COMBINED MODULE ANNOUNCEMENT

FOR

ADVANCED RESEARCH PROJECTS AGENCY FOR HEALTH
BIOMEDICAL DATA FABRIC TOOLBOX

ARPA-H-MAI-24-01-01

AMENDMENT 01

November 14, 2023

AMENDMENT 01 is being issued to remove the requirement for Cooperative Agreement Stage 1 submissions to be submitted to Grants.gov. Additionally, Proposers requesting a Cooperative Agreement are NOT required to submit Forms 1 and 2 during Stage 1 submissions. Proposers requesting a Cooperative Agreement must use Attachment 2 – Cooperative Agreement Bundle (Volume 1) to complete Stage 1 Volume 1 submissions. See section 5.B of this Module Announcement for revisions noted in red font.

TABLE OF CONTENTS

1. MODULE ANNOUNCEMENT OVERVIEW INFORMATION

FEDERAL AGENCY NAME: Advanced Research Projects Agency for Health (ARPA-H)

FUNDING OPPORTUNITY TITLE: ARPA-H Biomedical Data Fabric (BDF) Toolbox

ANNOUNCEMENT TYPE: Initial Announcement

FUNDING OPPORTUNITY NUMBER: ARPA-H-MAI-24-01-01

ASSISTANCE LISTING NUMBER: 93.384 Research and Development that accelerates better health outcomes for all Americans.

DATES: (All times listed herein are Eastern Time)

- ***Module Announcement DRAFT release date***: September 13, 2023
- ***Module Announcement DRAFT 2 release date****:* September 29, 2023
- ***Initial Questions & Answers (Q&A) due date***: September 20, 2023
- ***Module Announcement release date***: *October 20, 2023*
- ***Questions and Answers (Q&A) release date:*** *October 20, 2023*
- ***Proposal due date***: November 28, 2023

2. OPPORTUNITY DESCRIPTION

The Advanced Research Projects Agency for-Health (ARPA-H) is soliciting proposals for the ARPA-H Biomedical Data Fabric (BDF) Toolbox using a new Master Announcement Instruction solicitation strategy. For more information, please refer to the attached Master Announcement Instructions (MAI), which aim to provide proposers with the ability to scale the level of effort that they spend developing proposal materials with the magnitude of the effort that they plan to propose. The MAI introduces a tiered approach to proposal submission for small scale (BIT / BYTE), mid-scale (KILO/MEGA), and large scale (GIG/TERA) efforts. The ARPA-H BDF Toolbox solicitation described below is a Combined Module Announcement that is soliciting proposals at the BIT, BYTE, and KILO scales. This Module Announcement is issued under the Master Announcement Instructions (MAI), ARPA-H-MAI-24-01. All awards will be made in the form of an Other Transaction (OT) or Cooperative Agreement.

Specifically, ARPA-H is soliciting innovative proposals for research and development (R&D) in data integration and usability technologies. Proposed R&D should investigate innovative software approaches that enable revolutionary advances in the collection and usability of biomedical datasets that originate from thousands of different research labs, clinical care centers, and other sources of data in order to accelerate technical innovation across the health ecosystem. Specifically excluded are submissions that will primarily result in limited or only evolutionary improvements to the existing state of software capabilities to link datasets. Also excluded are approaches that artificially restrict data availability beyond what is necessary at a given level of data sensitivity.

A. INTRODUCTION

When effectively harnessed, biomedical data contributes to the development of next-generation treatments and cures. With the right data, researchers can construct a comprehensive and detailed picture of the relationships between human or pathogen molecular variations and potential disease states, and correlations between treatments and health outcomes, and among emerging public health events, risk factors, and transmission patterns. All together these insights have the potential to find new interventions more quickly and prevent and treat disease more effectively. Building such a comprehensive capability requires a disciplined approach to the integration of disparate data sources including, but not limited to, longitudinal patient data, treatment outcomes,

information about disease progression, clinical observations, genomics/proteomics/metabolomics or other "omics" data, imaging, and other foundational biomedical R&D experimental observations.

Through this announcement and potential future solicitations, the ARPA-H BDF Toolbox effort will create the software foundations that make it far easier and faster to integrate these disparate data sources in a meaningful way and make the data available to a broad set of R&D end users to accelerate the development of novel health technologies to improve health outcomes for all Americans. The ARPA-H BDF will make it possible to manage data across multiple systems, platforms, and clouds, while maintaining a consistent and comprehensive view of the data. The ARPA-H BDF focuses on the development of software technologies and tools that provide: 1) data integration from various sources and formats into a unified view, regardless of where the data resides; 2) data virtualization allowing users to access and work with data without knowing its physical location or structure; 3) a centralized approach to data management, including data provenance, quality, governance, and security; and 4) capabilities for analyzing and visualizing data, including those that use machine learning and artificial intelligence. This announcement seeks novel approaches to improve data fidelity, quality assurance, semantic integration, analysis, and visualization; ARPA-H intends to solicit technical solutions for the federated aspects at a later date.

Truly novel approaches are needed to solve familiar problems. This announcement seeks proposals to revolutionize data integration and usability approaches by making it less labor intensive to connect biomedical research data from thousands of sources, overcome barriers caused by incompatible data dialects, and increase fidelity and confidence in biomedical data so that it can more reliably be used to accelerate health technology innovation. The ARPA-H BDF tools will enable rapid integration of biomedical research data across thousands of research labs, hospitals, and centers. Straightforward applications of existing commercial capabilities are out of scope. Instead, technical solutions should identify relevant commercial products, describe the current limitations of those commercial products and the state of the art more broadly, and clearly explain how the technical approach will significantly go beyond what's possible today.

