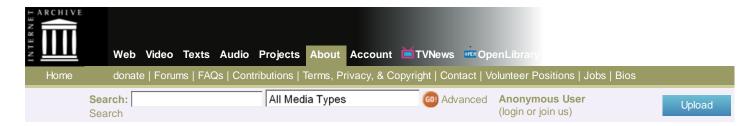
1/31/2013 6:55 AM http://archive.org/about/



## **Read More**

Why the Archive is Building an 'Internet Library'

**Future Libraries** 

Storage and Preservation

Related Projects and Research

Researcher Access

Server Statistics

**Archive Statistics** 

Job Opportunities at the Internet Archive

Usage Logs

## News [more]

Brewster Kahle's Internet Archive

Google and US
publishers reach
settlement over book
scanning, though
authors' suit remains

Internet Archive houses all TV news since 2009

Internet Archive collects, streams TV news archives

Seagate Helps Preserve Internet's Past

All the TV News Since 2009, on One Web Site

Internet Archive
Launches TV News

Internet Pioneer, Former Banker Behind Newest Credit Union

Internet Archive Turns Up the Speed With BitTorrent

Alexandria 2.0: One
Millionaire's Quest to
Build the Biggest Library
on Earth

Member

## **About the Internet Archive**

The Internet Archive is a 501(c)(3) non-profit that was founded to build an Internet library. Its purposes include offering permanent access for researchers, historians, scholars, people with disabilities, and the general public to historical collections that exist in digital format. Founded in 1996 and located in San Francisco, the Archive has been receiving data donations from Alexa Internet and others. In late 1999, the organization started to grow to include more well-rounded collections. Now the Internet Archive includes texts, audio, moving images, and software as well as archived web pages in our



collections, and provides specialized services for adaptive reading and information access for the blind and other persons with disabilities.

### Why the Archive is Building an 'Internet Library

Libraries exist to preserve society's cultural artifacts and to provide access to them. If libraries are to continue to foster education and scholarship in this era of digital technology, it's essential for them to extend those functions into the digital world.

Many early movies were recycled to recover the silver in the film. The <u>Library of Alexandria</u> - an ancient center of learning containing a copy of every book in the world - was eventually burned to the ground. Even now, at the turn of the 21st century, no comprehensive archives of television or radio programs exist.

But without cultural artifacts, civilization has no memory and no mechanism to learn from its successes and failures. And paradoxically, with the explosion of the Internet, we live in what Danny Hillis has referred to as our "digital dark age."

The Internet Archive is working to prevent the Internet - a new medium with major historical significance - and other "born-digital" materials from disappearing into the past. Collaborating with institutions including the <u>Library of Congress</u> and the <u>Smithsonian</u>, we are working to preserve a record for generations to come.

Open and free access to literature and other writings has long been considered essential to education and to the maintenance of an open society. Public and philanthropic enterprises have supported it through the ages.

The Internet Archive is opening its collections to researchers, historians, and scholars. The Archive has no vested interest in the discoveries of the users of its collections, nor is it a grant-making organization.

At present, the size of our Web collection is such that using it requires <u>programming skills</u>. However, we are hopeful about the development of tools and methods that will give the general public easy and meaningful access to our collective history. In addition to developing our own collections, we are working to promote the formation of other Internet libraries in the United States and elsewhere.

## Find out

- How to make a Monetary Donation to the Archive
- About our <u>announcement and discussion lists</u> on Internet libraries and movie archives

as well as our user forums

Future Libraries - How People Envision Using Internet Libraries

**From ephemera to artifact:** Internet libraries can change the content of the Internet from ephemera to enduring artifacts of our political and cultural lives.

1/31/2013 6:55 AM http://archive.org/about/



"I believe historians need every possible piece of paper and archived byte of digital data they can muster. The Smithsonian Institution sees the value, and has affiliated with the Archive to preserve the <a href="1996 campaign">1996 campaign</a> Web sites, official and unofficial."

Dan Gillmor, computing editor, San Jose Mercury News, 1 September 1996

Protecting our right to know: Most states have pre-Internet sunshine laws that require public access to government documents. Yet while the Internet has generally increased public access to information, states have just begun to amend those laws to reflect today's Internet environment. According to Bill Chamberlin, director of the Marion Brechner Citizen Access Project at the University of Florida's College of Journalism and Communications, such laws are being enacted "piecemeal, one state at a time," and cover information that varies widely in nature everything from "all public records" to specialized information such as education reports and the licensing status of medical practitioners. In the meantime, while public officials are posting more information on the Internet than their state legislatures require, there's little regulatory control over exactly what is posted, when it's taken off, or how often it's updated. This leaves a gap that online libraries can help to fill.

