# Digital Library Curriculum Development

# Module 8-b: Web archiving

(Last Updated:2008-11-24)

## 1. **Module Name:** Web archiving

## 2. Scope

This module covers the general ideas, approaches, problems and needs of web archiving to build a static and long term collection consisting of web pages.

## 3. Learning objectives

Students will be able to:

1. Explain the basic concepts and methods related to Web Archiving.
2. Understand and explain simple archiving strategies using some Web Archiving approaches and overcoming some difficulties existing in current web archiving.
3. Understand the necessity (for example, everyday losses of some web pages, etc.) and problems (for example, cultural, economic, and legal) of Web Archiving.

## 4. 5S characteristics of the module

- Stream: Web archiving collect data and ingest into Digital Library
- Structure: The archiving process should follow certain structures according to where the gathering process happens and the organization and storage of archived data should also involve some structures.
- Space: The physical storage for keeping this archived data. Also, the varied distributed locations to ensure no single physical event destroys all copies.
- Scenario: Process of collecting and saving web content
- Society: Individuals, groups and organizations involved in setting policy and in carrying out archiving policies

## 5. Level of effort required:

a. Prior to class: 4 hours for readings
b. In class: 2 hours

## 6. Relationship with other modules

Close connections with:

- 8-a: Preservation. The Web archiving module follows the preservation module. Students should know how web archiving supports the preservation of digital objects.
- 2-c (8-c): File formats, transformation, migration
- 9-e: Intellectual property
- 9-f: Cost/economic issues

## 7. Prerequisite knowledge required

- No prerequisite class is required
- Prior basic knowledge of how computer systems (hardware and software) and internet infrastructure work
- Understanding the fundamental ideas of digital preservation

## 8. Introductory remedial instruction

None

## 9. Body of knowledge

### I. Definition and Problems

A. Why archive the Web?

The Web is growing quickly, adding more than 7 million pages daily. At the same time, it is continuously disappearing. If we do not act to preserve today's Web, key parts of it will disappear.

B. What is to be collected?

The average Web page contains 15 links to other pages or objects and five sourced objects, such as sounds and images. If a Web page is the answer to a user's query, a set of linked Web pages is the answer to a user's query; a set of linked Web pages sufficient to provide an answer must be preserved. From this perspective the Web is like a reference library; that is it is the totality of the reference materials in which a user might search for an answer. The object to be preserved might include everything on the Web on a given subject at a given point in time.

C. Acquisition methods

The term "acquisition" designates the various technical means used to used to get the content into the archive. These methods can be roughly classified into three categories.

a) Client-side archiving

This option involves an archiving crawler or website copier, derived and adapted from search engine technology, providing a powerful tool for capture, in the hands of clients.

**Client-side archiving**

b) Transaction archiving

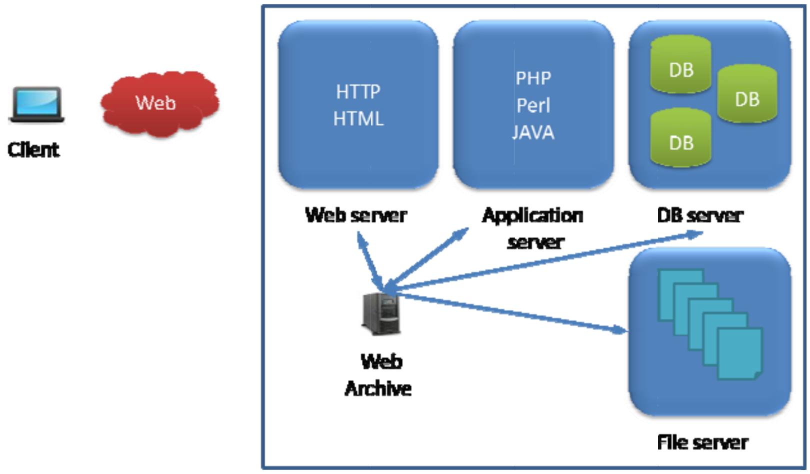
Transaction archiving, consists in capturing and archiving “all materially distinct responses produced by a website, regardless of their content type and how they are produced.”

