



Privacy Impact Assessment  
for the  
**Genomics Informatics System (GIS)**

**DHS/S&T/PIA-037**

**March 8, 2019**

**Contact Point**

**Timothy Stockwell  
System Project Manager  
National Biodefense Analysis and  
Countermeasures Center  
DHS Science and Technology  
(301) 619-5710**

**Reviewing Official**

**Jonathan R. Cantor  
Acting Chief Privacy Officer  
Department of Homeland Security  
(202) 343-1717**



## Abstract

The Department of Homeland Security (DHS) Science and Technology Directorate (S&T) National Biodefense Analysis and Countermeasures Center (NBACC) created the Genomics Informatics System (GIS) to analyze bio-forensic data for classified development projects and law enforcement casework. GIS processes deoxyribonucleic acid (DNA) and amino-acid sequences transferred to GIS from external sources. S&T uses GIS for the analysis of bio-forensic data that may include the DNA or amino-acid sequences of synthetic constructs, viruses, bacteria, plants, animals, and/or humans. GIS is not used to identify individuals, and the identity of an individual cannot be inferred from a DNA sequence or other information in the GIS database. S&T is conducting this PIA to analyze potential privacy risks in the collection, analysis, and storage of human DNA sequences by GIS.

## Overview

GIS is used for the analysis of bio-forensic data for classified development projects and law enforcement casework. GIS is operated within the National Bioforensic Analysis Center (NBFAC) laboratory program and used on a classified network. The NBFAC is designated by Homeland Security Presidential Directive-10 (HSPD-10)<sup>1</sup> to be the lead federal facility in the technical forensic analysis of materials recovered following a biological attack in support of the appropriate lead federal agency.

The NBFAC is one of two biological laboratory programs under the NBACC. The NBACC was established to fill critical, biodefense-related shortfalls in the nation's scientific knowledge of biological agents that could be used to cause harm to the public. The NBACC laboratory is a government-owned facility operated and managed as a Federally Funded Research and Development Center (FFRDC) by Battelle National Biodefense Institute, LLC (BNBI) for DHS S&T.

GIS conducts only DNA sequence<sup>2</sup> and amino acid sequence analysis, which cannot be used to identify individuals. It is not currently used for "DNA profiling," which is the process of determining an individual's DNA characteristics (which are as unique as fingerprints). GIS is only used for analyzing and processing data from classified projects, and is not used for unclassified casework; therefore, analysis conducted through GIS has a nexus to national security matters.

GIS periodically downloads from the National Institutes of Health (NIH) its publicly

---

<sup>1</sup> See Homeland Security Presidential Directive/HSPD-10: Biodefense for the 21<sup>st</sup> Century, available at <https://www.gpo.gov/fdsys/pkg/CPRT-110HPRT39618/pdf/CPRT-110HPRT39618.pdf>.

<sup>2</sup> A DNA sequence is a digital representation of a single strand of a DNA molecule, using the symbols A (for adenine), C (for cytosine), G (for guanine), T (for thymine), and N (for any or unknown - depending on the context of the DNA sequence, N could mean that any nucleotide is "allowed in the position" or it could mean that "the nucleotide at the position is unknown).



available database of known DNA and amino-acid sequences (the Known Sequence Database), which includes both human and non-human sequences.<sup>3</sup> The human sequences in the Known Sequence Database are identified as “human;” no associated identifying information is provided or available regarding who is the source of a human sequence. Non-human sequences may be identifiable in the Known Sequence Database by the type or identifier assigned to a particular species or type of animal, plant, bacteria, or virus. Genetically-modified sequences may also be identified. Known sequences are maintained in GIS using the NIH sequence identifier. The NIH sequence identifier for a human sequence does not contain any personal identifiers that would permit NBACC to link a human sequence to an individual.

Other Federal Government agencies submit unknown DNA and amino-acid sequences they obtain to NBACC for matching against the Known Sequence Database. These unknown sequences are DNA and amino-acid sequences collected by these agencies in connection with their own authorized responsibilities, and are considered to be classified information. For example, GIS may analyze these unknown sequences related to possible construction of a biological weapon of mass destruction. Human sequences submitted to NBACC could be human DNA found at the scene of constructing a chemical, biological, radiological, or nuclear weapon.

