Research Development Fund – Cover Page

Application Title: A Research Data Storage System to Accelerate Data-Driven Interdisciplinary Discovery Across Texas A&M

Lead contact for RDF Application:

Name	Donald F. McMullen
Department	High Performance Research Computing (HPRC)
Email address	mcmullen@tamu.edu
Phone number	979-458-8414

Key Participating Units: TAMU: HPRC, TAMU IT; AgriLife; Health Sciences Center; TEES; TAMUG;

Anticipated Request Amount ($): $1,500,000

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 services 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 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 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, a capability is available for individual units and research teams to purchase extensions to the facility for their own private use. The research data storage facility is a necessity for Texas A&M and will be operated through a partnership between HPRC and TAMU IT, with shared governance through representatives of the core and other campuses, AgriLife, TEES, and HSC.

The immediate deliverable is a storage facility with an initial capacity of 10 PB of "user facing" high speed storage coupled to an existing 2.5 PB archival storage system and is incrementally scalable in capacity and I/O throughput. The complex will be accessible through the TAMU campus network, via Internet2, and through the LEARN state research and education network to support regional, national and international collaborations within the Texas A&M research community. 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, as is envisioned for the recently approved Texas A&M Institute for Data Science.