INSTITUTE FOR DEFENSE ANALYSES

Considerations for Implementing a Defense Personnel Research Environment

Julie Pechacek
Alan Gelder
Amrit Romana
Ethan Novak
Kathy Conley
Cheryl Green
Dina Eliezer
P.M. Picucci
George Kennedy
Cullen Roberts

September 2018
Approved for public release; distribution is unlimited.
IDA Paper P-9254
Log: H 18-000375

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 Shelley Cazares, Nancy Huff, and Peter Levine for their review of this document; Hannah Acheson-Field, Nathaniel Latshaw, Allie Saizan, and Claire Summers for their assistance with stakeholder discussions; and representatives for each of the case studies for their insights. We additionally thank the numerous individuals who participated in the iterative development of user requirements for the Enterprise Data to Decisions Information Environment, including representatives from the five DOD-sponsored FFRDC Study and Analysis Centers (CNA's Center for Naval Analysis, IDA's Systems and Analyses Center, and RAND's Arroyo Center, National Defense Research Institute, and Project Air Force); Navy Manpower, Personnel, Training, and Education; Army Manpower and Reserve Affairs; Air Force Manpower, Personnel, and Services; OSD Cost Assessment and Program Evaluation (CAPE); National Guard Bureau; and many elements of OUSD(P&R), including Military Personnel Policy, Office of People Analytics, Defense Manpower Data Center, Military Community and Family Policy, and Transition to Veterans Program Office.

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 Paper P-9254

Considerations for Implementing a Defense Personnel Research Environment

Julie Pechacek
Alan Gelder
Amrit Romana
Ethan Novak
Kathy Conley
Cheryl Green
Dina Eliezer
P.M. Picucci
George Kennedy
Cullen Roberts

This page is intentionally blank.

Executive Summary

To build and manage its military, civilian, and contractor workforces, the U.S. Department of Defense (DOD) oversees an extensive research portfolio on military personnel policy. Analysts require accurate, detailed data on individual military members and civil servants to support a wide range of inquiries. The current data acquisition and preparation process requires a substantial investment of researcher time and expertise that is often duplicated across projects and organizations conducting personnel research—including offices within the military Services, the Office of the Secretary of Defense, and the Federally Funded Research and Development Centers (FFRDCs). A responsive data-hosting analytic environment could improve research quality and timeliness by rebalancing efforts away from data procurement and preparation, and toward analysis. Such an environment could also foster cross-organizational collaboration and modeling by providing a forum where authorized analysts could share and access programming code and documentation across organizational boundaries. The ability to reuse, modify, and combine models extends and enhances the value of individual research endeavors.

The Office of the Under Secretary of Defense for Personnel and Readiness (OUSD(P&R)) tasked the Institute for Defense Analyses (IDA) with engaging the defense personnel research community to identify user requirements for a new data-hosting collaborative analytic environment, known as the Enterprise Data to Decisions Information Environment (EDDIE). The user requirements in Pechacek et al. (2018; IDA NS D-9139) detail the desired structural, computational, data, and supporting capabilities for EDDIE. The requirements focus on the aspects of EDDIE that are procurable, and do not address the context and challenges impacting EDDIE design or operation. This report supplements the requirements document by providing contextual considerations for implementing a defense personnel research environment.

A. Methodology

User requirements for EDDIE were developed through iterative stakeholder engagement in roundtable discussions, written questionnaires, and multiple rounds of draft review and comment. The focal points of this engagement were two analyst stakeholder meetings, including as many as 60 analysts representing more than 20 organizations. The IDA team also surveyed industry best practices for data hosting and curation, and gathered lessons from similar government endeavors.

B. Findings from Analytic Community Engagement

Stakeholders uniformly expressed a need for more timely access to data. The current data acquisition process typically requires weeks or months, often with further delays after

a request is fulfilled due to errors or omissions in the data pull or due to insufficient metadata and documentation. Performing these extraction and data preparation tasks separately for each individual research project significantly delays the delivery of results, and requires DOD to pay multiple times for similar work.

Stakeholders also consistently emphasized their need for more information about the data they receive. This includes metadata on file and field descriptions, field coding conventions and category interpretations, changes in data field coding or definitions over time, crosswalks for linking data sources together, and details on the existing limitations of or gaps in historical data assets. The Defense Manpower Data Center (DMDC)—a primary provider of personnel data—cannot currently provide consistent compiled metadata and documentation on its data assets, due at least in part to limited resources and its ever-expanding mission to support real-time data requests (for example, in support of the TRICARE medical system, the Common Access Card (CAC) identification system, or Service member income verification). Clearer documentation represents a library science challenge: to catalog and curate the data and associated metadata, and then provide it to analysts in a transparent and responsive manner. If information on each data element is not captured and preserved in an easily transferable way, that knowledge is lost. The formation of a central library within a defense personnel research environment would provide a forum for recording, preserving, and transferring institutional knowledge on personnel data. Both data providers and experienced analysts could contribute to this forum.

EDDIE's value to the research community will largely depend on whether it facilitates timely performance of high-quality research. A large corpus of personnel data products, organized into well-documented relational databases that analysts can quickly access (or have quickly accessed on their behalf), would be a significant asset. To enable this functionality, EDDIE needs a high level of connectivity, large storage, rapid processing, and the ability to flexibly expand its computational capabilities. Because it is not possible to conduct all projects in an environment such as EDDIE, it should not attempt to become a one-size-fits-all tool to which all defense personnel analysis must conform. Transaction costs for interacting with EDDIE should be minimized. When an analyst must submit a request for service (e.g., file import, file export, or data request), it creates a transaction cost for both the research organization and the DOD approval authorities. Individually, these costs may be minor. But the more they enter an iterative research process, the more delays arise and compound. To prevent bottlenecks, an institution using EDDIE should be able to provide additional labor to ensure timely project completion. The requirements document includes provisions for institutions to designate individuals within their own institution, who—subject to completing any necessary training—may become authorized to perform many of these transactional functions.

Members of the defense personnel research community also expressed the need to streamline the human subjects review process, noting that it is not uncommon for a single

project to be subject to two or three seemingly duplicative or excessive reviews. These challenges are not unique to the DOD, and often arise when multiple institutions are involved in a single study. An emerging solution is to develop common processes across institutions' review boards and other oversight entities that allow for reciprocity. The DOD could develop similar processes for reciprocity following these successful models. Researchers also expressed the need for greater visibility into the overall human subject review process.

C. Case Studies

Other U.S. government efforts to enable data aggregation and collaborative research provide lessons for EDDIE. Elements that tend to contribute to success include a sustained commitment from senior leadership, appropriate authorities, and thoughtful design.

- **Advanced Distributed Learning (ADL) Initiative's** Total Learning Architecture project promotes a federated open architecture approach to distributed learning, to enable disparate learning systems to communicate and interoperate at the enterprise level. The ADL Initiative showcases challenges arising from practices unique to individual DOD components (e.g., cyber security policies), and the need for common architectures.
- **Defense Integrated Military Human Resources System (DIMHRS)** attempted to provide a single, unified personnel management and pay processing system that could be used across all military Services. It highlights the challenge of achieving consensus among multiple stakeholders, and the need for well-defined scope, pragmatic phasing, and sustained executive-level support.
- **General Services Administration's (GSA's) Data 2 Decisions (D2D)** is a cloud-based centralized repository for data and analytic tools. D2D has experienced challenges due to the absence of business rules for extracting, transforming, and loading data; D2D leadership recommends establishing standardized procedures for data entry, aggregation, and normalization.
- **Department of the Army's Person-Event Data Environment (PDE)** is a cloud-based analytic environment hosting defense personnel data that is used for operational support and research. PDE has struggled to meet users' needs for timely and easy access, and for adequate computing resources. Some users have left PDE due to these challenges and cumbersome security requirements. The PDE office urges close attention to initial architectural choices to ensure that the resulting environment and organizational priorities are aligned with user needs.
- **Quick Reaction Analysis Team (QRAT)** is composed of members from multiple FFRDCs, University Affiliated Research Centers, and National Labs. It is tasked by the Office of Net Technical Assessment with conducting quick response analyses that often involve collaboration between two to four research

institutions. QRAT does not have a central data repository, but demonstrates the benefits of enhanced collaboration between research institutions.

D. Governance

Stakeholders in the defense personnel research community represent a broad range of organizational types, each with its own research goals, internal structures, and external obligations. To accommodate this variety, the IDA team developed a concept for EDDIE governance to provide necessary oversight while preserving the research independence required for many community stakeholders to use this new resource.[1] The model governance structure includes high-level authorities and responsibilities for data access for defense-related personnel analyses, the development and financing of EDDIE, and the streamlining and oversight of the human subjects review process. It includes recommended language establishing the mission, function, membership, and responsibilities for a proposed EDDIE Oversight Committee. This committee would act as the authoritative governing body to establish and amend policies pertaining to EDDIE. It would manage EDDIE's ongoing operations, determine capability investments, promote information exchange for enhancing research related to DOD personnel, monitor relevant emerging technologies, and maintain financial accountability for EDDIE. We recommend that this Oversight Committee be led by three officials from OUSD(P&R), and include representative members from each of the military departments, DMDC, and the three organizations administering DOD-sponsored FFRDC Study and Analysis Centers (CNA, IDA, and RAND).

E. Use Standards

Increasing and improving the level of cross-organizational collaboration and peer review in modeling requires a high level of communication among research teams on analytic details. Ad-hoc code development practices frequently fail to provide analytic and programming products suitable for use by those outside the immediate development team, and can be difficult to scale as the audience increases. The adoption of quality scientific programming standards can provide efficiency and order in analytic coding, and make the resulting product useful beyond its development team. We present a set of best practices and use standards to support reusable code and fully reproducible results. We also recommend the adoption of a peer review process that includes evaluating code according to a checklist of use standards. For work performed in EDDIE, research sponsors should expect peer-reviewed code and data to be submitted as deliverables.

---

[1] The independent nature of FFRDC research provides significant value to the U.S. government and is central to the role of FFRDCs as established by Congress. The authors believe that appropriate EDDIE oversight can be compatible with FFRDC research independence for work performed in EDDIE, given appropriate EDDIE governance mechanisms.

F. Risk

Implementing a shared analytic environment involves trade-offs along multiple dimensions of risk. The environment's usefulness and adoption can be impaired by limited computational capabilities or the lack of an ongoing ability to modernize. Reasonable oversight and curation is needed to maintain data integrity and organization. Cybersecurity safeguards are critical to preventing data loss or misuse. Benefits from collaboration can be hindered by low rates of adoption. We suggest steps to mitigate these risks.

G. Recommendations and Cautions

The EDDIE researcher requirements include the following core features:

- **Analytic Environment.** EDDIE should be an environment in which users may access defense personnel data and perform complex analyses.
- **Data Access.** EDDIE should provide a broad corpus of personnel data and metadata.
- **Library.** EDDIE should support and contain a library with data dictionaries, metadata, public code, and a wiki collecting institutional knowledge.
- **Institutional Memory.** EDDIE should enable users to record information about data, code, results, and projects.
- **Workspaces.** EDDIE should provide access-controlled project workspaces.
- **Computing Resources.** EDDIE should provide sufficient computing capacity.
- **Analytic Tools.** EDDIE should provide users with research tools for conducting statistical, econometric, and predictive analyses.
- **Enable Collaboration.** EDDIE should enable collaboration in workspaces among individuals within and across institutions.
- **Import and Export Control.** EDDIE should support the import and export of data and analysis files. Exports may be limited to non-personally identifiable information (non-PII).
- **Human Subjects Review (HSR).** EDDIE should streamline the HSR approval process.

These elements contribute to enabling economies of scale across the community in data preparation and model development. The range of benefits include rapid and secure data access, increased vetting and dispersion of ideas, reproducibility and transparency of results, and greater responsiveness in providing actionable information to DoD leadership. Various combinations of these benefits may be achieved with lesser alternatives. However, lesser alternatives would also prolong costly inefficiencies in current practices.

This page is intentionally blank.

Contents

This page is intentionally blank.

1. Purpose and Scope

The U.S. Department of Defense (DOD) actively conducts and sponsors research on personnel policies to better manage and enhance the capabilities of its collective workforce of more than 3 million people—including the active and reserve components, the DOD civilian employees, and the supporting contractor labor force. Topics include the recruiting, training, readiness, compensation, retention, resiliency, workforce mix, and career management of America's all-volunteer military forces and supporting civil servants. Analysts from the Office of the Secretary of Defense (OSD), the Military Departments, the Federally Funded Research and Development Centers (FFRDCs), the Military Academies, and other organizations support these ongoing research needs with both quick response and in-depth investigations.

Analysts require accurate, detailed data on individual uniformed military members and civil servants to support a wide range of inquiries. Obtaining access to these frequently sensitive data, determining their suitability, and performing basic preparatory manipulations often consume major portions of the time and funding allotted to research projects. Data access impediments often hinder the DOD's efforts to obtain high-quality analyses within a relevant time horizon.

