April 2, 2018

Carrie D. Wolinetz, Ph.D.
Associate Director for Science Policy
Office of Science Policy
National Institutes of Health

Submitted electronically at: https://grants.nih.gov/grants/rfi/rfi.cfm?ID=73

Re: NIH Data Science Strategic Plan Request for Information

Dear Dr. Wolinetz:

The American Medical Informatics Association (AMIA) appreciates the opportunity to submit comments regarding the National Institutes of Health's (NIH) request for information (RFI) regarding its Data Science Strategic Plan. AMIA is the professional home for more than 5,500 informatics professionals, representing researchers, front-line clinicians, public health experts, and educators who bring meaning to data, manage information, and generate new knowledge across the health and research enterprise.

AMIA enthusiastically supports this Data Science Strategic Plan, including its five Overarching Goals. For the numerous and well-articulated reasons outlined in this Plan, the NIH must establish itself as a global leader in data science through dedicated stewardship and substantial financial support for data-driven discovery, translational research, workforce training and education, and rapid results dissemination.

However, **we are greatly concerned that this document omits any mention or reference to the science, study, and profession of health informatics, including clinical research informatics, biomedical informatics, and translational informatics.** Further, the increasingly vast amounts of data resulting from our national investment in electronic health records (EHRs), positions the field and application of clinical informatics to greatly impact the mission of the NIH and its data science strategy. In fact, several other domains of informatics become increasingly pertinent as our views of health expand to include behavioral, environmental, socioeconomic, and other data generated outside the realm of biomedical research, including imaging informatics, public / population health informatics, and consumer health informatics.

AMIA views the subdomain of biomedical data science as a natural constituent to the field of biomedical informatics.[1] Much like biostatistics, data science brings distinct methodologies, models, and tools that can be leveraged as inputs into the science of information – or informatics. Indeed, the January 2018 Issue of the *Journal of the American Medical Informatics Association* (JAMIA) was

[1] Kulikowski C, Shortliffe E, Currie L, et al. AMIA Board white paper: definition of biomedical informatics and specification of core competencies for graduate education in the discipline *Journal of the American Medical Informatics Association*. 2012 Nov-Dec; 19(6): 931–938. doi: 10.1136/amiajnl-2012-001053

dedicated to biomedical data science, wherein a series of articles explore the intersection between biomedical informatics and data science by highlighting ways that new and powerful data science tools can be linked with knowledge representation and clinical domain expertise to better understand health and disease.[2] Further, the 2018 AMIA Informatics Summit[3] included dedicated educational sessions, scientific panels, and posters to Data Science, in addition to Clinical Research Informatics, Translational Biomedical Informatics, and Implementation Informatics.[4] Not only were numerous NIH efforts highlighted,[5,6] dozens of grantees discussed several aspects of this Plan, including deep learning techniques,[7] using FAIR data practices,[8] and computational infrastructure necessary for personalized genomics.[9]

While we would not contend that informatics is data science, **AMIA strongly recommends that the finalized Plan clearly articulate how the NIH will integrate and leverage its existing biomedical informatics programs, grants, workforce training, and education efforts towards its data science strategy.** For example, the single largest NIH grant program – the Clinical Translational Science Awards – has provided more than $3 billion for clinical research, biomedical, and translational informatics over the last six years; the NIH Big Data to Knowledge (BD2K) grants have supported dozens of university informatics programs; and the recently funded Clinical Data to Health (CD2H) initiative will provide $25 million to five universities to developing standardized approaches and best practices to sharing data, including algorithms and other specialized tools, to address operational and institutional barriers.

In considering this strategic plan, AMIA members identified several key recommendations to ensure the NIH achieves its data science goals. Specifically, **AMIA recommends the NIH declare that all data generated through its grants must align with FAIR data practices.** On page 3 of the Plan, it is stated that, "this strategic plan commits to ensuring that all data-science activities and products supported by the agency adhere to the FAIR principles…" The NIH must go beyond adherence to FAIR principles and require that grantees also adhere to such principles as a condition of funding. Moreover, the NIH must develop policies that incentivize adherence to FAIR principles, and develop capacity to assist grantees in this endeavor. **To do this AMIA recommends NIH make Data Sharing Plans, including Genomic Data Sharing Plans, a "scorable" element of**

