9 Year-by-Year: From an Archive of the Internet to an Archive on the Internet

Michele Kimpton and Jeff Ubois

Internet Archive
michele@archive.org
jeff@archive.org

9.1 Introduction

Since its beginnings in 1995, the Internet Archive has pursued the long-term goal of providing universal access to all knowledge, within our lifetime.

Over the last 10 years, the Archive has engaged in projects in many different countries, for many different types of users, and worked to preserve Web pages, books, music, software, and moving images, and to make them accessible via the Internet.

Today, the Archive's Wayback Machine serves 70,000 unique visitors per day and 200 requests per second. Its 600 TB collection includes 50 billion Web pages, 30,000 books, 36,0000 sound recordings, 15,000 movies.

Hundreds of individual contributors have made this possible. So have technical advances in the computer industry, where system performance doubles every 12–18 months. In the archival community, 10 years is not a long time. But in the computer industry, 10 years is enough time to create a 100-fold factor of improvement: since 1997, the price of raw disk storage has dropped by more than 99%, from $180 to under 50 cents per gigabyte, and more than 25 million broadband connections have been added in the US alone.

It is by understanding the rate of improvement in computer based storage and access that it is possible to begin to credit what Raj Reddy of Carnegie Mellon University has described as "universal access to all knowledge."

9.2 Background: Early Internet Publishing

Even by the mid 1980s, it was clear that a change was coming to the world of electronic publishing. In the first half of the 1990s, as newspapers began to move from closed, proprietary databases to the Internet, the idea of the Internet as a library began to take shape. Internet publishing systems such as Wide Area Information Servers (WAIS) and Gopher were seen as complements to Web pages; the metaphor of the Internet as a book, with Web *pages*, referenced by Gopher servers that provided a *Table of Contents*, and an *index* produced by WAIS servers was offered as an alternative to the metaphor of the Internet as a "superhighway."

The launch of the AltaVista service in December, 1995 proved that all of the pages on the Web could be treated as a single collection, and indexed and made searchable for all users on the Net. What was not clear is how records of those pages would be maintained.

9.3 1996: Launch of the Internet Archive

The Internet Archive was formally incorporated in April, 1996, by Bruce Gilliat and Brewster Kahle.

By that time, broken links (404 errors) were a growing problem, and it was clear that most Web pages were short-lived. Some solution to this problem was needed, and a system for archiving Web pages before they vanished seemed like an obvious approach.

This led to an early design decision at the Archive about the collections policy: to be aggressive about collecting material that was in danger of disappearing, and to be opportunistic about collecting and digitizing items from the past, such as Usenet postings.

Still, as of 1996, there was little sense in the Internet community that loss of Web pages was a particularly important problem. Since the Web did not have much history, it was difficult to describe uses for expired pages.

To demonstrate the potential value of such pages, the Archive partnered with the Smithsonian Institution in Washington, DC to collect snapshots of the websites of all the 1996 Presidential candidates.[1] The tools for this project were not terribly sophisticated; they were essentially PC applications built to capture entire websites by following the links from the main page.

[1] See http://movie0.archive.org/96_Elections/index.htm

This data was eventually incorporated into the Smithsonian's presidential archive, which now includes pages from five political parties, and numerous candidates for president, ranging from Bill Clinton to Pat Buchanan. Many of the sites in this archive were shut down as candidates dropped out of the race.

Based on this success, the Library of Congress commissioned the Archive to create a focused online collection of the 2000 election, and renewed this request for the 2002 elections.

Also, in 1996, the Archive began its relationship with Alexa Internet, a for-profit company that began crawling and archiving the Web in November to provide data for a browser toolbar (plug-in) offering data about sites being viewed, and based on data gathered from other users, suggestions about related pages that might also be of interest. The Internet Archive still relies on Alexa Internet to provide data from crawls of the Web.

Two other developments from 1996 are worth noting.

The first is technical. In 1996, tape still offered a considerable price advantage over disk, and the Archive built its first generation infrastructure using tape storage robots, beginning with an ADIC 50. Despite generous contributions from leading vendors, this was ultimately to prove untenable; the access requirements posed by the Archive's users were simply too intense, and retrieval times were too slow. As Bruce Gilliat humorously says: getting a page could be done in few seconds… or days later.

