Appendix F: About Internet Archive

About the Internet Archive

"Universal Access to Human Knowledge."

The Internet Archive was founded in 1996 to be an "Internet Library," with the goal of archiving the worldwide web. New snapshots of the worldwide web – sites from dozens of countries in over 21 languages – are taken every two months and added to the Archive's public collection at the rate of 2 Terabytes per month. The cumulative collection now totals over 100 billion web pages from 1996 to the present.

In the late 1990s the collections were expanded to include books, videos, films, audio recordings and software. The Archive's collections now consist of nearly two Petabytes (one petabyte = one million megabytes) of data. The collections are available for viewing and download via its public websites at www.archive.org, www.nasaimages.org (NASA space imagery) and www.openlibrary.org (books).

The Archive is dedicated to open source principles, and has built its storage facilities and its software tools accordingly. The Archive is perhaps best known for the Wayback Machine, a user interface to the historical web archive that enables users to browse websites exactly as they appeared in the past. Many institutions now use the Heritrix web crawler developed and used by the Archive to conduct their own web harvests. The Archive is also known for its large-scale search capabilities – the ability to index and make fully text searchable collections of over half a million items as of late 2006 – using the NutchWAX application.

The Internet Archive operates 18 scanning centers throughout the United States and around the world for low cost, mass digitization of books and microfilm. The scanning centers utilize the Scribe, a high resolution scanning station designed by Archive engineers and operated by scanning staff hired and trained by the Archive.

In order to ensure the security and long-term preservation of content, the Archive has multiple preservation strategies, including storage of duplicate files on separate hard drives at its main Digital Repository, and the use of partial mirror sites around the world, including Amsterdam and Alexandria, Egypt, at the modern Bibliotheca Alexandrina. The storage architecture is simple, low cost and highly scalable.

The Archive acts as a technology partner for large institutions (federal, research, library, international) wishing to create large web collections via web harvesting. It also provides high speed, high quality book scanning services through multiple scanning facilities in North America and Europe, and digitizes microfilm, audio and video collections donated by individuals and partners. The Internet Archive had scanned over 400,000 books for free public access as of the end of 2008, hosts over one million digital books and over 20 million book records, and is a participant in the federally funded Million Books Project.

The Internet Archive, with 11 other international libraries and the Library of Congress, co-founded the International Internet Preservation Consortium (IIPC) in 2003. This

international group works to create standards and share best practices to ensure the preservation of ephemeral web content for future generations. The Archive also founded the Open Content Alliance, a collaboration of over 110 institutions working to build joint online collections.

The Internet Archive is funded through service fees from partners as well as grants and corporate partnerships. Selected supporters include the Hewlett Foundation, Alfred P. Sloan Foundation, Andrew W. Mellon Foundation, Gordon and Betty Moore Foundation, Microsoft and Yahoo. The Internet Archive is certified as a library in the state of California.