Michele Kimpton, Internet Archive

April 7, 2006

Written Response to Section 4, Section 108

TOPIC 4: Given the ephemeral nature of websites and their importance in documenting the historical record, should a special exception be created to permit the online capture and preservation by libraries and archives of certain website or other online content?

Internet Archive

IA has been archiving web pages for 10 years- and has made these pages available to the public since 2001. Currently in our repository, there are 60 billion pages, over 50 million websites from around the world. We archive approx. 2 billion pages per month. The majority of these pages have been donated by Alexa. We have tried to work with other companies to provide donations but have run into issues due the uncertainty of the law- so clarity would help us archive more content from a broader source of partners. We provide public access to these pages through chiefly our website, www.archive.org, where anyone can try and browse the web the way it was. We get several hundred thousand users per month requesting content via our website of archived web pages.

We also collect and archive material for other "cultural heritage" institutions around the world, like LC, NARA, Bnf, NLA. In all cases a copy of these collections is also maintained at the archive-

Lastly, we have recently launched a service to allow institutions to harvest, access and preserve websites through a web application. This service is available to non commercial "memory" institutions around world, to be able to archive important content but do not have the technical infrastructure or resources.

Unlike other content- the web is born digital and changes frequently-every 100 days on average-so if the content is not archived now it will be lost forever. Many documents change much more frequently

We allow the user to access the content by browsing the web as the way it was in time. So one can move from page to page seamlessly without broken links,

Putting limitations on what can be archived by type or by site-will prevent researchers and future generations from surfing the web the way it was. The web is a complex network of sites, linked together without boundaries. If a site is not publicly accessible- it is password protected, or has a login- we do not capture it.

We follow the Oakland Archive policy for access to web archive content. If a site owner would like to block access, he need only to put a robots.txt file on his site- and we will remove access to the website. This policy has been in place since 2001 and has worked successfully for over 50 million sites that are archived.

Web crawlers should identify themselves to the website owners upon capturing their site. This is a practice followed by Internet Archive, and is common among search engine companies.

Archived websites become available anywhere from 1 week to 6 months post archiving. We have found there is no issue to date of the "archived" website being preferred over the live site for browsing.

For users to distinguish archived web pages from a live web page they can look at the URL displayed in their browser and they will see archive.org followed by the archived date followed by the url of the website being displayed. Although putting a clear banner in front of an archived page makes good sense we have found it impractical and difficult to do consistently at large scale. Due to the numerous technologies and browsers for creating and displaying web pages- it is difficult to get something to work in every situation.

Conclusion:

IA was founded due to the fact that primary source content was disappearing off the web every day. At the time, in the US, none of the traditional memory institutions were saving the content. It was due to IA's foresight and innovation at the time, that billions of web pages from the 96 to now- no longer on the web,have been archived and preserved for future generations, in many cases being the only institution in the world with such holdings. It is important in the early days of the digital world to let innovative organizations, like IA, take part in leading and contributing to the digital llibraries and archives of the world.

Answers to specific questions:

Should a special exception be created to permit the online capture and preservation by libraries and archives of certain website or other online content? If so, should such an

exception be similar to section 108(f)(3), which permits libraries and archives to capture audiovisual news programming off the air
Yes of course, content is disappearing daily.

Should such an exception be limited to a defined class of sites or online content, such as non-commercial content/ sites (i.e., where the captured content is not itself an object of commerce), so that news and other media sites are excluded?
No, any website owner has the ability to block his site from being captured- if we choose sites selectively at this early stage- we will not capture the complete experience on the web where users can browse seamlessly from site to site.

Should the exception be limited to content that is made freely available for public viewing and or downloading without access restrictions or user registration?
Currently we archive any site which is not blocked by a robots.txt exclusion and not password protected.

Should there be an opt-out provision, whereby an objecting site owner or rights-holder could request that a particular site not be included? Should site owners or operators be notified ahead of the crawl that captures the site that the crawl will occur? Should “no archive” meta-tags, robot.txt files, or similar technologies that block sites or pages from being crawled be respected?

We follow the Oakland Archive policy, that allows a site owner to remove access from the archive, and prevent being archived by putting up a robots.txt exclusion on their website. The policy outlines an "opt out" approach where, if requested, we will expeditiously remove a site from access. In most cases we find when owners understand we are archiving the content for the library, they want to keep their site visable. We have implemented this policy since 2001 and works well.

The Oakland Archive Policy

Recommendations for Managing Removal Requests And Preserving Archival Integrity
School of Information Management and Systems, U.C. Berkeley
December 13 ? 14, 2002

Introduction

Online archives and digital libraries collect and preserve publicly available Internet documents for the future use of historians, researchers, scholars, and the general public. These archives and digital libraries strive to operate as trusted repositories for these materials, and work to make their collections as comprehensive as possible.

At times, however, authors and publishers may request that their documents not be included in publicly available archives or web collections. To comply with such requests, archivists may restrict access to or remove that portion of their collections with or without notice as outlined below.