[2] Ohno-Machado, L. Special Focus on Biomedical Data Science. *Journal of the American Medical Informatics Association*, Volume 25, Issue 1, 1 January 2018, Pages 1, https://doi.org/10.1093/jamia/ocx151
[3] 2018 AMIA Informatics Summit: https://www.amia.org/2018-informatics-summit
[4] Informatics Summit Agenda: http://bit.ly/2GoW3Ps
[5] Making NCBI's GEO Open Data FAIR and Useful: Translating Big Data into Precision Medicine with STARGEO, 2018 AMIA Informatics Summit http://bit.ly/2Ij8IUL
[6] The CTSA Program Center for Data to Health, 2018 AMIA Informatics Summit, http://bit.ly/2IgTmzV
[7] Deep Learning-Based Ventricle Segmentation in 3D MRIs, 2018 AMIA Informatics Summit, http://bit.ly/2IhFuoW
[8] Building Organizational Capability for Translating Knowledge into Practice: Using FAIR Principles in Diverse Ways, 2018 AMIA Informatics Summit, http://bit.ly/2IgSH1p
[9] Computational Resources for Personalized Genomics: High Performance Clusters and Bioinformatics Resources for Analysis and Functional Interpretation of Next-Generation Sequencing Data, 2018 AMIA Informatics Summit, http://bit.ly/2GpOVT7

American Medical Informatics Association | AMIA
4720 Montgomery Lane, Suite 500 | Bethesda, Maryland 20814

grant applications subject to the existing policies.[10, 11] **Further, NIH should include not just data, but software code and algorithms as required elements within the "rigor and reproducibility" section of grant applications.**[12] Data sharing has become such an important proximal output of research that we believe the relative value of a proposed project should include consideration of how its data will be shared. These kinds of changes will be force-multipliers for the NIH, as they will encourage broad improvements to how data are collected, shared, and reused. Making Data Sharing Plans scorable would enable those projects that prioritize systematic and strategic data sharing, through use of standards and accepted best-practice, to garner higher scores. By using the peer-review process, we will make incremental improvements to interoperability, while identifying approaches to better data sharing practices over time.

As a complement to making Data Sharing Plans scorable elements of NIH grant applications, **AMIA recommends that NIH leverage experts in biomedical informatics and data science to review applicants' data sharing plans.** By ensuring that expert reviewers possess the requisite skills and knowledge to differentiate between high-quality data sharing plans and low-quality plans, the NIH can greatly improve interoperability, research rigor, transparency, traceability, and reproducibility across its entire portfolio of grants. To better identify expert reviewers, **AMIA recommends NIH update the eRA profile templates to capture accomplishments in creating and/or contribute to useful public datasets and software**.

Establishing these requirements is necessary, yet insufficient to achieve FAIR principles. **AMIA recommends that the NIH support institutional change within academic promotion and tenure, which continues to rely on publishing journals, through both direct funding for scholars who create and/or contribute to useful public databases, knowledge repositories, and software tools, and through infrastructure investments.** As important steps the NIH can take in this regard, AMIA strongly supports efforts to develop "separate funding strategies, review criteria, and management" for databases, knowledgebases, and tool development, as articulated in the Plan. We also support NIH efforts to develop ways to link data to articles; cite data extracts and software tools; and recognize high-quality repositories. Funding and professional recognition for scholars who create and/or contribute to useful public databases, knowledge repositories, and software tools will be particularly important. But so too will investments in data infrastructure be necessary to facilitate data stewardship, curation, and maintenance.

To support the Plan's Overarching Goals, **AMIA recommends the final Plan elaborate the organizational strategy.** If there is to be no single operational center for data science within the NIH, there needs to be a well-articulated strategy to engage, coordinate, and manage various data

[10] National Institutes of Health, "NIH Data Sharing Policy and Implementation Guidance," March 2003 https://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm
[11] National Institutes of Health, "National Institutes of Health Genomic Data Sharing Policy," August 2014 https://osp.od.nih.gov/wp-content/uploads/NIH_GDS_Policy.pdf
[12] National Institutes of Health, "Principles and Guidelines for Reporting Preclinical Research," https://www.nih.gov/research-training/rigor-reproducibility/principles-guidelines-reporting-preclinical-research

American Medical Informatics Association | AMIA
4720 Montgomery Lane, Suite 500 | Bethesda, Maryland 20814

April 2, 2018

science efforts across the NIH. In particular, the new NIH Chief Data Strategist must have the dedicated authority and resources to compel coordination across the individual Institutes and Centers (ICs). Consistent with previous recommendations,[13] **AMIA recommends the National Library of Medicine (NLM) play a leading role to foster data science competencies, develop, or fund, data science tools / services, and otherwise be a central organizational hub for data science across the NIH.** Given the finalization of its own strategic plan,[14] the NLM is well positioned to (1) focus research on the "basic science" of data standards; (2) enable and improve open science and research reproducibility through research that will foster trust and assurance in the scientific process; and (3) build on its leadership in informatics education and training through cross-cutting and multidisciplinary programs.

Finally, **AMIA recommends the final Plan include discussion of comprehensive educational programs, not just training programs, for data science inclusive of both the discipline's technical dimensions and its ethical dimensions.** Underlying the decisions to instrument algorithms and data collection, to say nothing of decisions to share data, are ethical considerations. AMIA encourages NIH officials to ensure that ethics education is devoted to issues such as version control, annotation and documentation, fitness for purpose, and code recycling, among others.

