Federal Register Notice 86 FR 46278, https://www.federalregister.gov/documents/2021/08/18/2021-17737/request-for-information-rfi-on-an-implementation-plan-for-a-national-artificial-intelligence, October 1, 2021.

Request for Information (RFI) on an Implementation Plan for a National Artificial Intelligence Research Resource: Responses

National AI Research Resource

An RFI Response

Jared Freeman, Drew Leins, Aptima, Inc.
Niall Gaffney, Texas Advanced Computing Center

Corresponding author: [redacted]

1 September 2021

Introduction

This response to the NAIRR RFI is inspired by the authors' experiences planning and/or executing large-scale, distributed research programs in which AI and data for AI are the primary technical products. These programs are DARPA's Synergistic Discovery and Design (SD2), which aspires to accelerate the pace of exploration and production of biological and chemical constructs through AI, and NIH's Bridge to Artificial Intelligence (Bridge2AI), which is designed to develop flagship data sets

that advance AI applications to biomedical problems. The authors of this response are primes for evaluation (Aptima) and computing infrastructure (TACC) on SD2, and co-proposers on Bridge2AI.

Response to RFI Questions

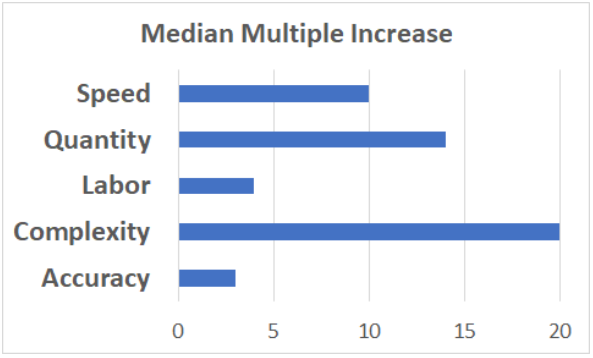
1.Roadmap Options

This section responds to question #1: What options should the Task Force consider for any of roadmap elements A through I above, and why?

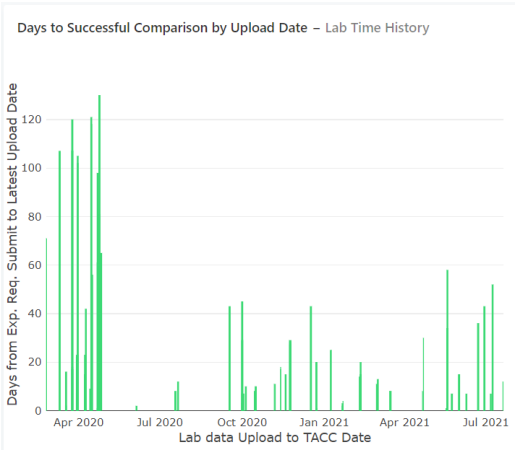
A. Goals

This section concerns: A. Goals for establishment and sustainment of a National Artificial Intelligence Research Resource and metrics for success;

The NAIRR should set as a goal transparently and measurably advancing science by creating technologies and data that are efficient, usable, used, and valued by an engaged community of researchers and application specialists. To achieve this goal, NAIRR should apply metrics and measurement technologies that quantify its progress. Specifically, NAIRR should:



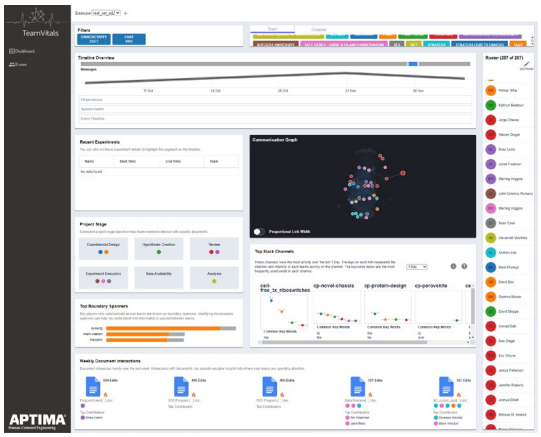
- **Apply metrics of scientific progress** -- Define classes of metrics that are relatively independent of scientific domain. Define measures that are simple to comprehend, such as multiplicative improvement over baseline. Solicit measurements both of baseline and of effects. In DARPA SD2, this strategy revealed striking effects of the program's sociotechnical system on speed, quantity, and accuracy of research or research constructs, increased complexity of constructs (e.g., synthetic circuits in biological systems), and reductions in human level of effort.



