INSTITUTE FOR DEFENSE ANALYSES

User Requirements for the Enterprise Data to Decisions Information Environment

Julie Pechacek
Alan Gelder
Ethan Novak
Amrit Romana
Paul Richanbach
Kathy Conley
George Kennedy
Cheryl Green

August 2018
Approved for public release;
distribution is unlimited.
IDA Document NS D-9139
Log: H 18-000246

INSTITUTE FOR DEFENSE ANALYSES
4850 Mark Center Drive
Alexandria, Virginia 22311-1882

The Institute for Defense Analyses is a non-profit corporation that operates three federally funded research and development centers to provide objective analyses of national security issues, particularly those requiring scientific and technical expertise, and conduct related research on other national challenges.

About This Publication
This work was conducted by the Institute for Defense Analyses (IDA) under contract HQ0034-14-D-0001, project BE-6-4311, "Next Generation Personnel Management Models and Modeling Environment Phase 2: Structuring the Environment and Preliminary Model Development," for the Office of the Under Secretary of Defense for Personnel and Readiness (OUSD (P&R)). The views, opinions, and findings should not be construed as representing the official position of either the Department of Defense or the sponsoring organization.

Acknowledgments
The authors thank Nancy Huff, Shelley Cazares, Karen Gordon, and Peter Levine for their thorough review of this document.

For More Information:
Dr. Julie Pechacek, Project Leader
jpechace@ida.org, 703-578-2858
ADM John C. Harvey, Jr., USN (Ret), Director, SFRD
jharvey@ida.org, 703-575-4530

INSTITUTE FOR DEFENSE ANALYSES

IDA Document NS D-9139

User Requirements for the Enterprise Data to Decisions Information Environment

Julie Pechacek
Alan Gelder
Ethan Novak
Amrit Romana
Paul Richanbach
Kathy Conley
George Kennedy
Cheryl Green

This page is intentionally blank.

Executive Summary

The U.S. Department of Defense (DOD) and the research institutions supporting it conduct extensive personnel analysis—both operational and research-oriented—to manage and enhance America's all-volunteer military forces. To provide rapid, secure access to defense-related personnel data in support of these analyses, and to enable cross-institutional collaboration, the Office of the Under Secretary of Defense for Personnel and Readiness (OUSD(P&R)) tasked the Institute for Defense Analyses (IDA) to identify and synthesize researcher requirements for a new data-hosting analytic computing resource to augment current capabilities resident at the Federally Funded Research and Development Centers (FFRDCs).

The requirements for this resource, currently known as the Enterprise Data to Decisions Information Environment (EDDIE), result from a collaborative process involving iterative conversation and feedback between the Study and Analysis Center FFRDCs—operated by the RAND Corporation, CNA, and IDA—and a selection of government analytic organizations identified by the project sponsor and the Requirements Development Team at IDA. The dialogue for generating these requirements considered research workflows and timelines, industry best practices, bureaucratic hurdles, and strategies to better support DOD research objectives. The requirements herein support the DOD Business Capability Acquisition Cycle (BCAC). Details on our methodology, case studies on other large-scale U.S. government data provision and analytic efforts, and recommendations for implementing EDDIE will be documented in a forthcoming paper.

The requirements outlined in this document pertain to the desired user experience. We do not attempt to translate these user requirements into technical architectures or detailed computational specifications for achieving that experience.

This page is intentionally blank.

Contents

This page is intentionally blank.

1. Introduction

The U.S. Department of Defense (DOD) and the research institutions supporting it conduct extensive personnel analysis—both operational and research-oriented—to manage and enhance America's all-volunteer military forces. Research topics commonly include recruiting, training, morale, readiness, compensation, retention, career management, organizational management and behavior, force resiliency, and transition from service, among others. To obtain accurate, timely, and rigorous examinations of topics affecting military personnel management, DOD relies heavily on analysts located in the Federally Funded Research and Development Centers (FFRDCs) and in government offices within the Office of the Secretary of Defense (OSD) and the military services. Their analytic efforts include both in-depth and quick-response projects.

To improve rapid, secure access to defense-related personnel data in support of these analyses, and to enable cross-institutional collaboration, the Office of the Under Secretary of Defense for Personnel and Readiness (OUSD(P&R)) tasked the Institute for Defense Analyses (IDA) to identify and synthesize requirements for a new data-hosting analytic computing resource. Currently known as the Enterprise Data to Decisions Information Environment (EDDIE), this resource is intended to supplement the existing data and computational resources of the defense personnel research community.

This document presents the results of a collaborative process to identify user requirements for EDDIE (occurring between November 2017 and June 2018). This involved iterative conversation and feedback between the Study and Analysis Center FFRDCs—operated by the RAND Corporation, CNA, and IDA—and a selection of government analytic organizations identified by the project sponsor and the Requirements Development Team at IDA. The dialogue for generating these requirements considered research workflows and timelines, industry best practices, bureaucratic hurdles, and strategies to better support DOD research objectives. The requirements herein support the DOD Business Capability Acquisition Cycle (BCAC). Details on our methodology, case studies on other large-scale U.S. government data provision and analytic efforts, and recommendations for implementing EDDIE will be documented in a forthcoming paper.

The requirements outlined in this document pertain to the desired user experience. We do not attempt to translate these user requirements into technical architectures or detailed computational specifications for achieving that experience.

This page is intentionally blank.

2. Assumptions

A. Assumptions

The requirements outlined in this document rely upon the following assumptions and aspirations of OUSD(P&R) in describing the purpose, scope, and contents of EDDIE.[1]

A1. EDDIE will supplement existing data and computing resources for the government and non-government analytic organizations supporting DOD. EDDIE will not replace existing channels of data access from the Defense Manpower Data Center (DMDC) or other DOD offices and organizations to computing environments approved for handling, storing, and processing Personally Identifiable Information (PII).

A2. EDDIE will enable rapid data access.

A3. By providing a secure analytic space, EDDIE will reduce the flow of data to unsecure data environments.[2]

A4. The initial data holdings in EDDIE will prioritize data held by DMDC. Other offices and organizations may contribute additional data.

A5. The EDDIE architecture will accommodate three access categories for data:

a) Immediate Access Data, which is listed in the EDDIE library catalog and is accessible to all users without additional authorization;

b) Restricted Access Data, which is listed in the EDDIE library catalog and is accessible only to users with specific authorization; and

c) Private Data, which is not listed in the EDDIE library catalog and is accessible only to users with specific authorization.

---

1 The initial dialogue for developing requirements for EDDIE began in an open-ended fashion, with no coordinated parameters for its underlying content, role, structure, or scope. The Executive Stakeholder meeting (2 November 2017), two Analyst Stakeholder meetings (16 November 2017 and 1 March 2018), meeting-related feedback, and the initial requirements document draft (15 February 2018) were conducted to engage in an open-ended discussion of needs and possibilities, under the assumption that the community was starting from a relatively blank sheet. This exploration enabled the sponsor to provide the IDA Requirements Development Team with firm guidance on the underlying assumptions on 5 April 2018.

2 Currently, some DOD offices that require access to disaggregated personnel information do not have adequate secure computing resources for processing and safeguarding that data.

A6. Files containing PII will not be exportable from EDDIE (rare exceptions, such as exporting PII back to the original data provider, may be accommodated).

A7. Files not containing PII may be extracted from EDDIE. Extraction will require a human check to ensure that extracted files do not contain PII.[3]

A8. EDDIE will not charge a use fee to users or to the institutions with which users are affiliated.

A9. The technological and administrative architectures chosen for EDDIE will flexibly accommodate multiple potential governance implementations, operational structures, and avenues for future expansion.[4]

A10. EDDIE will support research that is conducted both within single organizations and collaboratively across multiple organizations.

A11. EDDIE will provide computing resources; data repositories; a library containing data, metadata, analytic products, and code; and a streamlined process for research-related regulatory compliance.

A12. Detailed EDDIE governance practices and roles will be promulgated in a DOD Issuance following a future period of analytic community input similar to that which produced the initial user requirements documented herein.

A13. EDDIE will be an unclassified environment. However, noting A9, future expansions of EDDIE may include classified annexes.

B. Dependencies

D1. The completeness, relevance, and accuracy of the data that EDDIE hosts and facilitates analysis of depend on the data collection and preparation practices of the military departments, OSD, and DMDC.

---

[3] Automated systems that scan for social security numbers, bank account numbers, credit card numbers, dates of birth, and other information that is prone to identity theft exist. However, it is effectively impossible at this point to automate a system that is able to detect all forms of PII based on linking together multiple pieces of contextual information (especially when the independent elements seem innocuous).

[4] Flexibility in computing resources may be accomplished through the use of a cloud solution. Several commercial providers meet the security requirements to store and process data containing PII, as well as more sensitive information. Cloud solutions enable the user to modify capacity and type of computing resources on demand (such as doubling the number of processors or switching from central processing units to graphics processing units).

C. Constraints

C1. EDDIE must comply with the Privacy Act of 1974 (5 U.S.C. 522a) regarding the protection of PII, as well as DOD's stated policy on protecting PII in DOD Directive (DODD) 5400.11.[5] The protection of PII in EDDIE should also be informed by National Institute of Standards and Technology (NIST) SP 800-122.[6]

C2. EDDIE must comply with the appropriate Federal Cybersecurity Policies and Guidelines. These include DOD and NIST issuances. In particular, EDDIE must comply with NIST publication 800-53 (Revision 4) Appendix J[7] and DOD Instruction (DODI) 8510.01.[8]

C3. EDDIE must comply with DODI 8320.02[9] and 8320.07[10] on sharing data, information, and information technology (IT) services.

C4. EDDIE must comply with the Health Insurance Portability and Accountability Act of 1996 (HIPAA), Public Law 104-191; the HIPPA Privacy, Security, and Breach Notification Rules;[11] and DOD's policy on safeguarding protected health information (PHI) in DODI 6025.18[12] and DOD 6025.18-R.[13]

---

5 DODD 5400.11, *DOD Privacy Program*, 29 October 2014.

6 McCallister, Erika, Timothy Grance, and Karen A. Scarfone, *Guide to Protecting the Confidentiality of Personally Identifiable Information (PII)*, NIST Special Publication 800-122 (Gaithersburg, MD: National Institute of Standards and Technology, April 2010), https://csrc.nist.gov/publications/detail/sp/800-122/final.

7 Joint Task Force Transformation Initiative, *Security and Privacy Controls for Federal Systems and Organizations*, NIST Special Publication 800-53, Rev. 4 (Gaithersburg, MD: National Institute of Standards and Technology, April 2013), Appendix J, http://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-53r4.pdf.

8 DODI 8510.01, *Risk Management Framework (RMF) for DOD Information Technology (IT)*, 12 March 2014 (incorporating Change 2, 28 July 2017).

9 DODI 8320.02, *Sharing Data, Information, and Information Technology (IT) Services in the Department of* Defense, 5 August 2013.

10 DODI 8320.07, *Implementing the Sharing of Data, Information, and Information Technology (IT) Services in the Department of Defense*, 5 December 2017.

11 https://www.cms.gov/Outreach-and-Education/Medicare-Learning-Network-MLN/MLNProducts/Downloads/HIPAAPrivacyandSecurity.pdf

12 DODI 6025.18, *Privacy of Individually Identifiable Health Information in DOD Health Care Programs*, 2 December 2009.

13 DOD 6025.18-R, *DOD Health Information Privacy Regulation*, January 2003.

This page is intentionally blank.

3. Definitions

Repository	The authoritative source for a data product. For example, DMDC is the repository for the data product known as the Active Duty Master File.
Analytic Environment	A collection of computing, infrastructure, analytic tools, and personnel resources to receive, safeguard, host, manipulate, and support analysis of data from repositories and other sources.
Depository	A sequestered location within an Analytic Environment where source data from a repository is stored. For example, when an Analytic Environment receives a data product (such as the Active Duty Master file) from a repository (such as DMDC), it stores the data product in a Depository. A Depository can be partitioned between a Source Data Depository and a Derived Data Depository (that is, between the original source data and data derived directly from it). Multiple distinct Depositories may exist within a single Analytic Environment.
Direct Identifiers	Person-level data objects that immediately or nearly immediately identify a specific individual, including a person's name, social security number, DOD ID number, and date of birth.[14]
De-identified PII	Person-level data objects or records that do not contain Direct Identifiers, but that are still considered PII as some or all individuals in the data may be identifiable with reasonable effort.
EDDIE Anonymized Identifier	An identifier that is consistently assigned to each unique individual that is represented in any EDDIE data resource. It is consistent across all EDDIE data resources and is able to serve as a merge key. Moreover, it is constructed in such a way that it cannot be used to reverse-engineer, discover, or infer the values of any Direct Identifiers.

[14] Depending on the number of individuals represented, aggregations such as birth month (month-year) or birth quarter (quarter-year) may not "nearly or immediately identify a specific individual" and would therefore not be direct identifiers.