**Transaction Archiving**

c) Server-side archiving

The last type of acquisition method for Web archives is to directly copy files from the server, without using the HTTP interface at all.

Different pieces of information are obtained directly from servers. Generating a working version of the archived content and not only a back-up of files is the challenge of this method.



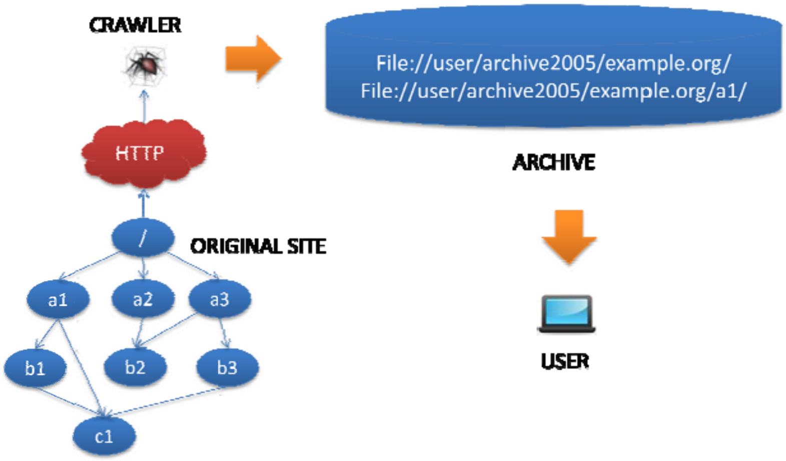
**Server-side archiving**

D. Organization and storage

After making a copy of a Web is a nontrivial task, we need to think about how to organize and store the archived data. Three strategies have been adopted so far for structuring Web archives.

a) Local Files System Served Archives

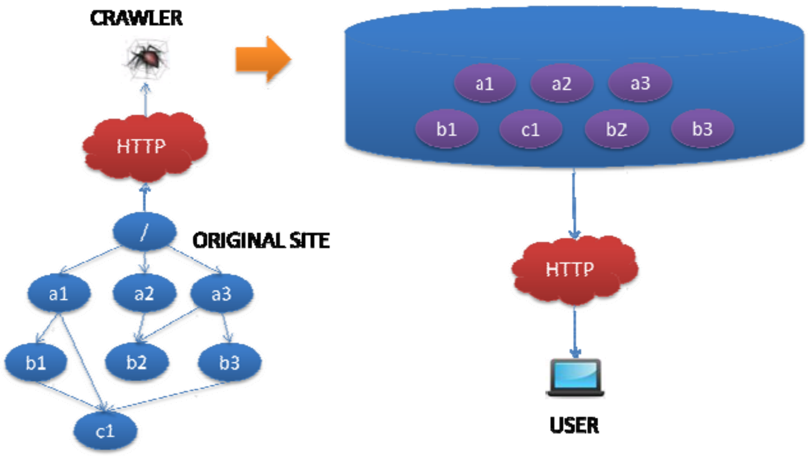
The first strategy is to create a local copy of the site's files and navigate through this copy in a similar way as on the Web.


CRAWLER
HTTP
/
ORIGINAL SITE
a1
a2
a3
b1
b2
b3
c1
File://user/archive2005/example.org/
File://user/archive2005/example.org/a1/
ARCHIVE
USER

**Local Files System Served Archives**

b) Web-Served Archives

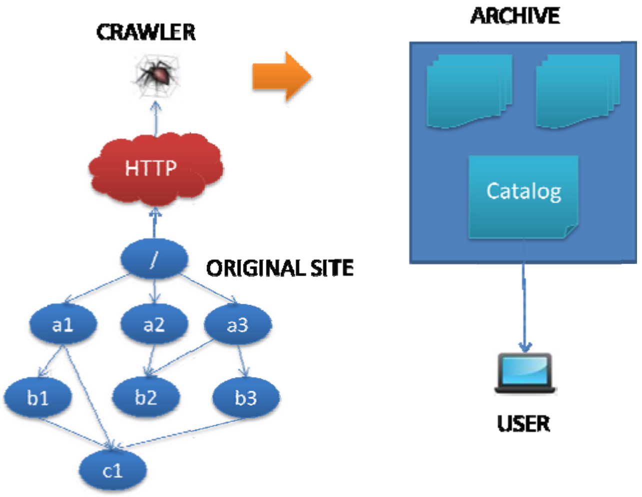
The second one is to run a Web server and serve content in this environment to user's browsers.