We are fully aware of the implications surrounding these recommendations, and AMIA stands ready to help ensure NIH has the requisite expertise to accomplish these worthy goals. We have provided detailed responses to this RFI in Table 1 of the enclosed document. Should you have any questions or require additional information, please contact AMIA Vice President for Public Policy Jeffery Smith at jsmith@amia.org or (301) 657-1291 ext. 113. We, again, thank NIH for the opportunity to comment and look forward to continued dialogue.

Sincerely,

Douglas B. Fridsma, MD, PhD, FACP, FACMI
President and CEO
AMIA

Peter J. Embi, MD, MS, FACP, FACMI
President & CEO
Regenstrief Institute
AMIA Board Chair

(Enclosed: Detailed Comments to NIH Data Science Strategic Plan RFI)

[13] AMIA Response to Notice Number: NOT-LM-17-006, Request for Information (RFI): Next-Generation Data Science Challenges in Health and Biomedicine, Nov. 1, 2017 http://bit.ly/2Im00Vv
[14] A Platform for Biomedical Discovery and Data-Powered Health: National Library of Medicine Strategic Plan 2017–2027, March 5, 2018. https://www.nlm.nih.gov/pubs/plan/lrp17/NLM_StrategicReport2017_2027.html

American Medical Informatics Association | AMIA
4720 Montgomery Lane, Suite 500 | Bethesda, Maryland 20814

April 2, 2018

Table 1: AMIA Comments to NIH Data Science Strategic Plan RFI

Goal	Objective & Implementation Tactics	AMIA Comments
Support a Highly Efficient and Effective Biomedical Research		**General AMIA Comments:** AMIA strongly supports efforts to create a Platform as a Service (PaaS) environment for cloud storage, computational, and related infrastructure services needed to facilitate the deposit, storage, and access to large, high-value NIH data sets. We encourage NIH to think in terms of ecosystem or environment when considering its strategy that will include many clouds and edge nodes, interoperable, interconnected, and resilient – rather than a single cloud. We also encourage NIH to consider developing policies and processes that will enable using commercial cloud solutions for enabling such an infrastructure, provided robust governance mechanisms are in place. We support plans to develop a phased operational framework that relies on the NIH Data Commons and NCBI as a coordinating hub. Where possible, NIH should leverage existing and proven environments to maintain and sustain publicly-funded data through deposition and knowledge repositories. Should a repository shutter or fail to meet its contractual obligations, it is important to have policies in place that will protect the data from being "locked-in" or lost.

American Medical Informatics Association | AMIA
4720 Montgomery Lane, Suite 500 | Bethesda, Maryland 20814

Goal	Objective & Implementation Tactics	AMIA Comments
	Objective 1-1 | Optimize Data Storage and Security Implementation Tactics: • Leverage existing federal, academic, and commercial computer systems for data storage and analysis. • Adopt and adapt emerging and specialized technologies. • Support technical and infrastructure needs for data security, authorization of use, and unique identifiers to index and locate data.	**Objective 1-1 Comments:** AMIA supports these Implementation Tactics. We view the NIH Data Commons is an important resource, and we understand that much of this plan relies on the promise of such asset. However, the NIH must approach its PaaS environment as an ecosystem of cloud environments, including commercially available cloud solutions, not simply a single central infrastructure. There is an opportunity to target something more like the internet rather than the electrical grid - a mesh rather than a utility. This would include many clouds and edge nodes, big and little, interoperable, interconnected, and resilient. To leverage the Data Commons effectively and achieve these Tactics, NIH must develop organization capacity and direct significant resources to understand the federal/academic/commercial computer system landscape and improve its knowledge of emerging and specialized technologies. AMIA recommends the NIH develop dedicated advisory groups and issue regular requests for information to help it develop detailed implementation plans.

American Medical Informatics Association | AMIA
4720 Montgomery Lane, Suite 500 | Bethesda, Maryland 20814

Goal	Objective & Implementation Tactics	AMIA Comments
	Objective 1-2 | Connect NIH Data Systems Implementation Tactics: • Link the NIH Data Commons and existing, widely-used NIH databases/data repositories using NCBI as a coordinating hub. • Ensure that new NIH data resources are connected to other NIH systems upon implementation. • When appropriate, develop connections to non-NIH data resources.	**Objective 1-2 Comments:** We encourage NIH to develop concurrent strategies to manage data resources developed through NIH ICs as well as non-NIH data resources. Both will be important and NIH should welcome opportunities to rely on trusted data resources outside its organizational boundaries. Where possible, NIH should leverage existing and proven environments to maintain and sustain publicly-funded data through platforms such as Dryad,[15] Dataverse,[16] Cancer Imaging Data,[17] Figshare,[18] Zendo[19] and BioCADDIE.[20] Consistent with previous AMIA recommendations on digital data repositories, we support development of metrics to evaluate the quality and fit-for-purpose of various repositories.[21] Key among these metrics should be consideration of the repositories sustainability and/or business model. Should NIH support existing, independently operated repositories, researchers depositing data need to be assured of their continued existence and availability. Additionally, NIH should prefer repositories that store the data in a non-proprietary (i.e. open) data format, and encourage them to adhere to best practices for data management (e.g. provide adequate metadata). Should a repository shutter or fail to meet its contractual obligations, it is important to protect the data from being "locked-in" or lost.