Specific Project Amendment (SPA) for Information	An informational filing submitted by a project member to obtain use of Immediate Access Data (as referenced in Assumption A5) and not subject to approval.
SPA for Concurrence	A request submitted by a project member to obtain use of Restricted Access Data (as referenced in Assumption A5). Request execution is subject to authorization.
Sponsor	The entity that provides the resources to pursue a research project.
Institution	An FFRDC, government office, or government organization that participates in the research process by conducting analyses, sponsoring analyses, or providing data for analyses.
Project Leader	The person who bears the primary responsibility for conducting the analysis for a project in accordance with the Sponsor's research objectives. This may involve forming a research team and coordinating the efforts of the research team to fulfill the research objectives.
User	An individual who uses an Analytic Environment for conducting research or analyses.
Folder	A directory or virtual workspace within a computing system for organizing and manipulating computing files. A folder is associated with an access control list that specifies which users have a given level of folder access (e.g., read only, read and write, or higher administrative privileges). Data files and other items in folders may either refer to (1) items that are stored on the disk space within that particular directory or (2) results of session-specific virtual databases, database instances, or queries that are approved for launch from that particular directory. A folder may also be interpreted as a version control system repository (such as a Git repository).

4. Requirements

A. EDDIE Overview and High-Level Requirements

The following overview establishes EDDIE's core features and functionalities. Many of the requirements covered in this high-level overview are elaborated on in later sections.

Table 1. High-Level Requirements

Analytic Environment	EDDIE must be an Analytic Environment: it must have sufficient computing capacity (e.g., CPU speed, parallel threads, RAM, disk memory, etc.), infrastructure, analytic tools (e.g., scientific computing software), and personnel resources to properly receive, safeguard, handle, and permit the analysis of PII sourced from DMDC and elsewhere. EDDIE's computing structures, implementation, governance, and management practices must enable its effective and efficient use as an Analytic Environment.
Data Access	EDDIE must provide defense personnel data to users in a consistent and convenient format, enabling users to focus primarily on analysis instead of on the process of obtaining and cleaning data.
Distinct Data Processing Areas	EDDIE must have the following distinct data processing areas: a Cold Room; a Staging Area; a Research Area; and an import and export control functional area.
Cold Room	EDDIE must contain a highly secure computational area and related staffing sufficient to safeguard Directly Identified PII, receive all PII entering EDDIE, prepare it for use by users, and place it in the Staging Area. Functions performed in the Cold Room include performing data merges that require the use of Direct Identifiers and other keys; stripping data of Direct Identifiers; creating EDDIE Anonymized Identifiers; and standardizing and labeling data.

Staging Area	EDDIE must contain a remotely accessible area suitable for holding and organizing data that is to be made available to analysts operating within EDDIE. This area must warehouse PII objects that do not contain Direct Identifiers (i.e., De-identified PII), which have been processed through the Cold Room; as well as non-PII objects which bypass the Cold Room. The Staging Area must be partitioned into three regions: Immediate Access Data, which is listed in the EDDIE library catalog and is accessible to all users without additional authorization; Restricted Access Data, which is listed in the EDDIE library catalog and is accessible only to users with specific authorization; and Private Data, which is not listed in the EDDIE library catalog and is accessible only to users with specific authorization. Further details are in the Cold Room and Staging Area section on p. 21.
Research Area	EDDIE must contain a remotely accessible, internally networked work area containing Institutional Folders (for FFRDC and government organizations), collaborative work folders (for working across institutions), and a library. Further details are in the Research Area section on p. 27.
Library	EDDIE must support and contain a Library including data dictionaries, metadata, public code, a wiki collecting institutional knowledge, and a repository of user profiles. Further details are in the Library section on p. 32.
Institutional and Collaborative Folders	Institutional and collaborative folders within the EDDIE Research Area must be able to hold a variety of subfolders for projects, institutional research and development, archives, operations (such as recurring report generation), and other purposes. Further details are in the Institutional and Collaborative Folder section on p. 27.
Folder Access	Within EDDIE, access to each institutional and collaborative folder and all subfolders must be granted or revoked under the authority of that institution and administered via an access control list controlled by that institution. Institutional access to EDDIE is subject to oversight, as described in the User Access section on p. 19. Folder privilege controls must support the ability to vary access within folders and subfolders on a user-by-user basis along all privilege dimensions: read only privileges, read and write privileges, and various levels of higher administrative privileges. Further details are in the Institutional and Collaborative Folder section on p. 27.

Import Control	EDDIE must support the import of PII and non-PII items as requested by users.[15] A mechanism for securely transmitting PII items from repositories and other analytic environments into the EDDIE Cold Room must exist. Non-PII items must be importable from outside of EDDIE into the networked portion of the EDDIE analytic environment within one business day of a user request. Further details are in the Import and Export Control section on p. 38.
Export Control	EDDIE must support the export of non-PII items as requested by users. Requests for exporting non-PII items from the EDDIE analytic environment must be fulfilled within two business hours.[16] This will require human review to verify that the requested item is not PII. Further details are in the Import and Export Control section on p. 38.
HSR Approval	EDDIE must streamline the Human Subjects Review (HSR) and Institutional Review Board (IRB) approval process for projects operating in EDDIE. Projects in EDDIE must not be subject to multiple human subjects reviews of the same information by different organizations. Further details are in the Human Subjects Review section on p. 39.
Institutional Memory	EDDIE must provide means for users to record information about data, code, results, and projects to facilitate the communication and preservation of information across users, project teams, and organizations.
Computational Resources	EDDIE must provide users with sufficient hardware computing capacity. All users have the ability to perform analyses through EDDIE's computing resources instead of relying on their own local systems. Further details are in the Computational Resources section on p. 35.
Analytic Tools	EDDIE must provide users with research tools to read, explore, and manipulate data; produce tabulations, graphs, and other visualizations of data; and build statistical and other models. Table 10 contains a list of software that must be included in EDDIE. Further details are in the Software section on p. 28.

[15] Users submitting import requests must state that they are authorized to import the data. This is especially relevant for data placed in the Immediate Access portion of the Staging Area.

[16] A major point of feedback from the analytic community was the need for a rapid capability to remove non-PII items from EDDIE. The analytic community noted that this was critical to maintain productivity in research processes that require high degrees of iteration (such as model development and model tuning) or integration with data, information, and models that reside outside of EDDIE. A rapid export capability is also crucial for facilitating quick response projects.

Team Workspaces	EDDIE must enable project teams to share data, code, and metadata in a workspace. Project leaders must be able to grant and revoke user access permissions (read only, read and write, and higher administrative privileges) within their team workspace. Institutional Access Managers implement changes to access but may delegate access changes within project folders to project leaders.
Enable Collaboration	EDDIE computing structures, implementation, governance, and management practices must enable collaboration in team workspaces among individuals both within and between institutions.

As depicted in Figure 1, data flows from various Repositories (e.g., DMDC) and other sources into Depositories within PII environments at FFRDCs, government offices, and other organizations, including EDDIE.[17] Figure 1 also illustrates the structure of EDDIE and its three primary sections: the Cold Room, the Staging Area, and the Research Area. The solid black lines depict the flow of data within EDDIE.[18] Import Control determines where the data will land after entry (i.e., the Cold Room, the Staging Area, or directly into a folder or folders in the Research Area), based in part on a determination of whether the data contain PII. PII containing Direct Identifiers enter EDDIE through the Cold Room portion of Import Control and land in the Cold Room. All items passing through Import Control are screened for viruses and malicious software. As described in the high-level requirements above (Table 1), the Cold Room is an air-gapped or otherwise highly secure space for receiving, storing, manipulating, and preparing PII containing Direct Identifiers.

The Staging Area stands between the Cold Room and the Research Area, and contains non-PII and PII that do not contain Direct Identifiers (i.e., De-identified PII). Once PII are stripped of Direct Identifiers and the data are otherwise prepared, Cold Room personnel extract the data from the Cold Room and place it in the appropriate portion of the Staging Area: Immediate Access, Restricted Access, or Private.[19] Non-PII may be placed directly in the Staging Area by Data Managers through the Import Control process. Users obtain Immediate Access or Restricted Access data from the Staging Area by filing the appropriate Specific Project Amendment (SPA) request. Users may need to provide additional information to access private data. A Data Manager pulls the approved data extract from the Staging Area and places it (or its working instantiation) in the designated

[17] An authority to operate, letter of certification, or similar credential endows FFRDCs, government offices, and other entities—including EDDIE—with the ability to operate a PII environment (that is, an analytic environment that is authorized to host, safeguard, manipulate, and analyze PII data).

[18] To parsimoniously represent Figure 1, the arrows represent major data flows, but not all possible data flows. The text illustrates in greater detail the specific nuances of data flows.

[19] The portion of the Staging Area to which each data object is assigned depends on the relevant contractual restrictions, if any. Unless there are extenuating restrictions, the default is to place data in the Immediate Access portion of the Staging Area.

folder in the Research Area for the user. Through Import Control, Data Managers may also scan and import non-PII and PII that does not contain Direct Identifiers directly to folders in the Research Area. Data Managers move data between and within Institutional and Collaborative Folders as allowed by the relevant access category and, if applicable, SPA approvals.

All analysis is conducted in the Research Area. Like the Staging Area, the Research Area contains non-PII and PII that do not contain Direct Identifiers (i.e., De-identified PII). The Research Area contains Institutional Folders (for each FFRDC and government office requiring one), Collaborative Folders, and a Library. The Institutional Folders contain subfolders that support projects, internal research and development, and other content, such as institutional tools, data objects, and operations. Project leaders have the ability to determine whether their project folder is searchable by users outside their institution, and whether their project folder is readable by users inside or outside their institution. As is currently the case, DOD research sponsors maintain ultimate discretion as to whether and when work they sponsor is released to a broader audience. Each user will have access to the Institutional Folder for the institution with which he or she is affiliated. Collaborative Folders host projects with teams drawn from multiple institutions, and have the same general features and privilege structures as Institutional Folders. Users will have access to analytic tools (e.g., software) and computational assets (e.g., RAM, processing threads, disk capacity, etc.) in the Research Area. In Export Control, Data Managers inspect all items requested for removal from EDDIE to determine whether they contain PII prior to moving them to a networked download location.

The Library contains data dictionaries, metadata, public code (such as programming scripts written and submitted by EDDIE users), a wiki collecting institutional knowledge, and a repository of user profiles. Users may submit items for inclusion in the Library. Data librarians work as curators, ensuring that items in the Library follow a standard format that can easily be understood by the community and is not overly burdensome to implement.

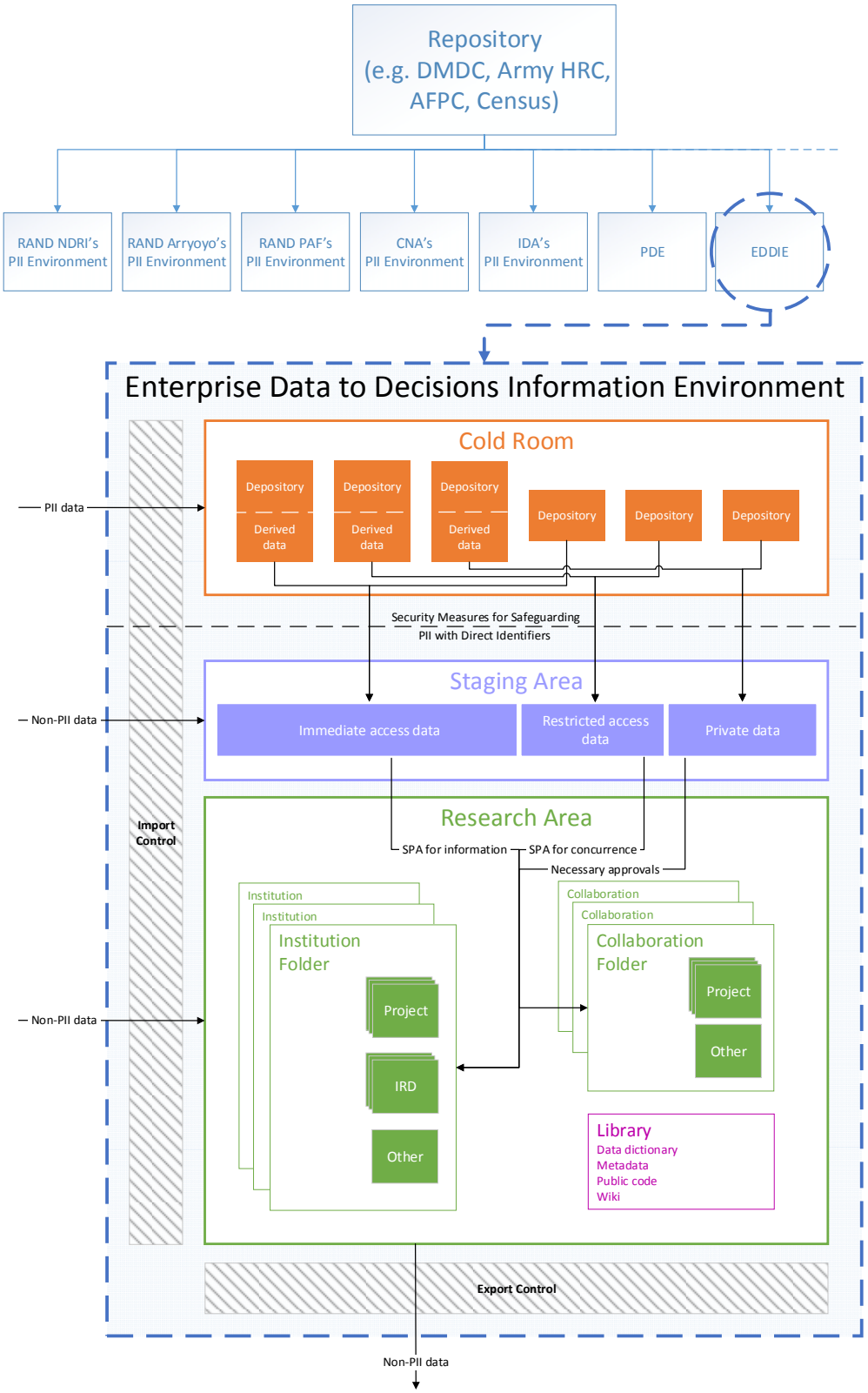

Figure 1. Repository, Depository and EDDIE Architecture

B. EDDIE Personnel Roles

The following table lists the personnel roles required to support EDDIE operations. These roles may be described in greater detail in other sections.

