follow the Terms of use, Privacy Statements, and PREDICT policies and procedures each time they log into the portal.

## Section 7.0 Access, Redress and Correction

The following questions are directed at an individual's ability to ensure the accuracy of the information collected about them.

### 7.1 What are the procedures that allow individuals to gain access to their information?

Registered users have a user profile. The profile contains contact information such as name, address, phone numbers, etc. Users have limited access to their profile - they can view and update the contact information but cannot change the roles they belong to. They can not access the profile of any other user. Users update their contact via a web form; the input is protected by an input validation scheme. All user-input text is tested for suspicious entries such as SQL keywords and scripting. Offending text is rejected and discarded.

PREDICT does not allow an individual to access or correct any data other than the data the individual has submitted him/herself.

General users (i.e., those who are not registered users) do not have access to any data and no information is collected about them until they submit an account request to become a registered user.

### 7.2 What are the procedures for correcting inaccurate or erroneous information?

An individual can submit a written request to amend his or her record. After the receipt of the amendment request, the PCC acknowledges in writing that it has either: (a) corrected any information which the individual asserts is not accurate, relevant, timely, or complete; or (b) informs the individual of our refusal to amend in accordance with the request, the reason for refusal, and the procedures for administrative appeal. If the individual disagrees with the PCC's refusal to amend the information and requests a review, the PCC will comply with requirements for a review pursuant to 5 U.S.C. §§ 552a(c)(4), (d) (3), and (d) (4).

### 7.3 How are individuals notified of the procedures for correcting their information?

Individuals are notified via e-mail that the request for Access to PREDICT and or the information hosted by PREDICT, has been denied. Denial of access to PREDICT and the information hosted by PREDICT can result from and is not limited to the following:

**Incorrect:**

- First and last name
- Address (street, city, state, zip code)
- Telephone number
- Email address
- Name of organization
- Title
- Researcher does not have the right criteria for requested data set.

In these cases the user is requested to provide updated/corrected information and re-submit his/her request.

### 7.4 If no formal redress is provided, what alternatives are available to the individual?

PREDICT provides redress to its users, see 7.1 and 7.3.

### 7.5 Privacy Impact Analysis: Please discuss the privacy risks associated with the redress available to individuals and how those risks are mitigated.

Privacy risks are mitigated through the opportunities to access, correct and seek redress regarding information in PREDICT.

## Section 8.0 Technical Access and Security

The following questions are intended to describe technical safeguards and security measures.

## 8.1 What procedures are in place to determine which users may access the system and are they documented?

The PREDICT Operations Manual is available on the PREDICT portal. It documents the operational process and procedures of the PREDICT initiative, including access to datasets. The process flow for PREDICT operations has been documented and reviewed by PCC personnel, consultants, data providers, and the DHS PREDICT project manager.

## 8.2 Will Department contractors have access to the system?

Yes, the entire PREDICT operational structure is handled by contractors and a volunteer community of data providers. PCC staff provides operation and maintenance of PREDICT. Contractors to DHS host the PREDICT datasets. Other contractors provide expertise regarding privacy, procedural, and audit issues. Only those contractors that have a valid need for access will be granted access to the PREDICT portal.

## 8.3 Describe what privacy training is provided to users either generally or specifically relevant to the program or system?

Personal information is gathered and maintained in accordance with pertinent Federal rules and regulations, including the Privacy Act of 1974, the E-Government Act of 2002, the Government Paperwork Reduction Act, and other relevant rules and regulations. Questions relating to privacy are forwarded from the PREDICT Information Systems Security Officer to the DHS S&T liaison to the DHS Privacy Office. Training is being conducted to ensure that individuals working with personal identifiable information are educated on its proper handling and use.

## 8.4 Has Certification & Accreditation been completed for the system or systems supporting the program?

Yes. The Certification and Accreditation (C&A) has been completed by the S&T CIO effective September 30, 2007.

## 8.5 What auditing measures and technical safeguards are in place to prevent misuse of data?

PREDICT project staff ensures audit trails and audit logs are recorded and retained in accordance with the Homeland Security Department Records and Information Management Program. In addition, auditing and intrusion detection capabilities are provided through firewalls and server logs, which alert the PCC staff. Proactive scanning and monitoring of logs and events on a daily basis assists in identifying incidents as early as possible to mitigate potential issues. An audit trail is maintained and stored to facilitate investigation of incidents. Incident reporting policies will include not only the DHS Information System Security Manager (ISSM), but also the DHS Systems Operations Center (SOC) and US-CERT where appropriate.

### 8.6 Privacy Impact Analysis: Given the sensitivity and scope of the information collected, as well as any information sharing conducted on the system, what privacy risks were identified and how do the security controls mitigate them?

#### 8.6.1 Administrative Data

There is the risk that inadvertent access to PREDICT administrative data containing PII may occur. To mitigate this risk, and in accordance with the DHS Sensitive Systems Handbook, the PCC ensures effective security controls and authentication. By enforcing system policies and settings and strong passwords, PREDICT project staff protects the privacy of PII and the integrity of the data itself. Requirements Identification: PREDICT project staff will include the following policies, practices, guidance, and legal requirements for this process: DHS 4300 - IT Systems Security - Sensitive Systems Pub - Vol I Part A; Federal Information Security Management Act of 2002; OMB Circular A-130 Appendix III, Security of Federal Automated Information Systems; Computer Security Act of 1987; OMB Circular A-11, Preparation and Submission of Budget Estimates; Presidential Decision Memorandum (PDD-63), Critical Infrastructure Protection; National Institute of Standards and Technology (NIST) Special Publication (SP) 800-55, Security Metrics Guide for Information Technology Systems.