[15] http://datadryad.org
[16] https://dataverse.harvard.edu
[17] http://www.cancerimagingarchive.net
[18] https://figshare.com
[19] https://www.zenodo.org
[20] https://biocaddie.org
[21] AMIA Response to NIH RFI on Metrics to Assess Value of Biomedical Digital Repositories, October 5, 2016 available at http://bit.ly/2i2XF5a

American Medical Informatics Association | AMIA
4720 Montgomery Lane, Suite 500 | Bethesda, Maryland 20814

Goal	Objective & Implementation Tactics	AMIA Comments
Promote Modernization of the Data-Resources Ecosystem		**General AMIA Comments:** This strategic plan rightly identifies the use of "dozens of different funding strategies to support data resources, most of them linked to research grant mechanisms that prioritized innovation and hypothesis testing over user serve, utility, access, or efficiency" as a primary culprit in the development of data silos with only local utility. Moreover, this funding approach has also perpetuated an award environment that favors unique and novel data repositories and data science tools (e.g. software, algorithms, etc.). AMIA strongly supports NIH efforts to develop separate funding strategies for databases and knowledgebases, as well as develop incentives for grantees to both utilize and donate to established/qualified knowledge- and databases. One strategy to consider is a line-item in grants to compensate maintainers of databases the same way there is a line-item for computers. In addition to improved funding strategies, the NIH must consider ways to leverage partnerships with associations, academia, journals, standards development organizations (SDOs) and industry to maximize the efficiency of data and software code generation, sharing, and reuse. We anticipate that some data and software code sharing requirements can be dictated through the NIH, such as dbGap, or through partnerships with academic journals (see Sequence Read Archive and Gene Expression Omnibus as examples of journal-based support). However, for the vast majority of data produced outside the lab – in the clinical setting and beyond the four walls of hospitals – new approaches and incentives will be needed. AMIA strongly recommends that NIH work with partner federal agencies to leverage the growing trove of observational data now available in EHRs across 6,000 hospitals and healthcare organizations. We reiterate our call to make Data and Software Sharing Plans scorable elements of all grants subject to the existing policy, subject to peer-review, so that all publicly funded research adheres to established best practices as determined by trained informatics experts.

American Medical Informatics Association | AMIA
4720 Montgomery Lane, Suite 500 | Bethesda, Maryland 20814

Goal	Objective & Implementation Tactics	AMIA Comments
	Objective 2-1 | Modernize the Data Repository Ecosystem Implementation Tactics: • Separate the support of databases and knowledgebases. • Use appropriate and separate funding strategies, review criteria, and management for each repository type. • Dynamically measure data use, utility, and modification. • Ensure privacy and security. • Create unified, efficient, and secure authorization of access to sensitive data. • Employ explicit evaluation, lifecycle, sustainability, and sunsetting expectations for data resources.	**Objective 2-1 Comments:** In addition to developing separate funding streams, review criteria, and management strategies for databases and knowledgebases, we also support explicit evaluation expectations for each repository type. These evaluation criteria must extend beyond the dimensions of "utility, user service, accessibility, and efficiency of operation of repositories." Specifically, AMIA recommends NIH add to these Implementation Tactics the need to establish policies that will provide a framework for grantees (and NIH itself) to evaluate the value of deposition repositories and knowledgebases. Consistent with recommendations offered in September 2016 to an RFI on Metrics to Assess Value of Biomedical Digital Repositories (NOT-OD-16-133),[22] AMIA recommends NIH develop a framework that captures dimensions of: 1. Utilization; 2. Impact; 3. Quality of service; 4. Governance; and 5. Data quality and data completeness. Additionally, AMIA believes it is important to acknowledge that repositories are only useful if they are comprised of quality data, including meta-data. Here too, the intersection between data science and informatics can be illustrated. Estiri, Stephens, Klann & Murphy (2018) describe an open source data quality assessment tool for evaluating and visualizing the completeness and conformance of EHR data repositories, which is an important step toward addressing challenges to integrating clinical data across distributed networks, as conceptualized in the Big Data to Knowledge initiative.[23]

American Medical Informatics Association | AMIA
4720 Montgomery Lane, Suite 500 | Bethesda, Maryland 20814