**Web-Served Archives**

c) Non-Served Archives

The third option is to re-organize documents according to different (non-Web) logic naming, addressing and rendering.



**Non-Served Archives**

E. Quality and completeness

a) Quality relates to an ideal scale of perfection in a specific area.

b) Completeness can be measured horizontally by the number of relevant entry points found within the designated perimeter and vertically by the number of relevant linked nodes found from this entry point.

F. Scope

a) Site-Centric Archiving

This type of archive, focused on a specific site, is mostly done by and for the creator of the site.

b) Topic-Centric Archiving

Topic Web archiving is becoming more and more popular, often driven by direct research needs. While working on a specific field and its reflection on the Web, many scholars have confronted the ephemeral nature of Web publication, where the lifespan of Web sites is inappropriate for scientific verification as well as for long-lasting referral.

c) Domain-Centric Archiving

Archive building also can be done based on location of content. This characterizes a third type of WA. “Domain” is used here in the network sense of the word or, by extension, in the national sense of the term, which is a combination criterion for targeting sites of a specific country.

## II. Methods and Approaches

- Approaches to Web archiving can be compared along several axes. Their scope, method, and level of quality can be different.

  Julien, M. (2005). Web Archiving Methods and Approaches: A Comparative Study. Library Trends, Vol. 54, No. 1, Summer 2005

  http://muse.jhu.edu/journals/library_trends/v054/54.1masanas.pdf

A. Scope: Web archiving today is either site-, topic-, or domain-centric.

  a) Site-centric: Site Archiving is mostly done by corporate bodies, institutions, or even individuals for limited archiving purposes. It does not entail collection building.

  b) Topic-centric: See I.f.a.2 above.

  c) Domain-centric: This Web archiving is not driven by content but by content location. “Domain” is used here in the network sense of the word or, by extension, in the national sense of the term.

B. Method: Projects also can noticeably differ with respect to the methodological approach they take for discovery, acquisition, and description of content.

  a) Automatic: Automation of these tasks enables a tremendous lowering of the cost per site archived. A crawler can “discover” and download millions of sites through link detection and following.

  b) Manual: Unfortunately, automation reaches some limits, and manual handling must be done in certain cases. Discovery, for instance, can be done manually or automatically. When done manually, it can be a specific activity or a by-product of other activities.

C. Quality: The quality of a Web archive can be defined by:

  a) Being able to render the original form of the site, particularly regarding navigation and interaction with the user.

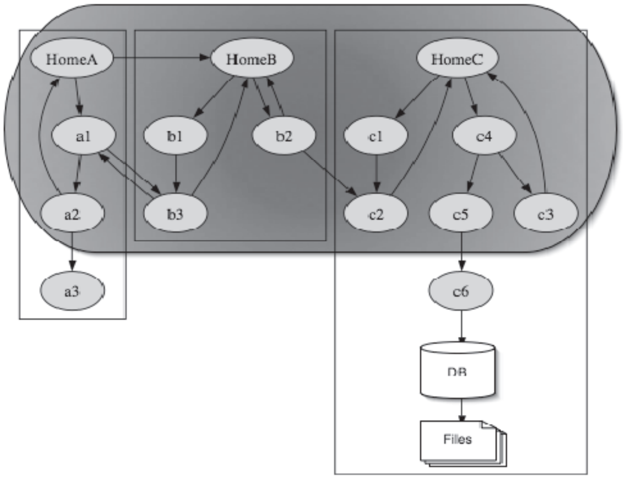
  b) The completeness of material (linked files) archived within a designated perimeter.