**Exercising our "right to remember":** Without paper libraries, it would be hard to exercise our "right to remember" our political history or hold government accountable. With much of the public's business now moving from paper to digital media, Internet libraries are certain to become essential in maintaining that right. Imagine, for instance, how news coverage of an election campaign might suffer if journalists had only limited access to previous statements that candidates had made in the media.

"The Internet Archive is a service so essential that its founding is bound to be looked back on with the fondness and respect that people now have for the public libraries seeded by Andrew Carnegie a century ago.... Digitized information, especially on the Internet, has such rapid turnover these days that total loss is the norm. Civilization is developing severe amnesia as a result; indeed it may have become too amnesiac already to notice the problem properly. The Internet Archive is the beginning of a cure - the beginning of complete, detailed, accessible, searchable memory for society, and not just scholars this time, but everyone."

Stewart Brand, president, The Long Now Foundation

**Establishing Internet centers internationally:** What is a country without a memory of its cultural heritage? Internet libraries are the place to preserve the aspect of a country's heritage that exists on the Internet.

**Tracing the way our language changes:** During the late 19th century, James Murray, a professor at Oxford University, built the first edition of the *Oxford English Dictionary* by sending copies of selected books to "men of letters" who volunteered to search them for the first occurrences of words and to trace the migration of their various meanings. Internet libraries could allow linguists to automate much of this extremely labor-intensive process.

**Tracking the Web's evolution:** Historians, sociologists, and journalists could use Internet libraries to hold up a mirror to society. For example, they might ask when different ethnic groups or special interests or certain businesses became a presence on the Internet.

"We don't know where this Internet is going, and once we get there it will be very instructive to look back."

Donald Heath, president of the Internet Society in Reston, Virginia

Reviving dead links: A few services - such as UC Berkeley's Digital Library

1/31/2013 6:55 AM http://archive.org/about/

> Project, the Online Computer Library Center, and Alexa Internet are starting to offer access to archived versions of Web pages when those pages have been removed from the Web. This means that if you get a "404 - Page Not Found" error, you'll still be able to find a version of the page.

Understanding the economy: Economists could use Archive data such as link structures - what and how many links a site contains - to investigate how the Web affects commerce.

Finding out what the Web tells us about ourselves: Researchers could use data on links and traffic to better understand human behavior and communication.

"Researchers could use the Archive's Web snapshots in combination with usage statistics to compare how people in different countries use the Web over long periods of time.... Political scientists and sociologists could use the data to study how public opinion gets formed. For example, suppose a device for increasing privacy became available: Would it change usage patterns?"

Bernardo Huberman, Xerox Palo Alto Research Center

"The Internet Archive has created a kind of test tube that allows a broad range of researchers to analyze the Web in ways that have never been possible before. What makes this type of research unique is that it often requires the fusion of traditional tools and techniques with new methods, and it results in the development of new theories, techniques, and metrics."

James Pitkow, Xerox Palo Alto Research Center

Looking back: With a "way-back machine" - a device that displayed the Web as it looked on a given date - historians and others would literally have a window on the past.

How would you use an Internet library?

Internet libraries raise many issues in a range of areas, including archiving technology, copyright, privacy and free speech, trademark, trade secrets, import/export issues, stolen property, pornography, the question of who will have access to the libraries, and more.

Below are links to projects, resources, and institutions related to Internet libraries.

Internet Libraries and Librarianship Archiving Technology

**Internet Mapping** 

**Internet Statistics** 

Copyright

Privacy and Free Speech

## Internet Libraries and Librarianship

Alexa Internet has catalogued Web sites and provides this information in a free service.

www.alexa.com

The American Library Association is a major trade association of American libraries.

www.ala.org

The Australian National Library collects material including organizational Web sites.

pandora.nla.gov.au/documents.html

The Council on Library and Information Resources works to ensure the well-being of the scholarly communication system.

www.clir.org

See its publication **Why Digitize?** at www.clir.org/pubs/reports/pub80-smith/pub80.html

The Digital Library Forum (D-Lib) publishes an online magazine and other resources for building digital libraries. www.dlib.org

**Attorney I. Trotter Hardy** explains copyright law and examines its implications for digital materials in his paper **Internet Archives and Copyright.** 

copyright\_TH.php

The Internet Public Library site has many links to online resources for the general public.