The ARPA-H BDF tools will overcome challenges associated with data integration and usability. Today, many data science efforts seek to leverage established technologies and operationalize data infrastructure platforms to meet FAIR[1] data principles. While established technologies can be engineered to accept data from more researchers, improve dashboards, and refine search capabilities, these technologies often fail to improve the quality, standardization, and timeliness of data availability for data collected across thousands of labs and hospitals. In addition, today's experimental software falls short of consistently collecting the fidelity of data provenance, calibration information, experimental protocol information, and context needed to reliably test for experimental reproducibility across different labs. Furthermore, established technologies are limited in their ability to integrate data from multiple sources and support intuitive multi-source exploration or data analysis, including through artificial intelligence/machine learning (AI/ML). Critical biomedical data usage by appropriate stakeholders can be impeded by access limitations, a lack of interoperability across hundreds of siloed data platforms, and a lack of robust, reusable methods to protect data privacy and security. At its crux, these limitations make it difficult to leverage data across different labs, hospitals, and centers, because each entity tends to manage data using incompatible biomedical dialects.

[1] The international research community has developed four guiding principles for data management and stewardship. These "FAIR" principles state that data should be findable, accessible, interoperable, and reusable. (The FAIR Guiding Principles for scientific data management and stewardship | Scientific Data (nature.com))

The ARPA-H BDF Toolbox project will advance data fabric capabilities to address the challenges around data integration, data usability, and incompatible biomedical dialects. Areas of technical emphasis include:

- Automated data collection to lower barriers to high-fidelity, timely access to computationally ready research data across labs and health record systems.
- AI-assisted curation to prepare, connect, and harmonize multi-source data for analysis at scale.
- Intuitive exploration to enable advanced, intuitive, human-centered data exploration and dashboards for use by diverse stakeholders and decision-makers across disciplines and health literacy levels.
- User testing to evaluate data fabric tools across diverse users – including researchers, clinicians, and patients – to create tools that will be enthusiastically adopted.

B. COLLABORATIVE MULTI-ORGANIZATIONAL CONTEXT

By improving data integration and usability, the ARPA-H BDF Toolbox effort seeks to accelerate the ability to develop next-generation health interventions, across many disparate disease areas, ultimately to accelerate integration of data across Government-funded research efforts, NIH Institutes and Centers, and beyond. These advances will further enable ARPA-H to leverage previous Government investment and build a model for health data science platforms that can be generalized across health disciplines.

Integrating biomedical data sources requires both novel software capabilities as well as access to existing data repositories. With that in mind, ARPA-H is working towards partnering with multiple institutes across the NIH as well as entities such as the Officer of the National Coordinator for Health Information Technology (ONC) and the Defense Advanced Research Projects Agency (DARPA) to integrate datasets, expand upon advanced research on data methods, and evaluate novel data fabric capabilities.

With cancer as an early use case, ARPA-H will partner with the Center for Biomedical Informatics and Information Technology (CBIIT) at the National Cancer Institute (NCI) to explore applications to cancer data. CBIIT is responsible for a range of biomedical informatics research, research support, data management and sharing policy implementation, and project management and administrative support. The ARPA-H BDF Toolbox research effort described here supports and advances both the mission of ARPA-H to "Accelerate better health outcomes for everyone by supporting the development of high-impact solutions to society's most challenging health problems," and the vision of NCI to "Maximize cancer data availability, usability, and utility through forward-thinking policies and processes." In addition to early use cases in the cancer domain, the ARPA-H BDF Toolbox project will also include use cases that generalize across disease areas.

The ARPA-H BDF Toolbox effort aims to leverage partnerships with other NIH Institutes and Centers to connect data siloes. Proposers to the ARPA-H BDF Toolbox project are expected to collaborate effectively with performers from partner organizations, including but not limited to the National Heart, Lung, and Blood Institute (NHLBI), National Center for Advancing Translational Sciences (NCATS), National Institute of Biomedical Imaging and Bioengineering (NIBIB), National Institute on Drug Abuse (NIDA), National Institute of Child Health and Development (NICHD), and the Office of the NIH Director (OD).

C. TECHNICAL AREAS (TAS)

The ARPA-H BDF Toolbox project is a 36-month effort with a 24-month base period and a 12-month option. Year 1 will focus on the project goals of establishing baselines and improving the accuracy, timeliness, and maintainability of applications for cancer and select use cases designed to advance capabilities for specific data types. Years 2 and 3 will add increasing attention to the objectives of generalizability and scalability. Metrics for each phase are described below in Section D. Proposals should address how the proposed research plan will meet associated technology metrics as well as the overall project metrics and milestones.

The project will employ a collaborative and iterative approach to development and integration, with performers in all TAs expected to work closely with other performers during all phases of development to ensure their research products are compatible and synergistic with other efforts. Proposers should describe how they intend to work with other teams performing in the same and different technical areas to promote integration and collaboration.

ARPA-H currently seeks innovative proposals to develop and advance software capabilities for the ARPA-H BDF Toolbox Project in the following TAs:

- TA1 Automated Data Capture: Lowering barriers to high-fidelity, timely data collection in computer-readable forms
- TA2 AI-Assisted Curation: Preparation for multi-source data analysis at scale
- TA3 Intuitive Exploration: Advanced and intuitive data exploration
- TA4 User Testing: Evaluating data usability by community members across disciplines and biomedical literacy levels

C.1. TECHNICAL AREAS (TA) ONE (1) THROUGH FOUR (4):

TA1 - AUTOMATED DATA CAPTURE: LOWERING BARRIERS TO HIGH-FIDELITY, TIMELY DATA COLLECTION IN COMPUTER-READABLE FORMS

Today, the data collection process includes a significant amount of manual manipulation with varying levels of adherence to data contribution requirements or requests. This is especially true when multiple types of data are aggregated across many stakeholders.

Performers in TA1 will research and develop innovative approaches to streamline and automate the collection of biomedical research data in a manner that captures far more context than is typical today. Data capture improvements should significantly increase the usability of research data downstream. Novel techniques that incorporate advances in ML, Natural Language Processing (NLP), and executable protocols to improve data collection fidelity are encouraged. Approaches that integrate with lab instruments, capture specific types of experimental data (e.g., sequencing, imaging, single cell measurements), or extract de-identified research data from Electronic Health Record (EHR) systems are of interest. This section first outlines proposal guidelines for data capture in lab settings and then discusses guidelines for EHR data.

