Application Title: ***A Central Research Data Store to Accelerate Data-Driven Interdisciplinary Discovery Across Texas A&M***

Lead contact for RDF Application: Honggao Liu, Director, **High Performance Research Computing (HPRC)**, honggao@tamu.edu, 979-845-2561

Key Participating Units:

Texas A&M Engineering Experiment Station (TEES); College of Engineering; College of Science; College of Geosciences; Health Science Center(HSC); College of Veterinary Medicine & Biomedical Sciences; College of Agriculture & Life Sciences; Texas A&M Transportation Institute (TTI); Division of Information Technology; AgriLife Genomics and Bioinformatics Service; Center for Geospatial Sciences, Applications and Technology (GEOSAT); International Laboratory for High-Resolution Earth System Prediction (iHESP)

Anticipated Request Amount: $2,229,755.00

Executive summary of the intended application to utilize Research Development Funds.

Research data in diverse environments such as universities, including Texas A&M, is often highly siloed, and difficult to find and access when needed. The quality of the storage systems used to hold research data across university labs and core facilities also varies widely creating an unknown risk of losing irreplaceable data sets. This proposal addresses both of these critical issues by developing and implementing a shared, scalable, high performance storage system built on the principles of **Findability, Accessibility, Interoperability and Reusability (FAIR)** that is also responsive to the needs of researchers to meet their data sharing obligations to their funding agencies. This centralized storage and data management system will support the diverse and geographically distributed Texas A&M research community with multiple access methods and file system protocols, and will include a novel metadata system that automates the extraction and use of basic and user-provided information. The proposed central data store will also provide security characteristics suitable for data with moderate confidentiality, controlled unclassified information, and export controlled information. Although the research data storage facility is intended to be a shared resource with shared governance, capability is available for individuals, units and research teams to purchase extensions to the facility for their own private use. The central research data storage facility is a necessity for Texas A&M and will be operated through a partnership between HPRC and Division of IT, with shared governance through representatives of the key units at TAMU, AgriLife, TEES, and HSC.

The immediate deliverable is a storage facility with an initial raw capacity of 19.2 PB of "user facing" storage coupled to existing 7 PB HPRC high performance storage systems, and is incrementally scalable in capacity and I/O throughput. The research data storage complex will be accessible through the TAMU campus network, via Internet2, and through the LEARN state research and education network to Brazos County PIs and their regional, national and international collaborators. The proposed facility offers several types of redundancy for maximum data assurance, and levels of security appropriate to a broad set of research projects. An open and extensible metadata management and search capability is integrated with the storage complex to make the contents highly findable and usable. Application- and community-specific data management services can be co-located with the complex and core metadata services to provide a means for annotation and re-use of data sets across disciplines. This facility will also support sandbox capabilities for moving experimental data management technologies and services into a production environment, i.e., a research laboratory for developing and evaluating new standards such as those expected to emerge from the Research Data Alliance. At a technology development level, the proposed facility will be in a critical position to support basic and applied research at the intersection of storage and data management technologies, and will significantly enhance the research capabilities of Texas A&M researchers in the fields of data-enabled science and engineering. Access to such a research infrastructure will significantly strengthen research proposals geared for federal funding.