www.ipl.org

**Brewster Kahle** is a founder of WAIS Inc. and Alexa Internet and chairman of the board of the Internet Archive. See his paper **The Ethics of Digital Librarianship** at

ethics\_BK.php

**Michael Lesk** of the National Science Foundation has written extensively on digital archiving and digital libraries. www.purl.net/NET/lesk

**The Library of Congress** is the national library of the United States. www.loc.gov

**The Museum Digital Library** plans to help digitize collections and provide access to them.

www.digitalmuseums.org

The National Archives and Records Administration oversees the management of all US federal records. It also archives federal Web sites including the <u>Clinton White House site</u>.

www.nara.gov

The National Science Foundation Digital Library Program has funded academic research on digital libraries. www.nsf.gov/home/crssprgm/dli/start.htm

National Technical Information Service (NTIS), U.S. Department of Commerce, Technology Administration. NTIS is an archive and distributor of scientific, technical, engineering and business related information developed by and for the federal government.

www.ntis.gov

**Network Wizards** has been tracking Internet growth for many years. www.nw.com

**Project Gutenberg** is making ASCII versions of classic literature openly available. <a href="www.gutenberg.org">www.gutenberg.org</a>

The Radio and Television Archive has many links to related resources. www.rtvf.unt.edu/links/histsites.htm

**Revival of the Library of Alexandria** is a project to revive the ancient library in Egypt. www.bibalex.org

The Society of American Archivists is a professional association focused on ensuring the identification, preservation, and use of records of

historical value.

www.archivists.org

The Royal Institute of Technology Library in Sweden is creating a system of quality-assessed information resources on the Internet for academic use.

www.lib.kth.se/main/eng

The United States Government Printing Office produces and distributes information published by the US government. www.access.gpo.gov

**The University of Virginia** is building a catalog of digital library activities. http://www.lib.virginia.edu/digital/

# **Archiving Technology**

The Association for Computing Machinery (ACM) computing and public policy page includes papers and news on pending legislation on issues including universal access, copyright and intellectual property, free speech and the Internet, and privacy.

www.acm.org/serving

The Carnegie Mellon University Informedia Digital Video Library Project is studying how multimedia digital libraries can be established and used.

www.informedia.cs.cmu.edu

The Intermemory Project aims to develop highly survivable and available storage systems.

www.intermemory.org

The National Film Preservation Board, established by the National Film Preservation Act of 1988, works with the Library of Congress to study and implement plans for film and television preservation. The site's research page includes links to the board's 1993 film preservation study, a 1994 film preservation plan, and a 1997 television and video study. All the documents warn of the dire state of film and television preservation in the United States.

lcweb.loc.gov/film/filmpres.html

The National Institute of Standards and Technology (NIST) posts IEC International Standard names and symbols for prefixes for binary multiples for use in data processing and data transmission. www.physics.nist.gov/cuu/Units/binary.html

The Text Retrieval Conference (TREC) encourages research in information retrieval from large text collections.

trec.nist.gov

# **Internet Mapping**

**An Atlas of Cyberspaces** has maps and dynamic tools for visualizing Web browsing.

www.cybergeography.com/atlas/surf.html

**The Internet Mapping Project** is a long-term project by a scientist at Bell Labs to collect routing data on the Internet.

www.cs.bell-labs.com/who/ches/map

The Matrix Information Directory Service has good maps and visualizations of the networked world. www.mids.org

Peacock Maps has maps of Internet connectivity.

www.peacockmaps.com

# **Internet Statistics**

**WebReference** has an Internet statistics page (publisher: Internet.com). webreference.com/internet/statistics.html

# Copyright

The Association for Computing Machinery (ACM) copyright information page includes text of pertinent laws and pending legislation. www.acm.org/usacm/copyright

**Tom W. Bell** teaches intellectual property and Internet law at Chapman University School of Law.

www.tomwbell.com

His site includes a graph showing the trend of the maximum US copyright term at www.tomwbell.com/writings/(C)\_Term.html

**Cornell University** posts the text of **copyright law** at <a href="https://www4.law.cornell.edu/uscode/unframed/17/107.html">www4.law.cornell.edu/uscode/unframed/17/108.html</a>

**The Digital Future Coalition** is a nonprofit working on the issues of copyright in the digital age. www.dfc.org

The National Academy Press is the publishing arm of the <u>national</u> academies.