Table 2. Personnel Role Requirements

Data Managers	EDDIE must empower specially trained individuals as Data Managers to move data within EDDIE and to provision data for users. Data Managers are responsible for **(1)** pulling data extracts from the Staging Area into subfolders within the Institutional and Collaborative folders; **(2)** moving submissions from the Institutional and Collaborative folders into the Staging Area and the EDDIE Library; and **(3)** conducting import and export control. Data Managers must be trained and authorized to determine whether files submitted for import or export contain PII. They must be authorized to export files from the EDDIE Research Area that do not contain PII; scan and import non-PII files and PII files that do not contain any Direct Identifiers from outside of EDDIE into the Staging Area, and the Institutional and Collaborative folders; scan and import non-PII files from outside of EDDIE into the EDDIE library; and securely submit PII imports with Direct Identifiers to the Cold Room. They may perform code review and quality control on products submitted to the Staging Area or EDDIE Library.[20] All imports, exports, and extractions from the Staging Area that Data Managers perform must be auditable. EDDIE must provide environment-wide Data Managers to serve all users. Additionally, EDDIE must permit institutions that heavily use Data Managers to nominate individuals within their organizations to be Institutional Data Managers. Nominations for Institutional Data Managers may be confirmed through a process set forth by the EDDIE Oversight Committee.[21]

[20] Code review may also be performed by EDDIE users. As stated later, Data Librarians oversee data quality control.

[21] Institutional Data Managers are intended to prevent limited EDDIE personnel resources from becoming a bottleneck hindering expeditious research. Demand for Data Managers at some institutions may ebb and flow in irregular ways. Instead of hiring extra government employees to maintain the capacity needed for surge demands, institutions can align their internal personnel resources to ensure that they have a sufficient number of Data Managers to meet their research commitments in a timely manner. Having data mangers within FFRDCs is also in keeping with the Federal Acquisition Regulation (FAR): "An FFRDC, in order to discharge its responsibilities to the sponsoring agency, has access, beyond that which is common to the normal contractual relationship, to Government and supplier data, including sensitive and proprietary data, and to employees and installations equipment and real property" (FAR 35.017).

	Both Institutional and EDDIE-wide Data Managers must be trained and authorized to perform their roles (to include any recurring training and auditing of their work). Institutional Data Managers must be able to perform all Data Manager roles for their Institution's users, folders, and projects, including accessing the Immediate Access portion of the Staging Area, the EDDIE Library, and performing Import and Export Control. Institutional Data Managers do not have access to Institutional Folders outside of their institution. Institutional Data Managers have access to the Collaborative Folders where researchers within their institution perform work. Institutional Data Managers have access to the data within the Restricted Access and Private portions of the Staging Area that has been approved for use in their institution. EDDIE-wide Data Managers have access to all Institutional and Collaborative Folders, as well as all portions of the Staging Area.
Cold Room Personnel	EDDIE must provide Cold Room personnel to process all files in the Cold Room and move data to the Staging Area. The team must have the ability to accomplish the Cold Room data manipulation, preparation, and curation tasks. To preserve separation of roles for human subjects protection purposes, Cold Room personnel must not be involved in research on data that they manipulate. Cold Room personnel may also serve as environment-wide Data Managers.
Data Librarian	EDDIE must staff Data Librarians who serve as liaisons between Data Managers, the data depository team, data providers, and analysts. Data Librarians maintain data use agreements (DUAs); oversee data quality control; proactively verify the quality and usability of merge keys for major data files in the Staging Area; curate all elements in the EDDIE library; and ensure the accuracy of metadata, data dictionaries, and the EDDIE Library Wiki.[22] They communicate, for instance, with data providers when data quality concerns arise and with the Cold Room personnel on issues related to merges.
Human Subjects Review (HSR) Coordination Team	EDDIE must provide an HSR coordination team to help move the study through the HSR process, and maintain a status dashboard to keep study members informed of where the study is in the review process. Further details are in the Human Subjects Review section on p. 39.

[22] Data Librarians are meant to oversee all of these functions, using any available tools, automation techniques, crowdsourcing endeavors, or other resources to perform these functions.

EDDIE Access Manager	EDDIE must provide Access Managers who: **(1)** have purview over access permissions to the Staging Area; **(2)** create and disable Institutional and Collaborative Folders (to include extending management privileges for Institutional and Collaborative Folders to Institutional Access Managers); and **(3)** may assist Institutional Access Managers, as needed, in managing access permissions.
Institutional Access Manager	EDDIE user institutions must have Institutional Access Managers who: **(1)** have purview over access permissions within the Institutional Folder; **(2)** distribute and collect security tokens; **(3)** track necessary trainings for EDDIE for individuals within their institution. The EDDIE Oversight Committee may set procedures to audit the performance of Institutional Access Managers.
EDDIE Liaison	EDDIE must staff a liaison function to **(1)** procure additional data resources for inclusion in EDDIE from both DOD and external sources; **(2)** advocate on behalf of EDDIE with data providers for improved data quality, seeking resolution to data problems and questions for objects held within EDDIE; **(3)** conduct education and outreach to data providers and users on behalf of EDDIE; and **(4)** help guide prospective and existing users and projects on how to (4a) enter and conduct analyses in EDDIE, (4b) satisfy HSR requirements, and (4c) obtain needed resources or assistance within EDDIE.
System Administrators	EDDIE must staff system administrators to perform routine maintenance and updates. This includes checking for available updates to tools and packages on at least a monthly basis and making them available to EDDIE users.
Help Desk Team	EDDIE must staff a dedicated Help Desk team to provide responsive IT helpdesk support, as well as a central routing and follow-up point on all requests for assistance.
EDDIE Oversight Committee	EDDIE must staff a governing body to establish and amend policies pertaining to EDDIE; manage EDDIE's ongoing operations; and direct capability investments for EDDIE.
EDDIE Director	EDDIE must staff a director who is responsible for **(1)** the oversight and management of EDDIE personnel and day-to-day EDDIE operations; **(2)** ensuring that any training requirements set forth by the Oversight Committee for users, Data Managers, Access Managers, and other roles in EDDIE are fulfilled; **(3)** executing other duties as set forth by the Oversight Committee and in the EDDIE Charter.

C. Governance Structure

This section describes the basic features required to implement a governance structure to operate and sustain EDDIE. A detailed governance structure will be promulgated in the future, as described in assumption A12.

Table 3. Requirements Enabling Implementation of Governance Structures

EDDIE Charter	EDDIE must have a charter promulgated in a DOD Issuance that details its mission, function, and oversight. At a minimum, the charter must empower an Oversight Committee to manage the operations of EDDIE; establish policies and procedures related to EDDIE; establish a mechanism for amending those policies and procedures; establish a method for collecting and resolving user needs and requests; establish a method for adjudicating disputes which may arise; establish working groups on specific issues as the Oversight Committee deems necessary; and solicit, receive, and rank proposals for funded efforts that are designed to enhance EDDIE's capabilities.
EDDIE Oversight Committee Membership	EDDIE's oversight committee must include the following roles and individuals: **Chair**, filled by the primary OUSD(P&R) research sponsor; **Deputy Chair**, filled by the EDDIE Director (a high-level Office of People Analytics (OPA) representative); and **Members**, filled by representatives from the Military Departments, DMDC, and the Study and Analysis Center FFRDCs supporting the DOD (RAND, IDA, and CNA).
EDDIE Executive Committee	The Chair and the Deputy Chair of the EDDIE oversight committee form the EDDIE Executive Committee, which must be able to approve membership on the oversight committee for other major data providers and institutions; make final funding and source selection decisions for funded efforts that are designed to enhance EDDIE's capabilities; and fulfill all roles and responsibilities enumerated in the EDDIE charter.
EDDIE Development Program	EDDIE must have funding dedicated to support investment efforts that sustain, augment, enhance, and evolve EDDIE's capabilities and usefulness.

Examples of projects that may be funded by the EDDIE Development Program include recording institutional knowledge about data available in EDDIE; generating or curating metadata for inclusion within the EDDIE Library; populating EDDIE with previous studies, models, and tools (such as those that predate the creation of EDDIE and have been approved for release by research sponsors); providing informed and timely answers to user forum questions; and sustaining FFRDC representation on the Oversight Committee.

D. User Access

Table 4. User Access Requirements

Institutional Access to EDDIE	EDDIE must have a process whereby FFRDC and government organizations submit requests for Institutional access to EDDIE (including the creation of an Institutional Folder). The approval timeline for establishing institutional access must not take longer than three weeks from the time of the institution's request. Obtaining and maintaining institutional access must depend on the institution's adherence to clearly defined criteria. These criteria will be established by the Oversight Committee.[23]
User Access to EDDIE	Individuals desiring user access to EDDIE must obtain that access through an Institution with Institutional access to EDDIE. Institutions receiving Institutional access to EDDIE must determine which individuals affiliated with their organization may obtain user access to EDDIE through their institution. General criteria for obtaining user access may be established by the Oversight Committee. The vetting of individuals for becoming EDDIE users should include a path for individuals who do not have a security clearance.[24]
User Training	Any trainings that users are required to complete in order to obtain and maintain access to EDDIE (such as human subjects protection training) must be clearly specified by the Oversight Committee and necessitate no more than 10 hours annually to complete.[25]
Multifactor Authentication	User access to EDDIE must require multifactor authentication. One factor must be a security token. For FFRDC users, the security token should not be a Common Access Card (CAC).[26]

[23] Some example criteria for institutional access may include an agreement that institutional users comply with personnel controls necessary to satisfy NIST Special Publication 800-171 (Revision 1) and NIST Special Publication 800-53 (Revision 4), Appendix J, as well as a stipulation that all the institution's users complete specified human subjects protection training.

[24] Not all researchers in FFRDC Study and Analysis Centers have security clearances. Alternative vetting paths might include such things as a public trust clearance or some equivalent to the U.S. Census Bureau's Special Sworn Status.

[25] Most commonly accepted training programs for human subjects research in the social sciences take no longer than 10 hours to complete. The 10-hour limit is for standard user access. Some data in the Restricted Access portion of the Staging Area may require additional training.

[26] Since DOD offices are often limited in the number of CACs they can sponsor for non-DOD employees, the fraction of researchers at FFRDCs that have CACs is smaller than the fraction who have security clearances. If a CAC is deemed necessary for accessing EDDIE, a mechanism should be established for granting CACs to researchers who need to use EDDIE.

Institutional Maintenance of User Permissions Rosters	Institutional Access Managers must be empowered to maintain the access control roster for their respective institutions' users, folders, and subfolders, including **(1)** establishing read, read/write, and higher administrative privileges; **(2)** distribution and collection of security tokens as needed; **(3)** tracking of user trainings and removal of individuals whose training has lapsed; and **(4)** any addition or removal of individual users as needed. Institutional Access Managers may delegate management of project folder access permissions to project leaders.[27]
New User Accounts	Requests directed to the EDDIE Help Desk Team to establish a new user account must be fulfilled within two business days.[28]
User Account Changes	Requests directed to the EDDIE Help Desk Team to implement changes to user account permissions must be fulfilled within one business hour.
Remote Access	Users must be able to access EDDIE remotely from workstations provide by their institutions.
Access Prerequisites	Subject to any general criteria established by the Oversight Committee, determining and granting user access must be the prerogative of the organization sponsoring the user's access. To permit exploratory analysis in Institutional Research and Development (IRD) Folders (see Table 8) and to permit the use of the Library, user access must not be predicated on currently having work that requires access to an EDDIE project folder. Like all research performed within the DOD and its associated FFRDCs, any findings from exploratory analysis must traverse proper DOD approval channels prior to any public release.
Access Boundaries	Managers of project folders must be able to determine whether their project folder is searchable by users outside their institution, and whether their project folder is readable by users inside or outside their institution. The default is for users to have read access to all items within the Institutional Folder (for their institution) and for all project folders to be searchable by all EDDIE users. (See "Folder Permissions" in Table 8.)

[27] As stated earlier, the EDDIE Oversight Committee may set procedures for auditing the performance of Institutional Access Managers.

[28] The two business days is solely for implementing the new user account. It does not include any prerequisite training or vetting requirements for becoming an EDDIE user.

