



December 3, 2021

Susan K. Gregurick, Ph.D.  
Director, Office of Data Science Strategy (ODSS)  
National Institutes of Health,  
9000 Rockville Pike,  
Bethesda, Maryland 20892

Re: NIH Request for Information (RFI): Search Capabilities across the Biomedical Landscape for NIH-wide Data Discovery

Director Gregurick:

The American Medical Informatics Association (AMIA) is pleased to provide input on NIH's request for information (RFI) on search capabilities across the biomedical landscape for NIH-wide data discovery.

AMIA is the professional home for more than 5,500 informatics professionals, representing frontline clinicians, researchers and public health experts who bring meaning to data, manage information, and generate new knowledge across the health and healthcare enterprise. As the voice of the nation's biomedical and health informatics professionals, AMIA plays a leading role in advancing health and wellness by moving basic research findings from bench to bedside, and evaluating interventions, innovations, and public policy across settings and patient populations.

NIH is seeking comments and observations on factors which may have led to limited uptake or impact of promising approaches to enabling data and information discovery. As informatics professionals, we both understand and are supportive of the possibilities of bringing together real-world data (RWD) to improve data discovery and the biomedical research system at large. However, we must caution that absent a careful plan and broader feedback-seeking process, the trajectory of current efforts to build nationwide RWD infrastructure risks undermining the existing internal research infrastructures of informatics at individual institutions. As the framework for NIH-wide data discovery and reuse evolves, these internal data infrastructure needs for research must be supported, rather than undercut.

We additionally note a troubling trend where ever-increasing expectations are pushed upon an ever-decreasing number of NIH-grant recipients to build, maintain, and – through science and rigorous evaluation – continuously improve informatics resources that will ultimately be used by

those who are not the experts who do this work. Further, as industry gets involved in building these data enclaves, we fear that the benefits will neither accrue universally nor equitably across the biomedical research space. Such a data infrastructure is not only unsustainable, but will ultimately lead to a – or perhaps exacerbate a current – data discovery divide.

AMIA thus believes that government can and must provide the foundation and funding for organizations and researchers to build out pieces of a national data enclave that are relevant to them. We envision that such an enclave will:

- Have a thorough business plan;
- Be supported nationally as a shared public good;
  - Industry users should be subject to certain data contribution requirements
- Include robust privacy protections;
- Allow users who do not have the resources to help build the enclave to avail themselves of the data in an equitable way, and critically:
- Include democratically, equitably funded informatics expertise from across the participating institutions in development and ongoing maintenance of the data enclave.

As mentioned, there is existing informatics infrastructure that will be critical for depositing data and/or metadata into this centralized enclave, which is why the resultant research should continue to be supported.

Finally, we urge NIH to consider pursuing the national data enclave through ARPA-H, should it be established. AMIA recently noted<sup>1</sup> the “potential for ARPA-H to be a convener that allows big team, big data science to occur in a fully-realized way,” while also observing that the “pandemic has demonstrated what is possible when universities, non-profits, governments, and commercial entities come together to rapidly achieve impossible and impactful scientific advances.” This opportunity should thus not be missed.

Thank you very much for considering our comments on this important issue. Should you have any questions or require additional information, please contact Scott Weinberg at [scott@amia.org](mailto:scott@amia.org) or 240-479-2134. We thank ODSS for the opportunity to comment and look forward to continued dialogue.

Sincerely,



Patricia C. Dykes, PhD, RN, FAAN, FACMI  
President and Chair, AMIA Board of Directors

---

<sup>1</sup> <https://amia.org/news-publications/amia-2021-arpa-h-statement-gretchen-purcell-jackson>

Program Director Research  
Center for Patient Safety, Research, and Practice  
Brigham and Women's Hospital

*Enclosed: Additional AMIA member comments on other RFI questions*

Questions	AMIA Response
<p>Comments on challenges and recommendations for capabilities to help researchers make their data more <a href="#">FAIR</a>-Findable. Approaches that minimize additional burden, for example by capturing dataset metadata in the course of existing activity – e.g., during primary data analysis.</p>	<p>Although it is currently required to make study data more FAIR, there is no feasible, universal, and sustainable way to actually achieve this goal, no matter the intentions of the researchers. Most of the study data ends up in researchers’ private folders. It is not practical for others to reuse the data or aggregate existing data sets, not to mention all the rules and regulations to access data by someone outside of one’s organization. To have a central data repository or a central data registry can be a solution.</p>
<p>Comments on the application of the data discovery catalogues, schema, and metadata standards for greater findability of research and clinical data, considering anticipated opportunities and challenges. Suggestions for manageable yet useful metadata schemas that can be broadly applied.</p>	<p>Participants in the National COVID Cohort Collaborative (N3C) should be consulted about their experience(s) and lessons-learned. Involving broader stakeholders, including, but not limited to large medical centers, teaching hospitals, small clinical research teams, and CTSA Program awardees is key, as well.</p>
<p>Current experiences and expectations of basic and clinical researchers, including potential use cases that underpin data and information discovery. Experiences with specific metadata schemas and search systems.</p>	<p>Whenever a new research project is under conception by a new investigator or a PhD student, a central data repository or data registry should be consulted to identify existing projects, data collected, and contacts. This would be similar to researchers currently searching for existing data in PubMed for literature during the conception and design of a new research project.</p>
<p>Comments on technology challenges and opportunities for appropriate technological frameworks that can support shared functionality across diverse environments and diverse scientific domains. Are there existing technologies that can be built upon?</p>	<p>An enclave similar to that of N3C may be an appropriate intermediate solution. However, N3C users must be consulted on ease of use and any desirable design improvements.</p>
<p>Within the biomedical research ecosystem what are the appropriate types of organization to provide discovery as a sustained (24x7), continuously updated, discovery capability for the community. What are the essential characteristics in such an organization? Where do those capabilities exist? Where do they need to be developed?</p>	<p>Ideally, a small network of well-established and respected academic medical centers may be contracted to provide “honest broker” services.</p>