When analysts receive data, the data are often raw and require considerable cleaning and preparation prior to conducting any analytic work. Multiplied across many research organizations, the data acquisition and preparation process entails multiple layers of redundant effort, which might be reduced under a restructured information management, curation, and distribution scheme. Capabilities enabled by a data-hosting collaborative analytic environment—including collection of and access to metadata on files and fields—can improve research quality by rebalancing the effort of research teams from data procurement to analysis. A library of community-vetted and curated data and metadata would minimize the uncertainty around the reliability of various data elements. Moreover, such an environment could facilitate cross-institutional collaboration and peer review.

The need to protect sensitive personnel data has historically contributed to the isolation of analysts within their respective organizations, which dampens cross-organizational collaboration. At present, the community lacks a secure means for gathering and sharing detailed analytic information among its members. The Office of the Under Secretary of Defense for Personnel and Readiness (OUSD(P&R)) seeks to improve

personnel research quality and timeliness by developing mechanisms that increase cross-organizational communication, collaboration, vetting, and dispersion of ideas.

To further these goals, OUSD(P&R) tasked the Institute for Defense Analyses (IDA) with engaging the analytic community supporting defense personnel research to identify and synthesize user requirements for a new data-hosting collaborative analytic resource, currently known as the Enterprise Data to Decisions Information Environment (EDDIE). The EDDIE Requirements Document outlines the structural, computational, data, and supporting capabilities needed to implement an analytic environment for the defense personnel research community.[2] These user requirements for enabling defense personnel research are the result of multiple rounds of oral and written interaction with and feedback from numerous analyst stakeholders, collected between November 2017 and June 2018.

This report supplements the EDDIE Requirements Document by detailing our methodological approach for generating the requirements, and by highlighting various perspectives and insights gleaned from our engagement with a broad community of analyst stakeholders. We provide contextual considerations for implementing a defense personnel research environment, and present case studies on other U.S. government endeavors to provide central points of data access or mechanisms for collaborative research. These case studies offer valuable lessons learned, and include such initiatives as the Army's Person-Event Data Environment (PDE), the Defense Integrated Military Human Resources System (DIMHRS), and DOD's Advanced Distributed Learning (ADL) Initiative. In particular, the ADL Initiative's governance document (DOD Instruction 1322.26) provides a helpful model for framing EDDIE governance structures. Based in part on the example set by the ADL Initiative, we suggest features for an EDDIE governance charter. While governance pertains to the administration and ongoing oversight and development of EDDIE, we also include use standards for how analysts might operate within EDDIE. These include best practices and expectations for such things as coding, documentation, and peer review. One goal of these use standards is to facilitate modular model development. We conclude by enumerating some risks that need to be considered in structuring and building EDDIE and summarizing our recommendations for EDDIE. We also identify alternatives to those enumerated in the EDDIE Requirements Document that may be cost-effective means to meet some (but not all) of the DOD's objectives for a new defense personnel research environment.

[2] Julie Pechacek, Alan Gelder, Ethan Novak, Amrit Romana, et al., *User Requirements for the Enterprise Data to Decisions Information Environment*, IDA Document NS D-9139 (Alexandria, VA: Institute for Defense Analyses, August 2018).

2. Methodology

Developing user requirements for EDDIE was an iterative process of engaging stakeholders in the defense personnel research community through meetings, roundtable discussions, structured conversations with subject matter experts, written questionnaires, and multiple rounds of draft review and comment.[3] This process also included surveying industry best practices for data hosting and curation, as well as lessons learned from similar government endeavors. As this process unfolded and issues began to crystalize, the OUSD(P&R) sponsors were able to provide guidance on the direction they desired for EDDIE.

The IDA team and the OUSD(P&R) sponsors jointly identified many of the stakeholder organizations to involve in this process, including the FFRDC Study and Analysis Centers supporting DOD personnel research, the Defense Manpower Data Center (DMDC), and several analytic organizations from OSD and the Military Departments. The number of participating organizations grew steadily throughout the process as word spread. IDA had a dual role in this process: to coordinate and synthesize user requirements from the analytic community and to represent the equities of IDA as a potential EDDIE user. To allow the core IDA team to remain neutral in its coordination role, a firewalled group of analysts represented IDA as users.

Community engagement began in earnest with an executive stakeholders meeting on November 2, 2017. This meeting brought together the leadership of several stakeholder organizations to introduce an initial vision for EDDIE and invite participation in the collaborative requirements development process. OUSD(P&R) sponsors and the IDA team invited these leaders to send analysts from their organizations to the subsequent meetings and submit feedback during comment periods.

An analyst stakeholder meeting followed two weeks later, with roughly 40 analysts from more than a dozen organizations in attendance. This meeting focused on eliciting baseline requirements for EDDIE and included small group breakout sessions on a few

[3] The following organizations provided input to this process: the five DOD-sponsored FFRDC Study and Analysis Centers (IDA's Systems and Analyses Center, CNA's Center for Naval Analysis, RAND's Arroyo Center, National Defense Research Institute, and Project Air Force); Navy Manpower, Personnel, Training, and Education; Army Manpower and Reserve Affairs; OSD Cost Assessment and Program Evaluation (CAPE); Air Force Manpower, Personnel, and Services; National Guard Bureau; and many elements of OUSD(P&R), including Military Personnel Policy, Office of People Analytics, Defense Manpower Data Center, Military Community and Family Policy, and Transition to Veterans Program Office.

specific topic areas: governance and administration, user requirements, and data requirements. We conducted breakout sessions as guided discussions, focusing on several distinct points within the respective topic areas. We invited participants to provide written feedback to questionnaires that spanned each of the topic areas. To get a sense for how existing research workflows might map into EDDIE, together with the tools and capabilities they would need in order for EDDIE to be a useful resource, we also invited participants to submit narrative use cases.

Based on discussions at the analyst and executive stakeholder meetings and extensive written feedback, we crafted a draft requirements document for EDDIE and broadly circulated it in February 2018 for review.

The next analyst stakeholder meeting was on March 1, 2018, after the stakeholder organizations had an opportunity to review the draft requirements. This meeting included roughly 60 analysts from more than 20 organizations. With a draft in hand, this meeting allowed for more specific discussion on the structure and capabilities of EDDIE. Six small group breakout sessions discussed the various sections of the draft. These included breakout sessions on governance and organizational structure; DOD data policy; risk factors and mitigation; user types, needs, and support requirements; data and information management; and tools, models and organization. These discussions and further written feedback following this meeting significantly helped to illuminate and refine the direction for the EDDIE requirements document.

Until this point, the dialogue for developing requirements for EDDIE had proceeded in a largely open-ended fashion, with no coordinated parameters for its underlying content, role, structure, or scope. The community was operating under the assumption that it was designing EDDIE from a relatively blank sheet. This exploration and maturing dialogue enabled the OUSD(P&R) sponsor to provide firm guidance on the underlying assumptions of and aspirations for EDDIE. These assumptions are provided in the EDDIE requirements document. By this time, the Office of People Analytics (OPA) had developed a statement of need for EDDIE to support the DOD Business Capability Acquisition Cycle (BCAC).[4] It describes the need for easily accessible and well-documented data that can be used to quickly provide analysis to decision makers. It also mentions the need for a collaboration environment in which analysts in the DOD, federal agencies, and external stakeholders can and reuse analytic models.

To produce an actionable and technically detailed requirements document that was consistent with the guideposts provided by the OUSD(P&R) sponsor, the IDA team engaged in further detailed discussions with representatives from CNA, IDA, and RAND in May 2018. The FFRDCs had submitted voluminous and often conflicting feedback throughout the requirements development process. However, this last round of discussions

[4] OPA, *Enterprise Data to Decisions Statement of Need/Problem*. Draft, April 2018.

ultimately resulted in a high degree of consensus for the desired structure of EDDIE. This became the basis for another draft of the requirements document. The final draft incorporates further feedback from the sponsor and analysts at the FFRDCs.

This page is intentionally blank.

3. Findings from Analytic Community Engagement

The analytic community supporting defense personnel research is diverse—both in terms of the DOD components and organizations that they support and in terms of their strengths and challenges. However, as the IDA team engaged this broad community through a series meetings and feedback periods, we identified several common concerns and aspirations for developing a defense personnel research environment.

A. Data

A primary concern with the current paradigm for the vast majority of the community is a need for *more timely access to personnel data*. Requests can take weeks or months to fulfill, even for less sensitive personnel data. When data is finally received, analysts frequently discover errors in the data pull. Some fields or time periods may be omitted. Some fields may appear to be mislabeled (e.g., a label of "gender" affixed to a field with floating point numbers with hundreds of unique values). In other cases, the data pull may be correct, but it is not accompanied by a dictionary for deciphering how the fields are coded (e.g., what the values "A" through "L" each signify within a particular data field). Alternatively, the analysts may have a dictionary, but it does not correspond to the values in the data (e.g., the dictionary contains values for "A" through "G" but the values in the data extend from "A" to "P"). Getting the data, ensuring that it is the correct pull, and ensuring that the values in the data can be deciphered is typically a process that involves repeated interactions with the data provider over extended weeks and months, consuming the resources of both the research organization and the data provider. Analysis and results, meanwhile, are delayed.

Expediting the data acquisition process is a two-fold endeavor that must address both the form of the data and the availability of the data. First, there is a library science problem of properly cataloging and curating the data and all the associated metadata and documentation. DMDC, as a primary provider of personnel data, currently lacks a significant amount of compiled metadata and documentation on its own data assets. Information on specific data assets is more likely to be resident in the heads and computers of a few subject matter experts than in compiled, easily sharable formats. There is concern in the analyst community that, without a clear catalog of this information, the analysts may not be asking the proper questions or framing analyses to take advantage of the data that

do exist. Moreover, inasmuch as information on data is not captured and preserved in an easily transferable way, it is perishable. The formation of a central library within a defense personnel research environment would provide a forum for recording and preserving institutional knowledge on personnel data. Both data providers and analysts who have experience working with various data assets could contribute to this forum.

The second aspect of facilitating the data acquisition process is to *make a large corpus of personnel data readily available to analysts within a secure space*. Ideally, the data would be curated to a point where analysts could submit a point-and-click order of data fields and coverage dates. A model example of this is the Integrated Public Use Microdata Series (IPUMS), hosted and maintained by the University of Minnesota, which contains anonymized person-level census data. IPUMS allows analysts to select the data fields and time periods they desire and then automatically compiles an extract of the data for the analyst. Although that level of organization, automation, and curation may be an unachievable gold standard, having a core set of DMDC data organized into a relational database that analysts can quickly access (or have accessed on their behalf) is essential to enabling timely and scalable data access. Each data field in each DMDC data file also needs sufficient documentation so that analysts who are unfamiliar with the data can learn how the data was coded and how it was constructed (if it was derived from other data fields), where the data originated from (e.g., did it originate with a military Service?), what processing was done to the data from the time DMDC received it (e.g., how were missing values in the original data handled?), and any known or suspected issues with the data.

Since data field codes and definitions have changed over time, it is important to the community to *have a place where the numerous idiosyncrasies of the different data files can be recorded*. Such a store of institutional knowledge—including crosswalks for changes in data field coding and definitions over time—is critical to effectively executing and interpreting studies that require longitudinal data. The community also desires to have ready access within EDDIE to *longitudinal data* (2001 has become a somewhat arbitrary cutoff date for data requests from DMDC; the community desires earlier data whenever possible to facilitate longitudinal studies, even if it is less curated).[5]

Changes or updates to the data should be made through a system of *version control* to enable the development of a standardized, authoritative data source for the community. This also facilitates reproducibility of results and will greatly reduce duplicative efforts of data preparation and curation that each analytic organization routinely undertakes.

The various DMDC personnel data files capture different pieces of information about individuals. The research community would benefit considerably from a unique *person-level identifier that is consistent across data files* to enable files to be merged and combined

[5] In some cases, organizations besides the data provider have maintained perhaps better longitudinal data records than the data provider. They may be useful resources for populating early data within EDDIE.

in analyses. This identifier need not be a social security number or other commonly used and sensitive marker. A scrambled, encrypted identifier would enable advanced analysis without the risk of compromising personal information.

No matter how richly populated the data in the environment is, some studies will inevitably require additional data. Some data owners may permit the additional data to come into EDDIE, while others will not. In some cases, unique data arrangements will be needed to address complicated sensitivity and legal issues. The FFRDCs have existing data use agreements in place with DMDC and other data providers within and outside of the DOD, which may fit some projects better than EDDIE does. As a result of these and other issues, some research will not be possible to conduct in an analytic environment such as EDDIE. Accordingly, the analytic community strongly recommends against an effort to make EDDIE a one-size-fits-all tool to which all defense personnel analysis must conform.[6]

Beginning at the November 2, 2017 stakeholder meeting, and continuing to the present, the primary OUSD(P&R) sponsor has recognized this complex research environment and stated that *EDDIE is intended to be an additional resource for the community, not a replacement for existing resources*. EDDIE provides value not as a replacement to existing analytical capabilities and data use agreements, but rather as a supplemental tool that provides a forum for collaborative model development, a library of institutional knowledge, and a core set of rapidly accessible, well-curated data (obviating the need to assemble such data from the ground up in every case).