One important area of innovation is the capture of high-fidelity data from experimental laboratories and other laboratory settings. Innovative solutions that lower the barrier to high-fidelity data collection throughout experimental planning, protocol execution, and data transfer stages are encouraged. Strong proposals will include automated or semi-automated workflows

that capture experimental hypotheses and designs starting from electronic lab notebooks or lab instrumentation software to capture raw data (e.g., 'omics, imaging, single cell, spatial 'omics) as well as associated metadata at runtime in standardized output format for specific data types. In addition, performers should test their new data collection software in three or more different labs or comparable sources and demonstrate data transfers into a common portal module.

Strong proposals will outline capabilities that focus specifically on sub-groups of data types or instruments. Approaches that collect information about experimental protocols, experimental intent, metadata, calibration information, and provenance are strongly encouraged. Proposers should aspire to develop solutions that automatically capture all information needed for downstream analysts to understand the strengths and weaknesses of the data collection approach, accurately assess appropriate and inappropriate uses for the data, and test for cross-lab reproducibility. Technical capabilities will take into consideration existing APIs, middleware, and/or electronic data capture systems to eliminate barriers to collecting biomedical research data by standardizing and semi-automating the data collection processes. TA1 will reduce the time and manual effort required to collect experimental data in a manner that can be easily shared with other researchers and stakeholders. The data collection capability should include features such as an auto-populated dashboard upon data upload that will enable researchers to evaluate adherence to the FAIR Data Principles.

Proposers who have novel data collection approaches for academic labs or other types of laboratories should describe not only how the innovative data collection approach works but also how it integrates with lab equipment, how it goes beyond existing commercial capabilities, and how the proposed prototypes will be tested in at least three additional, different labs or comparable sources. Proposal teams should include those who are developing or extending novel data capture technologies and labs willing and able to evaluate the new data capture technologies. Teams evaluating new technologies should be staffed appropriately to perform beta testing and provide feedback to the technology developers on the user experience and effectiveness of the software prototypes.

Proposals should also consider how the data will be transferred from geographically distributed labs into a common location. Proposers should develop easy-to-use data submission modules that enable researchers to submit specific data types from multiple labs as well as describe how their modules could integrate with modules from other TA1 performers into a larger submission portal. Approaches to incentivize increased data sharing are also of interest, including but not limited to mechanisms that reduce the manual labor required to submit data to a common portal, real-time submission dashboards, documentation, interactive chat-bot guidance, automatic data validation, detection of missing fields, mapping to well-established data models, and training for data submitters. Proposers are encouraged to provide a scorecard to the data submitter about the FAIRness of the data (i.e., how well does it conform to standard file formats, data dictionaries, and established data models), and to provide lab-side software that makes it easy for researchers to capture the information needed to score well by these metrics. Moreover, proposers should consider how to create communication mechanisms for data analysts to request additional metadata for previously submitted data and for future experiments. Data analysts should also be able to communicate potential data quality issues to the submitter using the dashboard.

Any proposals that aim to streamline extraction of research data from EHR systems are strongly encouraged to collaborate with the NCATS National COVID Cohort Collaborative (N3C) program and expand their existing pipelines for data extraction and syntactic analysis, and semantic data harmonization for EHR and clinical data. Proposals should describe approaches to

map EHR, clinical data, and other relevant data types into compatible data models. Wherever possible, proposers should leverage N3C data-providing sites and NIH research centers such as NCI Cancer Centers to test proposed data collection capabilities. Proposers should focus on ways to passively extract more value from existing EHR resources. Proposals that significantly change the way data is collected during clinical trials will be considered out of scope.

All TA1 efforts should propose explicit metrics to assess whether aggregate research datasets are representative of patient populations. Such metrics could provide evidence to drive future research funding toward patient populations underrepresented in current research. For example, N3C data-providing sites in underrepresented communities could offer comprehensive scorecards that include measures of representativeness to address diversity and inclusion issues.

Data exchange ultimately depends on people's interest levels and incentive structures, so performers are invited to include up to a page on creative incentive structures that could accelerate the adoption of TA1 technologies. Performers are also encouraged to identify places where open standards would accelerate innovation, including but not limited to open standards for laboratory instruments.

TA1 efforts may consider questions of security and privacy such as privacy preserving methods for access to data across federation boundaries. While it is anticipated that the majority of the security and privacy advancements will occur in TA5, we recognize that to be effective, privacy and security controls also need to be implemented at the point of data gathering.

TA2 - AI-Assisted Curation: Preparation for multi-source data analysis at scale

Today, significant manual data curation is required to prepare multi-source data for advanced analysis, data exploration, 'omics analysis, and machine learning. Furthermore, today's curation methods lack capabilities to assess sampling bias when data is aggregated across sources. In addition, comprehensive approaches are needed to assess how well research data represents the population of patients affected by a particular type of cancer or other disease.

Proposers are encouraged to develop innovative approaches that expand the suite of open-source data curation capabilities to address these limitations. Further, proposed tools should enable automated mapping to common data models, data transformations, and curation of multi-source data by using innovative methods such as NLP and large language models (LLMs) to improve mapping to data standards.

Proposers should describe how their methods can enable uniform processing, quality assurance, sample bias detection, 'omics processing, and normalization of multiple data types in preparation for advanced multi-source analyses, data exploration, application of AI techniques, and population-scale assessments.