    Graphically, completeness can be measured:

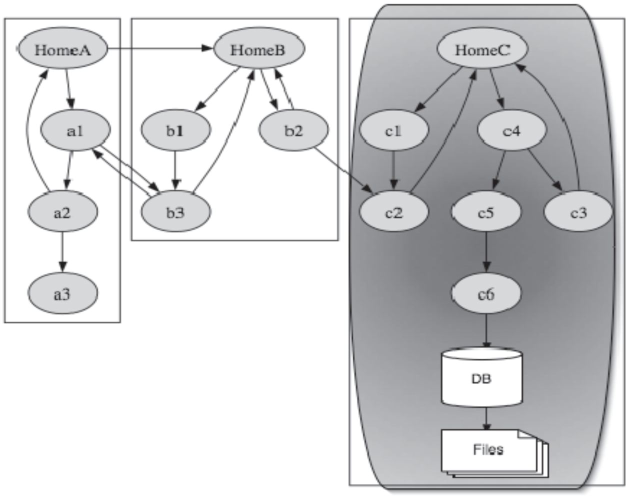
    i. Horizontally: by the number of relevant entry points (site home pages) found within the designated perimeter.

    ii. Vertically: by the number of relevant linked nodes (links can direct the user either to another site or to elements of the same site) found from this entry point.

- Archiving is called “extensive” when horizontal completeness is preferred to vertical completeness. Conversely, archiving is called “intensive” when vertical completeness is preferred to horizontal completeness.



**Extensive Archiving (Shaded Area)**



**Intensive Archiving (Shaded Area)**

## III. Difficulties and limitations

Lyman, P. School of Information Management and Systems University of California, Berkeley (2002). Archiving the World Wide Web. Building a National Strategy for Preservation: Issues in Digital Media Archiving. Council on Library and Information Resources Washington, D.C. and Library of Congress, Page 38-51. http://www.clir.org/PUBS/reports/pub106/pub106.pdf#page=42

A. Cultural problem

In the past, important parts of our cultural heritage have been lost because they were not archived. The hard questions are how much to save, what to save, and how to save it

B. Technical problem

Every new technology takes a few generations to become stable, so we do not think to preserve the hardware and software necessary to read old documents. A Web archive must solve the technical problems facing all digital documents as well as its own unique problems.

a) Information must be continuously collected, since it is so ephemeral.
b) Information on the Web is not discrete; it is linked.

Consequently, the boundaries of the object to be preserved are ambiguous.

C. Economic problem:

a) The economic problem is acute for all archives. Since their mission is to preserve primary documents for centuries, the return on investment is very slow to emerge, and it may be intangible hence hard to measure.

b) Web archives will require a large initial investment for technology, research and development, and training— and must be built to a fairly large scale if used continuously to save the entire Web.

D. Legal problem:

The Web is popularly regarded as a public domain resource, yet it is copyrighted; thus, archivists have no legal right to copy the Web.
Recently it is not preservation that poses an economic threat; it is access to archives that might damage new markets. Finding a balance between preservation and access is the most urgent problem to be solved.

E. Access problem: Access is a political as well as a legal problem.

Interested parties:

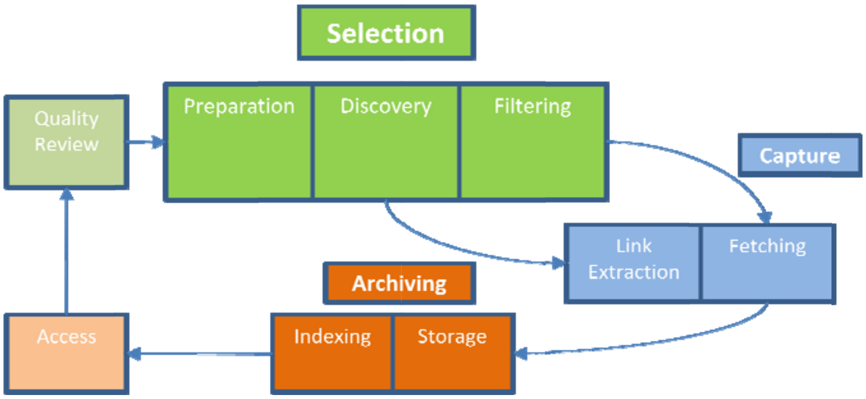
- For librarians and archivists, the key issue is to ensure that historically important parts of the documentary record are preserved for future generations.
- For owners of intellectual property rights, the problem is how to develop new digital information products and to create sustainable markets without losing control of their investments in an Internet that has been optimized for access.
- The constitutional interest is twofold: the innovation policy derived from Article I, Section 8 of the U.S. Constitution (“progress in the useful arts and sciences”), and the First Amendment.
- The citizen’s interest is in access to high-quality, authentic documents, through markets, libraries, and archives.
- Schools and libraries have an interest in educating the next generation of creators of information and knowledge by providing them with access to the documentary record; this means access based on the need to learn rather than on the ability to pay.