- **Apply metrics of system efficiency** -- Instrument systems that extract, transform, and load (ETL) data. Identify long latencies and cyclical processing, diagnose the root causes, and revise these data and their metadata to improve quality or revise the systems to make them more robust to poorer quality data. In DARPA SD2, this strategy helped the program to improve its systems, its data products, and the pace of discovery and design.
- **Apply metrics of use and utility** -- Instrument NAIRR systems to measure use (e.g., downloads, pulls, pushes) of the products it offers. Elicit feedback (e.g., through user surveys) concerning

other aspects of those products. In DARPA SD2, such data were used to profile technical products (data, software) with respect to comprehensibility (documentation), accessibility (e.g., public posting), utility/value, and usability (e.g., support required to modify the product or put it to use). The traditional TRL levels fail to capture these attributes.

- **Apply metrics of organization & community evolution** -- Instrument NAIRR collaboration platforms and conduct periodic surveys to capture data concerning its composition (e.g., professional and cultural diversity of personnel), organizational structure, process, member engagement, and climate. Perform computational and statistical modeling to identify positive and negative trends relative to best practices of organizational science. Identify correlational and causal effects of these factors on S&T productivity. Reward laudatory trends. Intervene to address worrisome trends, relative to the organization charter. Identify and leverage emergent, informal “boundary spanners” (individuals who are central to multiple groups, and emergent communities whose work has high value (figure). DARPA SD2, the USAF, and Army have applied these techniques in training and operations, and have funded and government-owned technologies to automate them.



B. Ownership & Administration

This section concerns: B. A plan for ownership and administration of the National Artificial Intelligence Research Resource, including: i. An appropriate agency or organization responsible for the implementation, deployment and administration of the Research Resource; and ii. A governance structure for the Research Resource, including oversight and decision-making authorities;

- **Ensure equity, diversity, and inclusion** -- Administration of NAIRR should commit in its charter, realize in its membership, and codify in its processes a commitment to Equity, Diversity, and Inclusion. Measures of these attributes should be established (see Governance, above) to assess the baseline, current state, and evolution of the organization with respect to EDI. For example, inclusion can be assessed from records of participation in meetings, chat, and collaborative documents. Diversity can be assessed by comparing the distribution of members over target attributes (e.g., race, socio-economic status, profession) to norms (e.g., the distribution of the US population, of the scientific population). Equity can be assessed through analysis of the distribution of power, measured as roles, control of resources, etc.

C. Governance & Oversight

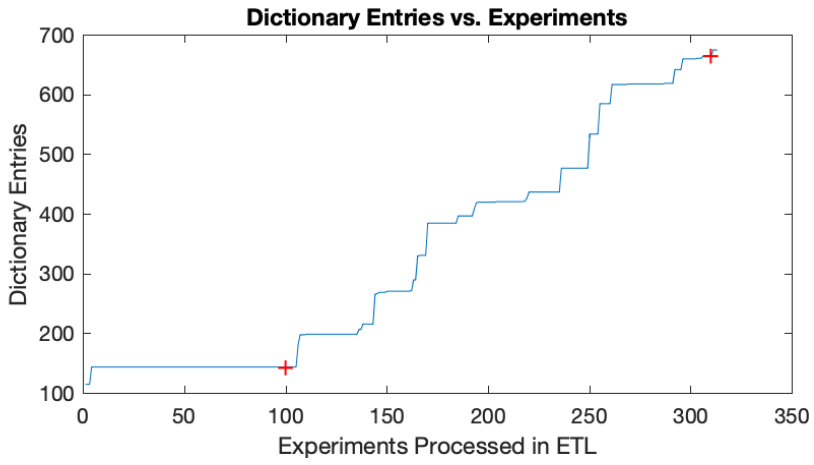
This section concerns: C. A model for governance and oversight to establish strategic direction, make programmatic decisions, and manage the allocation of resources;

- **Design governance for productivity** -- A Steering Committee should be established to set the goals of the NAIRR and assess its progress. Its members should include the administering organization, leads for major groups within NAIRR, and key representatives of the Accountability Council (below).
- **Design governance for accountability** -- An Accountability Council should be established whose members are sampled from the population of data providers (e.g., human research participants, the National Archives for historic records, bio-medical materials developers), data processors (e.g., data scientists, developers of ETL and analytic tools), consumers of derivative data (e.g., physicians, physicists, economists), and ethicists. This panel should identify issues in the conceptualization (e.g., intended use, sampling), generation, storage, processing, and use of AI data and tools; set policy regarding these; and review potential violations of that policy. As noted by Dr. Francis Collins, Director of NIH, in testimony to Congress in 2021, inclusive governance raises public trust and enhances recruitment and retention to research efforts.

D. Capabilities

This section concerns: D. Capabilities required to create and maintain a shared computing infrastructure to facilitate access to advanced computing resources for researchers across the country, including provision of curated data sets, compute resources, educational tools and services, a user-interface portal, secure access control, resident expertise, and scalability of such infrastructure;

- **Invest in metadata design** -- Support collaborative design of metadata that describe data and the context in which they were created, such as the experimental participants and/or materials, procedures, analyses, and relationship to published findings. In DARPA SD2, such metadata stores were correlated with, and likely causal, to dramatic increases in the quantity of biological research (analysis by Jacob Beal, BBN, for SD2; see figure).



