NATIONAL SCIENCE FOUNDATION
2415 EISENHOWER AVENUE
ALEXANDRIA, VIRGINIA 22314

NSF 20-085

Dear Colleague Letter: Pilot Projects to Integrate Existing Data and Data-Focused Cyberinfrastructure to Enable Community-level Discovery Pathways

May 19, 2020

Dear Colleagues:

Many research communities supported by the National Science Foundation (NSF) are challenged by the need to manage, integrate, access, and use ever larger and diverse scientific datasets to conduct research. NSF accordingly invests in the creation of a wide range of data-focused cyberinfrastructure (CI) tools, resources, and solutions for use by the various disciplinary communities that it supports in order to transform data into knowledge and discovery.

Scientific endeavors are increasingly collaborative, cross-disciplinary, and convergent. NSF thus recognizes the importance of promoting holistic CI approaches to address the growing and evolving data lifecycle and workflow challenges both within and across research fields. This holistic view is articulated in NSF's recent vision [1] for an *agile CI ecosystem*, and is predicated on harmonization, integration and interoperability among CI resources, tools, services and expertise to achieve accessible, seamless and flexible end-to-end discovery pathways that drive new thinking and enable transformative discoveries.

Through this Dear Colleague Letter (DCL), NSF encourages proposals to the Cyberinfrastructure for Emerging Science and Engineering Research (CESER) program [2] within the Office of Advanced Cyberinfrastructure for **pilot** projects that bring together researchers and CI experts to develop the means of combining existing community data resources and shared data-focused CI into new integrative and highly performing data-intensive discovery workflows that empower new scientific pathways. Aims of such pilot projects can include, but are not limited to:

- improving the end-to-end process of accessing, integrating and transforming research and education data to knowledge and discovery for one or more communities;

- creating new workflows and new usage modes to address multi-disciplinary and cross-domain scientific objectives;
- addressing emerging community-scale scientific data challenges such as real-time, streaming and on-demand data access; data discovery through knowledge networks and intelligent data delivery; enabling access to data with privacy concerns; and data fusion, integration and interoperability;
- enhancing the performance and robustness of community-scale data integration and discovery workflows such as through automated curation, end-to-end performance monitoring, provenance tracking, and means of assuring data trustworthiness; and
- federating learner data to empower innovative assessment tools for large-scale modeling of learning gains.

NSF welcomes submissions that address these project goals in **all areas of science and engineering (S&E) research and education supported by NSF in all directorates**. The following directorates have selected special areas of interest for proposals:

- **Directorate for Biological Sciences (BIO)**: Proposals are encouraged on integration of existing diverse spatial data resources from site-based to continental-scaled datasets for biological features on land and water. Such work would enable novel discovery of local, regional and continental biological systems. Importantly, CI advances required for this integration might include spatial data resources from ground-based studies, unmanned aerial vehicles (UAVs or drones), the National Ecological Observatory Network (NEON) Airborne Observation platform (AOP), satellites, and other existing spatial data sources. Encouraged activities include, but are not limited to, addressing bottlenecks in CI that inhibit such multi-modal integration and analysis to achieve the full integrative potential of the data.
- **Directorate for Computer and Information Science and Engineering (CISE)**: Proposals are encouraged on enabling accessibility, integration, discovery, management and analysis of community-scale research data in support of identified science objectives or use cases, such as those related to wireless networking, cybersecurity, machine learning applications, and related performance monitoring and assessment. CISE particularly welcomes proposals that aim to enable community-scale integration of, and access to, data sets from individual investigators and/or those generated by industry to enable new research and discovery broadly for the CISE and other research communities.
- **Directorate for Education and Human Resources (EHR)**: Proposals are encouraged that bring together CI experts and STEM education researchers to explore approaches to merging fine-grained "digital footprints" of learner interaction within disparate digital educational resources, tools, and platforms into a federated suite of interoperable data repositories. Existence of such a federated system would enable deep probing of individual conceptual learning gains that can support both generalizable conclusions

about populations of learners and the creation of personalized learning opportunities. Such a corpus of data, including assessments of learning gains, can also be leveraged to inform not only the improvement of existing educational resources and tools, but also the creation of new ones. Proposers are also encouraged to envision the creation of a services layer residing on top of the data.