Proposers should explore and support enhanced multi-source data preparation to maximize the immediate utility of direct data submissions, especially approaches that advance capabilities for (i) quality control (e.g., data integrity, data completeness, format); (ii) data de-identification (e.g., removal of personally identifiable information (PII) in text and images); (iii) mapping submitter data to standard dictionaries and data models; and (iv) detection, communication, and mitigation of sampling biases across multi-source data. These steps should be performed in a streamlined process flow assisted by ML, NLP, and other advanced mapping approaches along with human

curation. Comparative research that converges on process flows to minimize the amount of human effort required to curate data while maximizing curation accuracy is encouraged. Data preparation research will leverage and expand upon resources such as standard dictionaries, open data models, and comparative statistics about patient demographics.

While TA1 performers should address quality assurance for data captured in a common way across several labs, TA2 should focus on more significant semantic interoperability challenges. Specifically, strong TA2 proposals will describe novel approaches to semantically link data that differs due to substantially different data collection strategies, differences in data type, or differences in nomenclature across diseases, disciplines, or organizations.

For 'omics, imaging, single cell, and/or spatial 'omics data, TA2 proposers are encouraged to describe prototype workflows for processing raw data to generate secondary and tertiary results that are much smaller in size and easier to manipulate and interpret, thus lowering computational barriers for users. Many approaches have been deployed for genomics data harmonization; however, no comparable workflows exist for other 'omics data such as proteomics, single cell 'omics, microbiome, metabolomics, and spatial 'omics. Research and development of such harmonization workflows for raw data will enable the creation of uniform datasets that minimize biases (e.g., batch effects, inconsistent analysis pipelines), and provide technical consistency and standardization of data allowing for maximal use and interoperability with other datasets. Harmonization software should be open source so that those using the data can inspect which data preparation steps have occurred prior to analyzing the data.

TA3 - INTUITIVE EXPLORATION: ADVANCED AND INTUITIVE DATA EXPLORATION

Currently, many users face a steep learning curve to find, access, explore, and analyze biomedical research data, especially multi-modal datasets, as the data are stored in siloed data repositories, and existing tools are designed for those with informatics skillsets. By expanding capabilities to serve citizen scientists, researchers from many disciplines, advocates, clinicians, and patients, TA3 will accelerate progress toward the long-term goal of precision medicine.

Proposals are encouraged to explore, propose, design, architect, develop, test, and implement revolutionary approaches and tools to change how users query biomedical data. The current state-of-the-art is faceted search where users select a series of filters to sort and organize search results to create an artificial cohort of patients with a particular clinical phenotype. This method is very useful in narrowing search results but is limited by predefined categories of metadata. Typically, these data have been organized in structured query language (SQL) databases or more recently in document-oriented, graph-based, and schema-flexible data structures. Novel alternatives to traditional search using pre-defined categories are desired to create more intuitive and flexible approaches to find and explore relevant data.

Novel data exploration approaches are of interest, including those that use an appropriate combination of the latest data storage technologies combined with LLM approaches to provide a highly flexible text- and speech-based query interface to improve data usability by all cancer community members, and as well as community members in fields other than cancer, regardless of expertise. Proposers should explain how their approach will enable users to explore and make connections across multiple data types, including but not limited to multi-omics, imaging, longitudinal, and single cell datasets. Proposers are strongly encouraged to maximize the use of existing and emerging cloud-based technologies from the major cloud service providers. Strong proposals will go well beyond faceted queries and present natural language questions to the

system regarding any of the data it contains and receive expert results in return. Results ideally would be further augmented by scorecards for ML readiness and other critical metrics such as data completeness, integrity, bias, and representativeness of the patient population.

Proposers are encouraged to design, architect, develop, and implement intuitive AI-powered dashboards, easy-to-use data exploration tools, and data visualization techniques to accelerate scientific discovery in a manner that maximizes the meaningful use of biomedical data. The key to data usability is understanding where and what data are available and obtaining enough context to understand how the data can be analyzed, explored, or used. The data dashboards should provide real-time summaries of all data with a biomedical data ecosystem, including statistics by data type, disease type, number of patients, and other metrics. Proposals should outline strategies to increase the number of diseases supported over the course of the project and describe how capabilities will be expanded to enhance scalability and support an increasing diversity of users over time.

TA4 - USER TESTING: EVALUATING DATA USABILITY BY COMMUNITY MEMBERS ACROSS DISCIPLINES AND BIOMEDICAL LITERACY LEVELS

For the data fabric tools to have maximal impact, the ARPA-H BDF Toolbox capabilities must be usable by all key stakeholders in the research community to promote diversity, equity, and inclusion in data and data use. Therefore, the progress of the proposed work will be evaluated to assess the impact and adoption of the capabilities across disciplines and biomedical literacy levels (e.g., patient, family member, clinician, basic scientist, clinical scientist, translational scientist, and population scientist). In early phases of the project, user feedback should be collected from cancer researchers across an increasingly diverse set of disciplines. In later phases, all testing should expand beyond the cancer community to patients, clinicians, and other members of the health ecosystem.

Proposals to TA4 should provide detailed descriptions of developmental testing for all application, database, and software development, including unit and acceptance testing. Acceptance testing should include user testing by key stakeholder groups from the target research community. Together, the Performer and Government will identify members of key stakeholder groups to form test panels to perform iterative and final user acceptance testing for continuous feedback on the development. This iterative user engagement process will measure how intuitive and easy it is to submit, explore, and analyze the data by each stakeholder group. These evaluations shall collect key metrics to assess data usability by the stakeholder groups. Proposals should describe how this testing will be supported, including but not limited to user support, test scenario development, technical test hosting, test support, and result collection, analysis and reporting, and development of approaches, plans, schedules, and resource requirements for corrections, changes, and/or improvements. Additionally, Performers in this TA will assist the ARPA-H BDF Toolbox Project leadership with the development and management of use cases to support seamless integration of ARPA-H BDF tools across tasks as well as alignment with common use cases.

