Conceptualizing academic storage for collaborative science production: Patterns of use and governance

Melissa Cragin San Diego Supercomputer Center University of California, San Diego mcragin@sdsc.edu	Santiago Núñez-Corrales National Center for Supercomputing Applications University of Illinois at Urbana-Champaign nunezco2@illinois.edu

Abstract

Research that addresses Grand Challenges most frequently involves Big Data and computational resources distributed across a diverse spectrum of cyberinfrastructure services. Among these, data storage for access, sharing, and transfer constitute a growing and significant part of the tasks required for the conduct of research. As data-enabled research increases in complexity and scales of data generation, scientific data infrastructures (SDIs) develop to respond to a wide variety of needs across scientific and scholarly communities. Sociotechnical analysis of these functions and infrastructure could contribute to design and management, and improve their impact for (scientific) discovery and knowledge production.

As research policy has evolved and Open Science movements spread, there has been extensive study of data practices and incentives, data management services and repositories. However, in the context of active use data[1], it is difficult to find sociotechnical studies on the roles, growth, and governance of data storage and transfer systems designed for scientific production cycles. Growing evidence from practice and observation suggests that structural, organizational, and policy gaps exist between the various service centers across the U.S. data landscape, leading to disjointed interactions or disconnected data ecosystems. Further, shifts in technology and local management decisions introduce new complexities for researchers. Our work aims at gaining greater understanding of the interplay between researchers, projects, and generic storage services that facilitate data access and transfer.

This paper is in a developmental phase, as we describe and analyze the development and use of the Open Storage Network, and assess more broadly models of academic storage services and management regimes. The presentation and paper will describe the following components

1. ***Graphical typology on patterns of data use and pattern.*** We present a typology of use patterns abstracted from the literature, and refined via case analysis. The goal of this typology was to inform and facilitate "linkage" between anticipated user needs and the design and implementation of services; and, more generally, to ensure simplicity, functionality, and resilience of national scale research infrastructures.

[1] By active use here we intend, "Data that are currently relevant to scientific, scholarly, or policy communities, and continue to be used to produce new discoveries or understanding. Active data is not archival, and may not be changing in real time but endure through 'snapshots from a live data stream'." OSN General Policy.

2. ***Emerging organizational models of storage (local, regional, national).*** We discuss the impact of various scales of administrative and technological organization involved in the operation and management of data storage systems, as well as their implications for scientific collaboration, infrastructure governance and production-level operation.
3. ***OSN as a case: illuminating challenges in science production.*** The Open Storage Network (OSN) is an NSF- funded pilot project developed to test and provide a national distributed storage and transfer service. It is intended as a means to solve existing data sharing and access challenges that are not easily solved by a single institution, and which are found broadly across science and scholarly projects and communities.