Office of the Director of National Intelligence

2010 Data Mining Report

For the Period January 1, 2010 through December 31, 2010

Office of the Director of National Intelligence
2010 Data Mining Report
January 1, 2010 through December 31, 2010

I. Introduction

The Office of the Director of National Intelligence (ODNI) provides this report pursuant to Section 804 of the *Implementing the Recommendations of the 9/11 Commission Act of 2007*, entitled *The Federal Agency Data Mining Reporting Act of 2007* (Data Mining Reporting Act).

A. Scope.

This report covers the activities of all ODNI components from January 1, 2010 through December 31, 2010. Constituent elements of the Intelligence Community (IC) will report their activities to Congress through their own departments or agencies.[1]

B. Reporting Requirement.

The Data Mining Reporting Act requires "the head of each department or agency of the Federal Government that is engaged in an activity to use or develop data mining shall submit a report to Congress on all such activities of the department or agency."[2] This is an annual requirement. Under the Act, "data mining" is defined as:

> "… a program involving pattern-based queries, searches or other analyses of one or more electronic databases, where —
> (A) a department or agency of the Federal Government, or a non-Federal entity acting on behalf of the Federal Government, is conducting the queries, searches, or other analyses to discover or locate a predictive pattern or anomaly indicative of terrorist or criminal activity on the part of any individual or individuals;
> (B) the queries, searches, or other analyses are not subject-based and do not use personal identifiers of a specific individual, or inputs associated with a specific individual or group of individuals, to retrieve information from the database or databases; and
> (C) the purpose of the queries, searches, or other analyses is not solely— (i) the detection of fraud, waste, or abuse in a Government agency or program; or (ii) the security of a Government computer system.[3]"

When focusing on individuals or groups, the ODNI typically uses analytic tools and techniques that rely on "personal identifiers of a specific individual, or inputs associated with a specific individual or group of individuals," such as link-analysis tools. Unlike the predictive, pattern-based technologies envisioned by the Act, these tools and techniques start with a known or

[1] Section 804(c)(1) of the Data Mining Reporting Act.
[2] Section 804(c)(1) of the Data Mining Reporting Act.
[3] Section 804(b)(1)(A) of the Data Mining Reporting Act

suspected terrorist, or other subject of foreign intelligence interest, and use various methods to uncover links or relationships between the known subject and potential associates or other persons with whom that subject has a "link" (a contact or relationship). Thus, such analytic tools and techniques do not fall within the statutory definition of data mining.

C. Format and Content.

The format of this year's report has been reworked for clarity and readability. Part II reports on those ODNI programs, if any, that meet reporting requirements of the Data Mining Act. Part III reports on non-data mining programs that are included in the interest of transparency, for the reasons explained below. For programs included under Part III, descriptions are provided in a continuous narrative form, with information that is generally responsive to the reporting elements of the Data Mining Reporting Act.

In this year's report, there are no programs included in Part II, and five programs included in Part III. Three of those programs were described in last year's report: two Intelligence Advanced Research Projects Activity (IARPA) programs – Knowledge Discovery and Dissemination (KDD) and Automated Low-level Analysis and Description of Diverse Intelligence Video (ALADDIN Video) – and one ODNI Chief Information Officer (CIO) program – Catalyst. The fourth program is IARPA's Automatic Privacy Protection (APP) program, which includes discussion of the follow-on Security and Privacy Assurance Research (SPAR) program, and the fifth is the National Counterterrorism Center's (NCTC's) program known as DataSphere.

Although these five programs are not "data mining" programs, information is nonetheless provided in the interest of transparency. As with last year's report, KDD and ALADDIN Video are included because technologies investigated by, or later developed from, those programs could be used to support data mining (such technologies would then be subject to reporting under the Data Mining Reporting Act). Catalyst and DataSphere are in development stages and, while such development does not currently include pattern-based functionality, both programs contemplate potential pattern-based functionality in future stages. The APP and SPAR programs are included to provide information on programs that may have applicability in enhancing security and privacy protections in data mining activities.

