Factors that affect frequency of capture in website archiving.

Adrienne Sonder

March 6, 2003

Library of Congress Election 2002 Collection

Frequency of capture:

"All websites are captured daily unless otherwise indicated. For hotly contested races or to capture breaking news, the Library may require that, for a limited period of time, certain websites be collected several times a day (every four hours) or on an hourly basis, in accordance with [information below]."

Frequency types are *hi, medium, low,* and *one time*. The *low* frequency type has three different possibilities.

Hi = Crawl a single page every 5 minutes to one hour for 24 hours, capturing 0 links out.

Medium = Crawl a website every 4 hours for one week, capturing 1 link level out.

Low = Crawl a website every 24 hours on an ongoing basis, capturing 1 link level out.

Low = Crawl a website every week on an ongoing basis, capturing 1 link level out.

Low = Crawl a website every month on an ongoing basis, capturing 1 link level out.

One time = Capture a website once, capturing a maximum of 1 link level out.

Archival Preservation of Smithsonian Web Resources

http://www.si.edu/archives/archives/dollar%20report.html#Capture

Frequency of capture:

"The time of capture depends upon the scope and frequency of changes to Web sites and HTML pages. With static Web and HTML pages substantive change is likely to be infrequent so an initial base line capture would be sufficient until there is a major redesign of the site and pages, at which time another 'historical' snapshot would be necessary. Between the initial base line capture and the redesign of the site and pages, minor changes, such as monthly event calendars, can be captured and written to a history log change file. The 'historical' snapshot and history log change file should provide adequate documentation of the content of static Web sites and HTML pages.

The capture of dynamic Web sites and HTML pages can be done in two different ways. The first involves capturing 'on the fly' each set of transaction parameters, the derived tables (database 'view'), and the database application software. The second way involves capturing 'on the fly' each set of transaction parameters and the entire database at specified intervals (e.g., daily) along with the database application software. A daily capture of the database presumes that it is updated at regular intervals, say, every day at 12:00 AM. Based upon today's technologies the infrastructure required for the capture of dynamic Web sites and HTML pages is likely to be substantial. Also the storage costs for preserving executable Web sites and HTML pages over time are likely to be equally substantial.

The issue of capturing dynamic SI Web sites and HTML pages is of little consequence because currently virtually all of the SI Web sites and HTML pages (on the order of 95%

or higher) are static. Over the next five years or so this may change as more dynamic Web sites and HTML pages are created but this is not an issue that needs to be addressed now."

MINERVA-Library of Congress

Papers & Presentations:

Which sites to collect?

- ?? Bulk-- select all within a certain category.
- ?? Selective—collect sites selected by a librarian.

How often to make snapshots?

- ?? Monthly, weekly, or depending on circumstances

	Selection	*Frequency*	*Content*
Internet Archive	**Bulk**	**Monthly**	**HTML + images**
Pandora	**Selective**	**Varies**	**All**
Kulturarw3	**Bulk**	**Sweeps**	**All**
Minerva	**Selective**	**Irregular**	**All**

The Library needs a mixed strategy:

1. Selective selection, for known important sites
2. Bulk selection for selected categories (e.g., .gov sites)
3. Bulk collection without selection for other materials

-LC's Web Preservation Project: a pilot for selecting, collecting, cataloging, and accessing archived websites, February 2, 2001 (PowerPoint format)

http://www.loc.gov/minerva/

Election 2002 web archive:

Scope: The scope of the Election 2002 Web Archive includes web sites associated with U.S. 2002 mid-term Congressional elections, gubernatorial elections, and mayoral elections in 15 major U.S. cities.

Collection Period:

July 1, 2002 through November 30, 2002

Acquisition Parameters

Depth - the complete web site, if possible.
Breadth - Follow links out to one external level.
Frequency - Candidate sites daily, all other sites as indicated below.