The second is legal. The legal implications of collecting Web pages aggressively and to serve them up on an "opt-out" basis, as the Archive began to do that year, were unclear. AltaVista's promotion of the robots.txt exclusion protocol was an important step because it shifted the "defaults" – Web page owners that wanted to avoid landing in a search engine index, or in the Archive had an easy way to opt out, or to remove pages they clearly owned (as proven by their ability to modify the root directory of the Web site in question). While robots.txt provided resolution of the removals issue for those who owned pages in the Archive, it did not resolve the question of removing pages owned by others. This removals issue is described later with reference to small conference at UC Berkeley that was initiated by the Internet Archive in 2002 (Ubois 2002).

9.4 1997: Link Structure and Tape Robots

Collection of Web pages, link data, and "usage trails," i.e., the choices made by millions of Web users as they moved from page to page, began in

earnest in 1997 with the release of Alexa's toolbar, a browser plug-in that helped users navigate the Web by providing information on the site being viewed, including suggestions about related sites.

The link data and usage trails gathered by Alexa functioned as a collaborative filtering system, highlighting the pages that the Internet community as a whole valued most highly. Links and clicks were essentially votes on the value of a given page.

The ability to automatically determine page value was closely related to another critical design decision.

Some of the largest and most successful Web "libraries" in 1997 were essentially catalogs of sites like Yahoo. But it was uncertain how scalable a manual cataloging approach could be over time. Might it be possible to eliminate manual cataloging and the entire selections process in favor of a "collect it all" approach, combined with user-generated metadata in place of a catalog?

The answer seemed to be yes, and the Alexa began crawling pages according to how much usage they received, as measured by data gathered from its browser toolbar. Pages visited most frequently by people were backed up first.

The Alexa crawler was set to make a snapshot of the Web every eight weeks, and that schedule is still in place, though the size of each crawl increased to 100 TB in 2004 from 1 TB in 1997.

The other issue facing the Archive in 1997 was whether to rely on tape or disk storage. Tape still had cost advantages over disk, but access was slow.

As noted in (Gray and Shenoy 1999) "tape, disk, and RAM have maintained price ratios of about 1:10:1000. That is, disk storage has been 10x more expensive than tape, and RAM has been 100× more expensive than disk."

But when access costs are used as the yard stick, disk as actually much cheaper: "a tape archive is half the cost per terabyte of disk storage, but tape does not provide easy access to the data. The cost per random tape access is about a hundred thousand times higher (100 accsses/s/1 K$ disk vs. 000.01 accesses/second/10,000$ tape) (Gray and Shenoy 1999).

9.5 1998: Getting Archive Data Onto (Almost) Every Desktop

Between 1996, when Netscape went public, and 1998, the Internet and Internet-related businesses became the focus of a worldwide shift in investment priorities. As billions of dollars poured into public markets,

venture funds, and Internet startups, the number of pages to archive was doubling every 3–6 months. The number of users on the Internet was also doubling every few months. Access was becoming a concern.

In an effort to provide access to its holdings, and to establish the Archive and Alexa Internet as part of the Internet's infrastructure, Alexa entered into contracts with Microsoft and Netscape to bundle its software into the Internet Explorer and Netscape browsers. This gave Alexa presence on 90% of the world's desktop computers, and whether users knew or not, it gave them access to data held by the Internet Archive.

For the Archive, the need to begin serving up data for tens of millions of users put a strain on the tape-based infrastructure.

By the end of 1998, two things were clear:

- Scaling up would require a move from tape to disk. As the number of access requests increased, the ability of tape robots to respond was proving to be inadequate;
- Manual collections policies were more expensive than disk space; that is, archiving based on hand cataloging particular websites was more expensive than simply archiving all accessible sites according to data gathered from end users by Alexa.

9.6 1999: From Tape to Disk, A New Crawler, and Moving Images

The commercial success of the Alexa service, determined in part by its presence on virtually every PC connected to the Internet, led Amazon to acquire Alexa in 1999. This ultimately led to changes in the structure of both organizations.

An important technical development in 1999 was the creation of a new crawler by Andy Jewel. The new crawler was better able to handle parallel processes for gathering Web data, and was manageable across multiple machines. This crawler enabled Alexa to filter out 16 billions URL to crawl 4 billions and to expand the breadth and depth of its crawl.