These unknown sequences are submitted to NBACC through an agency liaison who is assigned to, and is onsite at, NBACC. These sequences are assigned anonymized alphanumeric identifiers by the other Federal Government agency that submits them prior to submission to NBACC. This PIA only addresses the human DNA and amino-acid sequences that GIS uses for the analysis of bio-forensic data. The only human DNA and amino-acid sequences in GIS are those received from the other Federal Government agencies or downloaded from NIH in the Known Sequence Database.

Upon receipt of an unknown sequence through a liaison, NBACC uses GIS to characterize the sequence by first searching against the publicly available Known Sequence Database using the Basic Local Alignment Search Tool (BLAST) software.<sup>4</sup> The results of BLAST software sequence similarity searches are made available to human analysts who then develop a final report analyzing the search findings for each sequence in a case. A compiled final report is sent to the Federal

---

<sup>3</sup> The Known Sequence Database consists of two databases “nr” and “nt.” “Nr” is the non-redundant protein database (a collection of protein sequences with entries from GenPept, Swissprot, Protein Data Bank, Protein Research Foundation, Protein Information Resource, and National Center for Biotechnology Information Reference Sequence (RefSeq) project). “Nt” is the non-redundant nucleotide database (which contains entries from traditional divisions of GenBank, European Molecular Biology Laboratory, and the DNA Data Bank of Japan). Both databases are downloaded from the National Center for Biotechnology Information, which is part of the United States National Library of Medicine, a branch of NIH. A submitter of data to GenBank is allowed to submit data derived from sequencing human DNA. GenBank states, “If you are submitting human sequences to GenBank, do not include any data that could reveal the personal identity of the source. GenBank assumes that the submitter has received any necessary informed consent authorizations required prior to submitting sequences.” See <https://www.ncbi.nlm.nih.gov/genbank/>.

<sup>4</sup> BLAST finds regions of similarity between biological sequences by comparing sequences and calculating the statistical significance of a match. More information is available at <https://blast.ncbi.nlm.nih.gov/Blast.cgi>.



Government agency that submitted the unknown sequence that includes, for each input sequence, the alphanumeric identifier and the report of findings regarding the sequence, including whether it matches a sequence in the Known Sequence Database. If the input sequence matches one or more sequences in the Known Sequence Database, the report also includes the NIH assigned identifier(s) for the matching sequence(s). The report, together with the corresponding agency identifier, is provided to the agency's liaison, who is responsible for submitting the report and data to his or her agency. If an unknown sequence does not match against the Known Sequence Database, that federal government agency is provided an Unknown Sample Final Report. The NIH does not maintain associated identifying information on human sequences in the Known Sequence Database.

A GIS Report to a Federal Government agency on an unknown sequence provided by that agency could not be used to match a human sequence to an individual because the GIS Report only indicates whether the sequences are thought to be from a human source. If the sequence is not reported as from a human source, it addresses the likely biological source of unknown DNA and amino acid sequences (*e.g.*, animal, plant, bacterial, viral). The GIS Report also addresses indications of a sequence having been modified or engineered, and, potentially, the function of the particular sequence. The type of biological forensic data may be used by the Federal Government agencies that submitted sequences to NBACC for lead generation and bio-crime attribution.

After the final GIS Report is received and accepted by the submitting Federal Government agency, the corresponding casework data created by NBACC is deleted from GIS. Neither GIS nor NBACC maintains a copy of reports provided to the Federal Government agency.

NBACC does not directly receive any DNA or amino-acid sequences other than the downloaded Known Sequence Database and unknown sequences from other Federal Government agencies. NBACC reports on such sequences are returned to the submitting Federal Government agency through that agency's liaison.