Web site categories:

- **Primary** - selected races.
- **Candidate** - House; Senate; gubernatorial; major mayoral. Daily.
- **Party** - national level (all registered parties); state level (Democratic and Republican Party only). Weekly for authorized. Monthly for Party committee.
- **Interest Group** - financial contributors as registered by the Federal Election Commission; partisan groups; additional categories from the *Washington Information Directory*. Weekly to daily.
- **Press** - Selective alternative and specialized press sites. Some press sites limited to front page, editorial section and election section if identifiable. Weekly to daily.
- **Government** - House.gov; Senate.gov. Monthly.
  FEC.gov. Weekly
  state and territorial central government web sites. Monthly
  and election boards web sites. Weekly.
- **Civic** - National and state level non-profits with election or voting emphasis. Weekly to daily.
- **Political Portal** - Additional political portals may warrant daily.
  Page specific political portals may warrant daily.
- **Public Opinion** - Gallup.com; Harrisinteractive.com. Weekly to daily.
- **Miscellaneous** - identified by the Library based on recommendations from the Webarchivist.org and Center for Communication and Civic Engagement. Frequency determined upon selection.

Election Day - Approximately 1,800 web pages were captured on an hourly basis.

http://www.loc.gov/minerva/collect/elec2002/select.html

September 11 Web Archive:

"A website is downloaded using a mirror program. A snapshot is stored in an archive. At selected time intervals, additional snapshots are made."

Capture-Internet Archive

- Capture time period: September 11-December 1
- Internet Archive's crawling and indexing

-The September 11 Collection: Archiving an Emerging Web Sphere, Presented to the Association of Internet Researchers, Maastricht, Holland, October, 2002 (PowerPoint format)

http://www.loc.gov/minerva/

PANDORA-National Library of Australia
-an archive of selected Australian online publications, including Web sites
http://www.rlg.org/preserv/diginews/diginews5-2.html#feature1

***** The Australian Pandora project has a policy of selective collection. Librarians select Web sites that have particular interest in Australia and decide the frequency of capture for each site.

[From PANDORA Business Process Model]

http://pandora.nla.gov.au/bpm.html#enab

4.7.1 The gathering schedule

"The gathering schedule, or time interval between each successive capture, must be managed in its own right. The gathering schedule shows the date of capture for each electronic document and should give details of frequency, for example daily, weekly, monthly and the date of last gathering.

The gathering schedule does not necessarily match the date of issue of any electronic publication or its objects. PANDORA will capture publications at intervals appropriate to each publication. For example, a journal issued monthly may be captured quarterly or annually. PANDORA must also provide the capability to capture a publication on a one-off occasion such as the demise of that publication or when the publisher advises the Electronic Unit that it is about to undergo a change in its presentation.

It is anticipated that the reader will access the publisher's site for the latest issue of any journal and for any issues not yet captured by PANDORA. To encourage this, the

publisher's URL will be displayed above PANDORA's in the catalogue record and on the title entry page in the archive.

Publishers will be encouraged to advise the Library of the availability of new versions, changes to publishing intentions (including demise) and other circumstances, such as changes in schedule. When this occurs, it will be necessary to conduct a manual override of the automated gathering schedule.

For some publications, a one-off capture will be necessary. Examples include Commonwealth Government election results, and selected ephemera. Some online newspapers will be captured on a 9-monthly basis, in order to ensure a wider variety of material."

General selection guidelines:

4.4 Topical issues

"4.4.1 More inclusive selection guidelines will be applied to online publications on social and topical issues identified by the Manager, Australian Collection Development, for intensive collecting (eg. Aborigines, the environment, euthanasia, the Olympics, the centenary of federation). The intention is not to duplicate the print collections, but to complement them by providing the broader context.