INDEPENDENT TESTING AND EVALUATION (T&E)

An additional team may be engaged throughout the project to provide independent testing and evaluations on any TA on behalf of the Government. Responsibilities of this team may include designing processes to measure and assess the individual TA components against the project metrics and for conducting project-level assessments to compare performance with and without

the ARPA-H BDF toolbox components, pipeline, and systems. All performers on TA1-TA4 will be expected to work with the T&E team to ensure assessment objectives can be addressed in conjunction with interim and final metrics and project assessments.

NOTE: This solicitation is NOT soliciting for T&E proposals. T&E proposals received against this solicitation will be **discarded**.

C.2. USE CASE:

Proposers should describe the use cases and data types that they intend to use to drive technology development. Use cases should be selected to evaluate the generality of the proposed software innovations. The following types of use cases are highly encouraged: (1) cancer-related use cases that are relevant to the missions of multiple NIH Institutes and Centers (e.g., leukemia); (2) use cases that easily generalize across disease areas (e.g., the relationship between genomic indicators, treatments, and outcomes); (3) a progression of cancer and non-cancer use cases designed to increase the generality of the technical approach in a meaningful way; and (4) use cases that demonstrate the utility of bridging data siloes managed by different organizations. Proposers should elaborate on the motivating use cases and explain how those use cases will be used to demonstrate generalizability across disease areas or data types.

In addition to specifying the motivating use cases, proposers should elucidate the types of data that they intend to focus on during each year of effort. Strong proposals will outline a progression of increasingly complex challenges by either integrating across a broader set of data types, across more diseases, or both. Whenever possible, proposers should include quantitative metrics to communicate the dimensions along which they plan to extend the proposed technologies.

D. PROGRAM METRICS

The ARPA-H BDF Toolbox project is exploratory and intentionally ambitious to inform future investments. As such, performers should clearly communicate which ambitious goals they plan to focus on as well as strategies to compare their new approaches to commercial baselines. Strong proposals will clearly explain how the proposed approach will significantly go beyond current commercial capabilities (i.e., state of the art) and, if possible, proposers should include a baseline comparison with relevant commercial products to evaluate whether their research prototypes exceed today's commercial capabilities. In addition, the Government is interested in open-source capabilities to enhance data transparency, track data provenance, and accelerate research and innovation. Proposers are encouraged to identify areas where open-source capabilities are limited.

Proposals should indicate which of the following metrics the proposers intend to prioritize during the ARPA-H BDF effort. Some of the metrics below refer to a progression of cancer and non-cancer use cases, but performers should feel free to propose similar metrics for other types of use case progressions so long as they align with the use case guidelines outlined above. Proposals incorporating AI in any TA should address considerations for responsible AI, including data trust, accuracy, and reliability, as well as transparency, inclusiveness, and fairness, and should propose appropriate metrics as applicable.

Ambitious progress toward a few metrics is of more interest than comprehensive coverage of all metrics. Note that the milestones will vary based on the complexity and variability of research data proposed, so strong proposals will focus on expanding coverage for complex data types and explain what ambitious milestones make sense relative to the specific details of the proposed approach. Strong proposals will pursue aggressive metrics during Year 1 and Year 2, and then

include options to broaden real-world applications and mature capabilities into easily deployable forms in Year 3.

Technical Area	Metric	Baseline	Year 1	Year 2
TA1	Speed of adding new data	Set baseline in first 3 months; expected to be on the order of months.	Weeks	Days
TA1	Tracking data accessibility of standardized data types	Set baseline in first 3 months	70% accessible	95% accessible*
TA1	I Apply use cases to demonstrate effectiveness of the pipeline from data collection to data upload	Build pipeline	Test with 75% accuracy across multiple data sources	Test with 98-100% accuracy across at least three labs
TA2	Auto-mapping of data to standards	Set baseline in first 3 months	50% accurate	80% accurate
TA2	'Omics workflows	Computationally tractable support for genomic data	Support for two additional ‘omics data types	Support for five additional ‘omics data types
TA3	Integration of algorithms into pipeline	Set baseline with standard methods	Add 2-5 new methods adopted by >100 users	Add 5-10 new methods adopted by >250 users
TA3	Variety and numbers of users supported	Bioinformaticians and computational scientists; Baseline number of users of each type	3 additional types of cancer researchers (basic, clinical, and translational researchers); 2x the number of users over baseline	Broader cancer community including patients, clinicians, and patient family members; 5x the number of users over baseline
TA3	Accessibility of diverse data types	Set baseline using test cases and initial search functionality	Meaningful retrieval of more diverse multi-omic data types	Meaningful retrieval of multi-omic and clinical data types
TA4	User Acceptance Testing (UAT) results	Set baseline after the first UAT	70% of users performing successful analysis	90% of users performing successful analysis

*In general, making 80% of the data interoperable takes 20% of the time, and at some point, the team reaches diminishing returns because not all data is useful to people unfamiliar with the experimental paradigm. The 95% target encourages performers to pursue a stretch goal without leading them to spend time pursuing semantic interoperability targets that will never provide a benefit to health outcomes.

E. Program Structure and Integration

The ARPA-H BDF Toolbox Project will ultimately integrate promising prototypes into a data fabric capability. The relationship between the TAs is shown in Figure 1 and described in detail below:

- TA1 Automated Data Capture: Lowering barriers to high-fidelity, timely data collection in computer-readable forms
- TA2 AI-Assisted Curation: Preparation for multi-source data analysis at scale
- TA3 Intuitive Exploration: Advanced and intuitive data exploration
- TA4 User Testing: Evaluating data usability by community members across disciplines and biomedical literacy levels
- TA5 – [Future] – Data Fabric Integration: Integration of promising software solutions from TA1 – TA3 into a federated data fabric capability

TA1, TA2, and TA3 proposals should discuss how their proposed software solutions will support the development and evaluation of open data standards for their respective technical areas, including but not limited to open data standards[2] and APIs for laboratory instruments, executable protocols, and disease-specific research areas. More specifically, researchers should discuss the exploratory set of data fields that will be used during the software R&D phase as well as the process for identifying which aspects are important to include in a more general open data standard. Essentially, TA1, TA2, and TA3 should discuss how their work and collaboration with other performers will inform effective open data standards and APIs.