## **Section 1.0 Authorities and Other Requirements**

### **1.1 What specific legal authorities and/or agreements permit and define the collection of information by the project in question?**

The NBFAC conducts bio-forensic analysis of evidence from bio-crime or bioterrorism investigations. The NBFAC is designated by HSPD-10 Biodefense for the 21<sup>st</sup> Century to be the lead federal facility to conduct and facilitate the technical forensic analysis and interpretation of materials recovered following a biological attack in support of the appropriate lead federal agency.<sup>5</sup> NBACC adheres to federal law enforcement and evidentiary standards for the collection and analysis of evidence in connection with sequences provided by other Federal Government

---

<sup>5</sup> Available at <https://www.gpo.gov/fdsys/pkg/CPRT-110HPRT39618/pdf/CPRT-110HPRT39618.pdf>.



agencies.

Additional relevant authorities that allow DHS and other Federal Government agencies to collect and analyze data relevant to their Chemical Biological Radiological and Nuclear forensic and attribution responsibilities are as follows:

- Homeland Security Act of 2002 Section 302;<sup>6</sup>
- 18 U.S.C. Chapter 11B, Chemical Weapons;<sup>7</sup>
- 28 U.S.C. § 533, Investigative and Other Officials; Appointment;<sup>8</sup>
- National Security Presidential Directive 17 (NSPD-17) / Homeland Security Presidential Directive 4 (HSPD-4), titled: National Strategy to Combat Weapons of Mass Destruction;<sup>9</sup>
- Homeland Security Presidential Directive-10 (HSPD-10);
- Homeland Security Presidential Directive-5 (HSPD-5), titled: Management of Domestic Incidents;<sup>10</sup>
- Presidential Decision Directive 39, titled: U.S. Policy on Counterterrorism;<sup>11</sup>
- 28 C.F.R. § 0.85, Organization of the Department of Justice – General Functions of the Federal Bureau of Investigation;<sup>12</sup> and
- 31 U.S.C. § 1535, The Economy Act, Agency Agreements.<sup>13</sup>

## 1.2 What Privacy Act System of Records Notice(s) (SORN(s)) apply to the information?

Although DNA is inherently identifiable, sequence information and other records in GIS have no associated identifying information that makes DNA linkable to any specific individual. Further, DHS will not use DNA it receives for the purposes listed in this document and match it against any other information it holds in its systems DHS-wide that would make this information

---

<sup>6</sup> Available at [https://www.dhs.gov/sites/default/files/publications/hr\\_5005\\_enr.pdf](https://www.dhs.gov/sites/default/files/publications/hr_5005_enr.pdf).

<sup>7</sup> Available at <https://www.gpo.gov/fdsys/granule/USCODE-2011-title18/USCODE-2011-title18-partI-chap11B/content-detail.html>.

<sup>8</sup> Available at <https://www.gpo.gov/fdsys/granule/USCODE-2011-title28/USCODE-2011-title28-partII-chap33-sec533>.

<sup>9</sup> Available at <https://www.gpo.gov/fdsys/pkg/CPRT-110HPRT39618/pdf/CPRT-110HPRT39618.pdf>.

<sup>10</sup> Available at

<https://www.dhs.gov/sites/default/files/publications/Homeland%20Security%20Presidential%20Directive%205.pdf>.

<sup>11</sup> Available at <https://www.hsdl.org/?view&did=462942>.

<sup>12</sup> Available at <https://www.gpo.gov/fdsys/granule/CFR-2010-title28-vol1/CFR-2010-title28-vol1-sec0-85/content-detail.html>.

<sup>13</sup> Available at <https://www.gpo.gov/fdsys/granule/USCODE-2009-title31/USCODE-2009-title31-subtitleII-chap15-subchapIII-sec1535/content-detail.html>.



linkable to a specific individual. Any retrieval only identifies a category of DNA, such as human, plant, viral, or animal. Thus, DHS has determined that GIS is not a Privacy Act system of records and does not require a SORN.

### **1.3 Has a system security plan been completed for the information system(s) supporting the project?**

A System Security Plan was completed to acquire the Authority to Operate (ATO) for GIS in 2015.