4.4.2 Sites for selected events or on particular subjects may be sampled during a limited time period and gathered together in a collective entry (eg. election campaigns, Sports 2000, Publishers Sept. 2000-Dec. 2000)"

http://pandora.nla.gov.au/selectionguidelines.html#genGuide

Project Prism-Cornell
http://www.prism.cornell.edu/Default.htm

Anne R. Kenney et al., "Preservation Risk Management for Web Resources: Virtual Remote Control in Cornell's Project Prism," *D-Lib Magazine* 8,1 (January 2002), available at: http://www.dlib.org/dlib/january02/kenney/01kenney.html

[Members of Project Prism make the following recommendations for assessing longevity risk factors of web sites. Web sites that exhibit these factors may require more frequent capture.]

Monitoring a Web Page as a Standalone Object:

?? *Tidiness of HTML formatting*: Just as sloppy work habits reflect badly on an employee, untidy HTML is a reason for some unease about the management of a Web resource. While early versions of HTML had poorly defined structure, the recent redefinition of HTML in the context of XML (XHTML) has now formally defined HTML structure.[55] The TIDY tool makes it possible to determine how well an HTML document conforms to this structure, revealing the sophistication and care of the page's manager.[56]

?? *Standards conformance*: Data format standards, such as the popular JPEG image standard, change over time, sometimes making previous versions unreadable.[57] A monitoring mechanism could automatically determine whether a Web resource conformed to current standards. Conformance to open standards could also be considered. Arguably, Web resources formatted according to a nonpublic standard—for example Microsoft Word documents—may be a greater longevity risk than those formatted to public standards. On the other hand, industry dominance can privilege some proprietary formats over formats that are standard but not widely adopted, e.g., PNG.[58]

?? *Document structure*: Like HTML formatting, a document that manifests good structure, in the manner of a good research paper, may be more dependable than one that consists of text with no apparent order…

?? *Metadata*: The presence or absence of metadata tags conforming to standards such as Dublin Core may indicate the level of management.[60]

Monitoring a Web Page in a Hyperlinked Context:

?? *Out-link structure*: Consider a page that links to a number of pages on the same server, in contrast to another page that either has no out-links or only links to pages on other servers. Intuitively, the "intralinked page" may be more integrated into a site and at lower risk. Pages with no links at all might be considered highly

suspicious, having the appearance of "one-offs" rather than long-term Web resources.

?? *In-link structure*: An equal if not greater indicator of longevity risk is the number of links from other pages to a page and the nature of those links. Isolated pages, ones with no in-links, should be highly suspect. Ascertaining the absence of in-links in the Web context is hard, since it requires crawling the entire Web. Two more tractable and meaningful in-link measurements are:
   - *Intra-site links*—As noted, a page that is integrated into a Web site structure seems more trustworthy than one not pointed to by any pages on its site. It is possible to crawl that Web site—defined by stripping the page URL down to its root dns component—to determine if any page on that site links to the page in question.
   - *Hub links*—Kleinberg's HITS algorithm describes the method for finding authoritative Web resources relative to a specific query.[65] The presence or absence of links to a page from one or more of these authoritative Web resources might be an indicator of risk. In related work, we are developing methods for classifying Web pages automatically in collection categories, each of which is characterized by a set of authoritative pages on the Web. We could then initiate a Web crawl from these authorities and find direct or "transitive" links to a given page.

?? *Page provenance*: The URL of a Web page can itself provide metadata about the page's provenance and management structure. The host name often provides useful information on the identity (the "address") of the Web server hosting a page, and, less reliably, the name of the institution responsible for publishing the page. A top-level domain name can help classify a publishing organization by type (.edu, .gov, .com). Project PRISM will investigate the correlation between top-level domain name and preservation risks.[66] Also, the path name may provide clues about organizational subunits that may be responsible for managing a Web page or site…

?? *Link volatility*: Once the nature of the links to and from a page is determined, it is useful to compare changes in those links over time. If out-links are added or updated, a page is evidently being maintained and is at reduced risk. A decrease in in-links may indicate approaching isolation and should cause concern.

Monitoring a Web Site:

Comprehensive care of a Web site has to include:

?? *Hardware and software environment*, including any upgrades to the operating system and Web server, the installation of security patches, the removal of insecure services, use of firewalls, etc.

