The Digital Preservation Network

Digital technology has enabled unprecedented growth of knowledge in essentially all areas of scholarly activity. This knowledge, however, is inherently vulnerable and the academy has been slow to recognize or deal with the problem. Consider the Sloan Digital Sky Survey. Begun in 2000, the first two phases took eight years to complete, covered 25% of the sky, mapped 930,000 galaxies, released more than 100 Terabytes of data to the scientific community, and has resulted in more than 2,000 articles and 70,000 citations to date. These are valuable data by any measure. But what are the odds that they will remain available to future generations? Thin enough to make everyone uncomfortable. In 2008, the University of Chicago Library entered into a formal agreement with Astrophysical Research Consortium agreeing to assume responsibility for archiving the data. While that was clearly a positive step, the library funding for these preservation efforts expires in 2013. Data that took 8 years to collect and that have scientific value measured in decades have a preservation horizon that expires next year. While the Sloan Foundation may well see fit to continue funding the initiative beyond 2013, it might not

Or consider what would seem to be a simpler task—preserving digital copies of printed books. By far the most ambitious and successful venture in this domain is the HathiTrust consortium. Consisting of more than 50 institutions, HathiTrust was originally built to provide access to public domain works and a safe home within the academy for all of the digital copies that resulted from the library partnership with Google to scan more than 20 million books. From the outset, it has been a forward thinking preservation effort that leverages scale economies and data center resilience, with an ambition to expand to new forms of scholarship. In short, it is arguably the current "best in class" digital preservation effort. And yet, it remains vulnerable to single points of failure with the most notable being the current lawsuit by the Authors Guild that seeks removal of in-copyright works from HathiTrust control. The Authors Guild seeks the removal of these works even though it is clear that the copies are for preservation purposes only. The lawsuit was filed in September and will likely take years to resolve, during which time the ability of the HathiTrust to serve its function as a permanent dark archive of scholarly literature will be in doubt.

These two examples, one focused on data and one on "books," highlight the fact that while digital collections proliferate at network speed, they are typically not durable and remain susceptible to multiple single points of failure. Moreover, the emphasis in building these collections tends to be more on providing current users access rather than on preserving them for the future. Absent focused and coordinated effort, much of today's scholarship will be lost to future generations.

The Digital Preservation Network (DPN) seeks to rectify this situation by building upon the higher education community's current efforts and creating a federated preservation network, owned by and for the academy, which will provide secure digital preservation of the scholarly and cultural record. At the heart of DPN is a commitment to replicate the data and metadata of research and scholarship across diverse software architectures, organizational structures, geographic regions, and political environments. Replication diversity, combined with succession rights management, will ensure that future generations have access to today's discoveries and insights.

Implementation

In the DPN ecosystem, access-oriented repositories become aggregation points (contributing nodes) that can be leveraged to deposit their content into a finite number of replicating nodes with a specific focus on long-term preservation. The replicating nodes contain redundant, dark copies of all deposits. They form, in essence, a preservation backbone that complements that access functions of existing repositories and ensures that the data and metadata remain available to future generations of scholars.

Operationally, DPN will be loosely coupled with the replicating and contributing nodes to ensure that on-going diversity is maintained. It is a federation whose primary function will be to:

- Audit and verify the successful execution of the replicating node requirements.
- Develop and maintain a sustainable financial model that supports investment for creation of the ecosystem, and a rate structure for its long-term use.
- Provide a legal framework for holding succession rights.
- Construct process and policy to address real preservation challenges that result from unplanned events such as the dissolution of a contributing node organization.
- Develop and share long-term preservation practices and technologies.