### **1.4 Does a records retention schedule approved by the National Archives and Records Administration (NARA) exist?**

Records from the Known Sequence Database downloaded from NIH are maintained in accordance with General Records Schedule (GRS) 5.2, Item 20: Intermediary Records. Retention is temporary, and the data file can be destroyed when no longer needed for a business use. All unknown DNA or amino-acid sequence data processed through GIS is considered to be owned by the submitting Federal Government agency, not NBACC or DHS. That agency's records retention schedules apply to all such data it submits. That data, together with the final report and related casework, is provided to that agency, as the owner of such records. Then, at the direction of that agency, all casework data is removed from GIS. Copies of the accepted reports are not retained in GIS. The acceptance of a report generally takes several months. If the report is not accepted, NBACC/GIS works with the submitting agency to address any questions or issues that arise, until the report acceptance is secured.

### **1.5 If the information is covered by the Paperwork Reduction Act (PRA), provide the OMB Control number and the agency number for the collection. If there are multiple forms, include a list in an appendix.**

GIS is not subject to Paperwork Reduction Act (PRA) requirements because it does not collect information directly from members of the public.

## **Section 2.0 Characterization of the Information**

### **2.1 Identify the information the project collects, uses, disseminates, or maintains.**

GIS processes biological sequences, including DNA sequences and amino-acid sequences, and searches these sequences against databases of known sequences available from the NIH. DNA, or deoxyribonucleic acid, is the hereditary material in humans and almost all other organisms. The information in DNA is stored as a code made up of four chemical bases: adenine (A), guanine (G), cytosine (C), and thymine (T) in a double helix molecule. A DNA sequence is a digital



representation of a single strand of a DNA molecule, using the symbols A (for adenine), C (for cytosine), G (for guanine), T (for thymine), and N (for any or unknown). Human analysts use the search results to develop textual reports that characterizes the unknown sequences. As previously noted, GIS is configured for DNA sequence analysis, and not for DNA profiling. Neither a sequence from the Known Sequence Database nor an unknown sequence provides sufficient information that could be used to cross match that sequence against DNA profiling databases and derive personally identifiable information (PII) of a specific individual or create an identifiable individual's genetic profile.

DNA sequences are received with corresponding anonymized identifiers from the submitting agency. DNA sequence data may also be received from DNA sequencing devices located at NBACC. In the case of raw sequencing data received from sequencing devices, the sequences would be identified by sequencer-generated unique identifiers.<sup>14</sup> Classified casework is processed on GIS. A single human DNA or amino acid sequence cannot be definitively linked to a specific individual.

## **2.2 What are the sources of the information and how is the information collected for the project?**

Unknown DNA and amino-acid sequence data with corresponding anonymized identifiers are received from the submitting agency as computer readable files. NBACC relies on the submitting agencies to follow their own evidentiary and privacy process and procedures prior to any submissions to NBACC. DNA sequence data may also be received from DNA sequencing devices located at NBACC. Known sequence databases used to identify unknown sequences are databases “nr” – the non-redundant protein database (a collection of protein sequences with entries from GenPept, Swissprot, Protein Data Bank, Protein Research Foundation (PRF), Protein Information Resource, and National Center for Biotechnology Information Reference Sequence (RefSeq) project) and “nt” – the non-redundant nucleotide database (contains entries from traditional divisions of GenBank, European Molecular Biology Laboratory, and the DNA Data Bank of Japan). Both databases are downloaded from the National Center for Biotechnology Information, which is part of the United States National Library of Medicine, a branch of the NIH.<sup>15</sup> “Nt” and “nr” databases are available from NIH's publicly available site.

## **2.3 Does the project use information from commercial sources or publicly available data? If so, explain why and how this information is used.**

Public databases of known DNA and amino-acid sequences compiled by NIH are

---

<sup>14</sup> Raw sequencing data comes directly from a sequencing device. It may contain low quality bases or other sequencing inaccuracies. It is standard practice to trim low quality sequence from raw sequencing data and to assemble reads into longer and more accurate “consensus” sequences.