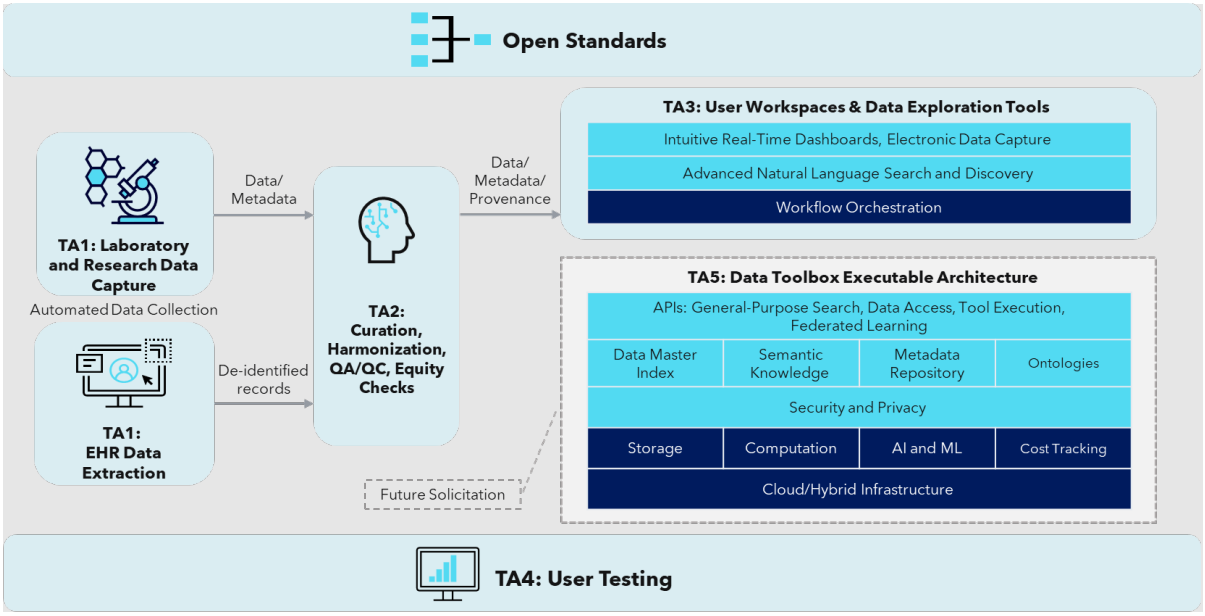

Figure 1. ARPA-H BDF Toolbox Technical Areas

The challenges addressed by TA1-TA3 exist to varying degrees across biomedical research domains. A key measure of success for the ARPA-H BDF Toolbox will be the integration of data

[2] In general, "open data standards" are developed cooperatively and democratically, are focused on availability and access, and are consensus-based to ultimately enable ways "data can be stored or exchanged for consistent collection and interoperability across different systems, sources, and users." https://resources.data.gov/

siloes and generalization across multiple disease types. Upon successful software development enabled from TA1-TA3, ARPA-H intends to solicit a fifth TA (TA5) at a later date to integrate the most promising software solutions developed in TA1-TA3 to create an executable ARPA-H BDF capable of generalizing across additional domains, diseases, and research communities. While TA1-TA4 will develop, evaluate, and refine proof-of-concept prototypes, TA5 will lay the foundation for broader application and maturation of the TA1-TA3 technologies and provide reference implementations of key mechanisms. TA5 will also identify and integrate best-in-class capabilities across biomedical domains to create an ARPA-H BDF that leverages the most advanced offerings from multiple research communities. This will enable generalized capabilities that can scale to a level where it could support all of HHS as well as ARPA-H's data infrastructure. It is important to note that TA5 will not create another data repository; instead, it is designed to provide an executable architecture that can be instantiated over any existing on-premises, cloud-based or new data repository, at scales from an individual lab or institution to a hyperscale cloud-hosted domain repository to make the data in that repository immediately available and usable across the Data Fabric ecosystem.

NOTE: This solicitation is NOT soliciting for TA5 proposals. TA5 proposals received against this solicitation will be **discarded**.

Multiple use cases, including ones focused on cancer, will be used throughout the project to exercise and evaluate the software tools and technologies being developed. Diverse use cases will be selected to evaluate the generalizability of software tools across disease domains and datatypes, which is a core element of the ARPA-H BDF project. While disease-specific domain subject matter expertise will be valuable on the performer teams, it is important to keep in mind that the focus of the ARPA-H BDF Toolbox project is software development and data applications, not the fundamental biology of the diseases in the use cases.

ARPA-H anticipates funding multiple technical approaches and performers for TA1-TA3 and making a single TA4 award. Proposers may submit multiple proposals. Each proposal may either address any combination of TA1, TA2, and TA3 or address TA4. A proposal may address any single TA or any 2- or 3-part combination of TA1, TA2, and TA3. While proposers may submit proposals for TA4 or for some combination of TA1-3, only one proposal will be accepted (i.e., performers who work on TA1-3 will not also work on TA4, and vice versa). The Government reserves the right to decide which, if any, are selected for award. If a performer is selected for TA4 award, the performer cannot be selected for the other TA(s) either as a prime or subcontractor. TA5 will be solicited in the future, and performers from TA1, TA2, TA3, and TA4 will be welcome to partner with TA5 proposers at that time.

F. SCHEDULE/MILESTONES

As noted above, the ARPA-H BDF Toolbox project is a 36-month effort with a 24-month base period and a 12-month optional extension. Year 1 will focus on the project goals of establishing baselines and improving the accuracy, timeliness, and maintainability of applications for cancer or disease-general use cases. Years 2 and 3 will add increasing attention to the objectives of generalizability and scalability.