It also cemented the decisions around the ARC file format used to store Web pages. Originally developed in 1996 by Mike Burner and Brewster Kahle, the ARC File Format Specification (Burner and Kahle 1996), which was designed to meet several requirements:

- The file must be self-contained: it must permit the aggregated objects to be identified and unpacked without the use of a companion index file;

- The format must be extensible to accommodate files retrieved via a variety of network protocols, including HTTP, FTP, news, gopher, and mail;
- The file must be "streamable": it must be possible to concatenate multiple archive files in a data stream;
- Once written, a record must be viable: the integrity of the file must not depend on subsequent creation of an in-file index of the contents.

While the specification does not require an external index of the contents and object-offsets, such an index greatly enhances the retrievability of objects stored in this format. Today, the Archive maintains such indices, and is seeking to standardize their format through the International Internet Preservation Consortium.[2]

1999 also marked a move beyond Web pages into other types of data. By 1999, storage prices had dropped to the point at which the Archive could begin collecting moving images. Through a partnership with Rick Prelinger of the Prelinger Archives, a project to digitize 1,000 films (for an ultimate cost of $160,000) and to begin to archive television news broadcasts began operating at the end of the year.

9.7 2000: Building Thematic Web Collections

By 2000, the Archive had achieved a level of technical stability. Acceptance of crawling data was routine, and the migration from tape to disk was long over.

Table 9.1. The Archive's Internet Collections as of March, 2000

Collection	Units	Size
Web (1996 to 3/2000)	1 billion pages	13.8 terabytes (TB)
FTP (1996)	50,000 sites	.05 TB
Usenet (1996-1998)	16 million postings	.592 TB

2000 was another election year in the US, and this time most of the electorate had Internet Access. It was clear to the political establishment that a presence on the Internet was vital to winning the election, and with this increased focus on politics online, the Internet Archive partnered with the Library of Congress to collect political sites.

This was the Archive's first project with the Library of Congress, and for many on staff at the Archive, marked a transition from experimental project into an established institution.

[2] http://netpreserve.org

The idea of providing access to preserved ephemeral works gained momentum, and the Moving Images Archive was released in 2000. Now under Creative Commons licenses, the Moving Images Archive consists primarily of films from the Prelinger Archive, a collection of over 1,900 "ephemeral" (advertising, educational, industrial, and amateur) films. The collection currently contains over 10% of the total production of ephemeral films between 1927 and 1987 in the USA, and is one of the most complete and varied collection in existence of films from these poorly preserved genres.

9.8 2001: Public Access with the Wayback Machine: The 9/11 Archive

Between March, 2000 and March, 2001, the Archive tripled the size of its holdings to a total of more than 40 TB. At that point, the Archive was growing by roughly 10 TB per month.

Table 9.2. The Archive's Collections as of March 2001

Collection	Units	Size
1996 – 3/2001	4 billion pages	40 TB
Election 2000 Archive	200 million pages	2 TB
Usenet: 1996 – 1998, 2000 – 3/2001	16 million postings	.5 TB
Archival movies: ca. 1903 to ca. 1973	360 movies	.5 TB
Arpanet: Historical documentation	5,000 pages	< .1 TB

But 2001 was a difficult year for many high tech organizations in the San Francisco area. The collapse of the stock market, the demise of hundreds of local companies, and the World Trade Center attack in New York all had an effect on the Archive's operations. In particular, the loss of high tech jobs in the San Francisco area made it easier to hire engineers, and archiving the events 9/11 provided a focus, as well as a test of the Archive's ability to handle moving images and to respond to events.

In early 2001, perhaps the main question facing the Archive was how best to provide access to the collection. Some data was served directly to the general public via the Alexa service, but direct access to the collections still required Unix programming skills.

Working under contract to the Archive, programmers at Alexa built the Wayback Machine, which serves up contents of the Archive based on URLs. On October 24, 2001, the Wayback Machine went live, offering access to more than 10 billion archived Web pages and 100 TB of data.

At that time, data was stored on Hewlett Packard and uslab.com servers running the FreeBSD and Linux operating systems. Each computer had about 512 Mb of memory and generally held just over 300 GB of data on IDE disks.