"The Digital Dilemma: Intellectual Property in the Information Age" <a href="http://www.nap.edu/html/digital\_dilemma/">http://www.nap.edu/html/digital\_dilemma/</a>

"LC21: A Digital Strategy for the Library of Congress" www.nap.edu/books/0309071445/html

**Pamela Samuelson** is a professor in the School of Information Management and Systems at UC Berkeley. info.berkeley.edu/~pam

Title 17 of US copyright code www.loc.gov/copyright/title17/

US Government Copyright Office www.loc.gov/copyright

# **Privacy and Free Speech**

The Association for Computing Machinery (ACM) free-speech information page includes the text of pertinent laws and pending legislation.

www.acm.org/usacm/speech

The Association for Computing Machinery (ACM) privacy information page includes the text of congressional testimony and links to other resources.

www.acm.org/usacm/privacy

The Benton Foundation Communications Policy and Practice Program has the goal of infusing the emerging communications environment with public-interest values.

www.benton.org/cpphome.html

The Center for Democracy and Technology works to promote democratic values and constitutional liberties in the digital age. <a href="https://www.cdt.org">www.cdt.org</a>

The Computers Freedom and Privacy Conference has a site

containing information on each annual conference held since 1991. <a href="https://www.cfp.org">www.cfp.org</a>

The Electronic Frontier Foundation works to protect fundamental civil liberties, including privacy and freedom of expression in the arena of computers and the Internet.

The Electronic Privacy Information Center, a project of the Fund for Constitutional Government, is a public-interest research center whose goal is to focus public attention on emerging civil liberties issues and to protect privacy, the First Amendment, and constitutional values.

www.epic.org

The Free Expression Policy Project is a think tank on artistic and intellectual freedom at NYU's Brennan Center for Justice. Through policy research and advocacy, they explore freedom of expression issues including censorship, copyright law, media localism, and corporate media reform.

www.fepproject.org

www.eff.org

The Internet Free Expression Alliance is an information and advocacy organization focused on free speech as it relates to the Internet. <a href="https://www.ifea.net">www.ifea.net</a>

The Internet Privacy Coalition aims to protect privacy on the Internet by promoting the widespread availability of strong encryption and the relaxation of export controls on cryptography.

www.privacy.org/ipc

The Privacy Page includes news, alerts, and links to privacy-related resources. Related organizations include the <u>Electronic Privacy Information Center</u>, the <u>Internet Privacy Coalition</u>, and <u>Privacy International</u>.

www.privacy.org

**Privacy International** is a London-based human rights group formed as a watchdog on surveillance by governments and corporations. <a href="https://www.privacy.org/pi">www.privacy.org/pi</a>

Please <u>suggest</u> other pages that may be appropriate here.

### Storage and Preservation

The Archive has two practical considerations in dealing with digital collections:

<u>How to store</u> massive amounts of data <u>How to preserve</u> the data for posterity

# **Storage**

Storing the Archive's collections involves parsing, indexing, and physically encoding the data. With the Internet collections growing at exponential rates, this task poses an ongoing challenge.

Our hardware consists of PCs with clusters of IDE hard drives. Data is stored on <a href="DLT tape">DLT tape</a> and hard drives in various appropriate formats, depending on the collection. Web data is received and stored in archive format of 100-megabyte ARC files made up of many individual files. <a href="Alexa Internet">Alexa Internet</a> (currently the source of all crawls in our collections) is proposing ARC as a standard for archiving Internet objects. See Alexa for the <a href="format specification">format specification</a>.

## **Preservation**

Preservation is the ongoing task of permanently protecting stored resources from damage or destruction. The main issues are guarding against the consequences of

1/31/2013 6:55 AM http://archive.org/about/

accidents and data degradation and maintaining the accessibility of data as formats become obsolete.

**Accidents:** Any medium or site used to store data is potentially vulnerable to accidents and natural disasters. Maintaining copies of the Archive  $\tilde{A}^ \hat{A}_2$   $\hat{A}_2$  collections at multiple sites can help alleviate this risk. Part of the collection is already handled this way, and we are proceeding as quickly as possible to do the same with the rest.

**Migration:** Over time, storage media can degrade to a point where the data becomes permanently irretrievable. Although <u>DLT tape</u> is rated to last 30 years, the industry rule of thumb is to migrate data every 10 years. We no longer use tapes for storage, however. Please take a look at <u>our page on our Petabox system</u> for more information on our storage systems.

**Data formats:** As advances are made in software applications, many data formats become obsolete. We will be collecting software and emulators that will aid future researchers, historians, and scholars in their research.

# Find out

How to get <u>free access to the Archive's Internet collections</u>
About our <u>announcement and discussion lists</u> on Internet libraries and movie archives

Terms of Use (10 Mar 2001)