While the FFRDCs and some analytic organizations within the DOD have invested in secure computing resources for analyzing various forms of controlled unclassified information (CUI), such as personally identifiable information (PII) and protected health information (PHI), other analytic organizations within the DOD have lacked the resources or the critical size to make that investment. *OUSD(P&R) views EDDIE as a means for better controlling the distribution and safety of its personnel data, while simultaneously making it more accessible*. If appropriately designed and implemented, the combination of security and accessibility within EDDIE may improve the incentive structure for providing data for research purposes. There is a false dichotomy between data security and data access that needs to be addressed. Secure systems can be scalable to allow access to many

[6] For instance, sensitive workplace climate studies on topics such as sexual assault and harassment may be most effective when they are conducted by an organization that is independent of the DOD (such as an FFRDC). Participants in such studies may only divulge truthful information if they can be assured that their individual responses will not be accessible by anyone in the DOD. It would therefore not be appropriate to place data from such studies on *any* government system. Survey responses are more fully contextualized when they can be linked to person-level administrative data from DMDC. Thus, to get the most value out of such studies for the DOD, the organization conducting the study must also have a local copy of the relevant DMDC files. Several existing research capabilities and programs would be jeopardized if all personnel research were required to conform to an environment like EDDIE.

authorized users. Computer endpoints and physical locations where the system is accessed can be vetted, as can the individuals who access it.

Without a dual objective of security and accessibility, it is far too easy for data providers to focus solely on security and simply lock up the data. But if the data are locked up and never used, they provide no value to decision makers.[7] If a data provider denies a data request for research, there is often little recourse for appeal. The research community would benefit from clearer, well-defined mechanisms for appealing and adjudicating data request issues within the DOD.

B. Interaction with the Environment

The value of EDDIE to the community will largely be determined by its ability to facilitate the timely performance of high-quality research. This includes having *strong computational capabilities*. Among other things, EDDIE needs a high level of connectivity (no latency issues), large storage, rapid processing, and the ability to scale and grow. The level of analyses in EDDIE will likely range from tasks that could easily be performed on a typical office computer to computations that require extensive random access memory (RAM) (e.g., more than 100 GB) and high-caliber GPU processing. Projects will often need to hold in memory and conduct analysis on large fractions of DMDC master files—hundreds of thousands of observations per month with dozens of fields per observation multiplied by a decade or more of monthly data. A one-size-fits-all computing allocation on a project-by-project basis will not suffice. Projects should be able to access computational resources according to their expected need. The frequency of projects requiring high-end computational resources will likely be intermittent, with a more constant demand at the mid and low end.

In addition to meeting computational needs, *transaction costs (including time delays) for interacting with the environment need to be minimized*. Data will need to be imported. Results, modeling code, and other non-PII items will need to be exported. Data requests within the environment may not be fully automated, requiring manual support in selecting the correct data extracts. Every action within the environment where the analyst needs to submit some kind of request creates a transaction cost. Individually, these may be minor. But the more they enter iterative research processes, the more room there is for delays and inefficiencies. To prevent work within EDDIE from becoming a string of bottlenecks that stymie production and deter users, there should be adequate resources to support any

[7] Analogous to the push in the intelligence community to move from a "need to know" to a "need to share" in the wake of September 11, 2001, there is a question of whether there needs to be a similar paradigm shift for data provision within the DOD—especially in light of emerging analytic techniques that can glean meaningful patterns and insights from vast quantities of data. These techniques have pushed the boundaries of what is meant by "minimal data requirements" for research. Analysts may not know *a priori* which data fields may be the most meaningful for the question at hand, so a fairly broad interpretation of "minimal" needs to be applied.

approvals and actions that analysts are not permitted to do by themselves. These resources should be scalable to support peak demand needs and ideally have some mechanism that allows for expedited service (e.g., a user could pay a premium to use an express service). If labor is in short supply and an institution using EDDIE is willing to pay for the labor needed to ensure timely support for their projects, then that should be an option. Based on this rationale, the EDDIE requirements document makes provisions for institutions to designate individuals within their own institution, who—subject to completing any necessary training—may become authorized to perform many of the transactional functions within EDDIE. This alignment of incentives between the operations of the analytic environment and the needs of its institutions will likely prove critical to EDDIE's long-term success.[8]

A core benefit of EDDIE is the establishment of a library for code, models, and data that the defense personnel research community can actively contribute to and draw upon. With this move toward greater openness, the community has also expressed concerns about maintaining an appropriate degree of privacy. As is currently the case, *research sponsors should maintain the ability to determine if and when a research product can be released to a broader audience*. Workspaces within EDDIE should permit collaboration between individuals within and across organizations, but access to those workspaces should also be controllable. Yet there should also be an expectation among analysts and sponsors to contribute research products to the public body of knowledge within EDDIE whenever possible.

C. Human Subjects Review

Before analyses can even begin, studies are taxed substantially in the meandering and laborious processes of obtaining access to data and navigating the human subjects review process. Anecdotal evidence suggests that these processes can consume as much as half of the study's time, budget, or both. Consequently, analyses are truncated and quality suffers. A major impetus behind EDDIE is to reduce the gap between when a research sponsor asks a question and when an analyst can begin to answer it. The research community broadly agrees that a more timely and simplified human subjects review process for work in EDDIE would contribute significantly to the quality and timeliness of relevant studies.

More so than for typical academic research, the DOD human subjects review process is a seemingly duplicative maze. Each of the military Services and many other DOD

[8] As trusted partners in conducting analyses for the DOD, the FFRDCs are a natural place to maintain resident employees to assist in performing transactional functions within EDDIE. It 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).

components maintain their own human research protection program or office. Studies that cross organizational boundaries within the DOD are frequently reviewed by each separate organization because reviews may not be recognized across organizations. Members of the defense personnel research community noted from experience that it is not uncommon for a single project to be subject to two or three seemingly *duplicative reviews*.[9]

The problem of duplicative reviews is not unique to the DOD and often arises when multiple institutions are involved in a single study. A growing solution is to develop common processes and procedures across reviewing entities that allow for reciprocity. One such example is the Harvard Catalyst Mast Reciprocal Common IRB Reliance Agreement, which has enabled Institutional Review Board (IRB) reciprocity across more than 20 distinct legal entities.[10] The DOD itself could develop similar processes and procedures that permit and encourage reciprocity.

---

[9] In reviewing defense personnel research, the National Academy of Sciences identified the same issue: "Reviews by multiple Institutional Review Boards can significantly slow down the research process and add months or years to the time it takes for researchers to have access to DOD data. This creates a serious problem for responding to policy needs in a timely manner." The National Academy of Sciences, Engineering, and Medicine, *Strengthening Data Science Methods for Department of Defense Personnel and Readiness Missions* (Washington, DC: The National Academies Press, 2017), 112.

[10] See Winkler, S.J., Witte, E., Bierer B.E. The Harvard Catalyst Common Reciprocal IRB Reliance Agreement: An Innovative Approach to Multisite IRB Review and Oversight. *Clin. Transl. Sci.* 8(1), 2015, 57–66.

4. Case Studies

The case studies presented in this chapter provide context for the EDDIE project by situating it within a growing stream of U.S. government-wide efforts to enable data aggregation, collaborative research, and sharing of results. While not all cases illustrate all of these characteristics, they provide a cumulative sense of the challenges that analysts commonly face, and sometimes overcome, as well as lessons that can be applied to the development of EDDIE. Each case study is based on structured conversations with individuals who have either participated as analysts, set policy, or managed a project that involved the sharing of data or research material. The cases are not exhaustive reviews of successes and failures; instead, they are focused on informing discussions of EDDIE requirements by highlighting choices that have been made in the past, along with the effects of those choices, where available.

To that end, each case begins with an overview of the problem that the government faced. We then describe the challenges encountered with implementing the initiative, as well as the results. Finally, each case includes lessons learned from the perspective of the participant, emphasizing the applicability of these lessons to the EDDIE requirements definition effort.

What emerges is an appreciation of the hurdles, both organizational and technological, of developing, fielding, and sustaining collaborative data and analysis capabilities and communities of practice—whether the goal is distributed learning, technical assessment, or improved personnel analytics. Elements that contribute to success are also apparent, including sustained commitment of senior leadership, appropriate authorities, and thoughtful design. In the fast-evolving field of data-intensive human capital research, these three elements will likely prove crucial.

A. Advanced Distributed Learning (ADL) Initiative

1. Overview of Problem and Key Stakeholders

The Department of Defense's ADL Initiative oversees distributed learning for DOD by developing policy, facilitating and conducting research on distributed learning technologies, and coordinating efforts across DOD's Military Departments and the U.S. federal government.[11] It reports to the Deputy Assistant Secretary of Defense for Force Education and Training within OUSD(P&R). One of the ADL Initiative's goals is to promote an open architecture for distributed learning that will enable disparate learning systems to communicate and interoperate at the enterprise level.[12] The ADL Initiative does not directly control data, nor does it consume data from data sharing entities. Rather, it exercises policy-level control of a data-driven enterprise. This case study demonstrates the challenges arising from unique practices (for example, cyber security policies) mandated by individual DOD components, and the need for common architectures and policies to address those challenges.

2. Challenges Encountered

The Office of Personnel Management's (OPM's) USALearning has been given government-wide statutory authority to support assisted acquisition of learning technologies and services. Following the lead of the DOD Chief Management Officer, the ADL Initiative promotes a greater use of USALearning across the DOD. The Military Departments are also collapsing multiple personnel and talent management systems into fewer systems that meet the OPM standard.

However, there is still a tendency within the DOD to develop singular distance learning programs—"shadow learning centers"—that are not integrated into the Military Departments' programs or OPM ecosystem. An ongoing effort is required to bring these programs into the federated architecture mandated by policy.

Another challenge is that policies specific to particular DOD components can inadvertently create barriers to consolidating data on learning measures. For example, cybersecurity policy in one Military Department may mandate a separate review process

---

[11] This case study is based on a discussion with a representative of the Advanced Distributed Learning (ADL) Initiative, Office of the Secretary of Defense, June 11, 2018. Further information about the ADL Initiative can be found on its website: http://www.adlnet.gov/.

[12] Raybourn, E.M., Schatz, S., Vogel-Walcutt, J., & Vierling, K. (2017). *At the Tipping Point: Learning Science and Technology as Key Strategic Enablers for the Future of Defense and Security*. In Proceedings of the Interservice/Industry Training, Simulation and Education Conference (I/ITSEC), Orlando, FL.

prior to adopting software developed by an outside organization. This defeats the purpose of developing open source software solutions to common problems and inhibits collaboration across the Military Departments.

The DOD also lacks a single business intelligence format that will enable the development of a system of systems for learning technologies. Lacking this common format limits the department's ability to adapt its training and education programs to meet changing workforce needs, as well as to track training in support of readiness estimates.

3. Results

The ADL Initiative's Total Learning Architecture project provides a roadmap for enabling greater interoperability across learning technologies. This work is currently being implemented, with empirical testing planned for the near future. A prominent feature of the Total Learning Architecture is a standardized web-based service, Experience API (xAPI), used for tracking experience.[13] Additionally, the ADL Initiative's network of collaborative relationships enable it to keep an active pulse on learning technologies being developed across the Military Services.

4. Lessons from Data Sharing

The ADL Initiative has benefitted from being able to integrate their efforts with OPM's USALearning, which provides assisted acquisition for purchasing systems, services, and cloud servers. The semi-structured data format of xAPI has enabled the ADL Initiative to better capture human learning data, which often lacks objective definitions. Depending on the nature of the data, xAPI may be a useful data specification within EDDIE. The governance structure for the ADL Initiative is formalized in a DOD Instruction, with up-to-date guidance on rapidly evolving technological standards published online. The "companion documentation" can "be easily updated independent of the [DOD] instruction."[14] The ADL Initiative governance structure provides a pattern for the proposed EDDIE governance structure in chapter 5.

---

[13] Office of the Under Secretary of Defense for Personnel and Readiness (OUSD (P&R)), the ADL Initiative website, https://www.adlnet.gov/tla/, accessed June 18, 2018. See also the Experience API website: https://xapi.com/.

[14] See "The ADL Initiative, DODI 1322.26 Reference, Background" at https://adlnet.gov/dodi/.