- **Directorate for Mathematical and Physical Sciences (MPS)**: The Division of Astronomical Sciences encourages proposals for pilot projects that will test a) approaches to common user-friendly interfaces optimized for data access and discovery by diverse inquiries, and b) ways to archive and curate the long tail of smaller value-added datasets. Innovative ideas in these and other areas are welcome, especially where they build on prior NSF support for astronomical data, software, and interfaces.

  The Division of Materials Research (DMR) encourages proposals that bring together materials researchers and cyberinfrastructure specialists to identify and engage scientific challenges that could be advanced through the innovative use of digital data, but for which barriers exist in combining existing data resources and data cyberinfrastructure into an effective data-intensive research approach. Data resources and data-focused computational tools, including those supported through: Cyberinfrastructure for Sustained Scientific Innovation (CSSI), together with Materials Innovation Platforms (MIP), DMR-supported National Facilities and Instrumentation (NaFI), Designing Materials to Revolutionize and Engineer our Future (DMREF) projects, Materials Research Science and Engineering Centers (MRSEC), and Partnerships in Research and Education in Materials (PREM). These form a network of data resources and computational tools to support a broad range of research that utilizes digital data. DMR seeks to fill the gaps in this cyberinfrastructure that hinder the effective use of digital data and computation to engage challenging problems at the frontiers of materials research, including the discovery of new materials in concert with the goals of the Materials Genome Initiative, and to stimulate the creation of new workflows and paradigms that will lead to the effective use of digital data and computation to accelerate materials research and transform the way it is done. A successful proposal will address an important and potentially transformative materials research problem through the application of a data-intensive research approach; it will identify barriers to combining existing data resources and data cyberinfrastructure, which include resources within and possibly outside the network outlined above, it will develop missing cyberinfrastructure to forge an effective data-intensive research workflow, and engage the problem.

  The Division of Chemistry (CHE) encourages proposals for pilot projects in response to this DCL that aim to explore the scientific and technical challenges facing big data standardization, storage, dissemination and repurposing in electrochemistry, including electrosynthesis, electrocatalysis, electrochemical sensor and battery development. The

research areas in electrochemistry have seen growing interest in recent years for their potential in renewable energy, chemical manufacturing, environmental remediation and sensing. The large volumes and heterogeneity of data generated in experimental and computational design of electrode materials, catalysts, and synthetic pathways provide a natural testbed to develop data workflows and identify common needs across different areas for full-scale deployment of data-driven CI in chemistry. Proposals are welcome for data CI pilot projects in electrochemistry that can serve as an initial proving ground for the development of wider chemistry community data CI, and that address best practices and scalable approaches for areas including, but not limited to, acquisition, storage, accessibility, dissemination, transparency and openness applied to both legacy and new large-scale data. The successfully competed pilot projects will facilitate exploration of new research areas, reuse of legacy data for emerging applications, (re)evaluation of new and previous findings, and comparison of data with future models. In preparing their proposals, proposers are encouraged to consult the findings and recommendations from the 2017 NSF-funded workshop on *Framing the Role of Big Data and Modern Data Science in Chemistry*[3] and the 2019 National Academies of Science, Engineering, and Medicine (NASEM) Workshop on *Advances, Challenges, and Long-Term Opportunities of Electrochemistry: Addressing Societal Needs - A Chemical Sciences Roundtable Workshop*[4].

- **Directorate for Social, Behavioral and Economic Sciences (SBE)**: Proposals are encouraged that focus on CI tools, resources, and/or solutions related to any of the directorate's core programs. Example areas of emphasis include:
    - Combination of diverse brain data - the collection of which varies in technology, methodology, and context – in order to make useable such data by a broader research community and to enable novel kinds of brain-behavior computational modeling.
    - Integration of heterogeneous human data – neural, cognitive, behavioral, economic, spatial – in ways that foster large-scale behavioral modeling. This could also incorporate networking tools into many aspects of human data analysis, including population migration and disaster response.
    - Linkage among social, behavioral, or economic datasets in ways that accommodate the imputation of ambiguous or nonexistent metadata and address the tradeoff between privacy and accuracy. Because of the transient complexity of human behavior, solutions are needed to account for data aggregation when researchers are unable to make observations or perform experiments more than once under identical circumstances. Specific challenges for solution include identifying co-factors, specifying models, sampling strategies, checking robustness, triaging data, and drawing causal inferences.