Goal	Objective & Implementation Tactics	AMIA Comments
	Objective 2-2 | Support the Storage and Sharing of Individual Datasets Implementation Tactics: • Link datasets to publications via PubMed Central and NCBI. • Longer-term: Expand NIH Data Commons to allow submission, open sharing, and indexing of individual, FAIR datasets.	**Objective 2-2 Comments:** The ability to version control data and code that is used to analyze the data is of utmost importance for FAIR analytics pipelines. In recent years it has been increasingly easy to manage software code used in analytics pipelines. Although online version controlled systems such as GitHub provide extensive support for storing and versioning source code, similar support for management of input data may be less readily available. To support reproducible science, researchers need easily-used and widely-available tools for storing and versioning both code and data. Although it is possible to upload and distribute large genetic datasets using NIH resources like SRA, there is currently no good mechanism for researchers to create immutable big data artifacts that can be securely stored on the cloud, and shared with other researchers via a stable URI similar to a PubMed ID for publications. AMIA recommends that the NIH endeavor to develop ways to version control data – and software code used to analyze data – for research it funds. This should include an analytical provenance system, so that non-destructive ways to keep an instance of the data in the state it was in during a given analysis can be developed - thus allowing for new or repeat analytics to come to the same outcomes.

[22] AMIA Response to NIH RFI on Metrics to Assess Value of Biomedical Digital Repositories, September 29, 2016: https://www.amia.org/sites/default/files/AMIA-Response-to-NIH-RFI-on-Metrics-to-Assess-Value-of-Biomedical-Digital-Repositories.pdf

[23] Estiri H, Stephens K, Klann J, Murhpy S. Exploring completeness in clinical data research networks with DQ^e-c. *Journal of the American Medical Informatics Association*, Volume 25, Issue 1, 1 January 2018, Pages 17–24, https://doi.org/10.1093/jamia/ocx109

American Medical Informatics Association | AMIA
4720 Montgomery Lane, Suite 500 | Bethesda, Maryland 20814

Goal	Objective & Implementation Tactics	AMIA Comments
	Objective 2-3 | Leverage Ongoing Initiatives to Better Integrate Clinical and Observational Data into Biomedical Data Science Implementation Tactics: • Create efficient linkages among NIH data resources that contain clinical and observational information. • Develop and implement universal credentialing protocols and user authorization systems that work across NIH data resources and platforms. • Promote use of the NIH Common Data Elements Repository.	**Objective 2-3 Comments:** A reality of clinical research computing is that many studies can use established data standards for some, but not all data collection needs. Often study teams are required to modify established standards and templates before or after collection to prepare the data for consumption by analytics teams for that single or group of studies. Further, the beneficiaries of standards-aligned study data are most often not the parties in the process who must perform the difficult task of maintaining semantically-aligned data throughout the study lifecycle or engage in the complex task of engaging the standards development community to adjust the standard to conform to the semantics required in a study. These kinds of technical modifications necessary to the utility of the study data across its lifecycle are typically unfunded and unplanned-for. Therefore, AMIA strongly recommends that additional funding be made available to studies which align towards a FAIR data strategy as a part of the data management / data sharing plan, and that supplemental funding be provided to accommodate this work proactively in the context of study execution. By making Data Sharing Plans scorable elements of all grants subject to the existing policy, NIH will "raise all boats" through the peer-review process, so that all publicly funded research adheres to established best practices as determined by trained informatics experts.

American Medical Informatics Association | AMIA
4720 Montgomery Lane, Suite 500 | Bethesda, Maryland 20814

Goal	Objective & Implementation Tactics	AMIA Comments
Data Management, Analytics, and Tools		**General AMIA Comments:** Again, AMIA applauds NIH for identifying the need to fund development of useful, generalizable, and accessible tools differently from hypothesis-driven research. As is the case with data repositories and knowledgebases, the NIH must also consider ways to identify good tools from poor tools, and foster an advancement pathway for those grantees who develop important tools. Likewise, we support the stated plan to encourage partnerships among public funders, academic grantees and private sector innovators. We support all of the Implementation Tactics outlined below, especially efforts to "promote community development and adoption of uniform standards for data indexing, citation, and modification-tracking (provenance)." We view these policies, technological strategies, and standards as the underpinnings for a new era of research and new type of researcher. We stand ready to partner with the NIH to provide expertise and avenues for trial to understand how to operationalize these tactics through academic journals (e.g. JAMIA and JAMIA Open).

American Medical Informatics Association | AMIA
4720 Montgomery Lane, Suite 500 | Bethesda, Maryland 20814

Goal	Objective & Implementation Tactics	AMIA Comments
	Objective 3-1 | Support Useful, Generalizable, and Accessible Tools and Workflows Tactics: • Separate support for tools development from support for databases and knowledgebases. • Use appropriate funding mechanism, scientific review, and management for tool development. • Establish partnerships to allow systems integrators/engineers from the private sector to refine and optimize prototype tools developed in academia to make them efficient, cost-effective, and widely useful for biomedical research. • Employ a range of incentives to promote data-science and tool innovation including "hackathons," prizes, public-private partnerships, and other approaches.	**Objective 3-1 Comments:** As part of this work, we recommend NIH look to foster a marketplace for the widening range of data visualization tools developed using taxpayer funds, and that steps be taken to prioritize open-source programs, categorizing these tools based on taxonomy.[24] We reiterate our above comments that it will be important that for NIH to make available avenues for tools to be published and cited, so they can be used for promotion criteria.[25]