- *Administrative procedures*, such as contracting with reputable service providers, renewing domain name registration, etc.

- *Network configuration and maintenance*, including load balancing, traffic management, and usage monitoring.

- *Backup and archiving policies and procedures*, including the choice of backup media, media replacement interval, number of backups made and storage location.

- *Physical location of the server* and its vulnerability to fire, flood, earthquake, electric power anomalies, power interruption, temperature fluctuations, theft, and vandalism.

Cassy Ammen, and Allene Hayes. *RLG DigiNews*, 5(2), April 2001.
http://www.rlg.org/preserv/diginews/diginews5-2.html#feature1

"For bulk collections, the usual practice is to take snapshots of every site with a regular frequency, e.g., monthly. When an archive follows a selective collection policy, it is possible to vary the frequency at which snapshots are collected. For example, a site for a special event might be collected daily during the event, but at less frequent intervals before and afterwards. The daily snapshots captured by the Internet Archive provide a fascinating record of the candidates' tactics in the days leading up to the election and during the Florida recount."

[Above article retrieved from http://www.cs.cornell.edu./wya/LC-web/]

Kulturarw3-National Library of Sweden
http://www.kb.se/kw3/ENG/Statistics.htm

Selection criteria: "We save everything found with the domain name **.se**. We also search for Swedish web sites among such top domain names as: **.org**, **.net** and **.nu** as there are several Swedish site owners among them. Naturally, we run the risk of missing several sites. For instance, we cannot reach educational institutions teaching Swedish language if they are part of an overseas university web server."

[Kulturarw3 does not perform selective archiving, but instead performs webcrawls, which have each lasted between 1-8 months. Each new webcrawl produces a new snapshot of a website. New webcrawls are performed within one month after the previous one has ended.]

Digital Archive for Chinese Studies-DACHS

http://www.sino.uni-heidelberg.de/dachs/intro.htm

Working Routines

"Depending on the material we have developed three different approaches for getting hold of relevant resources:

First of all we try to single out certain 'long term' topics such as China's relationship with the WTO, on which we are actively searching and collecting relevant material of all kind, making use of Internet search engines, newsgroups and mailing lists.

A second important focus are single events such as the September 11th terror attack or the NATO bombing of the Chinese embassy that cause heated discussions on the Internet. To capture such outbreaks of public opinion we are building up a check list of relevant discussion boards, newspapers, and Web sites, which will be worked through each time an important event happens. The result is a set of snapshots of relevant material covering a timespan of a few weeks before and after the event.

In addition to these two main approaches we also randomly collect fragments of public discourse that are believed by our researchers and informants to be of some relevance for current or later research and that neither belong to event related discussions nor pertain to one of our special collection topics.

Depending on these approaches and the kind of material we want to capture, we decide whether to apply regular downloads, irregular snapshots or single non-recurring downloads. Some categories such as single documents etc. clearly belong to non-recurring, complete downloads. On the other hand, discussion boards, some of them growing by hundreds or thousands of postings per day, can only be included in form of snapshots of a few week's discourse.

In the case of complete Web sites that we believe to be of major interest we will ensure automated download in regular intervals with additional downloads whenever we notice important changes or additions. In this we again depend on the help of our 'information network'."

Occasio: Digital Social History Archive

http://www.iisg.nl/occasio/#background

http://www.iisg.nl/occasio/Occasio-uk.PDF

The Digital Social History Archive is an archive of the news messages from the Association for Progressive Communications (APC), an international partnership of communication networks. The messages are sent to the archive as they are generated. Messages arrive to a dedicated server.

National Library of Canada-Electronic collections
http://www.nlc-bnc.ca/9/8/index-e.html

1.4 Versions/Editions

"1.4.1 The National Library does not necessarily collect every version/edition of all networked electronic publications collected. The frequency of capture will vary from comprehensive to representative and will depend on factors such as publication pattern, scope of changes, and the overall significance of the publication."