Other NSF science and engineering directorates and offices not listed above

(Directorate for Engineering, Directorate for Geosciences, Office of Integrative Activities [5] , Office of International Science and Engineering) encourage proposals in any of the disciplines they support.

Please note that the CESER program supports creation of CI for the benefit of research, but does not support activities aimed at conducting the targeted discovery-oriented research activities themselves including but not limited to creation of datasets or creation or maintenance of databases or repositories; proposers are discouraged from including these activities in their proposals beyond a limited degree of data collection that is necessary to verify the performance of the CI.

This DCL encourages proposals for exploratory pilot projects that specifically aim to integrate diverse existing scientific data sources and enable novel end-to-end discovery pathways through close collaboration of domain science researchers and CI experts and through sharing of approaches and solutions among the eventual supported projects (see below). This DCL is complementary to a number of existing disciplinary and cross-cutting NSF programs, such as the cross-cutting NSF CSSI program (https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=505505), which support a wide range of data and software infrastructure projects, and which may be alternative avenues for prospective proposers whose project ideas do not align with the cross-disciplinary and data-integration expectations of this DCL. Proposals previously submitted to, or otherwise designed for, those other CI programs may not be submitted to the CESER program in response to this DCL.

PI teams of successful proposals responding to this DCL should plan to engage regularly with one another, including attending one or more related workshops sponsored by NSF, to exchange approaches and envisaged solutions towards cross-domain harmonization and collective adoption of similar or compatible CI solutions.

HOW TO RESPOND TO THIS DCL

Proposers must follow the guidance and instructions provided under "GUIDANCE TO POTENTIAL PROPOSERS" in the CESER program description.

Per the CESER program description, in advance of submitting a proposal in response to this DCL, interested proposers are strongly encouraged to discuss their project idea with cognizant Program Directors in the CESER program and with the relevant NSF disciplinary research program(s). To initiate discussion of a project idea, prospective proposers are encouraged to send an email to CESERQueries@nsf.gov.

For planning purposes, projects funded through the CESER program are typically in the $300,000 to $1,500,000 budgetary range.

Proposals responsive to this DCL and received on or before July 1, 2020 will be considered for FY 2020 funding. Proposals responsive to this DCL and received after July 1, 2020 will be considered for potential funding in a future fiscal year, pending availability of funds.

The proposal title should begin with "Data CI Pilot:"

Awards for projects responsive to this DCL will be funded through OAC's CESER program with co-funding from the relevant directorate/office programs.

Proposals that fail to address the objectives and guidance described in this DCL and in the CESER program description may be returned without review.

Questions should be directed to CESERQueries@nsf.gov; do not contact the signatories to this DCL.

Sincerely,

Joanne S. Tornow
Assistant Director, Biological Sciences

Margaret Martonosi
Assistant Director, Computer and Information Science and Engineering

Karen Marrongelle
Assistant Director, Education and Human Resources

Dawn M. Tilbury
Assistant Director, Engineering

William E. Easterling
Assistant Director, Geosciences

Sean L. Jones
Acting Assistant Director, Mathematical and Physical Sciences,

Arthur W. Lupia
Assistant Director, Social, Behavioral, and Economic Sciences

Suzanne Iacono
Office Head, Office of Integrative Activities

Rebecca Keiser
Office Head, Office of International Science and Engineering

Manish Parashar
Office Director, Office of Advanced Cyberinfrastructure

REFERENCES

[1] Transforming Science Through Cyberinfrastructure, NSF's Blueprint for a National Cyberinfrastructure Ecosystem for Science and Engineering in the 21st Century, https://www.nsf.gov/cise/oac/vision/blueprint-2019/.

[2] CESER program description: https://nsf.gov/funding/pgm_summ.jsp?pims_id=505385.

[3] Workshop report link: https://www.nsf.gov/mps/che/workshops/data_chemistry_workshop_report_03262018.pdf.

[4]NASEM Report link: https://www.nap.edu/catalog/25760/advances-challenges-and-long-term-opportunities-in-electrochemistry-addressing-societal

[5] OIA will provide co-funding support for meritorious proposals from Established Program to Stimulate Competitive Research (EPSCoR)-eligible institutions.