E. Cold Room and Staging Area

Table 5. Cold Room and Staging Area Requirements

Cold Room Tasks	The Cold Room must be able to perform data preparation, data cleaning, derived data creation, and data manipulation tasks such as performing data merges that require the use of Direct Identifiers and other keys; stripping data of Direct Identifiers; creating EDDIE Anonymized Identifiers; and standardizing and labeling data.
Source Data	Source data for all PII in the Staging Area must always remain in the Cold Room and must be maintained in the state in which it was received.
Cold Room Data Version Documentation and Control	The Cold Room must maintain a clear version control system for data files, including a detailed record of the process used to obtain each derived data object from source data object(s). This record must be maintained for all versions and time periods of derived data objects, to support reproducibility and data improvement. For data items provided in the Immediate Access or Restricted Access Staging Areas, this information must be made available to users upon request.[29]
Source Data Key Preservation	Original data files must retain any Direct Identifiers received from the data source to enable data merges in the Cold Room.
Cold Room Security	The computational and personnel resources dedicated to the Cold Room must be sufficient to safeguard Directly Identified PII. The computers used to execute this function should be dedicated to this exclusive use. Example controls include granting physical and logical access to a limited number of people, and not networking Cold Room machines to any computer or device outside of the Cold Room area (i.e., establishing an "air gap").
Staging Area Data Storage	The Staging Area is a networked portion of the EDDIE analytic environment that must facilitate the movement of data between the Cold Room and the Research Area, and must provide storage adequate for all data made available to users. After data objects have been prepared and stripped of Direct Identifiers in the Cold Room, they are transferred to one of the following Staging Area partitioned regions: • Immediate Access Data, which is listed in the EDDIE library catalog and is accessible to all users (usage

[29] The rationale for this requirement is to maintain transparency in the processes used in the Cold Room to ensure a consistent scientific process.

	may be tracked and audited, but being an EDDIE user is sufficient authorization for access); • Restricted Access Data, which is listed in the EDDIE library catalog and is accessible only to users with specific authorization; and • Private Data, which is not listed in the EDDIE library catalog and is accessible only to users with specific authorization. Redundant storage of data in the Staging Area must be minimized.
Staging Area Extractions	Users obtain access to data in the Staging Area by filing the appropriate specific project amendment (SPA) request. Following approval of the request, a Data Manager pulls the approved data extract from the Staging Area and places it in the requested folder in the Research Area for the user. Once a SPA is approved, extraction requests for data present in the Staging Area must be completed and transferred to the requesting user within one business day. From the definition of "Folder" herein, a data extract may either be data that is actually stored on the disk space within a directory or it may be a virtual database, database instance, or query that is approved for launch from a directory. To maintain replicability for the latter, the data underlying the database instance must be preserved (that is, if the database is changed, the version of the data at the time of the initial query must be preserved).
Automated Specific Project Amendment (SPA)	The process for filing SPAs must be automated.[30] That is, for each major data file, users must be able to select the data fields they desire for the time periods or samples they desire. The interface would ideally be html-based with point and click functionality; the format of the IPUMS USA data request site may be used as a model.[31] The SPA request should be automatically generated through this process (noting the destination folder for the data, the requesting user, the data requested, and the time of the request). Automation should initially focus on high usage data files in the Immediate Access portion of the Staging Area and build out from there. The interface and resulting output should also be built with an architecture that could be extended to fully automating data requests—especially from the Immediate Access portion of the Staging Area. Ideally, the data request would be submitted, logged, and

[30] For example, see the data request submission system utilized by the Minnesota Population Center's IPUMS database (University of Minnesota: https://www.ipums.org/).

[31] See https://usa.ipums.org/usa-action/variables/group.

	the data extract or database instance of the data would automatically populate within the designated folder (without requiring a Data Manager to move it).
Staging Area Version Control	The Staging Area must maintain a clear version control system for data files and their contents, to include consistent file and field naming, labeling, and coding conventions. Records of these conventions must be maintained for all data objects, versions, and time periods made available at any point in the Staging Area to support reproducibility and data improvement. These metadata objects must be made available in the Library for all data items placed in the Immediate Access or Restricted Access Areas.
Data Usage Statistics	EDDIE must create usage statistics for the data objects in the Staging Area. These statistics should include how many times data elements were pulled into the research area and include a timestamp of the last extraction. These statistics, or other usage statistics that serve to inform users on which versions of data are most commonly used, must be incorporated into the available metadata.
Data Format	Data must be provided from the Staging Area in a format that allows users to perform data unpacking, loading, and transformation with the software listed in Table 10. Users must be able to specify the format of the data extract when requesting data from the Staging Area.
EDDIE Anonymized Identifier	All files in the Staging Area must be supplied with an EDDIE Anonymized Identifier that uniquely identifies an individual across all files in which that individual appears. This identifier must be constructed and applied in the Cold Room and must be unrelated to all other identifiers.
Geographic Identifiers	Where source data exist, EDDIE must include a rich collection of geographic fields for data objects placed in the Staging Area, including state, Metropolitan Statistical Area, Micropolitan Statistical Area, county, ZIP code, census tract, and latitude/longitude.
Broad Access to DMDC Data Elements	With the exception of Direct Identifiers, the EDDIE Immediate Access Staging Area must include all unclassified data fields in the files referenced in Table 6.
DMDC Data Frequency	DMDC data files must be included on a monthly or transactional basis in the Staging Area. For current data, updates to the monthly files included in the Cold Room and Staging Area should occur at least quarterly. Transactional files should include the broadest range of available dates for transactions.

DMDC Data Dates	To facilitate longitudinal analyses, all DMDC data files must extend to periods as early as possible (to the dates listed in Table 6 or the date of creation). Early files that contain incomplete data, data based on different definitions, or data with questionable accuracy must be included within EDDIE, with the areas of concern noted in the accompanying documentation. If DMDC has not curated an early file, but another organization (e.g., an FFRDC) has done so, EDDIE should seek to partner with that organization and include the curated early files in EDDIE.
Publicly Available Data	The Staging Area must be able to include data from publicly available sources commonly used in personnel research, such as the American Community Survey (ACS)[32] and the Current Population Survey (CPS).[33] EDDIE users must be able to submit requests for the inclusion of publicly available data in the Staging Area. Requests should be fulfilled within one working day. To the extent possible, the data must be accompanied by the appropriate metadata for library documentation.
Facilitate Incorporation of Data Submissions	EDDIE must facilitate incorporation of data submissions that contribute to its core functions from DOD, government, and FFRDC organizations. An organization submitting data must state that they are authorized to submit the data to EDDIE.
Establishing External Data Agreements and Flows	EDDIE should expand the data holdings available to users in the Staging Area by establishing data use agreements (or like arrangements) with organizations inside and outside of DOD, such as those listed in Table 7. This may include identifying and building on existing data use agreements between any of these organizations.

[32] https://www.census.gov/programs-surveys/acs/data/data-via-ftp.html.

[33] https://www.census.gov/programs-surveys/cps/data-detail.html.

Table 6. Data Files for Inclusion in the Staging Area

Master Personnel Files	Active Duty	1988 – current
	Reserve	1988 – current
	Civilian	1988 – current
Transaction Files	Active Duty	1988 – current
	Reserve	1988 – current
	Civilian	1988 – current
Contingency Tracking System (CTS)	Activations	2001 – current
	Deployments	2001 – current
Family Files	Active Duty	Earliest available – current
	Reserve	Earliest available – current
Pay Files	Active Duty	1991 – current
	Reserve	1988 – current
	Civilian	1995 – current
	Survivor	Earliest available – current
	Retired	Earliest available – current
Other Files	DEERS	2000 – current
	MEPCOM Recruiting Data	1988 – current
	Individual Training History	2002 – current
	PNL Extract	2000 – current
	WEX	1975 – current
	DHRA Force Status Surveys	
	Other DOD Personnel Files	

Table 7. Potential Organizations to Partner with for Data Agreements

Joint Advertising Marketing Research & Studies (JAMRS)
Defense Health Agency
Military Services
Defense Finance and Accounting Service
Defense Logistics Agency
Defense Threat Reduction Agency
Department of Health and Human Services
Department of Homeland Security
Department of Veterans Affairs
Census Bureau
Department of Education
Bureau of Justice Statistics
Bureau of Labor Statistics
Internal Revenue Service
Social Security Administration
Office of Management and Budget
Centers for Medicare and Medicaid Studies
Bureau of Economic Analysis
Agency for International Development
Federal Retirement Thrift Investment Board
State and local governments

F. Research Area

1. Institutional and Collaborative Folders

Table 8. Institutional and Collaborative Folder Requirements

Project Folders	At the request of a project leader, Access Managers must be able to create project folders within two business hours for projects that have been approved for inclusion in EDDIE. These project folders may be within their Institutional Folder or a Collaborative Folder. At a user's request, Data Managers must be able to export non-PII objects from project folders. Export requests must be fulfilled within two business hours.
Institutional Research and Development (IRD) Folders	Users must be able to operate within Institutional Research and Development (IRD) folders located in their Institutional Folder. Users must be allowed to use IRD Folders to conduct exploratory analysis (as such, work within IRD Folders must not require project descriptions). Full research projects must be conducted in project folders.[34] Users must be able to request the export of non-PII objects from IRD Folders.
Other Folders	Access Managers must be able to create additional folders for quick turnaround requests, operational analysis, or other purposes within two business hours. The Oversight Committee will determine the process for creating these folders.
Collaborative Folders	One or more users must sponsor each Collaborative Folder. After an EDDIE Access Manager creates a Collaborative Folder, management privileges for that folder may be assigned to one or more Institutional Access Managers at institutions affiliated with sponsoring users.
Folder Permissions	When requesting the creation of a folder, project leaders must be able to determine whether their project folder is searchable by users outside their institution, and whether their project folder is readable by users inside or outside their institution. Access Managers will assign the folder the requested categorization. Project leader will need to further specify a list of users with write (or higher administrative) privileges. For folders that have limited read privileges, project leaders will need to specify any additional users who will need read privileges.

[34] Exploratory analysis is useful, for instance, in identifying trends that would be worthwhile to explore in full research projects.

	There must also be a mechanism by which project leaders can request for an Access Manager to adjust both the folder's categorization and the detailed access control list of users at any time. Changes must be fulfilled within two business hours of the request.[35]
Folder Request Forms	To enable Access Managers to respond quickly and create folders, EDDIE must include folder request forms with a set of fields including project name; project leader; project description or statement of intent for Institutional Research and Development (IRD); categorization for whether the folder is searchable outside their institution, as well as readable inside and outside of their institution; and a list of authorized users that have read and/or write privileges. Access Managers must receive an email (or other automatically generated alert) when a new folder request form has been submitted.
Flexibility within Research Folders	Within their project, IRD, or other folders, users must be able to establish subfolder structures, naming conventions, version control mechanisms, and documentation standards that satisfy their needs.

2. Software

Table 9. Software Requirements

Research Tools	EDDIE must include software for conducting advanced analyses in statistics, econometrics, machine learning, qualitative analysis, etc. Users must be able to use these tools within their various Institutional and Collaborative Folders. A list of baseline research tools that must be included is in Table 10.[36] As user needs evolve, the list of baseline tools will also evolve
Collection of Packages	Python, R, Stata, LaTeX and other tools rely on numerous software add-ins commonly referred to as libraries or packages. EDDIE must maintain a standing and comprehensive collection of these objects and make them available for all users. If a user requires an additional package or a version of a package not contained in the current collection, a user may request its addition to the collection. EDDIE Help Desk Team must add requested packages within three business hours, pending a virus scan and security approval.

[35] As stated earlier, DOD research sponsors maintain ultimate discretion over whether and when work they sponsor is released to a broader audience.

[36] A few of the software tools listed in Table 10 are available for Windows, but not for Mac or Linux. If an exclusively Linux environment is selected for EDDIE, software such as Wine (https://www.winehq.org/) or CrossOver (https://www.codeweavers.com/) may permit the use of Windows software. Virtual machines may also be an option.

New Software Installation	Users must be able to request installation of software programs and packages. Because of licensing restrictions and software costs, some software may only be available to certain users or to certain project folders. EDDIE must fulfill software installation requests within one business day.
Customizable Environments	Users must be able to customize their work environments to support concurrent use of multiple versions of the same software (e.g., Python 2.7, Python 3.4, and Python 3.6).
User-Specific Software Licenses	If a project requires use of software that EDDIE does not provide, users must be able to bring their own license and use software licensed specifically to them within EDDIE, pending a virus scan and security approval. EDDIE must fulfill software installation requests (or inform the requesting user that the software does not satisfy the security requirements) within one business day.[37]
Software Security Requirements	Software installed in EDDIE must comply with appropriate security requirements. The Oversight Committee may provide rules and requirements regarding security standards and software installation.
Workflow Management System	To automate, track, and ensure that work requests in EDDIE are expedient, EDDIE must have an automated workflow management system. The system must be able to facilitate the workflow of new project requests; the human subject review process; metadata submissions; import and export requests; institutional access requests; software installation requests; helpdesk requests; Cold Room data merge requests; and all other workflows within EDDIE.[38]

[37] Security vetting procedures may take longer in some cases. When that occurs, the personnel evaluating the software should apprise the requesting user within one business day of the expected timeline and keep the user abreast of any progress or deviations from that timeline.

[38] A description of what constitutes a workflow can be found at: https://kissflow.com/workflow/what-is-a-workflow/.

