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

Response to RFI on an Implementation Plan for a National Artificial Intelligence Research Resource

Submitted by Maria T. Patterson, PhD

1. What options should the Task Force consider for any of roadmap elements A through I above, and why? [Please take care to annotate your responses to this question by indicating the letter(s) of the item (A through I in the list above) for which you are identifying options.]

B. If a truly democratized solution is established, an NAIRR should be completely reproducible at small scale by any entity.
C. Allocation of resources and the decision making authority is of high importance, especially if the NAIRR is a cross-disciplinary infrastructure.
H. Sustainability with partnerships with private sector like cloud service providers will be important but great care should be taken to prevent lock-in and reliance on any single provider or on any paid (and non open source) component.

2. Which capabilities and services (see, for example, item D above) provided through the NAIRR should be prioritized?

- Usage of *only free and open source* technology and software, to democratize participation when researchers move work in and out of the system and to remain agnostic and prevent lock-in to private-sector services
- Compute resources in a self-contained environment that can be entirely reproduced when deployed on other infrastructure (e.g., using Docker images)
- Access to a shared registry service for snapshotting compute environments (with embedded data / models or persistent identifiers / pointers to external data / models) that is version controlled (e.g., Docker registry)
- A science friendly user interface that could also be used in the exact same manner on a local laptop (e.g., JupyterLab)
- Interoperability with other research infrastructure and ability to freely move data and resources from system to system
- Data search systems that can be architected by any researchers for domain-specific usage over shared datasets (i.e., multiple ways to search data for different purposes should be a capability)

- Persistent identifiers for datasets that are agnostic to physical data location so as to modularly separate the technical system for ease of upgrading
- Tiered levels of data storage - ephemeral and frequently purged for personal sandbox space, shared longer-term collaborative spaces, highly-curated and ID-ed, searchable datasets
- Capability for any researchers to publicly publish literate-programmatic “papers,” or fully reproducible peer-reviewed publications that allow others to recreate data analysis in its entirety. This could also include partnerships with research journals.
- Streaming data pipeline execution capabilities

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?

For all human-centered research, regardless of whether or not researchers have collected publicly available data (e.g., increasingly popular social media mining, without explicit consent of individuals), an ethics review board (or “Institutional Review Board” IRB) should be utilized.

Equity, bias, fairness and transparency and accountability would be difficult to address in practice for AI without implementing a technical system for proactively monitoring models. While the adoption of “model cards” is one framework for documenting AI tools, this is a documentation solution that requires model providers to write this documentation, is only useful if other researchers are using the exact same model (limited applications in practice), and could still allow downstream users to adopt models or AI tools that are not appropriately used for their application. An overarching system that could continuously monitor “bias metrics” for evolving AI / ML models on different datasets would be a huge asset. An analogous service could be something like Kaggle submission boards with scoring and benchmarks.

Something similar to a “data donation” center where researchers could crowdsource voluntary contributions of donated data could allow for diversity of datasets and mitigate bias due to limited data on underrepresented populations.

Version controlling and timestamping models, data, publications, etc, similar to GitHub would be an asset.

4. What building blocks already exist for the NAIRR, in terms of government, academic, or private-sector activities, resources, and services?

A model for an interoperable, shared technical infrastructure that co-locates data, storage, and compute resources with common analysis tools is a "data commons." This has been successfully implemented for scientific researchers, collaborations, and data scientists and could be a framework for developing a national research resource focused on AI. (See "A Case for Data Commons: Towards Data Science as a Service," Computing in Science and Engineering, Grossman, Heath, Murphy, Patterson, and Wells, 2016 at https://doi.ieeecomputersociety.org/10.1109/MCSE.2016.92 or https://arxiv.org/abs/1604.02608, the Center for Translational Data Science https://ctds.uchicago.edu/datacommons, and Gen3 https://gen3.org/.)

5. What role should public-private partnerships play in the NAIRR? What exemplars could be used as a model?

Private entities such as cloud service providers could and should provide a role for burst capacity when researchers hit their quota limits or need additional resources that cannot be provided by the NAIRR but should not be relied on for any core services that cannot be exactly substituted with open source software. The NOAA Big Data Project and its Cooperative Research and Development Agreement with large cloud providers is an interesting model that democratizes core access to data at no cost but allows private entities (cloud providers) the ability to charge for additional services. Looking to the open source community and successfully models their like managed enterprise services could be useful.

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?

The "tragedy of the commons" risk that some researchers/collaborations/entities may dominate usage could be a problem. Quotas could be set within timeframes using a proposal system similar to a Time Allocation Committee process in the astronomical community https://www.noao.edu/gateway/tac/. Reviewing committees should be diverse and use best practices in peer review processes for removing bias from decision making.

Relying on private paid service providers (e.g., cloud service providers) for any component that would make it such that any researcher could not perform the exact same research on their own local machines (with appropriate resources) is an outright bad idea, both biased against small and underfunded institutions/organization and may also lead to vendor lock-in long term.