B. Defense Integrated Military Human Resources System (DIMHRS)

1. Overview of Problem and Key Stakeholders

Computerized data systems for personnel management and pay processing grew and developed within each of the military Services during the 1980s and 1990s.[15] Data standardization across the Services was limited, and business practices were largely Service- and component-specific. Operational deficiencies within and across these systems were exposed during the first Gulf War and established a clear case for the development of an integrated Service system.[16] The DOD's inability to settle on a single "best of breed" system led the Under Secretary of Defense for Acquisition and Technology to request recommendations from a Defense Science Board Task Force. In 1996, the Task Force recommended the creation of a cross-Service integrated personnel and pay system based on a commercial-off-the-shelf system. The Defense Science Board Task Force noted that:

> *The present situation, in which the Services develop and maintain multiple Service-unique military personnel and pay systems, has led to significant functional shortcomings (particularly in the joint arena) and excessive costs for system development and maintenance for the Department of Defense…. [DOD should] move to a single, all-Service and all-component, fully integrated personnel and pay system, with common core software.*[17]

Every organization with equity in the issue agreed with this recommendation, and it was approved by the Deputy Secretary of Defense in December 1996. The OUSD(P&R) formed a senior-level Executive Steering Committee, along with a mid-level Joint Integration Group comprised of representatives from the Military Departments, the Joint Staff, Defense Finance and Accounting Service, DMDC, OUSD(P&R) Manpower and Reserve Affairs, the OUSD Comptroller, OSD Program Analysis and Evaluation (PA&E) and the Assistant Secretary of Defense for Networks and Information Integration. The Department of Veterans Affairs (VA) also participated as an external stakeholder. In leading this effort, OUSD(P&R) attempted to establish and maintain support for DIMHRS from the full range of stakeholders. That proved to be a difficult task. DIMHRS ultimately failed because the Services were unwilling to relinquish control over their separate personnel systems and accept the common practices necessary to live under a single system.

---

15 This case study is based in part on a discussion with an individual who was closely involved in the requirements development process for DIMHRS, June 7, 2018.

16 U.S. Government Accountability Office, *DOD Systems Modernization: Maintaining Effective Communication Is Needed to Help Ensure the Army's Successful Deployment of the Defense Integrated Military Human Resources System,* GAO-08-927R (Washington, D.C.: Sep 8, 2008), accessed July 6, 2018, https://www.gao.gov/assets/100/95723.pdf, 1.

17 Ibid, 4.

In contrast with DIMHRS, EDDIE does not contemplate requiring changes to underlying personnel management processes. It will use data generated by those processes as they currently stand. The DIMHRS case study demonstrates the difficulty of achieving consensus among multiple stakeholders and the need for extensive stakeholder engagement to overcome disagreements

2. Challenges Encountered

a. Technical Challenges

As a purely practical matter, personnel management practices and data collection were not standardized prior to this effort. This resulted in the need to develop a set of common human resources information standards. While the choice of PeopleSoft as the commercial-off-the-shelf platform for the system imposed only limited substantive obstacles to system development (e.g., tracking military permanent change of station moves was particularly challenging for a commercial personnel management package), the corporate acquisition of PeopleSoft by Oracle resulted in uncertainty about continued product support, which would have been key to establishing initial operational capability for DIMHRS.

a. Equity/Stakeholder Challenges

Efficient functioning of the system entailed convincing the military Services to alter some of their business practices beyond just the standardization required by the common human resources information standards. For example, in the Navy and Marine Corps, individuals were paid twice a month. In the Army and Air Force, individuals had the option of being paid once or twice a month. The need to standardize such practices naturally led elements within the Services to hesitate before relinquishing full control of their business processes and practices. Further, there was a disagreement between the financial and personnel communities on whether to prioritize integration of pay or personnel systems. Competing stakeholder interests became visible to Congress and resulted in additional oversight. Program delays began to take their toll on even supportive stakeholders, resulting in increasing funding pressure. This was exacerbated by claims of poor communication between the DIMHRS program office and the Army with respect to DIMHRS capabilities, leading to increased Army concerns regarding the program's ability to meet their requirements.[18] Elements within the DIMHRS program office maintained that the Army's reluctance to alter business practices and to understand how DIMHRS could support those practices contributed to a gradual—and ultimately unachievable—expansion of system requirements.

[18] Ibid, 3.

b. Acquisition Process Challenges

The acquisition process itself provided opportunities for reluctant stakeholders to delay the project and to exercise influence in support of their bureaucratic preferences. Despite the initial direction from the Deputy Secretary to pursue a commercial-off-the-shelf solution, many delays were caused by reluctant participants in the Services or other organizations, who took advantage of the DOD acquisition process and the various oversight groups (including PA&E and OMB) to bog down the program in the coordination process, raise spurious issues to investigators that then had to be addressed, and repeatedly revisit decisions. These participants were often encouraged by the regulators to revisit issues that postponed final decisions on critical questions. The following vignettes demonstrate these kinds of delays.

- Operating on the Defense Science Board's 1996, recommendation, OUSD(P&R) initiated a commercial-off-the-shelf (COTS) evaluation study with the intent of selecting a product within six months. The contract award was pending by July 1997. However, PA&E stopped the selection process pending a formal Analysis of Alternatives to support the need for a COTS product. Approval for the COTS award to PeopleSoft was finally granted in March 2001—four years later.
- After the project had already gone through the Joint Integration Group and an Executive Steering Committee, a five-page Mission Need Statement was produced in October 1997 to support a Milestone 1 decision requirement. It took 13 months to complete Joint Requirements Oversight Council (JROC) coordination on this Mission Need Statement, which included addressing 48 pages of comments from a lower-level review. The entire process did not generate a single change to the Mission Need Statement.

3. Results

Based on the Army's initial assessment in 2005 that it could use DIMHRS with modifications, Army program officials were directed to proceed with DIMHRS acquisition. However, the reinforcing trends of bureaucratic delays and expanding requirements took their toll, and by September 2008, a GAO report had identified substantial Army concerns regarding the ability to meet their requirements.[19] By that point the Marine Corps had opted out, retaining their own internal Marine Corps Total Force System (MCTFS) despite having signed a Memorandum of Agreement to support DIMHRS development. The Navy had also elected to implement the MCTFS in 2006. Multiple postponements of the date for achieving initial operational capability (based on uncertainty as to DIMHRS' ability to meet Army requirements) ultimately resulted in a 2010 decision, announced by Secretary of Defense Robert Gates and the Chairman of the

[19] Ibid, 6.

Joint Chiefs of Staff, Michael Mullen, to cancel the program after an estimated half billion dollars in total expenditures.

4. Lessons for Data Sharing

Even for projects whose objectives are not controversial and whose need is demonstrated by obvious shortcomings of existing systems, initial buy-in from stakeholders and early high-level top cover do not guarantee long-term programmatic success. Lower-level parochial interests will spur resistance to implementation the more these data-sharing processes require relinquishing control over business practices. For projects that rely on the heavily bureaucratized acquisition process, there are ample opportunities for entrenched interests to delay project advancement.

For most of its history DIMHRS enjoyed significant support and backing within the DOD and the VA; however, this goodwill was ultimately insufficient to overcome the mutually reinforcing threats of program duration, cost, and requirements expansion. Compliance with the myriad of strictures of the acquisition process and with the needs of various oversight and regulatory elements necessitates significant and sustained program management expertise to balance length of process with adequate programmatic results. As time to implementation increases, program costs and stakeholder resistance to paying those costs tend to increase. Rising costs can in turn raise expectations for meeting narrow stakeholder interests, leading to even more requirements. Failure to meet those requirements, coupled with delays and further cost growth, can lead to program cancellation.

C. General Services Administration (GSA) Data 2 Decisions (D2D)

1. Overview of Problem and Key Stakeholders

GSA's Data 2 Decisions (D2D) platform is a cloud-based centralized repository for data, modern analytic tools, and analytic products.[20] Although data analytic capabilities have been resident in GSA for decades, the individual data systems and corresponding analytic tools have been decentralized. The D2D platform aims to replace GSA's traditional data systems and outdated business intelligence tools with a centralized data

[20] This case study is based on a discussion with a GSA official, May 9, 2018. The following additional background information on the D2D architecture was provided in an e-mail from the GSA official to Dr. David Chu, IDA President, on December 21, 2017:

"The D2D platform architecture represents a framework of abstract and loosely-coupled components such as Content/Document Repository, Business Process Management (BPM), Integration and Analytics. Each component may scale individually and collectively. GSA considers each component as an implementation black box that provides a specific feature or capability. This fundamental approach allows existing component to be switched with new implementations, new component to be added with ease and for clients to connect and use features without knowledge of how they are implemented. The architectural plug-in capabilities should allow the D2D platform to include new business domain additions without re-architecture, re-design or a major software development undertaking."

repository and access to modern analytic tools. Similar to EDDIE, D2D is a repository for data, an environment for conducting analyses, and a sharable library of research products. This case study demonstrates the need to establish business rules for data management.

2. Challenges Encountered

Shifting from decentralized data management to a centralized, federated data management environment posed a number of challenges for GSA. Although several entities within GSA have shifted their data systems to D2D, some organizations still maintain their traditional systems. The Federal Acquisition Service and Public Buildings Service are the largest divisions within GSA, each with their own data systems and traditional business intelligence tools. Given the distinct missions of these two organizations, there is little incentive to share data. To encourage a shift in thinking regarding data management practices, GSA has sponsored training initiatives to inform potential users of the benefits of D2D and the modern analytic tools it offers. D2D is an ongoing effort and many organizations within GSA (including the Federal Acquisition Service and Public Building Service) are currently shifting at least some of their data to the system. However, completely eliminating traditional data systems may not be feasible in the near term.

Data quality management is another challenge for the D2D platform. The data normalization process takes place in a decentralized fashion, occurring independently at each organization that contributes data. GSA lacks business rules to extract, transform, and load data, although efforts to establish data standards are underway. GSA is addressing D2D's data quality challenges by sponsoring a data management working group.

3. Results

D2D provides a cloud-based repository for data, along with analytic tools and reports for 7,000 users, mostly internal to GSA. Within D2D, the Data Science Virtual Desktop allows data scientists and analysts to access data and use analytic tools such as Tableau, MicroStrategy, R, and Python. Users of the Data Science Virtual Desktop can publish results and dashboards to the D2D portal for internal and external data consumers, subject to review by data stewards and an executive board. Business owners and data consumers can access analytic reports, products, and dashboards on the D2D portal. In addition to providing a data infrastructure and analytic tools, D2D also organizes a community of practice through working groups and training initiatives.

Although D2D stores data in a centralized manner, data is only shared on a need to know basis. Data stewards from each business line grant specific users access to individual data sets. It therefore appears that analysis of GSA data may still occur in silos by business line, and that cross-cutting research is not necessarily enabled by the system. However, D2D does facilitate greater communication across business lines through working groups. It also

promotes information exchange by providing reports and dashboards in a centralized location.

4. Lessons for Data Sharing

One recommendation from the D2D experience is to establish a Master Data Management Strategy (MDMS) at the onset of EDDIE system development. An MDMS establishes standardized procedures for data entry, aggregation, normalization, and analysis. GSA is in the process of creating a MDMS; however, it would have been beneficial to do so at an earlier stage. Currently, individual organizations within GSA have their own data management plans, with limited visibility to or oversight by the Chief Data Officer. Another recommendation is to define and implement data as a service early in EDDIE's development to standardize data sources and cleaning procedures.

D. Person-Event Data Environment (PDE)

1. Overview of Problem and Key Stakeholders

The Person-Event Data Environment (PDE) is a cloud-based analytic environment with extensive defense personnel data holdings that is used for Army operational support and research.[21] The Army Analytics Group (AAG) built PDE in conjunction with DMDC in 2006 to reduce barriers to data access and to bring together a broad range of personnel data from throughout the DOD. Since then, PDE has negotiated numerous data use agreements, with a goal to make data available "without the need for project-specific data use agreements."[22] The user base for PDE extends beyond the Army and has included analysts from throughout the DOD, the FFRDCs, and select academic organizations. User experience in PDE, however, has been less than desirable. The National Academy of Sciences reported that:

> …a slow and complicated approval process to gain access [to PDE], lengthy reviews for data import and export, limited computational capabilities, concerns about data quality and comprehensiveness, and concerns about data ownership rules pose a significant deterrent to utilizing the PDE. In addition, it is not clear

[21] This case study is based in part on a discussion with a PDE stakeholder at Headquarters, Department of the Army on June 5, 2018. Additional information was gathered in discussions with an AAG representative on September 20, 2017, and with a representative of the Army Training and Doctrine Command Analysis Center-Monterey on December 11, 2017.

[22] Vie, L.L., Griffith, K.N., Scheier, L.M., Lester, P.B., Seligman, M. The Person-Event Data Environment: Leveraging big data for studies of psychological strengths in soldiers. *Frontiers in Psychology*, 4, 2013, 1–7. The quotation is from page 2.

that the architecture scales up in such a way that it can serve all of OUSD(P&R)'s needs.[23]

Recognizing these and other challenges, the Army is reconsidering its strategic direction for PDE. A 2014 task force conducted by the Office of the Assistant Secretary of the Army for Manpower and Reserve Affairs concluded that "the Army must develop a framework to securely, ethically, and legally utilize Big Data in creating a more efficient and ready force."[24] Since no entity within the Army had the overall responsibility for Human Capital Big Data (HCBD), the task force proposed a strategy and implementation plan for building that capability on the pre-existing PDE repository. PDE is currently undergoing a revamp to meet the HCBD objectives.