When building a Web Archive the problems translate into three questions:

a) What should be collected?
b) How do we preserve its authenticity?
c) How do we build the technology needed to access and preserve it?

## IV. Selection for Web Archives

A. Selection phase



**The selection cycle**
**(Masanes, J., Web Archiving, Fig 3.1., page 71)**

a) The selection phase consists of three sub-phases

i. Preparation

ii. Discovery

iii. Filtering

b) Selection policy determines the type, extent, and quality of the resulting collection. (Simply applying policy for printed materials is not enough.)

B. Selection Policy

A general guiding policy is required for regular collection building. The benefits of having such a policy are the same as for printed material (Biblarz et al. 2001).

i. It reduces personal bias.

ii. It permits planning, identifies gaps in development, and ensures continuity and consistency in selection and revision.

iii. It helps in determining priorities and clarifying the purpose and scope.

a) Target and coverage

The policy has to describe the context, the targeted audience, the type of access, and the expected use of the collection.

b) Limitations

Web archiving encounters many technical difficulties such as the hidden web, streaming content, highly interactive content, etc.

C. Issues and concepts

a) Manual vs. Automatic Selection

The selectivity and the determinism of automatic tools take place at the levels of discovery and capture of material.

b) Location, Time and Content

Web references handle locations first, then objects. The relation between the object and its content depends on the publisher (technical ability and permanent use of resources). The temporal dimension has to become a core component of archiving processes because content update or removal can occur at anytime.

c) Hubs and Targets

i. Hubs contain referral information on other resources (targets).

ii. Targets provide content to be archived.

A hub also can be a target for a selection policy because it may contain content. Hubs are of interest for Web archiving because they can attest to relations like valuation, endorsement, etc., and are means for finding targets.

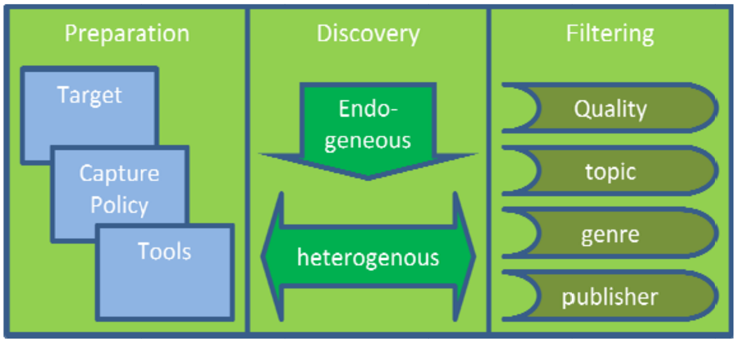
d) Entry point and scope

i. Entry Point (EP), also called 'seed', is defined as the first node from where a path to other documents will be found in a crawling process. Most EPs are usually hubs themselves, but not always.

ii. Scope can be defined as the extent of the desired collection, delimited by a set of criteria. When a new path is found from an EP, it is evaluated to check whether or not it fits in a scope.

iii. Criteria can be topological, thematic, based on genre, time, etc.