D. Protection of Privacy and Civil Liberties.

The ODNI Civil Liberties and Privacy Office (CLPO) works closely with the ODNI Office of General Counsel, ODNI components and the IC elements to ensure appropriate legal, privacy, and civil liberties safeguards are incorporated into policies, processes and procedures that support the intelligence mission. The CLPO is led by the Civil Liberties Protection Officer, a position established by the IRTPA. The duties of this Officer are set forth in that Act, and include: ensuring that the protection of civil liberties and privacy is appropriately incorporated in the policies of the ODNI and the IC; overseeing compliance by the ODNI with legal requirements relating to civil liberties and privacy; reviewing complaints about potential abuses of privacy and civil liberties in ODNI programs and activities; and ensuring that technologies

sustain, and do not erode, privacy protections relating to the use, collection, and other disclosure of personal information."[4]

The IC has in place a protective infrastructure built in principle part on a core set of U.S. Person rules derived from Executive Order (EO) 12333. This EO requires each IC element to maintain procedures, approved by the Attorney General, governing the collection, retention and dissemination of U.S. Person information. These procedures limit the type of information that may be collected, retained or disseminated to the categories listed in part 2.3 of the EO. Each IC element's Attorney-General approved U.S. Persons guidance is interpreted, applied, and overseen by that element's Office of General Counsel and Office of Inspector General. Violations are reported to the Intelligence Oversight Board of the President's Intelligence Advisory Board. In addition to EO 12333, IC elements are subject to the requirements of the Privacy Act, which protects information about U.S. citizens and permanent resident aliens that a government agency maintains and retrieves by name or unique identifier.

The IC's privacy and civil liberties protective infrastructure is bolstered also by guidance and directives issued by the Office of Management and Budget, including memoranda regarding the reporting of and response to incidents involving personally identifiable information and the minimization of Social Security Numbers.

Before any tool or technology could be used in an operational setting, the use of the tool or technology would need to be examined pursuant to EO 12333, the Privacy Act, and other applicable requirements to determine how the tool could be used consistent with the framework for protecting information about Americans and other U.S. Persons.

CLPO has been considering how advanced technologies, employed in accordance with proper laws and policies, enable sharing and use of information while protecting privacy and civil liberties. Such privacy-enhancing technologies (PETs) also prove useful in providing protections for data mining activities. CLPO is making available the results of its PET-related research to IC elements and other government offices with which CLPO collaborates. During 2010, CLPO began follow-on work to identify specific activities within the ODNI as to which PETs could be applied, as a means to better understand the practical utility, functionality, and potential of PETs for intelligence activities.

II. ODNI Data Mining Activities

The ODNI did not engage in any activities to use or develop data mining functionality in the reporting period.

III. Other Programs

The following programs are reported in the interest of transparency.

[4] National Security Act of 1947, as amended by the Intelligence Reform and Terrorism Prevention Act of 2004 (IRTPA), 50 § 403-3d.

A. CATALYST

The CIO manages the Catalyst program. During Fiscal Year 2010, Catalyst focused on pre-Phase A research activities, which did not include development of pattern-matching functionality. Catalyst Phase A alternative analysis and technology development activities are planned through Fiscal Year 2011. Pattern-matching functionality is contemplated for a later phase of development, currently scheduled to begin in 2013.

Catalyst will support IC information sharing and integration objectives as laid out in Intelligence Community Directive (ICD) 501 and the 2009 National Intelligence Strategy (NIS). In January 2009, the ODNI released ICD 501, establishing policies for discovery and dissemination/retrieval of intelligence and intelligence-related information. The NIS recommended making more information accessible or discoverable, while recognizing the need to protect aspects of sensitive reporting. The NIS highlighted the continuing difficulties analysts face in finding and accessing the universe of relevant data and information across large and diverse data sources, setting "Improving Information Integration and Sharing" as its fourth Enterprise Objective.

As the need for the sharing of intelligence data increases, the volume of data that analysts must interpret is expected to increase dramatically, further impeding efficient and effective access, discovery, and use of that data for analysis. Analysts will encounter data that was collected for different purposes, protected in different ways, and described using different terminologies. The Catalyst program was established to address information overload and improve inter-agency, multi-intelligence (multi-INT) information sharing by providing IC analysts with capabilities for entity disambiguation, correlation, and co-referencing.