Table 10. Software to Include in EDDIE's Baseline Tools

Software	Cost	Notes	Website
Scientific Computing			
Python 2	Free	Version 2.7.15 or latest	https://www.python.org/downloads/
Python 3	Free	Version 3.6.5 or latest	https://www.python.org/downloads/
Anaconda	Free	Each Python version should contain the Anaconda distribution of data science packages	https://www.anaconda.com/download/
R (MRO distribution)	Free	Version 3.5.0 or latest	https://www.r-project.org/
CRAN R Packages	Free	R should include all packages available from CRAN	https://cran.r-project.org/web/packages/available_packages_by_name.html
GNU Octave	Free	Open source alternative to MATLAB	https://www.gnu.org/software/octave/
JuliaPro	Free		https://juliacomputing.com/products/juliapro
ArcGIS	$		https://www.arcgis.com/features/index.html
GAMS	$		https://www.gams.com/
Mathematica	$		https://www.wolfram.com/mathematica/
SAS	$	Version 9.4 or latest	https://www.sas.com/en_us/software/sas9.html
Stata	$	30 perpetual licenses of Stata/MP 2-core (or higher)	https://www.stata.com/order/new/gov/network-licenses/dl/
StatTransfer	$		https://stattransfer.com/
World Programming System	$	SAS language compiler that is interoperable with R, SQL, and Python	https://www.worldprogramming.com/us/products/wps-analytics-platform
Qualitative Analyses			
QDA Miner Lite	Free	Qualitative Data Analysis software	https://provalisresearch.com/products/qualitative-data-analysis-software/freeware/

Software	Cost	Notes	Website
Editors and IDEs			
PyCharm	Free	IDE for Python	https://www.jetbrains.com/pycharm/
RStudio	Free	IDE for R	https://www.rstudio.com/
Notepad++	Free		https://notepad-plus-plus.org/
Atom	Free		https://atom.io/
Git	Free	Version control system	https://git-scm.com/
Database			
A database management system		Must facilitate the volume and types of DMDC and other data held in EDDIE	
SQLite	Free		https://www.sqlite.org/index.html
Scientific Writing			
Adobe Acrobat Reader	Free		https://get.adobe.com/reader/
Adobe Acrobat Pro	$		https://acrobat.adobe.com/us/en/acrobat/acrobat-pro.html
MiKTeX	Free		https://miktex.org/
Texmaker	Free	LaTeX editor (integration with MiKTeX)	http://www.xm1math.net/texmaker/
Forum and Workflow			
Stack Overflow for Enterprise (or alternative)	$	User forum for questions and answers	https://stackoverflow.com/enterprise
Box (or alternative)	$	Workflow management system	www.box.com/collaboration/relay-workflow
Office Suite			
MS Office (or alternative, such as WPS Office)	$	MS Access, Word, Excel, and PowerPoint or alternatives compatible with MS	

3. Library

Table 11. Library Requirements

EDDIE Library	EDDIE must have a searchable Library containing public code (such as programming scripts written and submitted by EDDIE users), data dictionaries, metadata, a wiki collecting institutional knowledge, and a repository of user profiles.
Best Practice Documentation	The documentation of all publicly searchable items in the library must follow industry best practices and standards. Publicly searchable items include metadata and data dictionaries on all data in the Immediate Access and Restricted Access portions of the Staging Area; all code, reports, and other objects stored in the EDDIE Library; subfolders within a user's Institutional Folder; and subfolders in other Institutional and Collaborative folders that have been marked as searchable. Best practices include using an eXtensible Markup Language (XML) schema to tag text and enable greater searching and linking capabilities.[39]
Public Code	The EDDIE Library must have the ability to store programming scripts and other code submitted by EDDIE users. Public code must be linked or tagged to any raw data it uses (such as input data files from the Staging Area), as well as to any data in the Staging Area or products in the EDDIE Library that it generates. Public code must be tagged with a title, author(s), affiliations, key words, phrases, versions, etc. Data Librarians will ensure that shared objects fulfill these requirements.

[39] Examples of best practices and standards for metadata documentation include:

- Dublin Core Metadata Initiative (DCMI)—established a widely-adopted core set of 15 metadata elements: contributor, coverage, creator, date, description, format, identifier, language, publisher, relation, rights, source, subject, title, and type.
- The Library of Congress' Metadata Encoding and Transmission Standard (METS)—provides an XML schema for documenting header, descriptive, administrative, file section, and other forms of metadata. See http://www.loc.gov/standards/mets/ and http://www.loc.gov/standards/mets/METSOverview.v2.html.
- Document, Discover and Interoperate (DDI)—provides a free XML schema that may be used in conjunction with METS for "describing social and behavioral science data and data in related domains," and which has been implemented in numerous data repositories and archives. See DDI's frequently asked questions page (http://www.ddialliance.org/resources/faq.html). Further description of DDI are available at http://www.ddialliance.org/ and http://www.ddialliance.org/training/why-use-ddi, with examples of repositories using the DDI standard at http://www.ddialliance.org/ddi-adopters. The METS Editorial Board has endorsed DDI and other XML schemas that may be used in conjunction with METS (http://www.loc.gov/standards/mets/mets-extenders.html).

Data Dictionaries and Metadata	The EDDIE Library must have data dictionaries and metadata on all data in the Immediate Access and Restricted Access portions of the Staging Area. To the extent possible, code should be used to scrape files and automate the production of data dictionaries and metadata. Data dictionaries and metadata must be tagged with a title, author(s), affiliations, key words, phrases, etc. Data Librarians will ensure that shared objects fulfill these requirements.
Automated Metadata Submission Forms	To enable consistent practices for metadata documentation that foster enhanced metadata curation and library searches, EDDIE must include automated metadata submission forms with a set of required fields and optional fields. Submission forms must be able to automatically convert each entry into XML for inclusion in the library. Forms may differ based on the type of submission (e.g., data, code, report).
Wiki	The EDDIE Library must store metadata in a wiki. Users must be able to collaboratively edit the contents of this wiki.
Search Feature	To leverage work done in previous analysis, EDDIE must include a search feature that searches for items across folders. Users can search across folders for which they have search access permissions. The search feature must include capability to search by item name, title, author(s), affiliations, key words, phrases, etc. When a user searches for an object, the search feature should return the item, its last edit date, and the number of users that have pulled the object, and users should be able to sort the results by each of these fields.
User Forum	The EDDIE library must include a forum that enables EDDIE users to submit questions, problems, comments, and suggestions about the data to Data Librarians, data providers, or the greater EDDIE community. The user forum must enable users to tag objects from the staging area or research area within conversations. The user forum must send email notifications to users who follow specific conversations. A user will follow a specific conversation if they contribute to the conversation, if they are an author of one of the tagged objects, or if they have opted to follow conversations on research topics associated with the tagged object.

User Profiles	The EDDIE Library must include a repository for users to create optional personal profiles, including at a minimum the user's name, contact information, education and credentials, institutional affiliation, analytic interests, data interests, and publications within and external to EDDIE.[40] Users must also be able to tag research topics that they would like to automatically follow in User Forum conversations.
Copying Library Items	EDDIE users must be able to copy objects from the Library into folders within the Research Area. To the extent possible, users should be able to copy the objects in a variety of formats (e.g., txt, pdf, csv).

[40] Given the large number of organizations that support defense personnel research (analytic organizations throughout the military services and the Office of the Secretary of Defense, as well as the FFRDCs), and given that many of these organizations are siloed, basic researcher profiles may ease the process of developing cross-organizational peer review and collaboration.

G. General Environmental Resources

1. Computational Resources

Table 12. Computational Resource Requirements

Expandable	EDDIE computing resources must be able to flexibly expand to meet a growing user base, escalating volumes of data, and increasingly computationally intensive analyses.
User Base	EDDIE must accommodate an initial user base of at least 500 people from DOD offices and FFRDCs.
Concurrent Usage	EDDIE must support a high degree of concurrent usage. At peak times, an estimated one-fifth of EDDIE users will use EDDIE concurrently.
Monitoring System to Track Usage	To enable the Oversight Committee and EDDIE Technical Support to track usage patterns, plan for growth, and minimize performance impacts from concurrent users, EDDIE must have a monitoring system to track computing resources used by projects and users. The monitoring system must also be able to audit usage; enforce security requirements; and enable all imports, exports, and user permission changes to be trackable and auditable.
Ability to Run Long Programs	EDDIE must support the ability for users to run programs lasting several days. EDDIE must not time these users out. EDDIE should send an email notification or text message every four hours for the duration a program is running on a user's account. This will allow users to check that programs are still running without needing to log in to EDDIE, and also remind users whose programs may be running by mistake to close the program.
Ability to Meet Surge Demands	EDDIE must have the ability to meet surge computing demands. From the user's perspective, EDDIE's computing performance must not decrease by more than 25% during peak load times (e.g., an analysis requiring 60 seconds to complete under EDDIE's median computing demand conditions should never take longer than 75 seconds to complete).
Scheduling and Queuing	If a scheduling and queuing system is used to smooth surge demands for large computing runs, EDDIE must have an automated system (such as automated email alerts) that notifies users when their runs are submitted to the queue, when they start and stop, and the computation resources used.

User Interface	The EDDIE user interface must have an intuitive, familiar feel. It should also be responsive, with no time lag between when commands are typed and when the system responds (i.e., no delay between keystrokes and the display of text). EDDIE users must also be able to view the same EDDIE session from multiple monitors simultaneously.
Graphical and Parallel Computing	EDDIE must support EDDIE graphical and parallel computing.
Computing Specifications	EDDIE must have computing resources that produce a user experience (with respect to computing speed, threads, throughput, RAM, and disk capacity) that meets or exceeds the user experience produced by the technical specifications described in Table 13. Note: The requirement here is for computing capability rather than precise specifications for particular servers. Elastic cloud computing, for instance, may provide a means to satisfy these computing capabilities.

Table 13. EDDIE Computing Resources

	Minimum	Ideal
Quantity of servers	8	12
RAM per server	256 GB	512 GB
Storage per server for Research Area	30 TB	60 TB
Storage per server for Cold Room and Staging Area	Sufficient for data holdings	
CPU cores per server	32	64
GPU cores per server (GGDDR5)	4	16
Simultaneous multithreading (two threads per core)		
64-bit Operating System		

2. Technical Support

Table 14. Technical Support Requirements

Round the Clock Availability	EDDIE must be available and operable to users 24 hours a day, 7 days a week. Maintenance outages must be communicated to all users no less than five business days in advance and be scheduled outside of peak use periods (8:00 AM to 8:00 PM Eastern Time, which covers the 8:00 AM to 5:00 PM work day throughout the continental U.S.). Outages must be limited to no more than once per week and for no more than six hours at a time.

Central Touchpoint	EDDIE must provide users with a central touchpoint for all forms of help. This assistance function must be able to route all help tickets and requests within 30 minutes to the appropriate party (e.g., IT support, system administrative support, and help from Data Managers, Data Librarians, Access Managers, the HSR coordinating team). Live support must be available in the central touchpoint function from 8:00 AM to 8:00 PM Eastern Time (Monday through Friday, excluding federal holidays), with on-call support during off hours (for emergency support, such as unscheduled system outages, security breaches, etc.); and it must be accessible in multiple ways, including phone and email. To support rapid IT support, the central touchpoint function may be organized within IT support.
Real Time IT Support	EDDIE IT support must address problems in real time with a goal of resolving each problem within two business hours.
IT Support Availability	EDDIE IT support must be available from 8:00 AM to 8:00 PM Eastern Time, with on-call support during off hours.
IT Support Tasks	EDDIE must support ongoing and responsive helpdesk capabilities for problems such as system access issues, connectivity issues, permission issues, password resets, requests for additional computing resources and disk space, software installation, and other technical support issues.
Remote Capability for IT	To enable better troubleshooting in difficult technical situations, IT support must be able to remotely access user sessions within EDDIE.
System Administrative Support	EDDIE must support ongoing and responsive high-level system administration functionality sufficient to sustain an advanced research environment, including support for Linux, Hadoop, Spark, and Windows environments; implementing hardware, operating system (OS), and software updates, upgrades, transitions, and security patches in a manner which sustains operations and system functionality; system outage recovery and repair; backups; and other advanced technical support functions. System administrators must be capable of ensuring ongoing cybersecurity compliance consistent with Constraint C2.
Ability to Track IT Support Performance	EDDIE must have the ability to track IT support performance to promote effectiveness. This must include the ability to track how long it took to close a ticket from the time it was opened, as well as the ability to request feedback from users as to how their requests were handled.

3. Import and Export Control

Table 15. Import and Export Control Requirements

Import and Export Control	EDDIE must have means for Data Managers to inspect, regulate, and approve or deny the flow of files entering or exiting EDDIE. Files must be quarantined until a Data Manager can determine whether they contain PII.
File Import and PII determination	EDDIE users must be able to submit requests for files to be imported into EDDIE by a Data Manager and indicate the user's judgement of whether the items requested for import are non-PII, PII with no Direct Identifiers, or PII with Direct Identifiers. Data Managers must make their own determination of the PII status within one business day of the user submission. Items that are non-PII or PII without Direct Identifiers must be imported within one business day. Import requests for PII items with Direct Identifiers must be fulfilled within five business days.
Import of PII with Direct Identifiers	All import files containing PII with Direct Identifiers must be imported through EDDIE's Cold Room.
Import of non-PII and PII with no Direct Identifiers	Import requests must have an appropriate destination for the import files (this may be an Institutional Folder, Collaborative Folder, or the Staging Area). Data Managers will place import files containing non-PII and PII with no Direct Identifiers in the requested destination. If the import files contain data to be made available in the Staging Area, the import request must be accompanied by related metadata for inclusion in the EDDIE library catalog.
File Export	EDDIE users must be able to submit requests for non-PII items to be exported for download outside of EDDIE from their Institutional or Collaborative folders, and from the EDDIE Library. Export requests must be fulfilled by a Data Manager within two business hours, from 8:00 AM to 8:00 PM Eastern Time. Items approved for export are placed in a network-accessible location for pickup.
Exportable File Designation	To facilitate the repeated export of files over time, files that have already been cleared as exportable must be able to be marked or stamped in such a way that (so long as the file remains unaltered) a data manager could export it again without further review.