Because issues of integrity and removal are complex, and archivists generally wish to respond in a transparent manner, these policy recommendations have been developed with help and advice of representatives of the Electronic Frontier Foundation, Chilling Effects, The Council on Library and Information Resources, the Berkeley Boalt School of Law, and various other commercial and non-commercial organizations through a meeting held by the Archive Policy Special Interest Group (SIG), an ad hoc, informal group of persons interested the practice of digital archiving.

In addition, these guidelines have been informed by the American Library Association?s Library Bill of Rights http://www.ala.org/work/freedom/lbr.html, the Society of American Archivists Code of Ethics http://www.archivists.org/governance/handbook/app_ethics.asp, the International Federation of Library Association?s Internet Manifesto http://www.unesco.org/webworld/news/2002/ifla_manifesto.rtf, as well as applicable law.

Recommended Policy for Managing Removal Requests

Historically, removal requests fall into one of the following five categories. Archivists who wish to adopt this policy will respond according to the following guidelines:

Type of removal request	Response
Request by a webmaster of a private (non-governmental) web site, typically for reasons of privacy, defamation, or embarrassment.	1. Archivists should provide a ?self-service? approach site owners can use to remove their materials based on the use of the robots.txt standard. 2. Requesters may be asked to substantiate their claim of ownership by changing or adding a robots.txt file on their site. 3. This allows archivists to ensure that material will no longer be gathered or made available. 4. These requests will not be made public; however, archivists should retain copies of all removal requests.
Third party removal requests based on the Digital Millennium Copyright Act of 1998 (DMCA).	1. Archivists should attempt to verify the validity of the claim by checking whether the original pages have been taken down, and if appropriate, requesting the ruling(s) regarding the original site.

	2. If the claim appears valid, archivists should comply. 3. Archivists will strive to make DMCA requests public via Chilling Effects, and notify searchers when requested pages have been removed. 4. Archivists will notify the webmaster of the affected site, generally via email.
Third party removal requests based on non-DMCA intellectual property claims (including trademark, trade secret).	1. Archivists will attempt to verify the validity of the claim by checking whether the original pages have been taken down, and if appropriate, requesting the ruling(s) regarding the original site. 2. If the original pages have been removed and the archivist has determined that removal from public servers is appropriate, then the archivists will remove the pages from their public servers. 3. Archivists will strive to make these requests public via Chilling Effects, and notify searchers when requested pages have been removed. 4. Archivists will notify the webmaster of the affected site, generally via email
Third party removal requests based on objection to controversial content (e.g. political, religious, and other beliefs).	As noted in the Library Bill of Rights, ?Libraries should provide materials and information presenting all points of view on current and historical issues. Materials should not be proscribed or removed because of partisan or doctrinal disapproval.? Therefore, archivists should not generally act on these requests.
Third party removal requests based on objection to disclosure of personal data provided in confidence.	Occasionally, data disclosed in confidence by one party to another may eventually be made public by a third party. For example, medical information provided in confidence is occasionally made public when insurance companies or medical practices shut down. These requests are generally treated as requests by authors or publishers of original data.
Requests by governments.	Archivists will exercise best-efforts compliance with applicable court orders Beyond that, as noted in the Library Bill of Rights, ?Libraries should challenge censorship in the fulfillment of their responsibility to provide information and enlightenment.?
Other requests and grievances, including underlying rights issues, error correction and version control, and re-insertions of web sites based on change of ownership.	These are handled on a case by case basis by the archive and its advisors.

Addendum: An Example Implementation of Robots.txt-based Removal Policy at the Internet Archive

To remove a site from the Wayback Machine, place a robots.txt file at the top level of your site (e.g. www.yourdomain.com/robots.txt) and then submit your site below.

The robots.txt file will do two things:

1. It will remove all documents from your domain from the Wayback Machine.
2. It will tell the Internet Archive?s crawler not to crawl your site in the future.

To exclude the Internet Archive's crawler (and remove documents from the Wayback Machine) while allowing all other robots to crawl your site, your robots.txt file should say:

```
User-agent: ia_archiver
Disallow: /
```

Robots.txt is the most widely used method for controlling the behavior of automated robots on your site (all major robots, including those of Google, Alta Vista, etc. respect these exclusions). It can be used to block access to the whole domain, or any file or directory within. There are a large number of resources for webmasters and site owners describing this method and how to use it. Here are a few:

? http://www.global-positioning.com/robots_text_file/index.html

? http://www.webtoolcentral.com/webmaster/tools/robots_txt_file_generator

? http://pageresource.com/zine/robotstxt.htm

Once you have put a robots.txt file up, submit your site (www.yourdomain.com) on the form on http://pages.alexa.com/help/webmasters/index.html#crawl_site.

The robots.txt file must be placed at the root of your domain (www.yourdomain.com/robots.txt). If you cannot put a robots.txt file up, submit a request to wayback2@archive.org.

For further information, please contact jeff - at - archive - dot - org.