Catalyst will provide IC intelligence professionals the ability to discover intelligence information correlated from the IC's vast data holdings. The Catalyst program will address analytic information overload and enable IC enterprise correlation of entity information. The three primary goals for Phase A of Catalyst's development are to: (1) systematically derive entity information from across vast information holdings and share it using consistent standards and processes that ease the time and effort required to exploit the data upon receipt; (2) provide a means of correlating entity data acquired from disparate sources to enable more rapid discovery and understanding of the data; and (3) provide an enterprise-wide system that allows a user to query for a given person or organization of interest using fragmentary intelligence to discover all that the IC knows about that entity and access that portion of intelligence information for which the user is authorized to access.

In its end-state, Catalyst will enable data fusion/analytic programs to share disparate repositories with each other, to disambiguate and cross-correlate the different agencies' holdings, and to discover and visualize relationship/network links, geospatial patterns, temporal patterns and related correlations. Data access controls will be in place to safely exchange essential data elements amongst authorized IC personnel with appropriate credentials. If an analyst does not have access to view a record, the system will provide the analyst with the necessary information to enable that analyst to efficiently request access to the critical information needed.

A key component of Catalyst will be the National Information Correlation Service (NICS). The NICS will maintain ingested source data, data resulting from Catalyst Data Preparation Services (CDPS) processing, assertions made by humans, and information derived through further processing and correlation analytics. CDPS services support systematic data preparation, extraction, and transformation for the NICS. The NICS correlation analytics will include entity disambiguation and automated knowledge inferencing, and will allow users to find attributes and relationships for entities of interest.

The data to be used in the Catalyst program is from IC sources. The NICS enables its data correlation services, through the application and exclusive use of the metadata associated with IC shared, multi-intelligence data sources. As such, the NICS entity correlation processing does not contain the substance or content of the external data sources, only the metadata associated with the data sourced. In implementing Catalyst, each agency's system will be stand-alone, and will reside on its own network. Catalyst will apply the metadata associated with the data provider's data source and use that metadata in its services to make it discoverable and accessible to the entire community. It is anticipated that Catalyst data sources will be modified and changed as the program moves forward.

As stated, pattern matching functionality for Catalyst is contemplated, but has not yet been designed and is, therefore, not currently in development. Such functionality is anticipated in a later phase of development. Currently, functional requirements for pattern matching are expected to be developed in the latter half of 2012. It is anticipated that the Catalyst program's pattern matching functionality will be primarily resident in the NICS data correlation services component. For example, such functionality might enable the use of pattern matching to identify a set of entities for further review by analysts, in response to a threat stream.

ODNI will provide additional detail in a future report when and if such functionality is developed.

CIO works closely with CLPO, as well as the ODNI Office of General Counsel (OGC), to build in privacy and civil liberties protections in its activities, and will do so with respect to the design of any pattern matching functionality. Safeguards include policies and training on the rules regarding the collection, retention, and dissemination of U.S. Person information, guidance on Constitutional limitations, and technical integration of oversight and compliance measures on access and use of data (e.g., PKI-enabled attributed based access controls, audit processes, data export management controls).

B. DATASPHERE

NCTC uses tools that do not fit the definition of data mining as defined by the Data Mining Reporting Act. NCTC uses network analysis tools to discover relationships between known and suspected terrorists (KST) and their associates. These activities are conducted using the personal identifiers of a KST as the starting point. Additionally, NCTC uses entity resolution tools to efficiently discover all data in its holdings pertaining to a given individual. The subjects of NCTC's entity resolution tool include persons already known to have a terrorism nexus, persons seeking entrance into the United States, and persons being assessed for a nexus to terrorism.

NCTC is currently developing a tool – DataSphere – which, once completed, will enhance data fusion and entity resolution, as well as discovery of unknown relationships. DataSphere, enables analysis of the activities of terrorists such as their communication networks and travel. The goal of DataSphere is to provide analysts with a tool to aid in the discovery of unknown terrorism relationships and the identification of previously undetected terrorist and terrorism information. DataSphere development work in 2010 does not fit the definition of data mining as defined by the Data Mining Reporting Act because the data set loaded into the tool is comprised entirely of terrorism information. Queries and searches in DataSphere, thus, begin with known identifiers (such as identifiers of known or suspected terrorists or terrorist groups) based upon threat-based reports. However, it is contemplated that Pattern-matching functionality using data not yet known to be terrorism information will be included in future development phases.