4. Human Subjects Review

One goal of EDDIE is to reduce inefficiencies in the Human Subjects Review (HSR) process. This may entail changes to DOD policies and procedures. Procuring and establishing the EDDIE infrastructure, support personnel, and other requirements should not be delayed or postponed while these policy changes are resolved. However, EDDIE should satisfy the following requirements.[41]

Table 16. Human Subjects Review: High-Level Requirements

Streamlined HSR Process	EDDIE must have a streamlined HSR process designed for quick turnaround. Specifically, EDDIE must facilitate the HSR approval process depicted in Figure 2.
EDDIE HSR Coordination Office (EHSRCO)	EDDIE must have a responsive HSR coordination office that is empowered to shepherd projects in EDDIE through the HSR process and approve projects for inclusion in EDDIE.
Sanctioned IRBs	To eliminate redundant IRB approval processes, EDDIE must establish clear criteria under which an IRB is sanctioned to issue determinations on projects in EDDIE. Determinations issued by a sanctioned IRB must be accepted by the EDDIE HSR Coordinating Office (through an IRB reciprocity agreement or other similar mechanism).[42,43]

[41] Analysts conducting research on DOD personnel-related topics often spend a significant portion of study time navigating the human subjects review process. Even for routine studies following generally accepted protocols, these obstacles can delay studies for months, consume research budgets, and compress the time for actual analyses. Furthermore, it is not uncommon for a single project to be subject to duplicative Institutional Review Board (IRB) and Human Research Protection Office (HRPO) reviews. This is especially true for projects that involve multiple military services or multiple DOD offices, because reviews may not be recognized across organizations.

[42] The problem of redundant IRB requirements is not unique to the DOD and often arises when multiple institutions are involved in a single study. In recent years, there has been a growing push to reduce duplicative IRBs through reciprocity agreements. One such example is the Harvard Catalyst Master Reciprocal Common IRB Reliance Agreement, which established common processes and procedures to enable IRB reciprocity across more than 20 distinct legal entities. The DOD, which has numerous OSD and Service offices that are involved in human subjects review, could mirror the Harvard Catalyst and other successful frameworks for streamlining the IRB review process. See Winkler, S.J., Witte, E., Bierer B.E. (2015). The Harvard Catalyst Common Reciprocal IRB Reliance Agreement: An Innovative Approach to Multisite IRB Review and Oversight. *Clin. Transl. Sci.* 8(1), 57–66.

[43] The DOD Department of Navy previously created an "Addendum to the Department of Health and Human Service's Federal Wide Assurance (FWA) for the Protection of Human Subjects." This addendum contained a listing of all DOD-wide and component-specific regulations for human subjects research. Institutions that already had an FWA and were consenting to follow DOD regulations could sign this addendum. Following this example, a sanctioned IRB for EDDIE may be an IRB that has signed such an addendum. That may be a first step toward a full IRB reciprocity agreement.

EDDIE IRB	EDDIE must have a centrally designated IRB that can review projects, specifically when an Institution does not have an established HSR process including an IRB relationship. The EDDIE IRB must adhere to the sanctioned IRB criteria.
Exempt Determination	EDDIE must permit a two-track process accomplishing HSR, as illustrated in Figure 2. If an Institution approved for Institutional Access in EDDIE has an HSR process, EDDIE HSR will rely upon the Institution's HSR determination of whether (1) a project constitutes research on human subjects, (2) a project is exempt from human subjects review, and (3) any related IRB determinations or criteria. The Institution's HSR process must issue a memo to EHSRCO stating the determination. EHSRCO must approve projects providing Institutional HSR approval memos for inclusion in EDDIE within one business day. If a user does not have access to an HSR process at an approved Institutional Access Institution, the user may submit the project to EHSRCO for review. EHSRCO must then determine whether (1) the project constitutes research on human subjects, or (2) the project is exempt from human subjects review. EHSRCO must then issue a memo stating the determination, and approve the project to proceed in EDDIE as appropriate to the memo determination. This entire process must take place within two business days.
IRB Determination	EDDIE must permit a two-track process for research that qualifies for expedited or full board IRB review. • An institution must be permitted to send non-exempt projects involving human subjects to a sanctioned IRB, and the determination of that IRB must be accepted by EHSRCO (through an IRB reciprocity agreement or other similar mechanism). If the sanctioned IRB approves the project, the IRB approval is submitted to the EHSRCO. EHSRCO then files the IRB approval. If Research Regulatory Oversight Office (R2O2) review is required, EHSRCO sends the IRB approval and other required project documentation to R2O2 within one business day. Otherwise, EHSRCO approves the project for inclusion in EDDIE within one business day. • Users must be able to submit non-exempt projects involving human subjects to the EDDIE IRB for review. If R202 review is required, EHSRCO sends the EDDIE IRB approval and other required project documentation to R2O2 within one business day of EDDIE-IRB approval. Otherwise, EHSRCO approves the project for inclusion in EDDIE within one business day.

Minimum Statutory HSR Review by R2O2	For EDDIE projects that qualify for expedited or full board IRB review, and that also require R2O2 review, R2O2 shall subject these studies to no more than the minimum statutory requirements for human subjects review. If R2O2 does not approve the project, then the user must be able to modify the project and begin the EDDIE HSR approval process again. If R2O2 approves the project, EHSRCO files the determination and approves the project for inclusion in EDDIE within one business day.
IRB Umbrella Protocols	To the greatest extent possible, EDDIE HSR process must permit and facilitate IRB Umbrella Protocols.[44]
EDDIE HSR Process Acceptance by Data Providers	Contractual relationships establishing data provision terms between EDDIE and data providers should make every effort to stipulate that the data provider will defer to the EDDIE HSR process.
Other EHSRCO Tasks	EHSRCO must assist users and projects in navigating any and all HSR hurdles and criteria. It must maintain an HSR project status dashboard for all projects moving through the HSR process to track performance statistics and keep study members informed of progress (this dashboard may be the Workflow Management System, listed as a requirement in the Software section).

[44] An umbrella protocol (or umbrella project approval) streamlines the human subjects review process by enabling an IRB proposal to be submitted for a class of studies, rather than for each individual study. They are used when multiple studies employ similar methodologies for interacting with human subjects, pose similar risks to human subjects, and draw on a common pool of human subjects. Umbrella protocols may be a means to facilitate blanket HSR coverage for classes of tasks within IRD Folders and other types of folders (such as quick response folders).

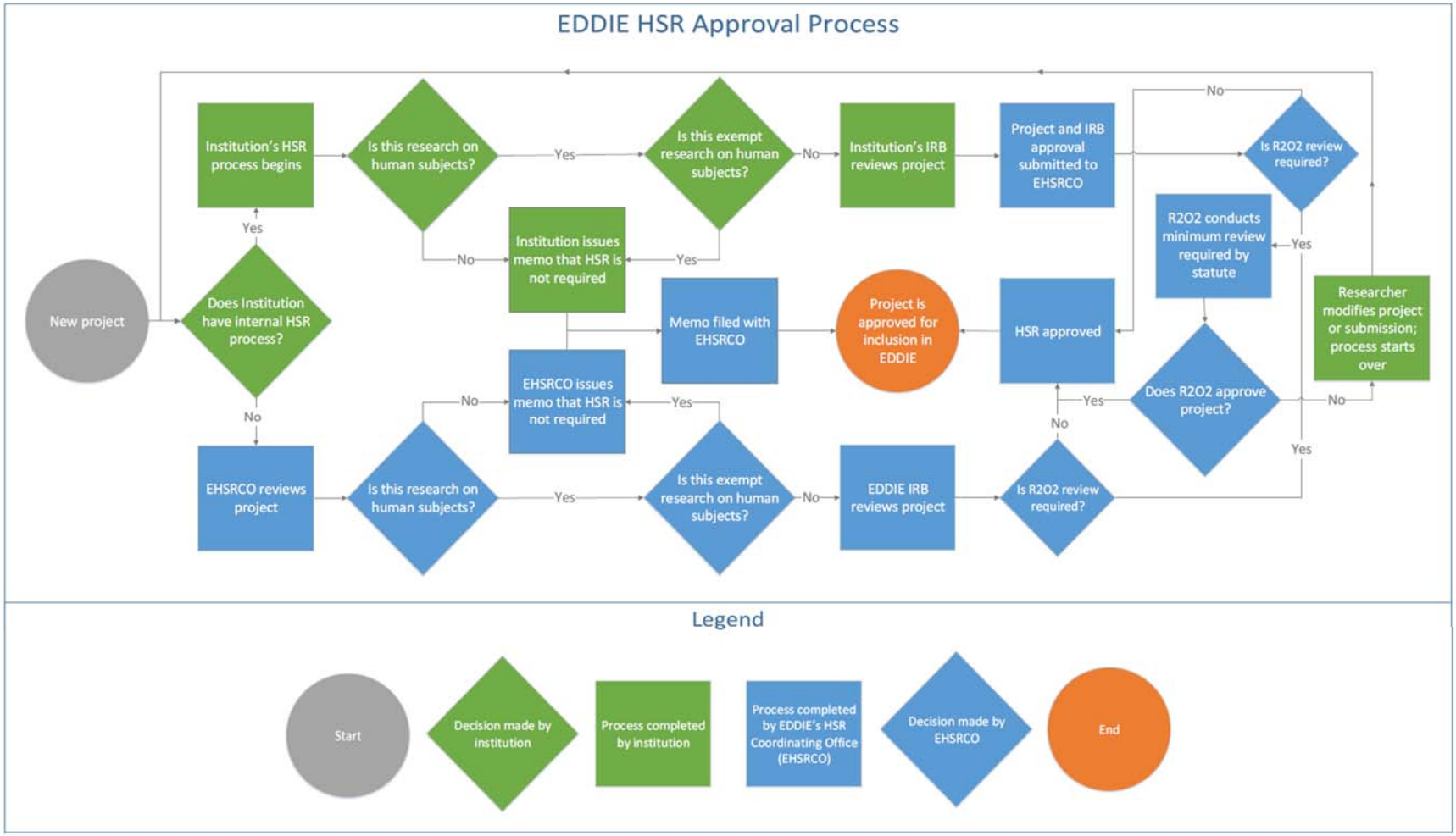

Figure 2. EDDIE HSR Approval Process

H. Summary of Request and Response Timelines

Timelines for completing various tasks within EDDIE are scattered throughout this requirements document. We collect them here in Table 17 for ease of reference. We also provide some brief context into the generation of these timelines.

Many analyst stakeholders in the defense personnel research community have experience working in data hosted environments that reside outside of their institution (the Army's Person-event Data Environment (PDE) is a prominent example). A common caution and concern voiced by stakeholders is that the benefit of providing access to data in a secure environment can be negated with excessive barriers to entering and operating within the environment. The research process can be a highly iterative, and each step that research teams cannot perform by themselves introduces delays and potential inefficiencies. Delays are compounded each time a particular step is repeated and can be particularly costly to research budgets and timelines.

The intent of the time expectations herein for fulfilling various requests is to apply lessons learned from other data hosted environments to ensure that EDDIE is responsive enough to capably facilitate quick turnaround work. A further consideration in developing these time expectations is that EDDIE may not always have sufficient personnel to fulfill requests at the pace or with the priority that institutions require. Institutional Access Managers and Institutional Data Managers have been specifically added to the design of EDDIE to permit institutions to align their own resources (in terms of personnel time that they are paying for) with the research commitments that they are responsible for executing.

As the operation of EDDIE unfolds, it will be within the purview of the EDDIE Oversight Committee to revisit these timelines and change them as necessary.

