

Appendix G: Preservation Architecture

Internet Archive Preservation Architecture

The Internet Archive proposes to store, maintain, migrate and preserve fully functional digital files, metadata and enabling software for the *Omeka Commons* content in its Data Repository along with over two petaBytes of additional public collections.

The benefit of hosting the *Omeka Commons* files in the shared Archive repository environment is that all the files will be subject to the same preservation policies, technology platform migrations and evolving standards that are applied to the Internet Archive's general collections. Savings resulting from scale and cumulative experience result in low operating costs, and the Archive's status as a 501(c)3 not for profit organization ensures a low overhead burden. This low cost structure enables the Archive to offer very low all-in lifetime hosting costs.

The Digital Repository utilizes open source clustered storage technologies developed by the Internet Archive and refined over 12 years of operational experience. The storage system is managed with an open source set of applications in a configuration the Internet Archive has dubbed "The PetaBox." (See Figure 1.)

The Internet Archive utilizes the PetaBox architecture in its main Data Repository in San Francisco and in a second repository site in Redwood City, California. Within the PetaBox, where files are stored permanently, duplicate files are stored on separate drives in separate servers to protect against server or disk failure. Moreover, changed or updated files are saved separately so that an original copy always exists.

The use of open source software for managing the PetaBox architecture enables the engagement of a wide community of open source developers for future software support. To facilitate the adoption and use of this software more broadly, the PetaBox management application is being refined and consolidated into a downloadable, replicable solution for use by other institutions under a grant from the Gordon and Betty Moore Foundation. A vital community for long-term support of the management software is as important as the use of robust hardware preservation strategies in ensuring the long-term preservation of content.

Facilities

The Internet Archive has a main Data Repository in San Francisco, with total raw storage capacity of 3.7 PetaBytes. A second location, in Redwood City, California, houses the Archive's 100 billion pshr Web Collection. Storage clusters with partial mirrors of the Archive's collections are in Amsterdam, Holland and Alexandria, Egypt. Discussions are being held with the California Digital Libraries for a storage partnership and with several potential partners in Asia for mirror sites.

The Archive has a bandwidth pipeline of 5 Gbps which is often used to full capacity in support of public traffic to and downloads from the digital collections. This pipeline will likely be increased to 8 Gbps in CY 2009 due to increased traffic.

The Archive also uses Internet and Internet 2 backbones for data transmission, including a point-to-point data line from the Archive's Data Repository to Internet 2. These multiple paths to the Internet protect against access interruptions to the digital collections.

Technology Migration

All of the Internet Archive's activities are centered on hosting, archiving and providing web access to digitized materials. The Archive therefore has in place numerous processes for risk mitigation and continuation of service.

1. The Internet Archive has created new access formats from time to time. This migration would be performed on *Omeka Commons* content along with other collections hosted by Internet Archive.

An example of this sort of migration was the Internet Archive's development of monochrome PDF files for scanned books. The Alfred P. Sloan Foundation indirectly funded the Archive to create files that would work well on a current print-on-demand system. The Archive used this opportunity to recode all books being scanned into this new format as well as recode thousands books that had already been scanned because of the expected interest in having them printed in proper format on the newer print-on-demand technology.

Another example is the recoding of almost all movie files on the Archive to flash format as that became popular in 2007. Sixty thousand video were thus recoded and a new interface installed on the Archive website.

2. The Internet Archive's technical migration strategy is twofold: hardware and file access formats.

Hardware: Since its inception in 1996, the Internet Archive has migrated its collections three times – from tape to disk, and twice more to later generation disk systems. The hardware migration strategy is also based on open source hardware and generally accepted standards. Accordingly, technology migration only occurs to and from open platforms, and only when new standards have been adopted on a sufficient scale to ensure that they are relevant.

File Access Formats: The Internet Archive has migration policies for file access formats that are best described with an example: the moving images collection. The Internet Archive first stored digitized moving images in mpeg2 format, which was a widely accepted industry standard. Mpeg2 was not an easily accessible format for public use, however. The files were therefore migrated to mpeg4, a more accessible format. Today, with the popularity of YouTube and other public websites using Flash technology, the Internet Archive has developed methods to compute and present its moving image files using Flash.

Figure 1



PetaBox Row
Internet Archive Data Repository