- **Invest in multi-entry workflows** -- Support development not just of independent tools but of workflows, typically enabled by a framework for integrating tools in novel combinations that are fit for a specific purpose. Equally important is that such workflows enable users to access data at multiple points in processing and storage, to support users with different levels of technical skill and different research objectives.

E. Data Dissemination

This section concerns: E. An assessment of, and recommended solutions to, barriers to the dissemination and use of high-quality government data sets as part of the National Artificial Intelligence Research Resource;

- **Address regulatory restrictions** -- HIPPA, Institutional Review Boards, tribal law, and other regulations restrict the distribution of raw data concerning human subjects. Datasets should be clearly labeled with such restrictions, and restrictive documents (such as informed consent forms) should be made available. Access to restricted data should be conditioned on NAIRR review (e.g., that a local IRB has approved processing of human subjects data; that a researcher has tribal authorization to analyze data concerning an American Indian tribe). Mechanisms should be available through NAIRR to anonymize human subjects data, and agreements required that users will not reverse engineer human identification.
- **Document data** -- Data are always collected with intent, process, and scope. These are key elements to achieving leveraging any dataset to produce results. While there are many successful AI outcomes driven by data repurposing (e.g. twitter flu predictions) for any AI data or model to be accurately leveraged, these key elements of metadata must be well described and revealed. Data and models should be published with documentation of their intent, process, and scope. In a perfect world, one would also capture both successful insight as well as the failures in using any data or combining any data sources to both promote accurate outcomes from data and to help inform future data generation efforts of limitations their intent, process, and scope may have on its utilization in addressing future questions they or others may want the data to address.
- **Automate data documentation** -- Data dissemination relies on both data discovery, tagging, interpretation, and linkages to other data. This requires effective metadata behind the data. Currently most systems require well defined schemas, ontologies or constrained vocabularies which creates a significant burden on the human data producers, scientists, and archivists who must keep these “standards” evolving as the utility and linkages of the data evolve with changes and discoveries. Lack of support often leads to degraded utility of that data. Recent works have shown some well managed NLP/AI systems can create such linkages through published literature (c.f., DIVE by Xu, Gupta, Jaiswal, Taylor, and Lockhart, 2016). However, this has not been extended to the multitude of data sources available. Support must be given to automate the evolution of metadata systems to keep key data relevant and useful for improving outcomes from federally funded projects and programs.

F. Security

This section concerns: F. An assessment of security requirements associated with the National Artificial Intelligence Research Resource and its management of access controls;

- (No response).

G. Privacy & Civil Rights

This section concerns: G. An assessment of privacy and civil rights and civil liberties requirements associated with the National Artificial Intelligence Research Resource and its research;

- (No response).

H. Sustenance

This section concerns: H. A plan for sustaining the National Artificial Intelligence Research Resource, including through Federal funding and partnerships with the private sector;

- (No response).

I. Organizational Structure & Process

This section concerns: I. Parameters for the establishment and sustainment of the National Artificial Intelligence Research Resource, including agency roles and responsibilities.

- (No response).

2. Capability Prioritization

This section responds to question #2: Which capabilities and services provided through the NAIRR should be prioritized?

- (No response).

3. Ethical AI

This section responds to question #3: How can the NAIRR and its components reinforce principles of ethical and responsible research and development of AI, such as those concerning issues of racial and gender equity, fairness, bias, civil rights, transparency, and accountability?

- (No response).

4. Building Blocks

This section responds to question #4: What building blocks already exist for the NAIRR, in terms of government, academic, or private-sector activities, resources, and services?

- **Learn from model programs** -- NAIRR should study extant programs for models of governance, capabilities, and productivity. Among the candidates are:
    - DARPA SD2 (PM: Joshua Elliott) -- The Synergistic Discovery and Design (SD2) program developed data-driven methods to accelerate scientific discovery and robust design in domains that lack complete models.
    - NIH Bridge2AI -- The NIH Bridge to Artificial Intelligence (Bridge2AI) Program seeks to bridge the biomedical and behavioral research communities with the rapidly growing community of experts developing AI/ML models by producing flagship datasets that adhere to the FAIR principles (Findable, Accessible, Interoperable, Reproducible) and critically integrate ethical considerations in preparing data for computation.

5. Public-Private Partnerships

This section responds to question #5: What role should public-private partnerships play in the NAIRR? What exemplars could be used as a model?

- (No response).

6. Democratization

This section responds to question #6: Where do you see limitations in the ability of the NAIRR to democratize access to AI R&D? And how could these limitations be overcome?

- (No response).