DataSphere draws upon data in NCTC's data holdings. Upon acquisition, each dataset utilized by NCTC is tagged, as appropriate, under existing interagency agreements addressing the provision of the information, as well as pursuant to Attorney General-approved guidelines for the handling of U.S. Person information. Additionally, all datasets are appropriately marked for retention, U.S. Person content, operational sensitivity, and whether there is a "filter" caveat associated with data use. A filter caveat is any data that must be preprocessed, i.e., technically compared against datasets with a pre-existing terrorism nexus prior to only the "terrorism-linked" results being made available to an appropriate limited set of NCTC analysts.

DataSphere will be able to detect patterns in data that links individuals with events and actions including identifying a set of individuals that fit the parameters described in a threat intelligence report. These results will be complemented with other tools and manual analysis for validation purposes. Details regarding such functionality will be provided in a future report, once such functionality is designed and in development.

As DataSphere's pattern matching capabilities are developed and tested, NCTC will continue to work closely with CLPO, as well as OGC, to ensure appropriate privacy and civil liberties protections are taken into account.

The legal and policy foundation for DataSphere is based on IRTPA and the National Security of 1947, as amended. These statutory authorities establish NCTC as the "primary organization in the United States government for analyzing and integrating all intelligence possessed or acquired by the United States government pertaining to terrorism and counterterrorism, excepting intelligence pertaining exclusively to domestic terrorists and domestic counterterrorism." Further, NCTC shall also "serve as the central and shared knowledge bank on known and suspected terrorists and international terror groups, as well as their goals, strategies, capabilities, and networks of contacts and support." Moreover, EO 12333 requires each IC element to maintain procedures, approved by the Attorney General, governing the collection, retention and dissemination of U.S. Person information, thereby protecting the rights and privacy of U.S. Persons. These procedures limit the type of information that may be collected, retained or disseminated to the categories listed in part 2.3 of EO 12333. NCTC currently operates under Attorney General Approved Guidelines that govern NCTC's access, retention, use and dissemination of Terrorism Information contained within datasets.

NCTC has instituted the following measures to guard against improper use of data, which would also provide protection in the context of any future data mining activities:

- Mandatory training on the proper handling of U.S. Persons information for everyone who has access to such information. The training includes descriptions of specific uses of U.S. Persons data that are prohibited.

- Strict access control over all information about U.S. Persons limiting access to only those charged with seeking to identify a terrorism nexus.

- Oversight by the OGC and CLPO of all NCTC use, retention and dissemination of data about U.S. Persons. This oversight includes reviews of procedures for determining whether an individual has a terrorism nexus, whether information about an individual can be shared with partner organizations to determine if there is a terrorism nexus, and how decisions are made to use data about individuals determined to have a terrorism nexus.

C. IARPA Research Programs

The mission of IARPA is to invest in high-risk/high payoff research programs that have the potential to provide the United States with an overwhelming intelligence advantage over its future adversaries. It does not have an operational mission and it does not deploy technologies directly to the field. As a scientific research funding organization, IARPA does not use, nor does it expect to make use of, data mining technology. IARPA programs are by nature experimental and are designed to produce new capabilities. The end goal of an IARPA program is typically a proof-of-concept experiment or prototype of an entirely new capability. Due to their experimental nature, IARPA programs do not always achieve their end goals, and when they do, further steps are required to transform the results into real world applications. Any results from IARPA research programs that do get incorporated into future operational programs within the IC, or other parts of the United States government, will be subject to appropriate legal, privacy, civil liberties and policy safeguards.

As with last year's report, the KDD and ALADDIN Video programs below are reported in the interest of transparency, due to the potential that technologies investigated by, or later developed from, those programs could be used to support data mining (such technologies would then be subject to reporting under the Data Mining Reporting Act).

1. *Knowledge Discovery and Dissemination (KDD) Program*. The KDD scientific research program is an IARPA program, begun in 2009. A Broad Agency Announcement (BAA) for KDD was released on December 22, 2009 and KDD research contracts were awarded in September 2010.

The objective of the KDD program is to enable an analyst to utilize large, complex and varied data sets that he has not seen before to produce actionable intelligence in a timely manner. KDD tackles two significant technical areas: (1) how to quickly understand the novel data sets so that the contents can be correctly integrated with data sets that are already in use (this is termed

"alignment"); (2) how to construct automatic analysis tools that are able to work effectively across aligned data sets. KDD research results will be evaluated using realistic challenge problems throughout the program.