2. Challenges Encountered

When PDE was created, resources were limited, and it was difficult to forecast how the environment would grow. PDE was not developed with an overarching architecture. It is not an integrated system, making it difficult to navigate. Users have to go back and forth between a workstation portal and a separate portal that holds data catalogs and allows users to submit data requests. The two portals operate as separate web-based applications with different logins and security profiles.

Access requirements are complex, due to derived security requirements from data owners and other stakeholders, without a clear and holistic risk profile for the system and its holdings. Security practices have been implemented without considering the burden they impose on users. For approved users, logging into PDE requires four separate Common Access Card (CAC) authentication points, with a login success rate of roughly 50%. Accessing low-risk, non-medical, fully anonymized data sets requires a formal data request with IRB approval. No data are available in the environment for quick access.

Both the technical and institutional framework of PDE have created a challenge for adapting PDE to the HCBD objectives. Instead of having the luxury of building from scratch, PDE is undergoing a remodeling effort that is constrained by its legacy structure.

The HCBD strategy also entails a cultural shift among senior Army leaders toward a greater integration of data-driven insights into Army policies and practices. It is important to educate senior leaders who will be using the HCBD research about its value, applicability, limitations, and relevance.

---

[23] The National Academy of Sciences, Engineering, and Medicine, *Strengthening Data Science Methods for Department of Defense Personnel and Readiness Missions* (Washington, DC: The National Academies Press, 2017), 5.

[24] "Big Data: Opportunities and Challenges for Human Capital," Department of the Army White Paper, November 12, 2014, 3.

3. Results

PDE is operational, with roughly 100 users over the course of a year. Projects vary in size, but typically involve two to four analysts. Usage tends to fall into three categories: basic usage requiring standard data sets and minimal amounts of computer processing time; high-performance usage involving computationally intensive models; and usage demanding large amounts of support from PDE personnel (such as novice users on a short timeline or projects requiring the acquisition of non-standard data).

PDE protects data using a hierarchical structure. A staging enclave receives PII data files with direct identifiers, such as Social Security numbers. Before analysts see the data, PDE staff anonymize the data by removing direct identifiers.[25] Analysts work with the anonymized data and a suite of tools in an analysis enclave and may submit requests for non-PII results to be exported.

The Army's HCBD effort, which employs PDE as its data environment, is expected to be operational in the coming months. While its current user base is within the Army, once HCBD is fully operational, this resource will hopefully be available for broader use.[26]

4. Lessons for Data Sharing Environments

PDE and EDDIE both seek to provide a central repository for defense personnel data in a secure environment. As an initial foray in this sphere with more than a decade of operational experience, PDE offers numerous lessons for EDDIE—both good and bad. The architecture for EDDIE incorporates PDE's structure of having a highly secure area where PII with direct identifiers are housed and anonymized prior to being released to analysts (the Cold Room in EDDIE and the staging enclave in PDE). EDDIE likewise adopts PDE's notion of having tiered levels of data access.[27] These structural features help to provide appropriate safeguards in handling sensitive data. Other safeguards within PDE have been

---

[25] This includes replacing the Social Security number with project-specific identifiers, limiting birth information to the month and year of birth (as opposed to the day of birth), and other measures to bin highly identifying traits. Even though removing direct identifiers anonymizes the data to some extent, the data are still considered PII since some or all individuals within the data could likely be re-identified through reasonable deductive efforts.

[26] The original strategy for HCBD envisioned making "access to the Army's Big Data assets" available "to stakeholders outside of the Army (e.g., the Veterans Administration or other executive departments; taxpayers who file for access via the Freedom of Information Act (FOIA); or researchers at a university who are working toward scientific discovery)." Ibid, 3.

[27] As stated earlier, PDE has no data that is persistently available for quick access (i.e., data that any authorized PDE can automatically access). However, PDE has what are called *open*, *restricted*, and *private* datasets. Open data require a low-level of approval, while restricted data require a higher-level. Both open and restricted data are indexed in the data catalog, but private data are not listed. EDDIE follows this pattern with the notable exception that all authorized EDDIE users can automatically access the open data (referred to as immediate access data in EDDIE).

restrictive to the point that many users or potential users have been unable to navigate the red tape within a timeline that would make PDE a viable and responsive resource.

The RAND Corporation explored the usability of PDE in an extensive two-year study by conducting three projects within PDE that are representative of its defense personnel research portfolio.[28] The goal of the study was not to obtain specific results for the three projects, but to evaluate how conducive PDE was to facilitating the research process. Relative to their own internal data and computing resources, and their own administrative procedures, the study found PDE to be prone to delays and misaligned incentives. Transaction costs (as discussed in chapter 3) were at the core of many of these delays. Incentives were misaligned between RAND and PDE in the sense that if RAND were conducting the research internally, it could have employed the necessary resources to quickly and efficiently work through any transactions costs. Instead, RAND had to wait—sometimes for weeks or months—for issues to be resolved. Minimizing transaction costs and providing avenues for research institutions to avoid unnecessary delays will be critical to making EDDIE an environment that is conducive to timely research. AAG officials recommend carefully considering the initial architecture for EDDIE and ensuring that it will be able to evolve with the needs of the analysts.

A lesson from the HCBD initiative is to define a common vocabulary. HCBD customers, who are not data scientists, often do not use terms such as "big data" correctly. This is true even for those who are technologically skilled. Misunderstandings will likely create problems throughout the development of EDDIE if they are not addressed by establishing a common vocabulary at the beginning of the development process.

E. Quick Reaction Analysis Team (QRAT)

1. Overview of Problem and Key Stakeholders

In 2011, the Office of Net Technical Assessment (ONTA), in the Office of the Assistant Secretary of Defense for Research and Engineering, developed the Quick Reaction Analysis Team (QRAT) to fill a gap in their ability to conduct quick response analyses.[29] The QRAT is composed of members of multiple FFRDCs, University Affiliated Research Centers (UARCs), and National Labs. The primary participants are IDA, Georgia Tech Research Institute, Lawrence Livermore National Laboratory, Johns Hopkins

[28] The study ran from October 2015 to September 2017. PDE version 2.0 debuted in February 2016, and their analysis is based on that updated system. See Knapp, David, Beth Asch, Christine DeMartini, Teague Ruder, Janet Hanley, *Using the Person-Event Data Environment for Military Personnel Research in the Department of Defense: An Evaluation of Capability and Potential Uses* (Santa Monica, CA: RAND Corporation, 2018).

[29] This case study is based on a discussion with IDA researchers who participate in QRAT, April 20, 2018.

University Applied Physics Laboratory, and MITRE. On a recurring basis, ONTA tasks the QRAT with analytic questions that benefit from a collaborative approach. These efforts generally involve 2 to 4 research institutions and take 2 to 12 weeks to complete.

Unlike EDDIE, QRAT does not have a central data system or repository. The QRAT case study demonstrates the benefits of enhanced collaboration between research institutions—collaborative practices that might be mirrored in EDDIE.

2. Challenges Encountered

The QRAT has encountered several governance challenges. First, there is no set of standards in place for how QRAT operates. As a result, when ONTA leadership changes, the operational procedures for QRAT change. Second, there are no government employees working full time on QRAT. Several government employees are involved with QRAT on a part-time basis, but having at least one full-time government employee to manage communications and to provide continuity and funding would be helpful.

Data sharing is another challenge for the QRAT. Government offices often send data directly to the research organizations involved. Those organizations may also collect third party data and share with each other, but this tends to evolve in an ad hoc manner. Data are typically shared over email, and there is no central repository. While organizations are usually willing to share, there are difficulties in keeping datasets organized and in sync across organizations.

3. Results

There have been several successful QRAT efforts in which FFRDCs, UARCs, and National Labs have collaborated. Analysts across these organizations have grown to know and trust each other through frequent interactions (e.g., weekly meetings). Analysts are able to learn from others' expertise in common communities of interest. As a result, research products are of a higher quality. In cases where an organization does not have expertise internally, QRAT has enabled analysts to be more knowledgeable about where they might seek an answer, which in turn benefits project sponsors within the government.

4. Lessons for Collaboration

Clear standards are needed to harmonize different versions of data and to ensure that all collaborators can find the most recent (or authoritative) version. Frequent interactions between collaborators, as well as a steady stream of projects, have allowed long-term professional relationships to develop across participating research institutions. The QRAT concept works well when sponsors maintain objectivity as to the capabilities of each research institution and propose new projects during weekly meetings when representatives from each institution are available. Because no single institution has the capacity and expertise to accomplish all of the studies, collaboration is both necessary and beneficial to all parties.

This page is intentionally blank.

5. Governance

The success of the EDDIE initiative requires overarching governance, policies, and a shared plan for long-term operations and ongoing enhancement. Stakeholders in the defense personnel research community represent a broad range of organizational types, each with its own research goals, internal structures, and external obligations. To accommodate this variety, the IDA team developed a concept for EDDIE governance to provide necessary oversight while preserving the research independence required for many community stakeholders to use this new resource.[30] This chapter provides a model for the implementation of EDDIE governance through the issuance of a DOD Instruction (DODI) or other official policy statement. We base these suggested governance structures on DODI 1322.26, *Distributed Learning* (5 October 2017), which defines the governance structure for DOD's Adaptive Distributive Learning initiative.

The model language provided here would align the EDDIE core responsibilities and authorities with established OSD and military Service roles, and provide a draft charter laying out the roles and responsibilities of a proposed EDDIE Oversight Committee.

A. Proposed Responsibilities

1. Under Secretary of Defense for Personnel and Readiness (USD(P&R))

The USD(P&R):

1. Acts as the Secretary of Defense's lead for determining and establishing the extent and nature of defense-related personnel data access for analyses of DOD's active duty, reserve, civilian, and contractor workforces.
2. Acts as the Secretary of Defense's lead proponent for and steward of EDDIE policy, programs, and guidelines—including any accreditation and access requirements for using EDDIE—in accordance with all relevant laws, regulations and DOD policies governing PII, data sharing, cybersecurity and information technology systems.

[30] The independent nature of FFRDC research provides significant value to the U.S. government, and is central to the role of FFRDCs as established by Congress. The authors believe that appropriate EDDIE oversight can be compatible with FFRDC research independence for work performed in EDDIE, given appropriate EDDIE governance mechanisms.

3. Issues instructions and guidelines to implement EDDIE initiatives, and acts authoritatively to generate or modify policy for developing, managing, implementing, and evaluating EDDIE.

4. Monitors the implementation of this issuance and related programs; issues supporting guidance, as necessary.

2. Assistant Secretary of Defense for Manpower and Reserve Affairs (ASD(M&RA))

Under the authority, direction, and control of the USD(P&R), the ASD(M&RA):

1. Provides oversight for the execution, development, and financing of EDDIE, and is empowered to implement the vision and direction expressed by the USD(P&R) for EDDIE.

2. Serves as the Chair of the EDDIE Oversight Committee (or appoints a designee to serve as Acting Chair in his or her absence).

3. Assistant Secretary of Defense for Health Affairs (ASD(HA))

Under the authority, direction, and control of the USD(P&R), the ASD(HA):

1. Supports DOD personnel analyses by streamlining, simplifying, and accelerating the human subjects review process for studies in support of personnel analyses, in coordination with the EDDIE Oversight Committee.

4. Deputy Assistant Secretary of Defense for Military Personnel Policy (DASD(MPP))

Under the authority, direction, and control of the USD(P&R), the DASD(MPP):

1. Serves as First Deputy Chair of the EDDIE Oversight Committee (or appoints a designee to serve as the Acting First Deputy Chair in his or her absence).

2. Supports the USD(P&R) and the ASD(M&RA) in the execution of the above stated duties.

5. Director, Office of People Analytics (OPA)

Under the authority, direction, and control of the USD(P&R), the Director, OPA:

1. Serves as the Second Deputy Chair of the EDDIE Oversight Committee (or appoints a designee to serve as the Acting Second Deputy Chair in his or her absence).

2. Supports the USD(P&R) and the ASD(M&RA) in the execution of the above stated duties.

6. Director, Defense Manpower Data Center (DMDC)

Under the authority, direction, and control of the USD(P&R), the Director, DMCD:

1. Supplies defense-related personnel data in support of personnel analyses. This includes data, metadata, data libraries, and other data documentation relevant to understanding and analyzing the data.
2. Collaborates with the EDDIE community in supporting the curation of defense-related personnel data in support of personnel analyses.
3. Designates a representative to serve on the EDDIE Oversight Committee.

7. Secretaries of the Military Departments and the Chief of the National Guard Bureau

The Secretaries of the Military Departments and the Chief of the National Guard Bureau:

1. Support the ASD(HA) in efforts to streamline, simplify, and accelerate the human subjects review process for studies in support of DOD personnel analyses.
2. Designate representatives to serve on the EDDIE Oversight Committee.

