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

Melissa H. 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. While this is particularly true for "Big Science" (often characterized by large instrument-generated data utilized by distributed collaborations), it is increasingly true for research requiring access to heterogeneous multi-source datasets with finer granularity. 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 among 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. These emerge at the interface between the technical aspects of data storage and sharing, and the social elements of science making across data infrastructures. As diversity of initiatives and technologies continues to increase, the challenges at the interface only magnify; new types of solutions are needed to navigate it within such a dynamic environment. We concern ourselves here with three trade-offs that articulate these challenges: (1) ease of use vs. increasing CI complexity, (2) robustness vs. increase in service diversity and (3) versatility vs. increasing organizational interconnections and regulations. Our work aims at gaining greater

[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'."

understanding of the interplay between researchers, projects, and various services that facilitate data storage, transfer, and access.

This paper is in a developmental phase, as we work to describe and analyze the development and use of the Open Storage Network, and assess other evolving models of academic storage services and management regimes. The Open Storage Network (OSN) is a new, production-level service designed to address specific data storage, transfer, sharing, and access challenges in scientific communities. Intended to enhance data-driven research collaborations across universities, the OSN leverages existing NSF networking and computing investments. OSN comprises a robust and sustainable software stack on top of a distributed network of hardware resources (called “pods”), with a small and lean administrative footprint. The aim is to maximize existing science cyberinfrastructure to make datasets more easily available for a variety of scientific and scholarly communities, especially where sharing is currently confined to “sneakernet” approaches.

The presentation will describe the following components:

1. Graphical typology on patterns of data use and transfer. 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.

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.