Table 17. Estimated Time to Complete Requests

Request Type	Request Submitter	Request Approver	Implementer	Time to Approve / Implement
Data Access (Staging Area Extraction)				
Immediate Access Data (SPA for Info.)	User	Data Manager	Data Manager	Immediately / 1 day
Restricted Access Data (SPA for Concur.)	User	Data Manager, Data Owner	Data Manager	3 weeks* / 1 day
Private Data (Necessary Approvals)	User	Data Manager, Data Owner	Data Manager	3 weeks* / 1 day
Research Area Folders				
Folder Creation within Institutional & Collab. Folders	Project Leader, EHSRCO	NA	Inst. Access Manager, IT	NA / 2 hours†
Copying Non-PII Between Research Folders	User	Project Leader	Data Manager	1 hour / 1 hour
Copying Non-PII from Library to Research Folders	NA	NA	User	NA / Immediately
Copying Non-PII from Research Folders to Library	User	Project Leader, Data Librarian	Data Manager	1 day / 1 hour
User Access				
Institutional Access	Institution	Oversight Committee (or designee)	EDDIE Access Manager, IT	3 weeks / 1 day
New User Account	User	Inst. Access Manager	Inst. Access Manager, IT	2 days$ / 2 days
Folder Permission Changes	Project Leader	Inst. Access Manager	Inst. Access Manager, IT	1 hour / 1 hour

Request Type	Request Submitter	Request Approver	Implementer	Time to Approve / Implement
Scrubbing & Merging PII with Direct Identifiers				
Cold Room Merge	User	Data Manager	Cold Room Personnel	Immediately‡ / 1 day◊
De-identifying PII data	NA#	NA	Cold Room Personnel	NA / 1 day◊
Extracting PII which does not contain direct identifiers to staging area	NA#	NA	Cold Room Personnel	NA / 1 day
Technical Support				
Central Touchpoint Routing	Anyone	NA	Central Touchpoint	NA / 30 minutes
IT Support	Anyone	NA	IT	NA / 2 hours§
Software Installation	User	SysAdmin	IT, SysAdmin	12 hours / 12 hours∆
Package Installation	User	SysAdmin	IT, SysAdmin	2 hours / 1 hour∆
Import and Export Control				
Import – Non-PII	User	Data Manager	Data Manager	12 hours / 12 hours∆
Import – PII without Direct Ident.	User	Data Manager	Data Manager	12 hours / 12 hours∆
Import – PII with Direct Ident.	User	Data Manager	Cold Room Personnel	12 hours / 5 days∆
Export of Non-PII items	User	Data Manager	Data Manager	1 hour / 1 hour
Human Subjects Review				
Exempt Determination – After Institution's HSR	Project Leader	EHSRCO	EHSRCO	1 day / 1 hour
Exempt Determination – EDDIE's HSR	Project Leader	EHSRCO	EHSRCO	2 days / 1 hour
EDDIE IRB Determination – Expedited	EHSRCO	EDDIE IRB	EDDIE IRB	2 weeks / 1 hour

Request Type	Request Submitter	Request Approver	Implementer	Time to Approve / Implement
EDDIE IRB Determination – Full Board	EHSRCO	EDDIE IRB	EDDIE IRB	4 weeks / 1 hour
IRB approved submissions to R2O2	EHSRCO	NA	EHSRCO	NA / 1 day
EDDIE approval after R2O2 approval	EHSRCO	NA	EHSRCO	NA / 1 day

Notes:

* Contingent on data already being in the staging area and on the data owner's access approval process. Although the data owner's access approval process is outside of EDDIE's direct control, EDDIE should encourage data owners to have approval processes that are no longer than this.

† Contingent on EHSCRO already approving project, if necessary

$ Contingent on institution's user approval process

‡ Contingent on user having approved access to involved data elements

◊ Contingent on the state of the data; the process needs to begin within one day and Cold Room personnel should provide a more accurate estimate at that point

Process started automatically and immediately as PII with direct identifiers enter the cold room

§ Contingent on type of support required, but IT should follow up with users regarding their request and a more accurate time frame within two hours

∆ Contingent on any purchasing or licensing from requirements being satisfied and the data or software passing a virus scan

5. Use Cases

The scenarios presented in these use cases are formulated as narratives that describe representative research projects and highlight the range of activities that may be performed in EDDIE if the requirements in this document are satisfied. Each covers the objectives of the team, the research process, and the outcomes that define success. Completed, current, and anticipated research efforts articulated by the defense personnel research community inform each of the presented scenarios. Research scenarios beyond the scope of EDDIE are not included.[45]

A. Use Case 1: Research Requiring DMDC, Other DOD, and Publicly Available Data

1. Problem

A government research sponsor wants to understand the costs and impacts of potential policies designed to achieve an enlistment mix more representative of the geographic and demographic diversity of the nation's population. The project team tasked with conducting the research completes an initial assessment of the research question, develops a research plan, and identifies the general data required for the analysis. The team concludes that data will be required from DMDC, DOD's Joint Advertising Marketing Research & Studies (JAMRS), and the military services. The project also requires that DOD data be combined with publicly available socioeconomic and demographic data sets, including two Census Bureau products—the Current Population Survey (CPS) and the American Community Survey (ACS)—and the Integrated Postsecondary Education Data System (IPEDS), maintained by the Department of Education. This narrative assumes that, of these data resources, only the DMDC, ACS, and CPS data reside in EDDIE at the time the project begins.

[45] Research projects using sensitive data (for example, information on sexual assault or harassment) may preclude the use of government systems, and are therefore ineligible for inclusion in EDDIE. Likewise, PII will not leave EDDIE, so projects requiring the transmission of PII to third parties are ineligible for operation in EDDIE. The research environments operated by the FFRDCs separate from EDDIE will continue to support projects that cannot be conducted on government systems.

2. Project Goals

- Obtain access to and collect the required data sets in an efficient and timely manner (including granular geographic and demographic data on both current enlistees and the U.S. population)
- Clean and prepare the data sets for analysis, and permanently store the resulting files (together with metadata and the code for generating the files)
- Complete the analysis in a rigorous, reproducible, timely, and efficient manner
- Report the results to the sponsor in a written report and a briefing
- Publish the results, model, and associated tools for other analysts and government sponsors to access, review, and refine

3. User and Sponsor Objectives

Analysts in EDDIE wish to:

- Gain immediate access to data dictionaries for potential data sources
- Gain access to all data from DMDC that the analysts deem necessary to support this research with minimal administrative delay
- Request and gain access to data supporting the research from other DOD or federal sources in an efficient and timely manner
- Import data from appropriate public data sources with minimal administrative delay
- Merge PII and non-PII from multiple sources
- Clean and develop the data for analysis
- Find and reuse relevant data sets previously generated by analysts performing related work, as well as any corresponding code and documentation
- Use appropriate software tools to conduct the analysis
- Run analysis algorithms to completion within a reasonable time
- Generate charts and graphs to communicate results effectively
- Prepare and deliver final research products
- Enable other research groups to reproduce results
- Submit work for peer review

Sponsors wish to:

- Obtain high quality results in a timely fashion
- Ensure that research time and money are spent on the research itself and not on navigating administrative hurdles
- Control access to project work spaces and research results

4. Project Initiation Triggers

- Sponsor authorizes the research study
- Institution performing the research (government or FFRDC) assigns EDDIE-credentialed users to perform the study

5. Scenario

A project leader (who is an authorized EDDIE user) submits a request to bring the project into EDDIE. The project qualifies as research on human subjects. The project leader's institution submits the project to an EDDIE sanctioned IRB. The project is subject to expedited IRB review and is approved. The project leader submits the IRB approval to the EDDIE HSR Coordination Office (EHSRCO), which acknowledges the approval of the sanctioned IRB. EHSRCO determines that review by the Research Regulatory Oversight Office (R2O2) is not required. Consequently, EHSRCO immediately approves the project for work in EDDIE and communicates this information to the institution's Access Manager. The project leader sends the Institutional Access Manager a list of users who need access to the folder and their respective permissions. Within two hours, IT support creates the project folder and implements the access permissions designated by the Institutional Access Manager.

Analysts are able to access EDDIE from their office computers. Even prior to the creation of their project folder, they are able to review data dictionaries and metadata in the EDDIE library catalog—including information from users who have used the data previously—to identify the most relevant data fields to access. They file a SPA for information to gain access to the relevant data from the immediate access portion of the staging area. The Data Manager (or an automated process) then pulls the data listed in the SPA from the staging area into the project folder as a SQLite database, as requested by the analysts. The analysts write a SQL program to perform initial data processing operations (such as checking for problematic data formatting and duplicate entries). The results are saved to the project folder.

The analysts request additional data from JAMRS and the military services. Data requests from multiple organizations are typically time consuming and require extensive coordination. Although a stated goal of EDDIE is to facilitate data use agreements and expand its data holdings, the data from JAMRS and the military services are not yet in EDDIE at the time of this study. The EDDIE Data Librarians in their role as liaisons between analysts and data providers are actively involved in the negotiations for bringing the data from JAMRS and the military services into EDDIE. The negotiations for the JAMRS data successfully conclude after one month. The data from the services takes considerably longer, with negotiations with one component spanning six months or more. As the data providers approve the requests, the Data Librarian arranges for the data (which contains Direct Identifiers) to be securely transmitted to the EDDIE cold room.

Once the JAMRS data and data from two of the military services have been received, the research team requests that these data be merged with each other and with some relevant DMDC data on the person level. This requires the use of social security numbers, names, and dates of birth. The Cold Room personnel perform the merge in the cold room. During the merge process, the Cold Room personnel communicate with the research team to clarify some methodological details for how the team would like the merge to be performed (for instance, the Cold Room personnel may need to clarify how the research team would like duplicate records handled in the merge). Once the merge is complete, the Cold Room personnel prepare a copy of the merged data that removes any Direct Identifiers. They then securely transfer this copy into the private data portion of the Staging Area. A Data Manager then moves the de-identified data from the Staging Area into the project folder (as a SQLite database, as requested by the analysts). This merging process will be repeated when the remainder of the military service data is received. In the meantime, however, the project team needs to move forward with analyzing the data it has received.

The analysts now need to incorporate publicly available data sets from the Census Bureau (the ACS and CPS). They identify a script in the EDDIE Library that transforms a raw ACS file already stored in the Staging Area and produces a clean data file of socio-economic variables for 2006–2015. The analysts request the raw ACS file from the Staging Area, and a Data Manager moves it to the project folder. The analysts use the script from the EDDIE Library to produce the cleaned data file.

The analysts determine that the most appropriate software platform for conducting the statistical analysis is R, which is already installed in EDDIE for all users to access. However, RStudio, an open-source integrated development environment for R, is not installed in EDDIE. The research team submits a request for RStudio to be installed, and IT support installs it for them within one business day.

The analysts continue to work within the project folder, saving files and creating subfolders within the project folder as necessary. Throughout the research process the research team exports intermediate outputs to present to the sponsor and reviewers outside of EDDIE. Data managers review each request to export non-PII within two hours. In the days leading up to a major deadline, the team iterates rapidly, performing analyses and requiring exports of successive non-PII results at the rate of four times per day.

Once the project is over, the sponsor approves the project for public release. The major research products (the report, modeling, and code objects) are submitted to the EDDIE Library.

6. Measures for Success

- Streamlined human subjects review process
- Timely data access

- Data from external sources may be added as needed
- Data sets may be cleaned, prepared and saved with metadata tags supporting catalog search
- Existing data sets may be discovered through the catalog, with lineage, and retrieved for reuse
- Software needed for analysis activities is already installed or is available within one day of user request
- Access to project folder is controlled by the project leader
- Transactional time in waiting for Data Managers, Access Managers, Cold Room Personnel, and other EDDIE personnel is minimized (both in terms of the number of interactions that are needed and the time required for each interaction)

B. Use Case 2: Research Requiring Collaboration among Different Institutions

1. Problem

A government research sponsor wants to evaluate the impact of particular retirement incentives on changes in the level of voluntary separation. The project requires data that follows individuals throughout their military career for all ranks of service members—a time span covering at least 30 years. The sponsor approaches multiple FFRDCs because the required modeling and subject matter expertise is not resident in a single institution. The sponsor funds a project requiring the collaborative efforts of three FFRDCs that operate Study and Analysis Centers.

2. Project Goals

- All the goals from Use Case 1
- Share PII (with no Direct Identifiers) and code across multiple institutions

3. User and Sponsor Objectives

- All the objectives from Use Case 1
- Run algorithms that may take hours or days to complete on high-performance computers

4. Project Initiation Triggers

- Sponsor authorizes the research study to three FFRDCs
- Institutions performing the research assign EDDIE-credentialed users to perform the study

5. Scenario

This study will require expedited IRB approval. Two of the three participating FFRDCs use IRBs that are sanctioned for EDDIE. However, the third FFRDC has an outside IRB that it typically uses, but that IRB has not yet been sanctioned for EDDIE. Hence, the third FFRDC would need to go through the EDDIE IRB (which is itself an EDDIE sanctioned IRB). A goal of EDDIE is to reduce duplicative IRB reviews through IRB reciprocity agreements. We will assume here that the legal process of becoming an EDDIE sanctioned IRB permits the following reciprocity. Instead of having each FFRDC undergo a separate IRB review for participation in the same project, the leaders of this project from the respective FFRDCs decide to undergo a single cross-institutional review for this project through the EDDIE IRB. As the study moves through the human subjects review process, the EHSRCO keeps the project leaders abreast of its progress by keeping the HSR dashboard updated and providing further communication as needed. Within two weeks, the EDDIE IRB approves the study. There is an aspect of this study that requires the oversight of the R2O2. EHSRCO submits the project and the EDDIE IRB documentation to R2O2, which approves the project within three business days. Immediately thereafter, EHSRCO approves the project for work in EDDIE.

An EDDIE Access Manager creates a Collaborative Folder for this project. The project leaders from the respective FFRDCs determine that, to enable to greatest amount of administrative efficiency, the Institutional Access Managers at each of the FFRDCs should have certain management privileges for that folder. The EDDIE Access Manager then enables the Institutional Access Managers to have those privileges. IT support helps in implementing these changes as needed (responding to and attempting to resolve any requests within two hours).