B. Proposed EDDIE Oversight Committee Charter

The EDDIE Oversight Committee supports the ongoing implementation, assessment and enhancement of EDDIE. This charter establishes the mission, function, membership, and responsibilities of this organization.

1. Purpose

Subject to the direction and control of the USD(P&R), the EDDIE Oversight Committee acts as the authoritative governing body to establish and amend policies and procedures pertaining to EDDIE; manage EDDIE's ongoing operations; provide strategic direction and determine capability investments for EDDIE; promote resource and information exchange to enhance research related to DOD personnel; monitor emerging technologies and techniques relevant to EDDIE; and maintain financial accountability for EDDIE.

2. Mission

The mission of the EDDIE Oversight Committee is to ensure that the DOD personnel analysis research community is provided with a capable, accessible, effective, responsive, and secure research environment that can be used to further the quality and timeliness of DOD personnel research. This includes facilitating cross-institutional collaboration,

modular modeling development, peer review, transfer of institutional knowledge, and expeditious data access and regulatory approval processes.

The EDDIE Oversight Committee is charged with determining policies and procedures related to research access and use of data maintained by OUSD(P&R) and held in the EDDIE environment under the authority, direction, and control of the USD(P&R). The EDDIE Oversight Committee provides advice to the DOD personnel analysis research community with respect to the policies and procedures included in this issuance. It also:

1. Promotes collaboration among DOD entities and the research community supporting and benefitting from EDDIE.
2. Fosters information and resource sharing among DOD entities and the research community to maximize the return on EDDIE investments.
3. Advocates for adequate resources and funding to support EDDIE operations and development.
4. Establishes mechanisms for amending and streamlining EDDIE policies and procedures.
5. Establishes a method for collecting and resolving user needs and requests and for adjudicating disputes.
6. Monitors science and technologies to expand EDDIE's ability to take advantage of emerging capabilities.
7. Establishes EDDIE Oversight Subcommittees as needed to:
    a. Evaluate and recommend allocation of funding and resources to enhance EDDIE's capabilities.
    b. Investigate and report on the availability and technical capabilities of emerging technologies, tools, standards, and specifications.
    c. Promote and facilitate data use agreements and similar arrangements that extend EDDIE's data holdings on topics pertaining to DOD personnel (including data both within and outside of DOD).
    d. Conduct discussions and participate in working groups on technical, policy, and process related topics including, but not limited to:
        i. System and data security practices
        ii. Research policy and risk
        iii. Communication and community outreach
        iv. Operational data collection and measurement
        v. Human Subject Review policy and procedures

vi. Standards for user approval and conduct

e. Monitor and improve EDDIE performance.

f. Perform other needs and functions designed to enhance research related to DOD personnel.

3. Members

The EDDIE Oversight Committee permanent membership will consist of an Executive Committee and Core Members.

1. The Executive Committee will consist of
    a. Chair filled by the ASD(M&RA), or their designee[31]
    b. First Deputy Chair filled by the Director of MPP, or their designee
    c. Second Deputy Chair filled by the Director of OPA, or their designee
2. Eight core members comprised of:
    a. One representative from each Military Department
        i. Department of the Air Force
        ii. Department of the Army
        iii. Department of the Navy
    b. One representative from the National Guard Bureau
    c. One representative from DMDC
    d. One representative from each of the organizations administering one or more DOD-sponsored FFRDC Study and Analysis Centers:
        i. The CNA Corporation
        ii. Institute for Defense Analyses
        iii. RAND Corporation

Individuals will be selected to serve on the Oversight Committee by the organization they represent, subject to the approval of the Executive Committee.

Only members of the Executive Committee may cast binding votes; Oversight Committee core members may cast non-binding informational votes.

[31] Subject to the direction and control of the USD(P&R), members of the Executive Committee should be able to perform government essential functions, such as source selection for EDDIE investments. Designees should be individuals who can perform such functions.

4. Meetings

The Oversight Committee will meet quarterly, or more frequently as required.

Core members may nominate additional subject matter experts to attend EDDIE Oversight Committee meetings or to participate in subcommittees. The chair, deputy chairs, or core members may recommend other invitees for specific purposes.

Oversight Committee meetings are generally open to representatives from other federal agencies interested in collaborating on personnel analysis or EDDIE. However, the Executive Committee may designate a meeting (or a portion thereof) as closed. Outside groups may bring relevant matters forward to the Oversight Committee, who may address the issue directly or delegate it to review by a subcommittee.

5. Subcommittees

The Oversight Committee may establish a subcommittee as *standing* or *ad hoc*. The Oversight Committee will appoint a lead for each subcommittee. Subcommittees will meet as needed and will report to the Oversight Committee on a timeline determined by the Oversight Committee. Standing subcommittees will address topics that require continuous consideration. Ad hoc subcommittees will be convened as needed to address specific issues as they arise; some ad hoc committees may be convened on a semi-recurring basis to address issues that may need to be revisited at irregular intervals.

Examples of areas to be addressed by standing subcommittees may include:

1. Ongoing EDDIE investment, appropriations planning, and prioritization.
2. Systematic community outreach and communications.
3. Cybersecurity strategy, integration, and prioritization.

Examples of areas to be addressed by ad hoc subcommittees may include, but are not limited to:

1. Negotiating or updating data use agreements.
2. Establishing or updating institutional accreditation and training requirements for various roles within EDDIE.
3. Investigating emerging technologies that have implications for EDDIE.
4. Establishing or updating a dispute adjudication process and other policies and procedures pertaining to EDDIE.
5. Establishing or updating evaluation and feedback mechanisms for monitoring the performance, use, and effectiveness of EDDIE.

6. Responsibilities

All participants in the EDDIE Oversight Committee are responsible for attending meetings of the committee, as well as those of any subcommittees with which they have accepted an assignment.

a. EDDIE Oversight Committee Chair

Advises DOD leadership on current and proposed EDDIE initiatives, actions, and programs. Additionally, the Chair:

1. Facilitates creation of a shared vision and strategy for EDDIE, adhering to guidance from decision authorities.
2. Helps prioritize and guide EDDIE development efforts.
3. Approves, establishes, and retires subcommittees as needed.
4. Represents the interests of the Secretary of Defense regarding EDDIE to the Oversight Committee.

b. EDDIE Oversight Committee Deputy Chairs

Assist and advise the Chair in recommending and implementing EDDIE policies, guidance, best practices, and investments. Additionally, the Deputy Chairs:

1. Schedule, coordinate, and execute Oversight Committee meetings, including developing agendas, preparing papers and briefings, and documenting and disseminating meeting results.
2. Implement Oversight Committee recommendations and decisions.
3. Specific duties of each Deputy Chair are assigned at the discretion of the Executive Committee to best fulfill the responsibilities listed.

c. Core Oversight Committee Members

Assist and advise the Chair in recommending and implementing EDDIE policies, guidance, best practices, and investments. Core members also:

1. Identify and recommend capabilities, research initiatives, and best practices for possible adoption in EDDIE.
2. Represent the defense personnel research community in implementing and managing EDDIE. This includes facilitating communication between the EDDIE user community and the EDDIE Oversight Committee.
3. Recommend the organization, membership, topics, tasks, and priorities for subcommittees as required.

d. Subcommittee Members

Assist and advise the Oversight Committee through the execution of the subcommittee's designated function. Subcommittee members also:

1. Investigate specific topics of interest for EDDIE as assigned by the Oversight Committee.
2. Report on progress and findings of subcommittee efforts to the Oversight Committee.
3. Recommend courses of action to the Oversight Committee based on their findings.

6. Use Standards

OUSD(P&R) intends for EDDIE to provide a means for the military personnel research community to increase and improve its level of cross-organizational collaboration and peer review in modeling. Achieving this goal requires a high level of communication among research teams on analytic details, which is facilitated by users adopting and adhering to defined standards for scientific programming and documentation. This section recommends best practices for coding and documentation in EDDIE, suggests processes for peer review of research products shared in EDDIE, and makes recommendations for how sponsors might incentivize adherence to these standards and derive greater value from the resulting research.

A. Best Practices in Scientific Programming

Appropriate standards for scientific programming depend on the task at hand and the intended audience. If OUSD(P&R) desires to enable collaboration, reduce errors, promote peer review, and effectively build on earlier efforts, it should adopt standards to promote organization, version control, clarity, reliability, optimization, portability, and modularity in the resulting research products: code, documentation, project directories, and databases. Meeting these high standards will be costly in the short run, but necessary if EDDIE is to foster collaboration, store vast quantities of cleaned data, and promote the development and use of modular code for overlapping research objectives.

Adoption of these standards would represent a significant behavioral shift for many members of the military personnel research community. The intended users of EDDIE have extensive analytical research experience, but typically on small-team projects (one to six individuals) with short-term deliverables and no intended code sharing or reuse. Because of the limited scale of these projects, analysts have not generally employed the tools necessary for larger, shared efforts. To assist analysts in advancing beyond ad hoc coding methods, Gentzkow and Shapiro (2014), Wilson et al. (2014), and Wilson et al. (2017) each prescribe best practices for scientific programming.[32] We incorporate much of their

[32] Gentzkow, Matthew and Jesse Shapiro, "Code and Data for the Social Sciences: A Practitioner's Guide" https://web.stanford.edu/~gentzkow/research/CodeAndData.pdf. Wilson, Greg, Jennifer Bryan, et al. "Good enough practices in scientific computing." PLoS computational biology 13, no. 6 (2017): e1005510. Wilson, Greg, Dhavide A. Aruliah, et al. "Best practices for scientific computing." PLoS biology 12, no. 1 (2014): e1001745.

direction in the Use Standards Checklist (Table 2). Wherever this checklist seems unclear or its contents unfamiliar, we refer readers to the source articles.

This checklist serves two purposes. First, it summarizes essential dictums from the above authors. Second, it serves as a yardstick for adherence to use standards; when one EDDIE user reviews the work of another, he can quickly assess whether the other has adhered to these best practices.

In addition, programming should adhere to stylistic guidelines to promote readability, reuse, and review. We refer to Martin (2009) and Hunt and Thomas (2000) for general guidance. Table 1 provides language-specific style guides. The content of these guidelines provide detail for some otherwise general style prescriptions in the Use Standards Checklist.

Table 1. Language-Specific Style Guides for EDDIE

Language	Style Guide	URL
Python	*PEP 8 Style Guide for Python Code* (Rossum et al.)	https://www.python.org/dev/peps/pep-0008/
R	*The Tidyverse Style Guide* (Hadley Wickham)	http://style.tidyverse.org/
Stata	*Suggestions on Stata Programming Style* (Cox 2005)	https://www.stata-journal.com/sjpdf.html?articlenum=pr0018
MATLAB	*MATLAB Style Guidelines 2.0* (Johnson 2014)	http://www.datatool.com/downloads/MatlabStyle2%20book.pdf
Julia	*Julia Style Guide*	https://docs.julialang.org/en/v1/manual/style-guide/

B. Peer Review

Papers, code, and other research products that are submitted for public dissemination in EDDIE should be subject to a peer review process. Whenever possible, at least one reviewer should be from a different EDDIE-using institution than the one conducting the research. The review of code should include evaluating its adherence to the Use Standards Checklist in Table 2. If the code fails to meet any of these standards, or if the reviewers raise other substantive shortfalls, these should be addressed prior to public dissemination in EDDIE. If any points are not addressed, the code should be accompanied by comments indicating how it fails to meet these standards (the peer review report and completed Use Standards Checklist should suffice). Since external peer review helps to ensure research quality and validates the research products for incorporation in subsequent work, sponsors should be made aware of the current and future benefits. Finally, the quality of the peer review process and its results should be periodically assessed by the EDDIE Oversight Committee to ensure its continued adherence to high academic standards.

Table 2. Use Standard Checklist

Version Control	All code and data should use version control
	Version control should adhere to the Major.Minor.Patch convention[33]
	Raw data files should never be changed except to correct data entry errors; for such changes, the corrected raw data should have a new version number
	The version of the code and data used to create a final paper should be permanently stored
	Significant change to code requires a new version
	When data cleaning involves imputation, different imputation approaches entail a major version increment; code used for imputing data must always be available with the imputed data
Files and Directories	Place projects in eponymous directories. Names of files should likewise reflect their function (e.g., "merge_data.do")
	Separate files into input and output folders
	The following belong in their own subfolders: project text documents (includes a changelog); metadata; intermediate data files; results; source code for project-specific programs
	Make project directories portable
	A single command or script must be able to execute all code[34]
	A README file in the main project directory should point to all documentation. Documentation should include data origins, fields, and values
Database Management	Normalize data tables. All tables must feature unique, non-missing keys; all variables must be attributes of the key (not an attribute of an attribute)[35]
	Construct a second set of normalized files that include transformation of the original variables as required for analysis.
	As a final step, merge together the tables in the database to form the rectangular array on which the model is estimated. At this stage, the database should still have unique, non-missing keys, but will likely not be normalized

[33] Patch versions represent bug fixes; minor versions represent added functionality that is backwards-compatible; major versions represent functionality changes that are not backwards-compatible. See Preston-Werner, T. "Semantic Versioning 2.0.0" https://semver.org/.

