From storage to archive: lessons in infrastructure for digital science data

CHRIS JORDAN

Abstract: The Texas Advanced Computing Center is one of the nation's largest providers of cyberinfrastructure for computational science, and has petabyte-scale storage resources. TACC currently manages a tape archive system with over 3 petabytes of science data alongside 2 petabytes of online data for active research. The storage of this data, however, does not equate to the organization or accessibility of the data, particularly for external researchers who may wish to search large numbers of collections. Over the past year, TACC has embarked on an effort to research and develop comprehensive services for managing and preserving scientific data. This requires investigation of community needs, policy frameworks, and disciplinary best practices both in the sciences and in the digital preservation arena. It also requires efforts to integrate into scientific infrastructure at the local, national, and global scales, including the NSF's TeraGrid and DataNet programs. We will present the results of our ongoing research and development, including what we have learned about the needs of the science community for digital archive services, in Texas and around the world, and how we are attempting to support these needs in both the short and long-term. This platform presentation will provide information on results obtained in the last year, and future directions for research on developing technical infrastructure into a comprehensive archive system for digital science data.

About the author:

Chris Jordan has a wide variety of experience with Information Technology as it is used for everything from Visual Media to Computational Science. A graduate magna cum laude of the University of New Mexico, he has spent the past 7 years supporting computational science and other technology-assisted research at major universities, and the past 3 years focusing on issues of access to and preservation of research data in all domains. He has been Adjunct Faculty at the University of New Mexico, Data Reliability Manager for the San Diego Supercomputer Center, and Vice President of a small new media startup. He is currently responsible for issues of data infrastructure and architecture at the Texas Advanced Computing Center. He also acts as lead for the architectural design and development of data cyberinfrastructure for the National Science Foundation's TeraGrid project. His areas of expertise include large-scale data movement, management, and preservation, parallel file systems, and the integration of high-performance storage systems into diverse computational facilities.