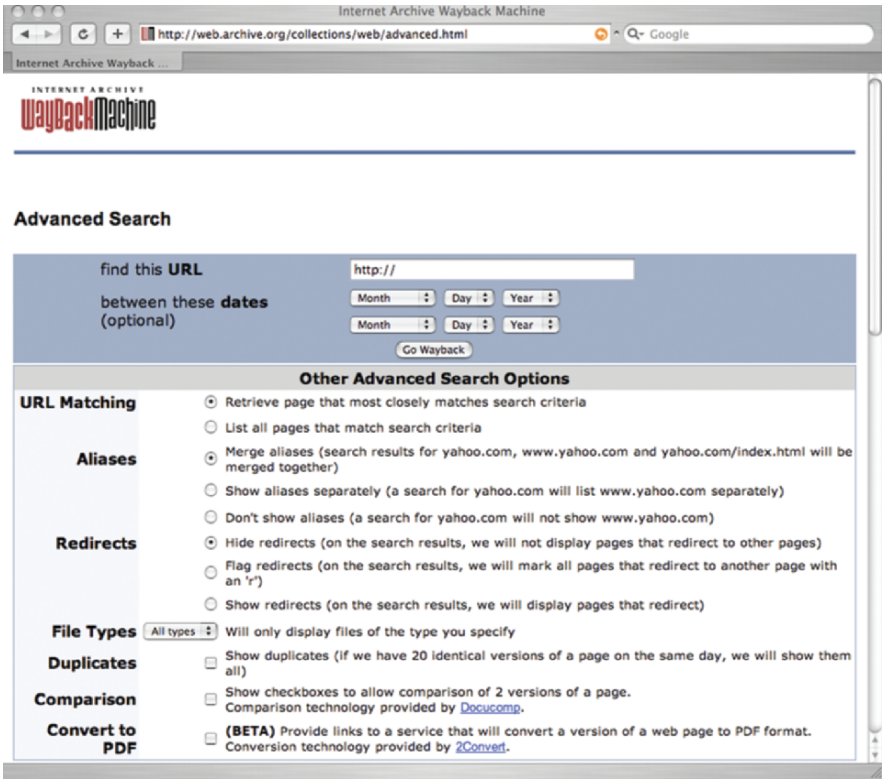

Fig. 9.1. Current search options for the Wayback machine

The other major project of 2001 was the September 11 Archive. Working with the Library of Congress, the Archive collected Images from over 30,000 selected websites from September 11, 2001 to December 1, 2001, and hundreds of hours of broadcast news footage.

9.9 2002: The Library of Alexandria, The Bookmobile, and Copyrights

The Archive undertook five major projects in 2002 to expand the breadth of its collections, the options for accessing those collections, its partnerships with other organizations, and its role in policymaking.

The first and largest was the creation of a mirror site at the Library of Alexandria in Alexandria, Egypt. Servers and more than 100 TB of data, valued at more than $5 million, were shipped to Egypt and installed in time for the Library's grand opening in April.

The second major project was the creation of the Internet Bookmobile, designed to show how the combination of electronic scans of books, print on demand technology and a satellite network connection could effectively fit a million-book library in the back of a van. Partnering with Carnegie Mellon in the summer of 2002, the Million Books Project (MBP) was launched, aiming to digitize at least one million books and offer them free-to-read on the Internet. With encouragement from Suzanne Mubarak, the Archive began building a bookmobile in the U.S., and working with others in India and Kenya to clone its basic prototype.

The third major project was policy related. On September 30, 2002, in an effort to raise public awareness of important copyright policy issues, Internet Archive's Bookmobile embarked on a cross-country journey to print and deliver free books. The Bookmobile parked and printed books at the United States Supreme Court building where, on October 9, the Justices heard arguments in Eldred vs. Ashcroft, a landmark case that decided how many books would be part of theBookmobile's digital library and all other digital libraries in the US Unfortunately, Eldred was defeated and the copyright extension went into effect, but the Bookmobile project blossomed and eventually became its own non-profit. Currently the Government of India is building 25 bookmobiles for use throughout India.

The fourth area of activity involved the creation of the Archive's first book and music collections. In June, the first book collections were put online; in August, the Live Music Archive, a collection of concert performances that could be downloaded legally, went online.

The fifth major project was the launch of the International Children's Digital Library, in partnership with the University of Maryland, and supported by Library of Congress, NSF, IMLS, Kahle/Austin Foundation, Adobe Systems Inc., the Markle Foundation, and Octavo. The ICDL was and is focused on the inherent promise of the Internet to provide direct and global access to quality content for children.

