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

DISCLAIMER: Please note that the RFI public responses received and posted do not represent the views and/or opinions of the U.S. Government nor those of the National AI Research Resource Task Force., and/or any other Federal agencies and/or government entities. We bear no responsibility for the accuracy, legality, or content of all external links included in this document.

Stanford | LIBRARIES

31 August 2021

RFI Response: National AI Research Resource, an RFI from the National Science Foundation and The Science and Technology Policy Office of the President of the United States of America

We believe an ecosystem for computing and data infrastructure for AI researchers and students should include librarians and library practices to do the critical work of acquisition（licensing, purchasing, accepting locally built or mashed-up data sets）, curating, managing, preserving, and providing means of discovery for data, models, and any other residual outcomes of AI research.

This response addresses Component D of the roadmap: the 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.

1. What options should the Task Force consider and why ?

In response to Component D, the Task Force should recognize that research libraries and archives are well-positioned to address the growing need for solutions for management and discovery for the large-scale datasets required for AI research. Libraries have for centuries been supporting the needs of research and have evolved in response to the changing needs of researchers. The AI community has struggled, and too often failed, to address concerns of privacy, protection of intellectual property, transparency, and democratization, all of which are core values that libraries and library practices have been developed to address. Indeed, many research libraries are already providing these services and support for data users at universities.

2. Which capabilities and services provided through the NAIRR should be prioritized ?

A shared computing and data infrastructure should be built on the foundational principles of libraries, archives, and museums: that information must be not just preserved, but discoverable; not just discoverable, but deliverable; not just deliverable as bits, but readable; not just readable, but understandable; and not just understandable, but usable（OAIS Reference Model https://public.ccsds.org/Pubs/650x0m2.pdf）.

The specific capabilities to be prioritized include:

- Application Program Interfaces (APIs). This is an initial step in lowering the barrier to access; necessary but not sufficient.
- Data sub setter for search and selection. Provide interactive filtering and selection based on metadata and statistical measures of a dataset.
- Delivery of data at scale. A pipeline from query and selection (as in 1.2) to delivery.
- Roundtrip derivatives from Machine Learning work to digital preservation and make them available, in context, via the delivery platform.
- Text extraction. Expand existing processes like OCR to include Handwritten Text Recognition and speech to text. Just as machine learning based OCR has transformed discovery, similar techniques can make a range of other materials, including audio and video, available to content navigation.
- Vectorization/Feature extraction pipelines. Representing text and images numerically makes it possible to perform meaningful analytics on this derivative form of the work and is a necessary first step for applying machine learning algorithms.
- Use library experience with knowledge bases and authorities to extend the value of named entity extraction (NER). Connect entities stored in records across databases and archives, identifying and providing means for verification and disambiguation.
- Build on the library's investment in RDF data structures to manage objects and their relationships. Once a content object's entities or metadata descriptors have been linked to authorities and knowledge bases, this becomes a basis for linking content across collections.
- Make the machine-actionable data discoverable nationally based on adoption of the DDI model.
- Apply computational analysis to make the machine-actionable data discoverable. The RDF graph is easily integrated into a machine learning workflow to produce analytics and visualizations to support curation as well as discovery.
- Bring the library model of subject specialists to contextual data analysis. Subject specialists as curators can make use of new quantitative views on collections to better understand their characteristics, gaps, and distributions to serve researchers.

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?

Business and research practices that prioritize optimization and efficiency over equity and a contextual, historical understanding of the data feeding the models are inherently discriminatory. The fact that AI, emerging as it has from computer science, lacks a grounding in ethical principles of data collection, management, and use should not be a surprise. Researchers across the natural sciences, social sciences and humanities rely on libraries to record provenance, provide context, make data sets discoverable, preserve, and perform maintenance, and manage data. While those actions are not neutral, they take place within social and technical systems that take into account the changing and contested nature of information. They give researchers the tools to explore and interrogate questions of fairness, bias, and accountability.

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

In the increasingly ethically challenged world of machine learning, libraries are not only ideally positioned but have an obligation to bring their considerable experience and expertise to bear as stewards of training data -- its creation, documentation, preservation, and reuse. Libraries have long been the centers for information and education in their communities, and those shared resources and the commitment to their care and preservation is a core value of what libraries do.

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

The non-commercial service model of libraries encourages resource sharing and consortial agreements, making it possible for libraries to provide access to collections beyond what they already offer. Those offerings are wholly democratic to begin with, as they are a shared resource for all members of its community. Libraries do not seek profit as a core mission, and do not face the conflicts of interest that private businesses may have. Any member of the community of a library has access to it, and that access is far cheaper and wide-ranging than what most individuals could hope to attain for themselves. We aim to serve everyone for free and leverage collective resources to do so. Libraries also observe restrictions and limitations to use based on negotiated terms and government regulations, recording data use agreements to specific data sets when necessary.

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?

Though libraries increasingly provide access to digitized content, systems are still, by and large, oriented to access to one object at a time or on a collection by collection basis. Platforms that can manage these data collections and their derivatives, link features across them, and provide tools for analysis, discovery, selection, and delivery are necessary. Shared collections repositories such as HathiTrust, which have provided access to collections through programs such the Emergency Temporary Access Service to enable print collections to be unlocked during COVID-19 lockdown restrictions, or who have provided hosting for groups like the Technical Report Archive and Image Library to digitize and make accessible federal government technical reports which might otherwise be lost or inaccessible, are key partners to libraries and can provide models for cross-object/cross-collection access.

Thank you for your attention to this response. We stand ready to discuss this response, which is submitted by the Stanford Libraries and is likely not the only response that will be submitted by agencies and individuals at Stanford.

Respectfully submitted,

Michael A. Keller
  Vice Provost and University Librarian
  Director of Academic Information Resources
Stanford University

Stanford Libraries:
101 Green Library
Stanford, CA 94305-6004
U.S.A.