ARPA-H BDF Toolbox Project events will include a kick-off meeting and a minimum of six (6) Principal Investigator (PI) meetings, held every 6 months over the 36-month period of performance. In addition, at the discretion of the Government, hack-a-thons and connect-a-thons may also be held every 6 months; if so, they will be associated with and co-located with the PI meetings. Proposers should plan and budget for the attendance of appropriate and relevant personnel at all

events. Relevant personnel may vary by event type; however, best practice is to assume that hack-a-thons and connect-a-thons should be attended by everyone likely to contribute to the objectives, and PI meetings should be attended by everyone with significant roles in the project who could contribute to, or benefit from, the discussions at the meetings. For rough budgeting purposes, proposers should assume the locations of events will alternate between metro Washington DC and Los Angeles, CA with PI Meetings lasting two days and hack-a-thons/connect-a-thons lasting one week. The Government also anticipates making visits to performer sites at least once per year, which should be planned and budgeted for as 1.5-day events.

Performers are expected to submit monthly status reports, quarterly progress reports, and a final report. The specific number and types of reports will be specified in the award document but will include at a minimum quarterly technical and monthly financial status reports. The reports shall be prepared and submitted in accordance with the procedures contained in the award document and mutually agreed at or before award. A final report that summarizes the project and tasks will be required at the conclusion of the resulting award.

G. Policy Conformance, Agile Development, Open Standards, and Intellectual Property

Proposers will be expected to adhere to all relevant Government laws and policies applicable to data and information systems and technologies, including but not limited to:

- Common IT Security Configurations
- Federal information technology directives and policies
- Section 508 of the Rehabilitation Act of 1973 (29 USC 794d) as amended by P.L. 105-220 under Title IV (Rehabilitation Act Amendments of 1998)
- HHS OCIO Policy for Information Technology (IT) Enterprise Performance Life Cycle (EPLC)

The Government has embraced the use of Agile frameworks to deliver software capabilities in a consistent and repeatable manner to improve the services delivered. As new frameworks evolve, ARPA-H will become more diverse in response to business needs and impacts. The scaling of these frameworks is necessary for the continued delivery of successful projects throughout the organization. Performers are expected to follow an Agile approach that incorporates an iterative, incremental delivery model throughout the period of performance. Iterative delivery means that performers deliver work frequently (sprints) rather than all at once. Incremental means that they deliver it in small packets of end-to-end functionality which are usable. This iterative and incremental development model is modeled around a gradual increase in feature additions and a cyclical release and upgrade pattern. The outcome of the subsequent iteration (sprint) is an enhanced working increment of the product. This is repeated until performers accomplish the required functionalities. The Government also recognizes that in some instances a broader framework than pure Agile is required and would draw upon practices from both the Scale Agile Framework (SAFe) and the Unified Process.

In concert with ARPA-H and partners, proposers should address innovative solutions to design, architect, develop, test, and implement, and develop ARPA-H BDF tools and associated open standards as described in the TAs. It is expected that all performers will work together to converge on open standards and APIs to ensure interoperability across prototype capabilities.

The ARPA-H BDF Toolbox will emphasize creating and leveraging open-source technology and architecture. Intellectual Property rights asserted by proposers are strongly encouraged to be aligned with open-source regimes. A key goal of the project is to seed the establishment of a sustainable open-source ecosystem for biomedical data. Thus, it is desired that all non-commercial software (including source code), software documentation, and technical data generated by the project is provided as deliverables to the Government with open-source or unlimited rights, as lesser rights may negatively impact the potential for this biomedical data ecosystem to become self-sustaining. Open-source code is highly encouraged using permissive, business-friendly open-source licenses such as CC-BY, BSD, MIT, Apache 2.0 or similar. Approaches that inhibit this objective are not desired and would adversely affect the project goals and objectives.

H. PERFORMER COLLABORATION/ASSOCIATE CONTRACTOR AGREEMENT (ACA)

The ARPA-H BDF Toolbox will be developed by a number of "performers" that includes contractors and subcontractors, to include those with deep knowledge of key data assets as well as those selected through this announcement or through complementary funding mechanisms at partner organizations. Therefore, it is expected that all performers will interact and work collaboratively with other performers in developing the data fabric architecture leveraging existing platforms, standards, and tools wherever appropriate, using open, timely, and effective communication, information exchange, and reporting.

Performers across all partner organizations will attend common meetings and technical exchanges to advance data fabric technologies, bridge across data siloes, and move toward common use cases that cross disease boundaries.

The intent is for collaboration exchanges to focus on common areas of interest relevant to the cancer domain (e.g., pediatric cancers such as leukemia cross the interests of NCI, NICHD, and NHLBI), use cases that generalize across disease areas (e.g., the relationship between genomic indicators, treatments, and outcomes), or areas where there is a significant potential to bridge data siloes.

Consistent with the rapid prototyping model used by ARPA-H, the ARPA-H BDF Toolbox Performers evaluate and de-risk a set of technologies that then contribute to the development of a reusable, easily deployable data fabric to maximize the usability of research data for researchers, patients, and clinicians while reducing the human effort needed to generalize data fabric capabilities across multiple diseases. Multiple performer teams will collaborate to develop innovative tools and methods that can contribute across the TAs with some of these tools and methods being adopted for implementation into an executable architecture.

To facilitate the open exchange of information described above, performers will have Associate Contractor Agreement (ACA) language included in their award. Each performer will work with other ARPA-H BDF Toolbox performers to develop an ACA that specifies the types of information that will be freely shared across performer teams. The open exchange of scientific information will be critical in advancing the software research required to achieve the ARPA-H BDF Toolbox objectives. The ACA will establish a common understanding of expectations to guide the open exchange of ideas and establish a collaborative foundation for the ARPA-H BDF

Toolbox project. Each performer will also work with other performers to converge on open standards and APIs to ensure interoperability across prototype capabilities.