At the end of 2002, the Archive led an effort to ensure the integrity of digital archives by standardizing the criteria under which materials might be removed or made unaccessible. In a meeting at UC Berkeley, representatives from the Archive met with other digital librarians to evolve the Oakland Archive Policy (Ubois 2002),[3] which details procedures for disclosing removals of materials as required by law, or as requested by site owners and others (see also Chapter 1 (Masanès 2006)).

[3] See see http://www.sims.berkeley.edu/research/conferences/aps/removal-policy.html

9.10 2003: Extending Our Reach via National Libraries and Educational Institutions

In 2003, the Archive continued to reach out to national libraries and educational institutions around the world. With the International Internet Preservation Consortium (IIPC), the Archive began working closely with participating organizations on new standards and a new open source crawler.

In July, 2003, the Archive helped to launch the International Internet Preservation Consortium, a group of 12 national libraries that agreed to work on developing standards, tools and policies to to acquire, preserve and make accessible knowledge and information from the Internet for future generations everywhere, promoting global exchange and international relations. To achieve this mission, the IIPC is working to accomplish the following goals:

- To enable the collection of a rich body of Internet content from around the world to be preserved in a way that it can be archived, secured and accessed over time;
- To foster the development and use of common tools, techniques and standards that enable the creation of international archives;
- To encourage and support national libraries everywhere to address Internet archiving and preservation.

The IIPC was chartered at the Bibliothèque nationale de France with 12 participating institutions. The members agreed jointly to fund and participate in projects and working groups to accomplish the goals of the IIPC. The initial agreement is in effect for 3 years. During that period, the membership is limited to charter institutions.

In 2003, the Archive received significant outside funding from outside organizations, including, the Hewlett and Sloan Foundations, and began work on a series of special collections.

Further declines in the cost of disk storage and internet bandwidth led the Archive to make a standing offer of "unlimited bandwidth, forever, for free" to organizations and individuals with digital materials.

This offer led to a partnership with Etree, an all-volunteer organization founded in 1998 to enable free, legal trading of recordings of live music concerts. As a result of the partnership with Etree, the Archive now hosts more than 15,000 live music concerts.

To support growing needs for both storage and bandwidth, the Archive opened a new data center in San Francisco. The new data center was connected to the Internet with a 1 Gbps link and housed more than 1,500 commodity PCs all running Linux.

9.11 2004: And the European Archive and the Petabox

In 2004, the Archive began migrating data to its third generation of hardware, known as the Petabox. Based on rack-mounted commodity hardware and the Linux operating system, the Petabox design offered RAID storage for roughly $2,000 per terabyte, or $2 million per petabyte.

The first installation of the new design was in Amsterdam, at the newly formed European Archive, an institution intended to serve the needs of the European community that Internet Archive supported with other european partners in its first year of existence. The installation in Amsterdam is growing to provide a mirror to the collections in Alexandria and San Francisco. Creating a network of independent institutions around the world, each capable of operating independently, will help to prevent catastrophic losses of information.

Also in 2004, the International Internet Preservation Consortium launched Heretrix, an open-source, extensible, Web-scale, archival-quality Java-based Web crawler. Now in use by many heritage institutions worldwide, Heretrix supports multiple different use cases including focused and broad crawling.

Collection development in 2004 took major steps forward through the hiring of additional staff, completion of book and movie scanning projects, donations of data from other institutions.

9.12 The Future

Advances in computing and communications make it possible to cost-effectively store every book, sound recording, movie, software package, and public Web page ever created, and provide access to these collections via the Internet to students and adults all over the world.

As stated in the Universal Declaration of Human Rights, Article 19 “Everyone has the right to freedom of opinion and expression; this right includes freedom to hold opinions without interference and to seek, receive and impart information and ideas through any media regardless of frontiers.”

For the years ahead, the Archive’s mission is clear: to create a new Universal Library that makes all knowledge easily available to every man woman and child around the world.

References

Burner, M. & Kahle, B. (1996). The ARC File Format

Gray, J. & Shenoy, P. (1999). Rules of Thumb in Data Engineering. *Microsoft Technical Report*, *MS-TR-99-100*

Masanès, J. (2006). Web archiving: issues and methods. In J. Masanès (Ed.), *Web Archiving*. Springer, Berlin Heidelberg New York

Ubois, J. (2002). The Oakland Archive Policy. Recommendations for Managing Removal Requests And Preserving Archival Integrity