#### 8.6.2 Datasets

As noted in 1.7.8, the PREDICT datasets will be secured by their respective PCC data hosts and will only be accessed by approved researchers who have been provided with terms and conditions for accessing the datasets, downloading them, storing them, and destroying or returning them. The quality of the PREDICT datasets will be maintained through authentication and authorization controls, regular backups, PREDICT policies and procedures (including MOAs regarding the use and handling of the datasets), and audits of PREDICT operations and compliance with policies and procedures and operational terms and conditions.

The dataset will be downloaded to the sponsoring institution site where the researcher is working. The researcher is bound contractually to use the data only as specified and within the bounds of the research described in the application. In addition, the data provider may require an additional written agreement between the provider and researcher setting forth specific rights and restrictions regarding the use of the data.

## Section 9.0 Technology

The following questions are directed at critically analyzing the selection process for any technologies utilized by the system, including system hardware, RFID, biometrics and other technology.

### 9.1 What type of project is the program or system?

The PREDICT portal is a unique, custom web-based system developed by the PCC. The portal also includes COTS products as part of the system. The security of all of these products has been assessed during the system C&A.

### 9.2 What stage of development is the system in and what project development lifecycle was used?

The PREDICT system is operational.

### 9.3 Does the project employ technology which may raise privacy concerns? If so please discuss their implementation.

PREDICT uses the latest in technology to ensure that personal information, along with the rest of the system's data, is secured against unauthorized access. All communications with PREDICT are through encrypted means (SSL). There is no part of PREDICT that is accessible without first authenticating with a username and password. Physical access to the room where the PREDICT server is housed is controlled through key card access. All access to PREDICT is logged and maintained for future reference if necessary. The use of secure encryption protocols were chosen not only for personal information protection but also to ensure that all information within PREDICT is secured while being transmitted. Access to user information is restricted to very few personnel and logging of all access to that information is conducted within the servers.

## Approval Signature Page

Original signed and on file with the DHS Privacy Office

Hugo Teufel III
Chief Privacy Officer
Department of Homeland Security

# Appendix A – Privacy Statement

PREDICT is committed to protecting your privacy and developing technology that gives you the most powerful and safe online experience. This Statement of Privacy applies to the PREDICT Web site and governs data collection and usage of that site. By using the PREDICT website, you consent to the data practices described in this statement.

### Collection of your Personal Information

PREDICT collects personally identifiable information, such as your e-mail address, name, home or work address or telephone number.

There is also information about your computer hardware and software that is automatically collected by PREDICT. This information can include: your IP address, user name, browser type, and access times. This information is used by PREDICT for the operation of the service, to maintain quality of the service, and to provide general statistics regarding use of the PREDICT Web site.

PREDICT encourages you to review the privacy statements of Web sites that you may choose to link to from PREDICT so that you can understand how those Web sites collect, use and share your information. PREDICT is not responsible for the privacy statements or other content on Web sites outside of the PREDICT web site.

### Use of your Personal Information

PREDICT collects and uses your personal information to operate the PREDICT Web site and deliver the services you have requested. PREDICT may also contact you via surveys to conduct research about your opinion of current services or of potential new services that may be offered.

PREDICT does not sell, rent or lease data it collects to third parties. PREDICT may share data with data hosts, data providers, and review board members of the PREDICT community to help us deliver requested PREDICT services. These parties are prohibited from using your personal information except to provide these requested services, and they are required to maintain the confidentiality of your information.

PREDICT will disclose your personal information, without notice, only if required to do so by law or in the good faith belief that such action is necessary to: (a) conform to the edicts of the law or comply with legal processes served on PREDICT or the site; (b) protect and defend the rights or property of PREDICT; and, (c) act under exigent circumstances to protect the personal safety of users of PREDICT, or the public.

### Use of Cookies

The PREDICT Web site requires the use of non-persistent (session-based) "cookies" during an active session for proper operation of pages for registered PREDICT users during that one session. The purpose of the cookies is to identify a user to the PREDICT web server only when accessing content within the PREDICT domain. PREDICT does not use “ad” cookies to track user activity. PREDICT does not retain any cookie data. You have the ability to accept or decline cookies. Most Web browsers automatically accept cookies, but you can usually modify your browser setting to decline cookies if you prefer. If you choose to decline cookies, you may not be able to fully experience the interactive features of the PREDICT services.

### Security of your Personal Information

PREDICT secures your personal information from unauthorized access, use or disclosure. PREDICT secures the personally identifiable information you provide on computer servers in a controlled, secure environment, protected from unauthorized access, use or disclosure.

### Changes to this Statement

PREDICT will occasionally update this Statement of Privacy to reflect feedback or changes in operations or policy. PREDICT encourages you to periodically review this Statement to be informed of how PREDICT is protecting your information.

### Contact Information

PREDICT welcomes your comments regarding this Statement of Privacy. If you believe that PREDICT has not adhered to this Statement, please contact PREDICT at PREDICT-Contact@RTI.org. We will use commercially reasonable efforts to promptly determine and remedy the problem.