The Archipol project
http://www.archipol.nl/english/project/projectplan.html

The Archipol project involves the archiving of web sites produced by political parties in the Netherlands Functional description

Archiving standard

Archiving standards can be based on two methods: frequent integral archiving and continued archiving of modifications. The first approach involves downloading entire sites at specified times and the second approach involves copying all the modifications made to a downloaded site and writing them to a log file. It is also possible to use a method that lies between these two extremes.

State Library of Tasmania-Our Digital Island
http://odi.statelibrary.tas.gov.au/About/selpolicy.asp#top

Scope: The State Library collects and preserves Websites which are considered to have permanent value in recording or reflecting cultural, social, political and other processes, activities and achievements in Tasmania and Websites which are considered significant to the historical development of the World Wide Web in Tasmania.

2.5 In these guidelines, the level of commitment given to individual Websites within the context of the Library's specific collecting categories is described using the following terminology and is undertaken with consideration to any limitations imposed by copyright legislation or other practicalities:

2.5.1 *comprehensive* coverage entails capture of all updates or Webpages published in a selected Website, with the depth of coverage extending to all internal Webpages and the scope of coverage extending to primary, secondary and tertiary external Webpages. It is recognized that this level of commitment can be undertaken in relatively few instances;

2.5.2 *selective* coverage entails capture of key updates or Webpages published within a selected Website, with the depth of coverage extending to all internal Webpages and scope of coverage limited to significant primary and secondary external Webpages;

2.5.3 *representative* coverage entails capture of occasional updates or individual Webpages published in a selected Website, with the depth of coverage limited to significant internal Webpages and scope of coverage restricted key primary external Webpages;

2.5.4 *snapshot coverage* entails capture of individual Webpages in depth and scope sufficient only to provide a sample of the Website.

….

3.1 The State Library differentiates in the level of commitment given to the preservation of individual Websites according to each of its specific collecting objectives: to archive government information; to preserve Tasmania's documentary heritage; and to trace the historical development of the World Wide Web in Tasmania.

3.2 State Government Information

3.2.1 State government information which, if it were in print format, would be subject to depository distribution and which has been mandated for Web-based publication is comprehensively archived the State Library. Examples of comprehensive archived State government information include the following types of Websites and URLs:

?? Annual reports of State government departments and agencies, e.g.,

Annual report of the Department of Premier and Cabinet
<http://www.dpac.tas.gov.au/annualreport/1996-1997/index.htm>

Annual report of the Department of Education, Training, Community and Cultural Development
<http://www.tased.edu.au/corpdiv/anreport/96-97.htm>

?? Authorized State legislation, e.g.,

Tasmanian Consolidated Legislation Online
<http://www.thelaw.tas.gov.au>

3.2.2 State government information which is assembled, linked and issued for citizen information or education and which is not the product of government business is selectively archived the State Library. Examples of selectively archived State government information include the following types of Websites (and URLs):

?? State government departments, agencies and services homepages, e.g.,

Justice Tasmania
<http://www.justice.tas.gov.au>

?? Citizen information pages, e.g.,

Parliamentary Research Service Guide to Internet Information
<http://www.parliament.tas.gov.au/prsstaff.htm>

3.2.3. Government information published for public relations purposes is archived on a representative basis by the State Library. Examples of government information which is archived on a representational basis include the following types of Websites and URLs:

?? Event information, e.g.,

Tall Ships Website
<http://www.tdr.tas.gov.au/tallships.nsf>

?? Tourist information, e.g.,

Tullah
<http://www.tased.edu.au/tasonline/tullah>

3.3 Tasmanian Documentary Heritage

3.3.1 Tasmanian Heritage Websites archived on a comprehensive basis are those sites the information content of which is of prime significance. Examples of Tasmanian heritage information which is archived on a comprehensive basis include the following types of Websites and URLs:

3.3.1.1 Web-based periodical publications, e.g.,
The write stuff
<http://www.utas.edu.au/docs/ahugo/tws/index.html>

3.3.1.2 Websites which collectively relate to a single specific event or occasion, e.g.,

?? *1998 State Election*
<http://www.tas.gov.au/elections.htm>

3.3.2 Tasmanian heritage Websites archived on a representive basis are those which demonstrate significance and innovation of their design; and those whose information content is of secondary significance. Examples of Tasmanian heritage information which is archived on a representational basis include the following types of Websites and URLs:

?? Not-for- profit organisations 'databases, e.g.,

Arts Tasmania
< http://www.tased.edu.au/artstas>

3.3.3 Tasmanian heritage Websites archived on an occasional or snapshot basis are those which are subject to little change in content or, conversely, are highly ephemeral in content. Examples of Tasmanian heritage information which is archived on a snapshot basis include the following types of Websites and URLs:

3.3.3.1 Static content

?? *St. Helen's History Room*
http://www.focusontas.net.au/culture/histroom.html

?? *Tasmanian Olympic Council*
http://www.tased.edu.au/tasonline/tassport/toc

3.3.3.2. Ephemeral content, e.g.,

?? *22nd Burnie Cub Scout Pack*
http://www.smallbusiness.net.au/~2nd_burnie/

3.3.3.3 Personal websites. e.g.

?? *Daniel's Homepage*

<http://www.tassie.net.au/~dvella/index.html>

3.3.3.4 Advertising sites, e.g.,

?? *Bass Bakery*

< http://www.ontas.com.au/bass_bakery>

3.3.3.5 Information of secondary significance, such as biographical and genealogical information presented, e.g.,

Grandfather's Grandfather: The story of Daniel Blackwell and his Descendants
<http://www.ozemail.com.au/~kemoon/Danielb.html>

3.4 Historical Development of the World Wide Web

3.4.1 Historical development of the World Wide Web in Tasmania is traced through the archiving, on a highly selective basis, of seminal Websites. Examples of historical or seminal information which is archived on a highly selective basis include the following types of Websites and URLs:

1. Seminal or proto-type Websites, e.g.,

   Tour of Tasmania
   <http://www.tased.edu.au/tot>

3.4.2 Tasmanian Websites which demonstrate unique developmental significance and innovation of design, or whose content is essentially non-Tasmanian, are archived on snapshot basis. Examples of innovative Tasmanian Websites which are archived on a snapshot basis include the following types of Websites and URLs:

3.4.2.1 Highly innovative Websites

Imagine It
<http://www.tasmall.com.au/imagine_it>

3.4.2.2 Websites which use the Web environment as the foundation for works of the artistic imagination e.g.

David McDowell's
<http://toolshed.artschool.utas.edu.au/~mcdowell>

3.4.2.3 Sites which are essentially non-Tasmanian, e.g.,

The Unofficial Susan Ivanova Home Page
<http://www.tassie.net.au/~pmorse>

3.4.2.4 Tasmanian Websites which serve merely to promote non-Web-based publications or which replicate physical-format digital publications are archived on a snapshot basis but monitored to assess their development as publications in their own right. Examples of promotional Tasmanian Websites which are archived on a snapshot basis include the following types of Websites and URLs:

Island Magazine
<http://www.tased.edu.au/tasonline/island/island.html>

Siglo
<http://www.utas.edu.au:80/docs/siglo>

3.5 General Considerations

In selecting any Website for archival treatment, consideration is given the following factors:

3.5.1 Websites and webpages are assessed, archived and preserved on their own merits regardless of National Library or Pandora Project holdings;

3.5.2 Websites are assessed for preservation regardless of the availability of print-based equivalents, though Websites which merely duplicate printed documents might be preserved at a selective or representative level only;

3.5.3 Tasmanian Websites which serve primarily as gateways to non-Tasmanian Websites are not viewed as equally significant as Tasmanian Websites which serve as content sites;

3.5.4 In the organization of the *Our Digital Island* Website, priority is given to comprehensively archived Websites above those Websites which are selectively and representational archived and identified as having incomplete information.