<sup>15</sup> More information is available at <https://www.ncbi.nlm.nih.gov/guide/dna-ma/>.



downloaded from NIH into GIS, and are used to characterize the unknown DNA and amino-acid sequences received. In this PIA, referencing a human DNA or amino acid sequence as “Known” does not mean that either NBACC, GIS, or NIH can relate that sequence back to a specific individual or has or maintains records from which to link the NIH identifier to an individual (or retrieve a specific individual’s sequence). “Known” means that NIH knows whether the DNA or amino acid sequence is human or non-human (and, if non-human, whether it is animal, plant, bacteria, virus, etc., and potentially the type of animal, plant, bacteria, or virus). Consequently, having access to an NIH sequence identification number for a human sequence does not provide a means through which to retrieve or find identifying information about the individual who is the sequence’s source. Neither the NIH public database nor GIS contain related identifying information on either known or unknown sequences.

## **2.4 Discuss how accuracy of the data is ensured.**

NBACC ensures that data received are encoded as appropriate for protein sequences or DNA sequences. For NIH data, the ftp site provides MD5 checksum files<sup>16</sup> for each partition of the “nt” and “nr” databases that are used to ensure data is downloaded without errors. Human subject matter experts review the accuracy of matches when developing the text for the final reports. GIS does not extract DNA from samples or sequence DNA molecules. GIS functions as a data processing system.

## **2.5 Privacy Impact Analysis: Related to Characterization of the Information**

**Privacy Risk:** There is a risk that there may be an unauthorized disclosure of the DNA sequences or identifiers maintained in GIS.

**Mitigation:** The risk is mitigated through physical and administrative controls administered by NBACC. Access to human DNA or amino-acid sequences in GIS is only permitted through DHS-owned workstations located on the classified network, physically located in a secure area at NBACC. To access these workstations and facility, an individual must be an authorized GIS system user. GIS system users must have the appropriate security clearance/suitability, and a “need-to-know.” GIS system users maintain up to date training for Information Technology Security Awareness, Safeguarding National Security Information (NSI), and Privacy at DHS. Users have also read and signed the Rules of Behavior that clearly outlines what information access is appropriate. The risk is further mitigated because GIS can only be accessed by its Internet Protocol (IP) address, which is classified, which limits hacking into the GIS database containing human DNA and amino-acid sequences.

Identifiers assigned to known human sequences in GIS are assigned by NIH. The NIH

---

<sup>16</sup> MD5 (Message Digest algorithm 5) is an algorithmic hashing function that ensures that a file has not been changed as a result of a faulty file transfer, disk error, or other actions.



assigned identifiers do not link to any individual's PII, and NIH does not collect PII with respect to those sequences. GIS can only analyze and compare an unknown sequence to the Known Sequence Database and determine that a particular unknown sequence matches a particular known sequence. The "linkage" between the two sequences can only be reported as NIH sequence [X] matches unknown sequence [Z]. GIS does not store or otherwise have access to either NIH or the submitting agency's data that could link the "matched sequences" as being the DNA sequence of a particular individual. Moreover, neither a sequence from the Known Sequence Database nor an unknown sequence provides sufficient information that could be used to cross match that sequence against any other DNA profiling databases to derive PII of a specific individual or create an identifiable individual's genetic profile.

## Section 3.0 Uses of the Information

### 3.1 Describe how and why the project uses the information.

GIS compares DNA and amino-acid sequences to databases of known DNA and amino-acid sequences. GIS then outputs a report that lists scores and sequence alignment details when matching sequences are found.<sup>17</sup> These metrics include length of match, percent identity, locations of matching regions on the unknown sequence and known sequence, probability that the match occurred by random chance, and a description of the known sequence. A subject matter expert reviews the match information generated by BLAST software for each unknown sequence and develops a human readable report summarizing the analysis. Currently, GIS does not focus on analyzing human DNA, but instead characterizes any DNA or amino acid sequence to determine the most likely biological source of the sequence (animal, plant, bacteria, virus, etc.), determine if there is evidence of modification or engineering, and determine the function of the sequence. Bio-forensic data is used by the other Federal Government agencies for lead generation and bio-crime attribution.