In evaluations of research teams' prototypes, the KDD scientific research program will utilize real-world, classified data sets that are large and complex. KDD researchers' work will be evaluated in the context of challenge problems using these data sets. The challenge problems will not be problems that require data mining technology as defined by the Act. The data sets used by researchers will be highly varied, and may include, for example, regional biographic data, incident reports, translated newspaper articles, etc. The use of all data sets will be consistent with all U.S. laws and regulations.

For more information on the KDD Program, see http://www.iarpa.gov/solicitations_kdd.html

2. *Automated Low-level Analysis and Description of Diverse Intelligence Video (ALADDIN Video) Program.* The Automated Low-level Analysis and Description of Diverse Intelligence Video (ALADDIN Video) scientific research program is an IARPA program that is expected to award research contracts in early 2011. A BAA for ALADDIN was released on June 10, 2010.

The objective of the ALADDIN program is to enable an analyst to query large video data sets to quickly and reliably locate those video clips that show a specific type of event. The ALADDIN program will research technologies designed to automatically search large numbers of video data files for analyst-defined events of interest and direct the analyst to those video data files that are likely to contain occurrences of those events. ALADDIN's technologies, if successful, will help to automate a triage process that is currently performed largely manually by analysts. Although this is not "data mining," technologies that result from ALADDIN research could, potentially, be applied by operational organizations to support capabilities that involve pattern recognition.

ALADDIN addresses three significant technical areas: (1) High-speed processing of large amounts of video clips to extract information that can later be used to support queries about each clip's contents; (2) Generation of effective queries from small sets of example video clips and a textual description; (3) Robust query processing that identifies the clips of interest and summarizes the rationale for their selection. ALADDIN research results will be evaluated by IARPA and the National Institute for Standards and Technology (NIST).

The ALADDIN program will use video data files in its research and evaluations that are acquired by NIST for its annual, international, Video Retrieval video search technology research program (TRECVID). TRECVID sponsors public evaluations of video and multimedia search technologies that are open to worldwide participation. ALADDIN performers will participate in these evaluations to demonstrate objective progress in their research. The data collections used in the TRECVID evaluations are made available to all participants through an evaluation participation agreement that stipulates that the TRECVID data collections are to be used for research purposes only. The TRECVID data is collected using a rigorous process that protects privacy.

For more information on the TRECVID Program, see http://www-nlpir.nist.gov/projects/trecvid

For more information on ALADDIN, see http://www.iarpa.gov/open_solicitations.html

3. *Automatic Privacy Protection (APP) Program.* The 2009 ODNI Data Mining Report included a discussion of the APP program that, while not specifically focused on data mining, may enhance security and protect privacy in data mining activities.

In the 2010 reporting period, APP research achieved two complementary goals. First, it developed secure distributed private information retrieval (PIR) protocols that permit an organization (Client) to query a cooperating data provider (Server) so that the Server cannot learn about the query posed or the results returned. At the same time, the Server is assured that only information relevant to the query is returned to the Client. These protocols add minimal overheads for computation and communication for simple queries and databases using a trusted third party. Second, APP demonstrated algorithms to compare database queries with privacy policies so that a Client's auditor can verify that only authorized queries have been submitted to the Server.

A follow-on program began with solicitation of proposals for Security and Privacy Assurance Research (SPAR, http://www.iarpa.gov/solicitations_spar.html) in December 2010. This research will build on the APP program's results to handle complex queries, large datasets, and three new information sharing architectures: publish-subscribe systems, message queue (mailbox) systems, and outsourced data storage systems. Furthermore, SPAR will develop policy compliance checking algorithms that work on encrypted queries, enabling the Server to verify that Client queries are authorized. In addition, the SPAR program will explore efficient homomorphic encryption techniques to implement queries on encrypted data.

If successful, the SPAR program will benefit the IC by securing and protecting the privacy interests of both those who search and those who own databases. The technology may enhance cooperative information sharing within the IC, and among governments and the private sector. Efficient private information retrieval may expand policy options for reconciliation of privacy and security concerns when information is shared. Furthermore, the program could enable both Clients and Servers to enforce restrictions on data mining activities.