American Medical Informatics Association | AMIA
4720 Montgomery Lane, Suite 500 | Bethesda, Maryland 20814

Goal	Objective & Implementation Tactics	AMIA Comments
	Objective 3-2 | Broaden Utility, Usability, and Accessibility of Specialized Tools Implementation Tactics: • Adopt and adapt emerging and specialized methods, tools, software, and workflows. • Promote development and adoption of better mobile-device and data-interface tools. • Support research to develop improved methods for using electronic medical records and other clinical data securely and ethically for medical research.	**Objective 3-2 Comments:** AMIA notes that there is a current shift ongoing in cloud computing, known as serverless or "Function as a Service (FaaS) architecture." This approach provides a piece of code that can be run through an API call, and we see this as a promising way to democratize availability of complicated functionalities. A specific example relevant to bioinformatics is described by Harwood (2017) and leverages this form of cloud computing to parallelize DNA sequence alignment.[26] We recommend that in addition to mobile-device and data-interface tools, NIH add advanced visualization techniques / mixed reality / GIS / VR / 3D printing / wearable devices and sensors utilizing available interoperable data standards.[27]

[24] B. Shneiderman, "The eyes have it: a task by data type taxonomy for information visualizations," in Proceedings 1996 IEEE Symposium on Visual Languages, 1996, pp. 336–343. http://ieeexplore.ieee.org/document/545307/
[25] By way of example: https://data.mendeley.com/
[26] https://hackernoon.com/using-server-less-architecture-to-massively-parallelize-dna-sequence-alignment-via-stdlib-and-343dd2d5aebf
[27] For example: www.web3d.org

American Medical Informatics Association | AMIA
4720 Montgomery Lane, Suite 500 | Bethesda, Maryland 20814

Goal	Objective & Implementation Tactics	AMIA Comments
	Objective 3-3 | Improve Discovery and Cataloging Resources Implementation Tactics • Promote community development and adoption of uniform standards for data indexing, citation, and modification-tracking (provenance).	**Objective 3-3 Comments:** This may be one of the most important, yet difficult aspects of this strategic plan. These policies, technological strategies, and standards will ultimately provide the underpinnings for a new era of research and new type of researcher. It is incredibly important that the NIH partner with organizations like AMIA who can provide expertise and avenues for trial (e.g. JAMIA and JAMIA Open) to understand how to operationalize these tactics through academic journals. We would also point to the worked planed by CD2H as providing possible strategies for this Objective.
Workforce Development	**General AMIA Comment:** This section of the strategic plan is, perhaps, the most incomplete. As mentioned in the body of our transmittal letter, AMIA believes there is great overlap between the fields of data science and health informatics – broadly defined. There is perhaps no better example than the Educational Resource Discovery Index (ERuDIte) developed through the BD2K Training Coordinating Center.[28] The ERuDIte project demonstrated how informatics concepts and methods applying information retrieval, natural language processing, and machine learning techniques are required to understand and assess the richness of online training materials and the constant evolution of biomedical data science – essentially using informatics to understand and develop data science education.[29] Across dozens of funded programs, there are literally thousands of informatics professionals funded through these programs, and it is likely that many of them possess the skills needed to help the NIH achieve the Goals articulated in this Plan. While we do not refute the notion that specific and targeted programs are needed, we simply ask that the NIH not look past what it already funds. In addition to training, AMIA strongly recommends that the final Plan emphasize education focused on the intersection of biomedical informatics and data science. AMIA strongly recommends NIH ensure that ethical considerations and dedicated ethics education is devoted to issues such as version control, annotation and documentation, fitness for purpose, and code recycling, among other issues.	

[28] https://bigdatau.ini.usc.edu/about_erudite

[29] Van Horn JD, Fierro L, Kamdar J, et al (2018) Democratizing data science through data science training. Biocomputing 2018: pp. 292-303. https://doi.org/10.1142/9789813235533_0027

American Medical Informatics Association | AMIA
4720 Montgomery Lane, Suite 500 | Bethesda, Maryland 20814