GIS is not used to identify a specific individual. GIS is used to determine if unknown sequences are human or non-human (other animal, plant, bacterial or viral) in origin. However, NBACC could be asked to use GIS for classified human forensic casework if requested by a submitting agency (if, for instance, human DNA is found at the scene of a crime). Comparing an unknown sequence against a known sequence may not create a one-to-one match. However, if sufficient human DNA is sequenced at specific marker locations, a DNA profile for the unknown sample could be generated and then used to identify a relationship to known DNA profiles.

GIS uses anonymous alphanumeric identifiers to link input biological sequence data (*e.g.*, unknown sequences) with results from automated searches against known biological sequence

---

<sup>17</sup> When a statistically significant sequence similarity match is found between two sequences, the length of the match in nucleotides, the percent identity (number of matching nucleotides divided by length of match expressed as a percentage), and the location of the matching sub-regions within the original sequences are output for a subject matter expert to review.



databases (e.g., the Known Sequence Database). Neither GIS nor NBACC is charged with or responsible for any task or responsibility involved in linking biological sequence data to an individual who is the source of the tested biological sequence data. Bio-forensic data is used by other Federal Government agencies for lead generation and bio-crime attribution.

**3.2 Does the project use technology to conduct electronic searches, queries, or analyses in an electronic database to discover or locate a predictive pattern or an anomaly? If so, state how DHS plans to use such results.**

No. DHS does not use the GIS database to discover predictive patterns or anomalies.

**3.3 Are there other components with assigned roles and responsibilities within the system?**

No other DHS components have access to GIS. Information security officers at NBACC and DHS have access to GIS. The submitting agency liaison does not have access to GIS.

**3.4 Privacy Impact Analysis: Related to the Uses of Information**

**Privacy Risk:** There is a risk that users may use human DNA or amino-acid sequences stored in GIS for unauthorized purposes.

**Mitigation:** The risk is mitigated. GIS is currently configured to only analyze unknown sequences to sequences downloaded from NIH. The system is currently unable to be used for other purposes. Use for any other purpose would require a software update to the system and the ingestion of data that the system currently does not accept. All individuals with access to GIS must sign the Rules of Behavior, which restricts unauthorized use of the system or data. Facility regulations also prohibit bringing in any data recording or storage device into the area that houses the GIS data, and documents may not be removed from this area; therefore, sequences could not be moved to another system for unauthorized use.

## **Section 4.0 Notice**

**4.1 How does the project provide individuals notice prior to the collection of information? If notice is not provided, explain why not.**

NBACC and GIS do not collect human DNA sequences directly from individuals, and therefore cannot directly give notice. To the extent that NIH collects a human DNA sequence in its Known Sequence Database, NIH is not associating an individual's identifying information with a sequence from the individual. Moreover, those uploading sequences into the NIH GenBank are instructed not to include additional identifying information when uploading the sequences. NBACC assumes that NIH complied with all required provisions of applicable law regarding



notice to the individual. Similarly, NBACC adheres to the federal law enforcement and evidentiary standards for the collection and analysis of evidence in connection with sequences provided by the submitting agency. With respect to unknown sequences collected by these agencies at crime scenes or as potential evidence in a criminal case, because the sequence is unknown, there is not an identifiable person to whom notice can be provided prior to collection.

## **4.2 What opportunities are available for individuals to consent to uses, decline to provide information, or opt out of the project?**

No data in the GIS database is collected by DHS directly from individuals. NBACC physically cannot link the information in GIS to individuals, and therefore cannot ensure that an individual's data is not used in the system in the event he or she may decline or opt out of its use. Responsibility for providing these opportunities, if required to be provided to individuals, is the responsibility of the agency collecting the information directly from the individuals.

## **4.3 Privacy Impact Analysis: Related to Notice**

**Privacy Risk:** There is a risk that GIS will process data that was attained without an individual's consent or that an individual is unaware that NBACC is using their information for sequence analysis.

**Mitigation:** This risk cannot be mitigated by DHS, which does not collect DNA sequences directly from individuals and has no way to identify the individual associated with a particular DNA sequence. This PIA provides some notice to the public of NBACC's use of NIH and submitting agency's data. NIH does not collect sequences directly from individuals and does not associate (or otherwise connect) human DNA sequences in the Known Sequence Database with the related individual and therefore cannot retrieve such sequences by the individual's name or other unique personal identifier. The known human sequences are not part of an NIH system of records subject to the Privacy Act.