[34] For reproducibility, only those results produced by this command or script should be included in the final paper.

[35] For instance, state population would be an attribute of a state, but not of a county in that state.

General Style	Write code so that people can read it
	Maintain consistent formatting and style in line with the language specific style guide
	Keep code short and purposeful. Factor long scripts (more than several hundred lines) into smaller functions; factor long functions (more than 80 lines) into subfunctions
	Order functions for linear reading (subfunctions should appear immediately after higher level functions that call them)
	Modular code receives parameters rather than hardcoding values (e.g., input file paths, scaling factors, fault tolerances)
	Break complicated algebraic code into pieces (algebra is difficult to read when expressed as code)[36]
Naming	Use distinctive function and variable names. Avoid having multiple objects whose names do not delineate how they are different
	Use descriptive function and variable names. Avoid abbreviations unless these will be both consistent and obvious for likely readers.
	Make logical (TRUE/FALSE) switches intuitive
Functions and Testing	Utilize functions from well-maintained libraries/packages whenever possible; otherwise, construct custom functions
	Use unit testing or assertions on custom functions.[37] These should include informative error messages. Check for errors that would cause difficult-to-detect problems, like infinite loops or incorrect but plausible answers
	Explicitly state dependencies, requirements, inputs, and outputs for custom functions
	Rarely use global variables in custom function; use local variables instead
	Whenever possible, supplement the documentation for custom functions with simple examples or test data sets
Commenting	Code should be self-documenting. Good comments do not substitute for confusing code. Avoid creating comments if they are unlikely to be maintained—internal inconsistency between comments and code can cause enormous confusion
	Document the purpose, design, interfaces, or reasons for a programming choice (e.g., using one algorithm over another). The code itself should speak to the actual implementation
	Embed documentation and comments close to the code so that they can be easily discovered and maintained. Provide a brief explanatory document at the start of every program.
	For programs that may be used by others, maintain a wiki or other shared list that tracks known issues, bug fixes, added functionality, etc.

[36] Comments linking to a document showing the relevant equations are desirable for complex calculations.

[37] Off-the-shelf unit testing libraries should be utilized when possible. Turn bugs into test cases. Parsimony is a virtue; avoid error checking for cases that add little functionality (e.g., testing that an input is not a string when the input is highly unlikely to be a string).

Optimization	Optimize software only after it works correctly
	Profile the program to identify computationally slow sections. Optimize bottlenecks, particularly if they likely to be used repeatedly
	For slower sections, it is permissible to store intermediate results as output files (e.g., storing parameter estimates from a structural model)
Attribution	Make programs citable. When applicable, make the license explicit
	Research products that use code created for other research products must cite the original code or papers
Modeling	When feasible, make models modular and flexible.[38]

C. Code and Data as Deliverables

Higher standards within EDDIE will result in higher fixed costs for initial research, but will yield broader payoffs to the defense personnel research community and higher quality results for the DOD. For work performed in EDDIE, research sponsors should expect peer-reviewed code and data used in the execution of the research to be submitted as deliverables by the project teams. For data accessible within EDDIE, the data deliverable may entail documenting the specific data files, fields, inclusion conditions, time periods, and versions used.

[38] For instance, since small changes to a maximum likelihood model require rewriting the likelihood function, subsequent work cannot easily build on models estimated via maximum likelihood. The simulated method of moments is more forgiving in this regard.

This page is intentionally blank.

7. Risk

This chapter lists a number of the typical risks associated with the use of a shared analytic environment data repository and ways in which those risks can be mitigated. Since this report does not address the technical implementation of EDDIE, these risks are framed at a high-level and are admittedly not exhaustive.

A. Limited Computing and Performance Risks

- **Risks**
    - Lack of computing resources to meet processor demand slows or denies service to some or all users.
    - Lack of adequate data storage impedes data ingestion or availability of needed tools.
    - Lack of scalable hardware limits the ability to expand to meet growing user demands (either in terms of a growing user base or growing computational needs for existing users).
    - Lack of modernization capabilities and resources limits performance as technologies evolve. This includes a lack of hardware to support new technologies (e.g., sufficient hardware to support GPUs), as well as a lack of a recapitalization process for routinely updating hardware.
- **Mitigations**
    - Implement the analytic environment with optimized hardware. This may include housing the data store in a cluster of powerful, dedicated servers with multiple multi-core processors; provisioning the servers with large RAM capacity; and using high-speed, solid-state hard drives for data persistence.
    - Partition the environment into separate sub-data stores hosted in high-performance machines that can be managed as a single instance using frameworks for distributed storage and processing, such as Hadoop.[39]

[39] "Welcome to Apache™ Hadoop®!," last accessed July 19, 2018, http://hadoop.apache.org/.

- Use, where appropriate, the equivalent of "sandboxes" for queries or processes that take substantial time to complete (e.g., tens or even hundreds of minutes). This approach means that the slow queries are executed against a subset of computation resources or during non-peak hours, leaving sufficient processing time for less intensive demands.
- Host the entire environment in a highly scalable cloud solution that can handle not only the data volume demands, but also the processing requirements. Cloud computing services also enable ongoing modernization of computing capabilities.

B. Data Quality Degradation Risks

- **Risks**
  - Lack of reliable data above a minimal quality threshold will negate and degrade the value of data in EDDIE.
  - Lack of reasonable oversight and curation of the data assets committed to EDDIE limits the ability for those assets to continue to be accessible and usable by users.
  - Lack of curated metadata and other pertinent documentation further limits the usability of the data assets. Poor data documentation and organization can turn a *data lake* into a *data swamp*.[40]
- **Mitigations**
  - Implement a life-cycle management strategy for the preservation and protection of all digital assets in the analytic environment, as well as a data governance process to ensure data quality.[41]
  - Build a conceptual information model to define the metadata needed to characterize data assets in the analytic environment. Associate the appropriate metadata with each piece of data.
  - Enable sufficient resources for data curation.
  - Explore automation techniques to reduce human error.

---

[40] Thor Olavsrud, "3 keys to Keeping Your Data Lake from Becoming a Data Swamp," June 8, 2017, https://www.cio.com/article/3199994/big-data/3-keys-to-keep-your-data-lake-from-becoming-a-data-swamp.html.

[41] See, for example, Earley, Susan, and Deborah Henderson. 2017. *DAMA-DMBOK: Data management body of knowledge.* An analysis on issues related to data quality can be found in Francisco L. Loaiza-Lemos et al., *Development of a Data Quality Framework for Creating and Maintaining Army Authoritative Data Sources*, IDA Document D-4275 (Alexandria, VA: Institute for Defense Analyses, March 2011).

- Maintain careful versioning of data.
- Define and implement metrics for data quality.
- Conduct periodic reviews of emerging technologies applicable to data quality maintenance.

C. Cybersecurity Risks

- **Risks**
  - Lack of sufficient firewalls and access limitations exposes defense personnel data and unreleased analyses to cyberattacks intended to damage or exfiltrate the environment's contents. It is DOD policy to treat information as a strategic asset and to protect it to the maximum extent possible.[42]
  - Lack of sufficient protocols and safeguards for analysts working within the environment exposes the environment's contents to insider threat.
  - Overly burdensome or poorly prioritized cyber defensive measures degrade the utility of the environment.
- **Mitigations**
  - Multiple government organizations have published guidance for mitigating cybersecurity issues. A list of relevant publications is in Table 3. In addition to DOD and Congress, these organizations include:
    - National Institute of Standards and Technology (NIST)
    - Committee on National Security Systems (CNSS)
    - Federal Information Processing Standards (FIPS)
  - In addition to the above, Executive Order 13800 (May 11, 2017) directs all federal agencies to use the NIST Framework for Improving Critical Infrastructure Cybersecurity.[43]
  - Given the sensitivity of data that will likely reside in EDDIE, if EDDIE were managed through a cloud provider, the provider would need

[42] DoDD 8000.1, *Management of the Department of Defense Information Enterprise (DOD IE)*, 17 March 2016.

[43] NIST, *Framework for Improving Critical Infrastructure Cybersecurity*, Version 1.1, 16 April 2018, https://doi.org/10.6028/NIST.CSWP.04162018 (see also https://www.nist.gov/cyberframework).

Federal Risk and Authorization Management Program (FedRAMP) certification for at least DOD Impact Level 4.[44]

- Data encryption, both while data is at rest and while in transit, can be helpful in reducing the potential damage associated with data exfiltrated during a cyberattack.
- Ensure that cybersecurity measures are weighed and implemented through an active risk management process and that implementation provides value added commensurate with the threats that the measures are intended to mitigate.

Table 3. Summary of Federal Cybersecurity Publications

Publication Number	Title
40 USC40 USC § 11331	Federal Information Security Management Act (FISMA) of 2002
NIST SP 800-37, Revision 1	Guide for Applying the Risk Management Framework to Federal Information Systems: A Security Life Cycle Approach
NIST SP 800-53, Revision 4	Security and Privacy Controls for Federal Information Systems and Organizations
NIST SP 800-30, Revision 1	Guide for Conducting Risk Assessments
NIST SP 800-171, Revision 1	Protecting Controlled Unclassified Information in Nonfederal Information Systems and Organizations
FIPS 199	Standards for Security Categorization of Federal Information and Information Systems
FIPS 200	Minimum Security Requirements for Federal Information and Information Systems
CNSSI 1253	Security Categorization and Control Selection for National Security Systems
DODI 8500.01	Cybersecurity
DODI 8510.01	Risk Management Framework (RMF) for DOD Information Technology (IT)
CJCSM 6510.01B	Cyber Incident Handling Program

[44] Impact Level 4 permits Controlled Unclassified Information (CUI), including PII and PHI. "If a higher level of protection [is] deemed necessary by the information owner, public law, or other government regulations," Impact Level 5 may be required. See sections 3.2.4 and 3.2.5 on Impact Levels 4 and 5 at https://iase.disa.mil/cloud_security/cloudsrg/Pages/ImpactLevels.aspx (last accessed 19 July 2018).

D. Community Buy-in Risks

- **Risks**
  - Lack of any of the following could deter analysts from using EDDIE:
    - An efficient protocol for obtaining user access
    - Timely access to data within EDDIE
    - Sufficient computing resources
    - Adequate personnel to quickly fulfill any tasks that the analysts may not be permitted to do themselves
    - An easy-to-learn, usable design for working in EDDIE
  - Lack of a critical mass of analysts working within EDDIE limits the benefits of collaboration and the extent of institutional knowledge that can be recorded and shared.
  - Lack of clear expectations from research sponsors for code and data documentation, as well as peer review, limits the extent to which use standards conducive to replicability, reuse, and modularity will be adopted.
- **Mitigations**
  - Establishing and maintaining procedures that facilitate rapid data access, minimal administrative hurdles, and sufficient computing resources will provide incentives for analysts to work in EDDIE.
  - Ongoing investments to improve the data assets and library within EDDIE likewise contribute to building a critical mass of users.
  - For sponsored research, requiring that code and data documentation are deliverables will facilitate a more rapid adoption of use standards.
  - Tutorials, examples, and other training resources could expedite and ease the transition to working effectively within EDDIE.

This page is intentionally blank.

8. Recommendations and Cautions

A. Problem Statement Priorities

The *Enterprise Data to Decisions Statement of Need/Problem* drafted by the Office of People Analytics within OUSD(P&R) highlights three high-priority capabilities for EDDIE: data management, optimized data analytics, and collaboration.

Data management includes the capability for users and decision makers to quickly acquire, access, securely store, validate, and analyze data. These data should be accurate, complete, consistent, and reconcilable across processes and organizations.

Optimized data analytics encompasses the capability for users to use analytic tools to derive actionable information from data. Users should be able to perform a variety of analytic techniques ranging from ad hoc queries to predictive analysis using modern software tools and high-throughput computing techniques.

Collaboration requires the ability to share data and analytic models across organizations. These work products should be preserved and documented in the environment, and available for reuse when appropriate. This level of collaboration will enable information sharing and peer review within the department and among federal agencies and external stakeholders to improve research quality and reduce redundant effort.

B. Summary of EDDIE Requirements

To meet these priorities, using the collaborative methodology described in preceding sections, the IDA requirements team developed the following high-level requirements.

- **Analytic Environment.** EDDIE should be an analytic environment in which users may access PII data and perform complex analyses.
- **Data Access.** EDDIE should enable users to focus primarily on analysis rather than data access or cleaning.
- **Library.** EDDIE should support and contain a Library, including data dictionaries, metadata, public code, a wiki collecting institutional knowledge, and a repository of user profiles.
- **Team Workspaces.** EDDIE should enable project teams to share data, code, and metadata in a workspace.