	Objective 4-1 | Enhance the NIH Data-Science Workforce Implementation Tactics: • Develop data-science training programs for NIH staff. • Launch the NIH Data Fellows program.	**Objective 4-1 Comments:** We wholeheartedly support this Objective, yet we wish to reiterate the need for NIH to view its current training, workforce, education programs as assets that can currently apply, or be modified to apply, to data science. For example, the following programs are currently funded by the NIH and have a focus on informatics that will likely develop the kind of workforce NIH is seeking: • Early Stage Development of Technologies in Biomedical Computing, Informatics, and Big Data Science[30] • NLM Express Research Grants in Biomedical Informatics • Data Science Research: Personal Health Libraries for Consumers and Patients • NLM Informatics Conference Grants[31] • Simulation Modeling and Systems Science to Address Health Disparities • NLM Career Development Award in Biomedical Informatics and Data Science[32] • University Biomedical Informatics and Data Science Research Training Programs[33] • Ruth L. Kirschstein NRSA Individual Predoctoral Fellowships[34] There are likely hundreds of informatics professionals funded through these programs that possess the skills needed to help the NIH achieve the Goals articulated in this Plan. While we do not refute the notion that specific and targeted programs are needed, we simply ask that the NIH not look past what it already funds. Additional resources for the NIH to consider on the issue of workforce to support data science workforce and training, include: • Informal inventory of competency and curricula[35] • Clinical and research informatics[36],[37],[38]

American Medical Informatics Association | AMIA
4720 Montgomery Lane, Suite 500 | Bethesda, Maryland 20814

Goal	Objective & Implementation Tactics	AMIA Comments
		• Imaging Informatics (radiology, pathology) • Pathology Informatics (like NIH data fellowship, pathology informatics focuses on the blend of disparate data systems, such as EHR, LIS, and imaging)[39]

[30] https://www.nlm.nih.gov/ep/grantsbcomp.html
[31] https://www.nlm.nih.gov/ep/GrantConf.html
[32] https://grants.nih.gov/grants/guide/pa-files/PAR-16-204.html
[33] https://www.nlm.nih.gov/ep/GrantTrainInstitute.html
[34] https://www.nlm.nih.gov/ep/NRSAFellowshipGrants.html
[35] Kulikowski CA, Shortliffe EH, Currie LM, et al. AMIA Board white paper: definition of biomedical informatics and specification of core competencies for graduate education in the discipline. J Am Med Inform Assn 2012;19:931–8. doi:10.1136/amiajnl-2012-001053
[36] Silverman H, Lehmann CU, Munger B. Milestones: Critical Elements in Clinical Informatics Fellowship Programs. Appl Clin Inform 2016;7:177–90. doi:10.4338/ACI-2015-10-SOA-0141
[37] Gardner RM, Overhage JM, Steen EB, et al. Core content for the subspecialty of clinical informatics. J Am Med Inform Assoc 2009;16:153–7. doi:10.1197/jamia.M3045
[38] Safran C, Shabot MM, Munger BS, et al. Program requirements for fellowship education in the subspecialty of clinical informatics. J Am Med Inform Assoc 2009;16:158–66. doi:10.1197/jamia.M3046
[39] McClintock DS, Levy BP, Lane WJ, et al. A core curriculum for clinical fellowship training in pathology informatics. J Pathol Inform 2012;3:31. doi:10.4103/2153-3539.100364

Goal	Objective & Implementation Tactics	AMIA Comments
	Objective 4-2 | Expand the National Research Workforce Implementation Tactics: • Enhance quantitative and computational training for graduate students and postdoctoral fellows. • Build on diversity-enhancing efforts in data science, such as the NIH BD2K Diversity Initiative. • Engage librarians and information specialists in developing data-science solutions and programs. • Employ data-driven methods to monitor workforce diversity.	**Objective 4-2 Comment:** AMIA fully supports these Implementation Tactics. To address the interdisciplinary nature of informatics, AMIA views the task of training and education as a set of three distinct imperatives,[40] and we believe a similar approach may be appropriate for data science: • Basic "informatics literacy" for all health professionals that goes beyond computer or HIT literacy. Literacy in informatics should become part of medical education, biomedical research, and public health training to give clinicians the skills needed to collect and analyze information and apply it in their practice. • Intensive applied informatics training to improve leadership and expertise in applying informatics principles to the collection and analysis of information and its application to health care problems. This level of training will ensure a supply of qualified professionals for the emerging roles of chief medical information officers, chief nursing information officers, chief clinical informatics officers, chief research officers, chief digital health officers, and similar roles. • Support for education professionals who will advance the science and train the next generation of informatics professionals in this developing and dynamic field of study. Beyond the technical aspects of data science education, AMIA strongly recommends NIH ensure that ethical considerations and dedicated ethics education is devoted to issues such as version control, annotation and documentation, fitness for purpose, and code recycling.

American Medical Informatics Association | AMIA
4720 Montgomery Lane, Suite 500 | Bethesda, Maryland 20814

Goal	Objective & Implementation Tactics	AMIA Comments
	Objective 4-3 | Engage a Broader Community Implementation Tactics: • Give citizen scientists access to appropriate data, tools, and educational resources (see text box "Citizen Science"). • Develop materials to train healthcare providers in data science-related clinical applications.	**Objective 4-3 Comment:** AMIA fully supports these Implementation Tactics. We note that access to EHR data remains a barrier, and NIH could look for ways to address these barriers, such as the CMS Blue Button 2.0 and MyHealthEData initiatives, and the NIH's Sync 4 Science initiative.