In the case of unknown sequences, NBACC assumes that the submitting agency, in exercise of its anti-terrorism or law enforcement investigation authorities, properly collects evidence, including DNA and amino acid sequences. When a submitting agency collects a DNA or amino acid sequence that is to be submitted to GIS/NBACC for processing, that agency generally does not know (1) whether the sample is human or non-human or (2) if human, the identity of the individual whose sequence it is.

Submitters of data to GenBank are allowed to submit data derived from sequencing human DNA. GenBank states, "If you are submitting human sequences to GenBank, do not include any data that could reveal the personal identity of the source. GenBank assumes that the submitter has received any necessary informed consent authorizations required prior to submitting sequences."<sup>18</sup>

---

<sup>18</sup> See <https://www.ncbi.nlm.nih.gov/genbank/>.



## Section 5.0 Data Retention by the project

### 5.1 Explain how long and for what reason the information is retained.

Data is retained in GIS for only as long as necessary to create a report. Intermediate working data that is casework is maintained by GIS until the final report regarding that casework is accepted by the submitting agency. At the direction of that agency, after the final report for casework is delivered and accepted, all corresponding casework data is deleted from GIS. It may take several months to accept the final GIS Report submitted by NBACC. Known Sequence data from NIH is maintained in accordance with GRS 5.2, Item 020: Intermediary Records.

### 5.2 Privacy Impact Analysis: Related to Retention

**Privacy Risk:** Data may reside on the GIS database longer than necessary because GIS does not have an associated records retention schedule or automated destruction of data.

**Mitigation:** This risk is not mitigated. After the final GIS Report for casework is delivered to the submitting agency and accepted, all corresponding casework data is manually deleted from GIS. There are no automated processes for removing casework intermediate data from GIS. However, neither a sequence from the Known Sequence Database nor an unknown sequence provides sufficient information that could be used to derive the identity of a specific individual or create an identifiable individual's genetic profile.

## Section 6.0 Information Sharing

### 6.1 Is information shared outside of DHS as part of the normal agency operations? If so, identify the organization(s) and how the information is accessed and how it is to be used.

GIS receives unknown biological sequences and associated anonymized identifiers from other Federal Government agencies. Automated search results of unknown sequences against the Known Sequence Database downloaded from NIH are provided to human analysts to apply their subject matter expertise to develop reports on each unknown sequence in a case for reporting back to a submitting agency. These agencies may use this data for lead generation or attribution of a crime. Unknown biological sequences are not downloaded from NIH.

### 6.2 Describe how the external sharing noted in 6.1 is compatible with the SORN noted in 1.2.

DHS has determined that GIS is not a Privacy Act system of records. GIS's purpose is to categorize a DNA sequence and to share that information with other federal government agencies.

### 6.3 Does the project place limitations on re-dissemination?

The forensic information generated for a submitting agency is the property of that agency.



GIS does not place limitations concerning re-dissemination of information by that agency.

## **6.4 Describe how the project maintains a record of any disclosures outside of the Department.**

Because the submitting agency owns the unknown sequences it provides to NBACC, conducting analysis of those sequences in GIS and then providing reports on the results to the submitting agency is not a reportable disclosure of records.

## **6.5 Privacy Impact Analysis: Related to Information Sharing**

**Privacy Risk:** Information may be disclosed to external entities, and possibly re-disseminated, for unauthorized purposes.

**Mitigation:** The MOA entered into with a submitting agency defines the relationship between the parties in administering the forensic programs (including the National Bioforensic Analysis Center, the Bioforensic Research and Development Program, and the Chemical Forensic Program, all under the management of S&T). The MOA outlines that the forensic information generated for the submitting agency is the property of that agency. Such information will be considered law enforcement sensitive, at a minimum, and, to the extent possible, exempt from the Freedom of Information Act. Final reports are classification marked as appropriate for each case, and transmitted by appropriate methods to the submitting agency.

