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

86 FR 46278 Request for Information (RFI) on an Implementation Plan for a National Artificial Intelligence Research Resource

Submitted by the Open Commons Consortium (OCC), a Division of the Center for Computational Science Research Inc. (CCSR), a 501(c)(3) not-for profit located in Chicago, IL.

Question 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.]

Question 1D. 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;

Response. We argue that such an shared computing infrastructure should be based upon a **data commons,** which is software platforms that co-locates: 1) data, 2) cloud-based computing infrastructure, and 3) commonly used software applications, tools and services to create a resource for managing, analyzing, integrating and sharing data with a community [1–3]. An example of a data commons is the NCI Genomic Data Commons [4] that is used by over 50,000 researchers each month and provides over 1 PB of data to the research community in an average month.

More information about data commons is contained in the Response to Question 4 below.

Question 1E. 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;

Response. As in the response to question 1D above, we propose that government datasets supporting the National Artificial Intelligence Research Resource should be made available to the public using a data commons. More specifically, we propose that a data commons with open FAIR APIs [5] be used as the foundation for make the data available to the public.

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

Response. We argue that data commons developed to support different geographic regions (**regional data commons**) provide a good foundation for the NAIRR.

An example of a regional data commons is the Chicagoland COVID-19 Commons, which is an instance of the Pandemic Response Commons that is operated by the Open Commons Consortium. The Chicagoland COVID-19 Commons contains data COVID-19 related data from the Chicagoland and Illinois region, including case and fatality data, vaccination data, clinical

data from COVID-19 patients provided by regional healthcare providers, SARS-CoV-2 strain data, health disparities data, and related data.

Multiple regional commons can be integrated together to form a national data ecosystem [6] to tackle an AI problem of interest, while still reflecting important regional differences in the data, as well as regional differences in how can be best be analyzed to serve its region. Multiple regional commons can also be used as a foundation for federated machine learning.

Importantly, regional data commons can better reflect and engage with the local community, which provides a basis for reducing bias and increasing diversity of the data it supports.

Although it can take a while to set up a regional data commons, once it is set up with the appropriate consortium governance agreements, data governance agreements and commons governance agreements, the regional commons can be quickly repurposed to collect new data types, support new projects, and respond quickly to emergencies, such as providing a data driven foundation for new public health emergences.

For this reason, prioritizing setting up regional data commons provides a good foundation for the NAIRR.

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?

Answer. An important component of the OCC Pandemic Response Commons approach has been to engage with the regional community through a Community Engagement and Outreach Working Group. This approach is possible because we are developing and operating a regional data common with close ties to the local and regional community.

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

Answer. Data commons developed and operated by the Open Commons Consortium are developing using **open source software** (Gen3, https:gen3.org); support **open and FAIR data** through Gen3's open APIs; support **reproducible research** through Gen3's use of containerizing workflows that access data with persistent opaque identifiers; and its consortium membership agreements specify that research results be published in **open access** journals whenever possible.

The Open Commons Consortium has standard and time-tested agreements for: i) Consortium membership and governance; ii) contributing data to commons (data contribution agreements); accessing and analyzing data from (data use agreements); iii) setting up and operating working groups around projects of interest; and related activities.

In short, data commons developed using the open source Gen3 software and operated by the not-for-profit Open Commons Consortium for consortia provide a good foundation for the NAIRR.

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

Answer. A good model might be to support multiple different types of public-private partnerships serving different roles in the NAIRR. As an example, CCSR supports the BloodPAC Consortium, a private-public partnership that was originally launched as part of the Cancer Moonshot that accelerates the development, validation and accessibility of liquid biopsy assays to improve the outcomes of patients with cancer. The BloodPAC Consortium is a consortium of over 50 member organizations, including universities, commercial companies, and USG agencies. The BloodPAC Consortium develops and operates the BloodPAC Data Commons to provide its members and the broader liquid biopsy community with a data driven approach to advance research through the open sharing of data and its analysis.

References

1 Grossman RL. Data Lakes, Clouds, and Commons: A Review of Platforms for Analyzing and Sharing Genomic Data. *Trends Genet* 2019;**35**:223–34. doi:10.1016/j.tig.2018.12.006

2 Grossman RL, Heath A, Murphy M, *et al.* A Case for Data Commons: Toward Data Science as a Service. *Comput Sci Eng* 2016;**18**:10–20. doi:10.1109/MCSE.2016.92

3 Heath AP, Ferretti V, Agrawal S, *et al.* The NCI Genomic Data Commons. *Nat Genet* 2021;:1–6. doi:10.1038/s41588-021-00791-5

4 Grossman RL, Heath AP, Ferretti V, *et al.* Toward a Shared Vision for Cancer Genomic Data. *N Engl J Med* 2016;**375**:1109–12. doi:10.1056/NEJMp1607591

5 Wilkinson MD, Dumontier M, Aalbersberg IjJ, *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 2016;**3**:160018. doi:10.1038/sdata.2016.18

6 Grossman RL. Progress Toward Cancer Data Ecosystems. *Cancer J* 2018;**24**:126–30. doi:10.1097/PPO.0000000000000318