[40] Perlin J, Baker D, Brailer D, et al "Information Technology Interoperability and Use for Better Care and Evidence: A Vital Direction for Health and Health Care." (2016) National Academy of Medicine. Washington DC. https://nam.edu/wp-content/uploads/2016/09/Information-Technology-Interoperability-and-Use-for-Better-Care-and-Evidence.pdf

Goal	Objective & Implementation Tactics	AMIA Comments
Stewardship and Sustainability		**General AMIA Comment:** AMIA supports these Implementation Tactics. We view NIH polices related to biomedical data science as the guardrails for the actions and behavior of actors engaging in biomedical data science. The kind and orientation of public policy developed by the NIH will determine whether stakeholders align practices with FAIR data principles; whether datasets are open and accessible; and whether security is prioritized. As articulated in our transmittal letter, AMIA believes there are a handful of "force multiplier" policies that NIH must prioritize in order for additional policies and activities to be successful. Below, we offer a handful of additional considerations meant to help NIH achieve stewardship and sustainability, including: • A need to develop a comprehensive strategy for articulating *how* to adhere to FAIR data principle, especially regarding the "interoperability" aspect of FAIR. This must include a focus on making sure that the metadata that describe scientific datasets are comprehensive and comprehensible; • Encouraging the professional development of a data science workforce through changes to NIH's eRA profile templates[41] to capture accomplishments in creating and/or contribute to useful public datasets and software; • Considering review process changes, such as multi-level reviews that separate review of applications from researchers[42];

[41] https://era.nih.gov/

[42] Guglielmi G. "Gender bias goes away when grant reviewers focus on the science," *Nature*, January 26, 2018 https://www.nature.com/articles/d41586-018-01212-0

American Medical Informatics Association | AMIA
4720 Montgomery Lane, Suite 500 | Bethesda, Maryland 20814

Goal	Objective & Implementation Tactics	AMIA Comments
	Objective 5-1 | Develop Policies for a FAIR Data Ecosystem Implementation Tactics: • Create rational and supportable data-sharing and data-management policies. • Promote development of community standards that support FAIR principles for data storage. • Develop model open-data use licenses to enable broad access to datasets. • Optimize security management and access policies.	**Objective 5-1 Comments:** AMIA wholeheartedly supports the FAIR principles articulated across this Plan. And we view these Implementation Tactics focused on important component pieces of the FAIR definition, save for the "I". The NIH needs to have an expansive view of standards and interoperability, bridging bio/genomic/imaging with clinical data using HL7 standards (e.g. FHIR, CCDA) and IHDSDO standards (e.g. SNOMED) among others (e.g. RxNorm, LOINC, etc.). We noted earlier that clinical and research communities will likely rely on different standards, and AMIA recommends NIH develop a strategy to translate / harmonize these different data standards. An important component to fostering a competent data science workforce will be in setting policies for citation, as mentioned previously, so that academia can acknowledge accomplishment. While this kind of acknowledgement may be outside the direct control of the NIH, one step the NIH can take directly is to update the ERA profile templates to capture accomplishments in creating and/or contribute to useful public datasets and software. Another step NIH can take on its own, is to make process changes to review criteria by further separating evaluation of candidate from evaluation of the research, in addition to the aforementioned changes to data sharing plans. Evidence suggests that there is bias in reviews when the CV is considered alongside the proposal,[43] and more steps must be taken to mitigate this bias. We know the NIH Center for Scientific Review takes the issue of bias in the review process very seriously, and we exhort continued action to address it.

[43] Guglielmi G. "Gender bias goes away when grant reviewers focus on the science," *Nature*, January 26, 2018 https://www.nature.com/articles/d41586-018-01212-0

Goal	Objective & Implementation Tactics	AMIA Comments
	Objective 5-2 | Enhance Stewardship Tactics: • Develop standard use, utility, and efficiency metrics and review expectations for data resources and tools. • Establish sustainability models for data resources. • Develop a reward and expectation system for investigators to make data FAIR and for ensuring open-source data-analysis tools are available.	**Objective 5-2 Comments:** AMIA has expressed prior support for NIH efforts to articulate the value of deposition repositories and knowledgebases.[44] We believe an evaluation framework for repositories would play a critical role in enabling research and promoting biomedical research rigor, transparency and reproducibility. We reiterate our position that in addition to the metrics discussed in the 2016 RFI, data quality and data completeness are important dimensions to assess. Beyond a framework, we recommend NIH develop guidance on standards for database quality so that distributed repositories are interoperable and forward compliant.

[44] AMIA Response to NIH RFI on Metrics to Assess Value of Biomedical Digital Repositories, September 29, 2016: https://www.amia.org/sites/default/files/AMIA-Response-to-NIH-RFI-on-Metrics-to-Assess-Value-of-Biomedical-Digital-Repositories.pdf

American Medical Informatics Association | AMIA
4720 Montgomery Lane, Suite 500 | Bethesda, Maryland 20814