- **Import and Export Control.** EDDIE should support the import of PII and non-PII items and the export of non-PII items as requested by users.
- **HSR Approval.** EDDIE should streamline the HSR and IRB approval process for projects operating in EDDIE.
- **Institutional Memory.** EDDIE should 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.
- **Analytic Tools.** EDDIE should 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.
- **Computing Resources.** EDDIE should provide users with sufficient hardware computing capacity.
- **Enable Collaboration.** EDDIE computing structures, implementation, governance, and management practices should enable collaboration in team workspaces among individuals both within and between institutions.

C. Cautions

Our extensive dialogue with the defense personnel research community revealed the significance of several key choices in shaping an environment that enables economies of scale across the community in the data preparation and model development process. The range of benefits include rapid and secure data access, increased vetting and dispersion of ideas, reproducibility and transparency of results, and greater responsiveness in providing actionable information to DOD leadership. Various combinations of these benefits may be achieved with lesser alternatives. However, lesser alternatives would also perpetuate the costly inefficiencies of current practices.

1. Alternative: Omit data hosting and management

Acquiring, tracking, combining, maintaining versions, and controlling access to data requires significant effort and investment. Constructing a collaborative space that does not include personnel data or its management would reduce the level of staffing required to support the environment and reduce the computing requirements at the EDDIE level, but replicate those expenses in other research organizations.

An environment without personnel data would still permit collaboration on model development, although at a significantly reduced level. The resulting environment might operate similarly to a GitHub or BitBucket site and would enable sharing of code, documentation, and institutional knowledge.

Implementing this alternative would not address one of the primary pain points the community is currently experiencing with respect to obtaining data access and performing redundant data management efforts. In addition, this alternative would impair the ability of analysts to collaborate quickly on modeling efforts because developing a model is intrinsically dependent on the underlying data. Data documentation and institutional knowledge about data loses its value as it drifts farther from the corresponding data. Separating the data from its metadata introduces challenges in keeping metadata current, relevant, and linked to the data it describes. Furthermore, separating the modeling code from the data on which it is built inhibits replication and peer review.

2. Alternative: Narrow library holdings

The structure of EDDIE reflected in the requirements document includes a library with data catalogs, data dictionaries, metadata, reusable code, and a user forum for asking questions about the data that others in the community can answer. Organizing, moderating, and maintaining the proposed library reflects a significant investment. Reducing the library's scope would reduce the initial development costs and the subsequent support costs.

However, removing the library or limiting its contents impedes the extent to which information about data is transferable across individuals and institutions, inhibiting EDDIE's core purpose of sharing of information between teams. Being able to review metadata and notes provided by data providers and analysts on the accuracy, completeness, consistency, and reconcilability of data significantly reduces the time required to understand and prepare data for analysis. The cost savings in analyst time likely outweighs the investment in developing a well-populated library. Such a library also insulates DOD from information losses due to staff turnover.

3. Alternative: Reduce automated features

A variety of subsystems requiring automation appear in the requirements document, such as an automated system for documenting and fulfilling data requests. Developing automated tools requires upfront investments. However, the return on these investments in terms of reduced labor costs, increased accuracy, and shorter request processing times can be substantial—especially for high-volume, error-prone, labor-intensive tasks.

4. Alternative: Limit software choices

The requirements document lists a number of software tools, including a number of proprietary, commercial programs with a variety of expensive and complicated licensing structures. Including only open-source software (perhaps supplemented by a few commercial software solutions) would still enable users to perform rigorous, complex, and meaningful analyses in the environment. As open-source analytical software like R,

Python, and Julia have matured, they have come close to or surpassed the capabilities of their commercial peers in analytical capability. In fact, many of the most advanced machine learning techniques are only possible in open-source software.

Excluding commercial software like Stata, Matlab, or SAS restricts the set of tools available to analysts when working in EDDIE. There is a time cost to analysts—and, by extension, to DOD research budgets—in identifying and learning alternative tools. Moreover, alternatives may not have equivalent functionality, imposing a cost in analytical rigor by not being able to choose the desired method or algorithm.

5. Alternative: Less responsive timeline expectations

Setting judicious expectations for the maximum time required to fulfill user requests helps to ensure that users are able to operate and deliver research products without undue delays. The requirements document quantifies time expectations for data requests, technical support, import and export control, and other actions. Relaxing these expectations to permit more time for fulfilling support requests would reduce the number of personnel required to staff the environment. However, making the system less responsive may not support the timeline of government offices requesting research to inform upcoming policy decisions. There are systems that analysts avoid using because the delays are not conducive to effective workflows.

Appendix A.
Illustrations

Tables

This page is intentionally blank.

Appendix B.
References

Apache. "Welcome to Apache™ Hadoop®!." Accessed July 19, 2018. http://hadoop.apache.org/.

Department of Defense, Office of People Analytics (OPA), "Enterprise Data to Decisions Statement of Need/Problem." Draft, April 2018.

Department of Defense. "Distributed Learning (DL)." DOD Instruction 1322.26. October 5, 2017.

Department of Defense. "Management of the Department of Defense Information Enterprise (DOD IE)." DOD Directive 8000.1. March 17, 2016 (Incorporating Change 1, July 27, 2017).

Department of the Army. "Big Data: Opportunities and Challenges for Human Capital," White Paper, November 12, 2014, 3.

Federal Acquisition Regulation (FAR). FAR 35.017, 2015.

Government Accountability Office. DOD Systems Modernization: Maintaining Effective Communication Is Needed to Help Ensure the Army's Successful Deployment of the Defense Integrated Military Human Resources System. GAO-08-927R. Washington, D.C.: US Government Accountability Office, September 2008. https://www.gao.gov/assets/100/95723.pdf, 1.

Information Assurance Support Environment/ "3 Information Security Objectives/Impact Levels." Last updated January 23, 2015. https://iase.disa.mil/cloud_security/cloudsrg/Pages/ImpactLevels.aspx

Hunt, Andrew and David Thomas. *The Pragmatic Programmer: From Journeyman to Master*. Stoughton, Massachusetts: Addison Wesley Longman, 2000.

Knapp, David, Beth Asch, Christine DeMartini, Teague Ruder, Janet Hanley. *Using the Person-Event Data Environment for Military Personnel Research in the Department of Defense: An Evaluation of Capability and Potential Uses.* Santa Monica, CA: RAND Corporation, 2018.

Lanzano, Rosemary. E-mail from Ms. Lanzano to Dr. David Chu, IDA President, on December 21, 2017.

Loaiza-Lemos, Francisco L., et al. *Development of a Data Quality Framework for Creating and Maintaining Army Authoritative Data Sources*. IDA Document D-4275. Alexandria, VA: Institute for Defense Analyses, March 2011.

Martin, Robert. *Clean Code: A Handbook of Agile Software Craftsmanship.* Westford, Massachusetts: Pearson Education, 2009.

National Academy of Sciences, Engineering, and Medicine. *Strengthening Data Science Methods for Department of Defense Personnel and Readiness Missions*. Washington, DC: The National Academies Press, 2017.

National Institute of Standards and Technology (NIST), "Framework for Improving Critical Infrastructure Cybersecurity" Version 1.1, 2018. https://doi.org/10.6028/NIST.CSWP.04162018

Office of the Under Secretary of Defense for Personnel and Readiness (OUSD (P&R)), ADL Initiative website, https://www.adlnet.gov/tla/, accessed June 18, 2018. See also the Experience API website: https://xapi.com/.

Olavsrud, Thor. "3 Keys to Keeping Your Data Lake from Becoming a Data Swamp," June 8, 2017, https://www.cio.com/article/3199994/big-data/3-keys-to-keep-your-data-lake-from-becoming-a-data-swamp.html.

Pechacek, Julie, Alan Gelder, Ethan Novak, Amrit Romana, Paul Richanbach, Kathy Conley, George Kennedy, Cheryl Green. *User Requirements for the Enterprise Data to Decisions Information Environment*. IDA Document NS D-9139. Alexandria, VA: Institute for Defense Analyses, August 2018.

Preston-Werner, T. "Semantic Versioning 2.0.0" https://semver.org/.

Raybourn, E.M., S. Schatz, J. Vogel-Walcutt, & K. Vierling. "At the Tipping Point: Learning Science and Technology as Key Strategic Enablers for the Future of Defense and Security. In Proceedings of the Interservice/Industry Training, Simulation and Education Conference (I/ITSEC)." (2017), Orlando, FL.

Vie, L.L., Griffith, K.N., Scheier, L.M., Lester, P.B., Seligman, M. "The Person-Event Data Environment: Leveraging big data for studies of psychological strengths in soldiers." Frontiers in Psychology, 4, 2013, 1–7.

Winkler, S.J., Witte, E., Bierer B.E. "The Harvard Catalyst Common Reciprocal IRB Reliance Agreement: An Innovative Approach to Multisite IRB Review and Oversight." Clin. Transl. Sci. 8(1) (2015), 57–66.

Wilson, Greg, Jennifer Bryan, Karen Cranston, Justin Kitzes, Lex Nederbragt, and Tracy K. Teal. "Good enough practices in scientific computing." PLoS computational biology 13, no. 6 (2017): e1005510.

Wilson, Greg, Dhavide A. Aruliah, C. Titus Brown, Neil P. Chue Hong, Matt Davis, Richard T. Guy, Steven HD Haddock et al. "Best practices for scientific computing." PLoS biology 12, no. 1 (2014): e1001745.

Appendix C.
Abbreviations

AAG	Army Analytic Group
ADL	Advanced Distributed Learning
ASD(HA)	Assistant Secretary of Defense for Health Affairs
ASD(M&RA)	Assistant Secretary of Defense for Manpower and Reserve Affairs
CAC	Common Access Card
CNSS	Committee on National Security Systems
COTS	Commercial-Off-the-Shelf
CUI	Controlled Unclassified Information
D2D	Data 2 Decisions
DASD(MPP)	Deputy Assistant Secretary of Defense for Military Personnel Policy
DIMHRS	Defense Integrated Military Human Resources System
DMDC	Defense Manpower Data Center
DOD	Department of Defense
EDDIE	Enterprise Data to Decisions Information Environment
FAR	Federal Acquisition Regulation
FFRDC	Federally Funded Research and Development Center
FIPS	Federal Information Processing Standards
GSA	General Services Administration
HCBD	Human Capital Big Data
HSR	Human Subjects Review
IDA	Institute for Defense Analyses
IPUMS	Integrated Public Use Microdata Series
IRB	Institutional Review Board
MCTFS	Marine Corps Total Force System
MDMS	Master Data Management Strategy
NIST	National Institute of Standards and Technology
ONTA	Office of Net Technical Assessment
OPA	Office of People Analytics
OPM	Office of Personnel Management
OSD	Office of the Secretary of Defense
OUSD(P&R)	Office of the Under Secretary of Defense for Personnel and Readiness
PA&E	OSD Program Analysis and Evaluation
PDE	Person-Event Data Environment

PHI	Protected Health Information
PII	Personally Identifiable Information
QRAT	Quick Reaction Analysis Team
RAM	Random Access Memory
UARC	University Affiliated Research Center
USD(P&R)	Under Secretary of Defense for Personnel and Readiness
VA	Department of Veterans Affairs
xAPI	Experience API (application programming interface)

REPORT DOCUMENTATION PAGE	Form Approved OMB No. 0704-0188

1. REPORT DATE (DD-MM-YY)	2. REPORT TYPE	3. DATES COVERED (From – To)
xx-09-2018	Final	

4. TITLE AND SUBTITLE

Considerations for Implementing a Defense Personnel Research Environment

5a. CONTRACT NO.

HQ0034-14-D-0001

5b. GRANT NO.

5c. PROGRAM ELEMENT NO(S).

6. AUTHOR(S)

Julie Pechacek
Alan Gelder
Amrit Romana
Ethan Novak
Kathy Conley
Cheryl Green
Dina Eliezer
P.M. Picucci
George Kennedy
Cullen Roberts

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 Paper P-9254
Log: H 18-000375

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

To build and manage its workforce effectively, the U.S. Department of Defense (DOD) oversees an extensive military personnel policy research portfolio. The DOD seeks to develop a new data-hosting analytic computing resource—the Enterprise Data to Decisions Information Environment (EDDIE)—to enable faster, broader access to defense personnel data throughout the research community supporting these efforts, and to foster cross- organizational collaboration and modeling. This report supplements the research community's user requirements for EDDIE documented in Pechacek et al. (2018; IDA NS D-9139). Here, we describe our methodology for collecting and synthesizing user requirements from the research community, highlight core findings, and give illustrative case studies of other government initiatives for providing data or collaborative resources. We observe that sustained commitment from senior leadership, adequate legal and regulatory authorities, and thoughtful design contribute to success in these programs. Additionally, we develop a model for EDDIE governance, provide use standards to facilitate modular modeling development in the environment, and document risks and recommendations to consider when acquiring and implementing EDDIE.

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	72	19b. TELEPHONE NUMBER (Include Area Code) (703) 571-0114

This page is intentionally blank.