Analysts access the folder by logging into EDDIE from their office computers. They review the data dictionaries and metadata in the EDDIE Library to identify the most relevant data fields to access. They file a SPA for information to gain access to the relevant data from the immediate access portion of the staging area. The data manager transmits the PII that does not contain Direct Identifiers from the staging area into the project folder. The analysts create a script to perform initial data processing operations, such as checking for problematic data formatting and duplicate entries, and save these results to the project folder.

While reviewing the metadata and information about the data in the wiki, the analysts learn that for a three-year period in the 1990s, there are documented concerns about the accuracy and completeness of some data fields in a particular data file. Upon further exploration, the analysts identify more data fields that may also be problematic. The user posts a message to the user forum asking if others have encountered problems with the specific fields of interest prior to 1997. Other analysts and the data providers respond with

their experiences and solutions using the data from that period. This conversation is attached to the metadata for other users to review in the future.

The analysts go through a lengthy process to clean data from 1980 to present. They observe that there are no cleaned versions of the data back to 1980 in the staging area. With permission from the sponsor, the project team requests that a data manager save a copy of the data in the staging area. They also post their code for cleaning the data in the library, together with a summary of their cleaning methodology and any rationale for particular decisions they made during the cleaning process. Sharing a cleaned version of the data will enhance the efficiency and timeliness of future projects.

Once the cross-institutional team acquires, reviews, and cleans the required data, they continue to work within the project folder and establish a subfolder structure that satisfies their needs. Analysts build a model to estimate responses to changes in the level of voluntary separation and retirement incentives. Because of the complexities of the resulting model, analysts need to run the model over the weekend. EDDIE has a process for readily facilitating long runs (such as a batch submission process or user sessions that do not terminate because of inactivity), so the team is able to run the model without interruption.

As the analysis nears completion, the team uses software available in EDDIE to prepare reports and presentations. The sponsor accesses and reviews the project results within the EDDIE environment and desires to circulate the results of the report among select offices. Once the project is over, the sponsor approves the project for public release. The project leader submits the written reports and code products from the project to the EDDIE Library.

6. Measures for Success

- All the measures of success from Use Case 1
- Analysts from multiple institutions who are collaborating on a project can undergo a single cross-institutional IRB review for the project
- Analysts from multiple institutions have access to the same data extracts and can conduct research collaboratively in a common, secure workspace

C. Use Case 3: Quick Turnaround Analysis

1. Problem

A government analytic office is tasked with analyzing reenlistment over the past ten years by service and occupation. The requesting office would also like to know about the impact of deployment on reenlistment rates. Reenlistment here refers to a service member's decision at the conclusion of their term of service to reenlist (a multi-year service

extension), exit, or extend their current contract for a short duration. The analysis needs to be completed within 72 hours.

2. Project Goals

- Obtain access to clean data sets to support the research in an efficient and timely manner
- Complete analysis in a rigorous, timely, and efficient manner

3. User and Sponsor Objectives

Analysts in EDDIE wish to:

- Quickly identify the relevant data files and data fields
- Gain access to data supporting the research from DMDC with minimal administrative delay
- Use data that requires minimal cleaning
- Use familiar software tools to conduct the analysis
- Easily generate charts and graphs to communicate results effectively

4. Project Initiation Triggers

- Government office authorizes the research
- Government office assigns EDDIE-credentialed users to perform the analysis

5. Scenario

A government office requires statistics on reenlistment over the past ten years by service and occupation. It tasks an analytic office with producing these statistics within 72 hours. This analytic office is accustomed to quick turn analytic tasks. It therefore already has institutional access to EDDIE, and a few of its employees are EDDIE users. Within its Institutional Folder, the office has an Operational Folder that is used for quick turn projects that do not qualify as research in the sense that the activity is not "designed to develop or contribute to generalizable knowledge" (45 CFR 690.102, Definitions). The knowledge is only intended to be used internally and not contribute to a greater scientific dialogue. For quick turn projects that qualify as research, their Institutional Folder contains a project folder that has received HSR umbrella project approval for a range of low-risk research activities that employ certain data and methodologies. So long as the research activities are within the bounds of the umbrella approval, no additional HSR review is needed and the research activities can begin immediately (the umbrella approval may incorporate some degree of auditing on a periodic basis to ensure that activities pursued under the approval are in keeping with the stated bounds). Based on additional context about the intended

dissemination of the final product, the analyst in charge of this task decides that the project folder with the umbrella HSR approval is a more fitting home than the Operational Folder.

Using the EDDIE Library, the analyst reads the documentation on the fields of interest within the DMDC master files and the Contingency Tracking System Deployment file and determines that they will support the project's goals. From the documentation, the analyst learns of differences over time and between services for some occupation codes, but also learns of a computer script developed by the community that standardizes these fields.

The project already contains the necessary data from the DMDC Active Duty Master File. The analyst loads the data into Python and computes the reenlistment rates, graphing these rates for the last ten years by service and occupation.

Deployment information from the most recent quarter is not already in the project folder, so the analyst requests updated deployment information from the staging area. A Data Manager transmits the updated data to the project folder within one business day.

The analyst uses the EDDIE Anonymized ID to merge the master file data with deployment information. The analyst then develops graphs to present reenlistment rates between populations that did not deploy, deployed once, or deployed more than once in the last ten years.

Following the completion of the analysis, the analyst requests a Data Manager to export the graphs from EDDIE. The Data Manager does so within two hours. The analyst transmits the results to the requesting office within the 72-hour timeline.

6. Measures for Success

- Existing data files and information on their fields may be discovered through the library and retrieved for use within the same day
- Software needed for the analysis is immediately available
- Minimal bureaucratic hurdles

This page is intentionally blank.

Appendix A.
Illustrations

Figures

Tables

This page is intentionally blank.

Appendix B.
References

CrossOver, accessed June 8, 2018, https://www.codeweavers.com/.

"Data," Current Population Survey, United States Census Bureau, last modified March 6, 2018, accessed June 8, 2018, https://www.census.gov/programs-surveys/cps/data-detail.html.

"Data via FTP," American Community Survey, United States Census Bureau, last modified June 17, 2018, accessed June 8, 2018, https://www.census.gov/programs-surveys/acs/data/data-via-ftp.html.

Department of Health and Human Services. *HIPAA Basics for Providers: Privacy, Security, and Breach Notification Rules.* Washington, DC: Centers for Medicare and Medicaid Services, 2016. https://www.cms.gov/Outreach-and-Education/Medicare-Learning-Network-MLN/MLNProducts/Downloads/HIPAAPrivacyandSecurity.pdf.

Department of the Navy. *Addendum to the Department of Health and Human Services Federal Wide Assurance (FWA) for the Protection of Human Subjects*. Washington, DC: Human Research Protection Program, 2012. https://www.med.navy.mil/bumed/humanresearch/Documents/HRPP/Resources/DONHRPPGuidance/DON%20HRPP-%20Addendum%20to%20FWA%20Addl%20Requirement_DEC2012.pdf.

DODD 5400.11. "DOD Privacy Program." Washington, DC: Department of Defense, 2014.

DODI 8510.01 (incorporating Change 2, 28 July 2017). "Risk Management Framework (RMF) for DOD Information Technology (IT)." Washington, DC: Department of Defense, 2014.

DODI 8320.02. "Sharing Data, Information, and Information Technology (IT) Services in the Department of Defense." Washington, DC: Department of Defense, 2013.

DODI 8320.07. "Implementing the Sharing of Data, Information, and Information Technology (IT) Services in the Department of Defense." Washington, DC: Department of Defense, 2017.

DODI 6025.18. "Privacy of Individually Identifiable Health Information in DOD Health Care Programs." Washington, DC: Department of Defense, 2009.

DOD 6025.18-R. "DOD Health Information Privacy Regulation." Washington, DC: Department of Defense, 2003.

Dublin Core Metadata Initiative (DCMI), accessed June 8, 2018, http://dublincore.org/.

"FAQ," Document, Discover and Interoperate (DDI), last modified 2018, accessed June 8, 2018, http://www.ddialliance.org/resources/faq.html.

"IPUMS-USA: vars by group," Minnesota Population Center IPUMS database, accessed June 8, 2018, https://usa.ipums.org/usa-action/variables/group.

Joint Task Force Transformation Initiative. NIST Special Publication 800-53, Revision 4, Appendix J. *Security and Privacy Controls for Federal Systems and Organizations.* Gaithersburg, MD, 2013. http://nvlpubs.nist.gov/nistpubs/ SpecialPublications/NIST.SP.800-53r4.pdf.

Library of Congress Metadata Encoding and Transmission Standard, last modified April 23, 2018, accessed on June 8, 2018, http://www.loc.gov/standards/mets/.

McCallister, Erika, Timothy Grance, and Karen A. Scarfone. NIST Special Publication 800-122, *Guide to Protecting the Confidentiality of Personally Identifiable Information (PII).* Gaithersburg, MD: National Institute of Standards and Technology, 2010. https://csrc.nist.gov/publications/detail/ sp/800-122/final.

"METS: An Overview & Tutorial", last modified March 30, 2017, Library of Congress, http://www.loc.gov/standards/mets/METSOverview.v2.html.

Miller, Neil. "What is a Workflow?" *KiSSFLOW*, April 21 2018. https://kissflow.com/workflow/what-is-a-workflow/.

Minnesota Population Center IPUMS database, accessed June 8, 2018, https://www.ipums.org/.

Ross, Ron, Patrick Vicuso, Gary Guissanie, Kelley Dempsey, and Mark Riddle. NIST Special Publication 800-171 (Revision 1), *Protecting Controlled Unclassified Information in Nonfederal Systems and Organizations*. Gaithersburg, MD: National Institute of Standards and Technology, 2016. https://nvlpubs.nist.gov/nistpubs/specialpublications/nist.sp.800-171r1.pdf.

Winkler, Sabune, Elizabeth Witte, and Barbara Bierer. "The Harvard Catalyst Common Reciprocal IRB Reliance Agreement: An Innovative Approach to Multisite IRB Review and Oversight." *Clinical and Translational Science* 8, no. 1 (2015): 57–66.

Wine HQ, accessed June 8, 2018, https://www.winehq.org/.

Appendix C.
Abbreviations

ACS	American Community Survey
BCAC	Business Capability Acquisition Cycle
CAC	Common Access Card
CPS	Current Population Survey
CRAN	Comprehensive R Archive Network
CTS	Contingency Tracking System
DCMI	Dublin Core Metadata Initiative
DDI	Document, Discover and Interoperate
DMDC	Defense Manpower Data Center
DOD	Department of Defense
DODD	Department of Defense Directive
DUA	Data Use Agreements
EDDIE	Enterprise Data to Decisions Information Environment
FFRDC	Federally Funded Research and Development Center
HRPO	Human Research Protection Office
HSR	Human Subjects Research
IDA	Institute for Defense Analyses
IDE	Integrated Development Environment
IPEDS	Integrated Postsecondary Education Data System
IRB	Institutional Review Board
IT	Information Technology
JAMRS	Joint Advertising Marketing Research & Studies
METS	Metadata Encoding and Transmission Standard
NIST	National Institute of Standards and Technology
OPA	Office of People Analytics
OS	Operating System
OSD	Office of the Secretary of Defense
OUSD(P&R)	Office of the Secretary of Defense for Personnel and Readiness
PHI	Protected Health Information
PII	Personally Identifiable Information
R2O2	Research Regulatory Oversight Office
SPA	Specific Project Amendment
XML	eXtensible Markup Language

This page is intentionally blank.

REPORT DOCUMENTATION PAGE	Form Approved OMB No. 0704-0188

1. REPORT DATE (DD-MM-YY)	2. REPORT TYPE	3. DATES COVERED (From – To)
xx-08-2018	Final	

4. TITLE AND SUBTITLE

User Requirements for the Enterprise Data to Decisions Information Environment

5a. CONTRACT NO.

HQ0034-14-D-0001

5b. GRANT NO.

5c. PROGRAM ELEMENT NO(S).

6. AUTHOR(S)

Julie Pechacek
Alan Gelder
Ethan Novak
Amrit Romana
Paul Richanbach
Kathy Conley
George Kennedy
Cheryl Green

5d. PROJECT NO.

5e. TASK NO.

BE-6-4311

5f. WORK UNIT NO.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Institute for Defense Analyses
4850 Mark Center Drive
Alexandria, VA 22311-1882

8. PERFORMING ORGANIZATION REPORT NO.

IDA Document NS D-9139
Log: H 18-000246

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

OUSD(P&R)
Pentagon

10. SPONSOR'S / MONITOR'S ACRONYM(S)

OUSD(P&R)

11. SPONSOR'S / MONITOR'S REPORT NO(S).

12. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The U.S. Department of Defense (DOD) and the research institutions supporting it conduct extensive personnel analysis to manage and enhance America's all-volunteer military forces. DOD desires a new data-hosting analytic computing resource—currently known as the Enterprise Data to Decisions Information Environment (EDDIE)—to augment current data and collaborative capacities. This document presents collaboratively-developed research user requirements reflecting the desired user experience and developed in a manner consistent with the DOD Business Capability Acquisition Cycle (BCAC). We do not attempt to translate these user requirements into technical architectures or detailed computational specifications.

15. SUBJECT TERMS

Requirements Generation, Machine Learning, Computation, Personnel Data

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NO. OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Mr. Lernes Hebert
U	U	U	U	74	19b. TELEPHONE NUMBER (Include Area Code) (703) 571-0114

This page is intentionally blank.