3. AWARD INFORMATION

Multiple awards are anticipated under this announcement; however, the number of proposals selected for award will depend on the quality of the proposals received and the availability of funds. Proposals selected for award negotiations will result in an award for an OT or Cooperative Agreement.

See Section 1.4 of the MAI, ARPA-H-MAI-24-01 for additional information on award information.

4. ELIGIBILITY

See Section 2 of the MAI, ARPA-H-MAI-24-01 for eligibility requirements.

5. MODULE ANNOUNCEMENT RESPONSES

A. PROPOSAL CONTENT AND FORMAT

This combined Module Announcement is soliciting Stage 1 Volume 1 proposals. Stage 1 Volume 1 proposals must contain the following document submissions:

- TECHNICAL & MANAGEMENT
- BASIS OF ESTIMATE (BOE)
- TASK DESCRIPTION DOCUMENT OR RESEARCH DESCRIPTION DOCUMENT
- ADMINISTRATIVE & NATIONAL POLICY REQUIREMENTS

ARPA-H anticipates BIT, BYTE, and KILO modules for the ARPA-H BDF Toolbox 24-month base effort. ARPA-H anticipates BIT and BYTE size submissions for the 12-month option period effort. NOTE: the base and option period shall be written as one proposal within the page limits stated below. Strong proposals will select a cost point that is commensurate with the scale and complexity of the proposed approach. ARPA-H expects that proposals for larger efforts will include more thorough technical descriptions, more ambitious milestones, and more details regarding metrics. Larger efforts should also create more mature or comprehensive capabilities that are more thoroughly tested and evaluated. Smaller efforts may be more exploratory or focus on a subset of the overall technical area, and they will be selected based on the uniqueness of the proposed effort within the overall portfolio.

If a Stage 1 proposal is selected for potential award, a proposer will be notified by the Government and required to submit a Stage 2 price/cost proposal for further consideration.

All proposals submitted in response to this announcement must comply with the content and formatting requirements of the bundles of attachments. Proposers must use the templates provided in the bundles associated with this announcement. Information not explicitly requested in the MAI or this announcement, applicable Bundles, may not be evaluated.

All submissions, including proposals, must be written in English with font type not smaller than 12-point font. Smaller font may be used for figures, tables, and charts. Content and formatting are disclosed in each Bundle of Attachments. Below is the page restriction for each Module category:

- **BIT Module** is ≤ $2,000,000: Volume 1 shall be limited to **7** pages.
- **BYTE Module** is > $2,000,000 ≤ $4,499,999: Volume 1 shall be limited to **15** pages.

– **KILO Module** is > $5,000,000 ≤ $10,000,000: Volume 1 shall be limited to **25** pages.

NOTE: A proposer submitting to multiple TAs (whether a combination of two (2) TAs or three (3) TAs) shall be limited to **35** pages.

B. PROPOSAL SUBMISSION INSTRUCTIONS

Stage 1 proposal submissions ~~requesting an OT~~ against this Module Announcements shall be submitted to the electronic Contract Proposal Submission (eCPS)[3], ensuring receipt by the date and time specified in Section 5.C. of this Module Announcement.

~~Proposal submissions requesting a Cooperative Agreement must submit Form 1 and 2 which are provided in the Cooperative Agreement bundle of Attachments to this announcement. Cooperative Agreement proposal submission shall be submitted to Grant.gov.~~

Proposers should consider the submission time zone and that some parts of the submission process may take from one business day to one month to complete (e.g., registering for a SAM Unique Entity ID (UEI) number or Tax Identification Number (TIN); see Section 5.2.1 of the MAI for information on obtaining a UEI and TIN).

C. PROPOSAL DUE DATE AND TIME

Proposals in response to this notice are due no later than 12:00 PM ET on **November 28, 2023**. Full proposal packages as described in Section 5.A and 5.B must be submitted per the instructions outlined in this Module Announcement and received by ARPA-H no later than the above time and date. Proposals received after this time and date may not be reviewed.

Proposers are warned that the proposal deadline outlined herein is in ET and will be strictly enforced. When planning a response to this notice, proposers should consider that some parts of the submission process may take from one business day to one month to complete.

6. PROPOSAL EVALUATION AND SELECTION

Proposals selected and evaluated in accordance with Section 4 of the MAI, ARPA-H-MAI-24-01.

7. ADMINISTRATIVE AND NATIONAL POLICY REQUIREMENTS

Section 5.2 of the MAI, ARPA-H-MAI-24-01 provides information on Administrative and National Policy Requirements that may be applicable for proposal submission as well as performance under an award.

8. POINT OF CONTACT INFORMATION

Questions about the MAI should be directed to the MAI Coordinator at:
Email: MAIQuestions@arpa-h.gov
ATTN: ARPA-H-MAI-24-01

Technical questions should be directed to:
BDFToolbox@arpa-h.gov
ATTN: ARPA-H-MAI-24-01-01

[3] electronic Contract Proposal Submission (eCPS) is a component of an integrated, secure system for electronic submission, capture, tracking and review of contract proposals. Be advised eCPS requires user registration to submit a proposal response (https://ecps.nih.gov/).

9. QUESTIONS & ANSWERS (Q&AS)

All questions regarding this notice must be emailed to the two points of contact noted in Section 8. Emails sent directly to the Program Manager, or any other address will be **discarded**.

All questions must be in English and must include name, email address, and the telephone number of a point of contact. ARPA-H will attempt to answer questions in a timely manner; however, questions submitted within 10 business days of the proposal due date listed herein may not be answered.

In concert with this Announcement, ARPA-H has posted Q&As for the ARPA-H BDF Toolbox Module Announcement and Master Instructions Announcement at SAM.gov and the ARPA-H BDF Toolbox website. ARPA-H encourages all proposers to review the Q&As provided before submitting additional questions to the respective email noted in Section 8. The Government may not answer repetitive questions already answered in the posted Q&As.