Any Freedom of Information Act request that is received is forwarded to the submitting agency for which it corresponds.

## **Section 7.0 Redress**

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

Forensic information generated for a submitting agency by NBACC forensic programs is the property of that agency. NBACC will forward any requests for disclosure of forensic information to the public under the Freedom of Information Act or the Privacy Act<sup>19</sup> regarding projects and programs to the corresponding submitting agency. That agency shall make a determination as to whether such information must be released and inform DHS of its determination.

### **7.2 What procedures are in place to allow the subject individual to correct inaccurate or erroneous information?**

GIS cannot link individuals to the data within the system. Thus, there are no procedures that

---

<sup>19</sup> See 5 U.S.C. § 552 and § 552a.



allow individuals to correct inaccurate or erroneous information in GIS.

### **7.3 How does the project notify individuals about the procedures for correcting their information?**

There are no procedures available to individuals to correct their information in GIS, therefore NBACC cannot notify individuals about such procedures.

### **7.4 Privacy Impact Analysis: Related to Redress**

**Privacy Risk**: GIS does not provide a method of redress for incorrect information within the system.

**Mitigation**: Although this risk cannot be mitigated, S&T has taken several steps to lessen its impact. GIS does not have the capability to identify an individual's data within the system, and therefore cannot respond to an access or redress request. GIS is purged of information after a report is made to the submitting agency, which limits the time inaccurate data could reside in the database. Moreover, NBACC does not make any law enforcement decisions regarding matches made within GIS. The owner of the data and decisional authority, the submitting agency, will have its own access and redress procedures that an individual may use.

## **Section 8.0 Auditing and Accountability**

### **8.1 How does the project ensure that the information is used in accordance with stated practices in this PIA?**

As part of the GIS Authority To Operate (ATO) accreditation, a System Security Plan was developed and reviewed by DHS accreditors. Third-party audits are performed as part of the ATO accreditation/re-accreditation process. The IP address for GIS, which is required to log into GIS, is classified. GIS system user password complexity, password expiration, and system logging requirements are based on the Security Technical Implementation Guides (STIGs) from the Defense Information Systems Agency. GIS logs the following events – Successful/Failed login, privilege escalation, tasks performed under privilege escalation, cron (recurring) jobs, bash jobs, every attempted network access, and external media connected to any GIS server. GIS system logs are reviewed periodically by the GIS system administrator. The GIS system servers are housed in a secure server room at NBACC. Classified workstations that can access GIS are all located in a secure office area at NBACC.

### **8.2 Describe what privacy training is provided to users either generally or specifically relevant to the project.**

All GIS system users must complete DHS S&T mandated annual Privacy at DHS: Protecting Personal Information Training, annual IT Security Awareness Training, one-time Safeguarding National Security Information (NSI) Training, and annual Insider Threat Awareness Training.



### **8.3 What procedures are in place to determine which users may access the information and how does the project determine who has access?**

GIS system users must possess the appropriate clearance/suitability, been “read in” to specific projects, and have “need-to-know” in order to access casework data on GIS. Two types of roles are 1) bioinformatics staff – who can log into GIS, transfer unknown sequence data from classified workstations to the GIS system server, perform searches against known sequence databases, and transfer search results back to classified workstations; and 2) analysis staff – who are subject matter experts that review the search results and develop text describing the characterization of the unknown sequences in order to develop a final case report for a submitting agency. Remote access to GIS is not allowed. External storage and communications devices are forbidden to be connected to GIS.

### **8.4 How does the project review and approve information sharing agreements, MOUs, new uses of the information, new access to the system by organizations within DHS and outside?**

NBACC enters into an MOA with any Federal Government agency that wishes to submit data for analysis. Any new MOA will be reviewed and approved by all appropriate DHS and S&T oversight offices.

## **Responsible Officials**

Timothy Stockwell  
GIS Project Manager  
National Biodefense Analysis and Countermeasures Center (NBACC)  
DHS Science and Technology (S&T)

## **Approval Signature**

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

---

Jonathan R. Cantor  
Acting Chief Privacy Officer  
Department of Homeland Security