D. Selection process



**The phases of the selection process**
**(Masanes, J., Web Archiving, Fig 3.3., page 82)**

a) Preparation

The objective of the preparation phase is to define the target, the capture policy, and implementation tools. This phase is important for the success of the whole process and should not be underestimated because of time and resources required for successful performance.

b) Discovery

The objective of the discovery phase is to determine the list of entry points used for the capture, as well as the frequency and scope of this capture.

i. Endogenous discovery (automatic collection) is made from the exploration of EP's and crawled page's linking environment, while exogenous discovery (manual collection) results from the exploitation of hubs, search engines, and non-Web sources.

ii. Heterogeneous discovery entirely depends on the type, quality and usability of the sources used.

c) Filtering

The objective of the filtering phase is to reduce the space opened by the discovery phase to the limits defined by the selection policy. Filtering can be done either manually or automatically. Manual filtering is necessary when criteria used for the selection cannot be directly interpreted by automatic tools. Evaluation criteria which can be used alone or in combination, for manual selection, are quality, subject, genre, and publisher.

## V. Copying Websites

A. The parsers

a) The HTML Core Parser

The goal of the HTML parser is to scan the page to collect links, analyze them and pass them back to the crawler. It is one of the two core components in a Web copying tool.

a. The simplified core automation: A linear scan of the HTML page data bytes → starting from the beginning → detecting starting tags (<) and recognizing the various HTML elements by their names

b. Two classes of items to recognize inside HTML tags

   i. Tag names (such as ‘img’ or ‘a’)

   ii. Tag properties (such as ‘href’ or ‘src’)

c. Tags can be split into two groups: tags that allows to embed resources (such as images), and tags that allow to navigate to other resources (such as hyperlinks).

b) The Script Parser

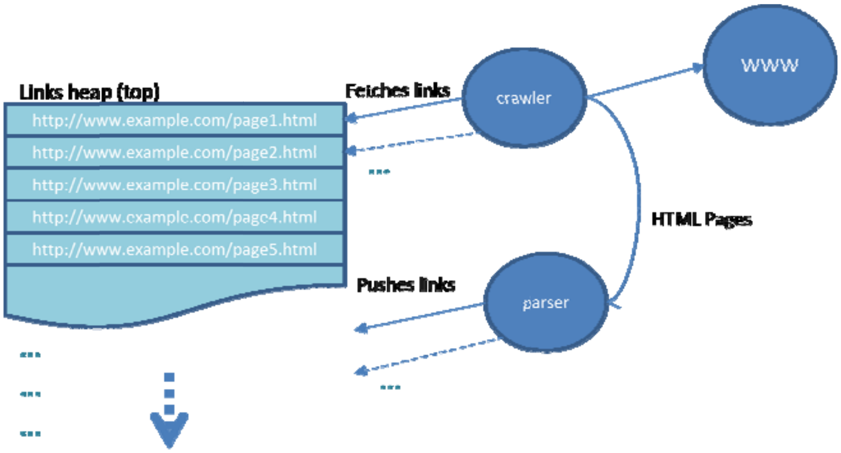
When Web pages are copied, specific scripting zones must be considered inside HTML pages, such as JavaScript. These scripting zones require specific parsing.

c) The Java Classes Parser

Binary formats such as Java classes rely on complex structures that cannot be modified easily. A reasonable strategy to handle Java classes is similar to the JavaScript heuristic, able to handle simple cases, not requiring a perfect binary Java Parser.

B. Fetching Document

One of the engine elements in the copying tool architecture is the robot responsible for gathering data from online websites (HTML pages, images, style sheets, etc.). While the parser is scanning pages, the crawler downloads data using multiple connections, dispatching ready files to the parser.


Links heap (top)
http://www.example.com/page1.html
http://www.example.com/page2.html
http://www.example.com/page3.html
http://www.cxamplc.com/pagc4.html
http://www.example.com/page5.html
Fetches links
crawler
www
HTML Pages
Pushes links
parser

**Crawler/parser interactions**
**(Masanes, J., Web Archiving, Fig 4.3., page 103)**

C. Handling Updates

a) Purposes of making regular copies of preserved websites.

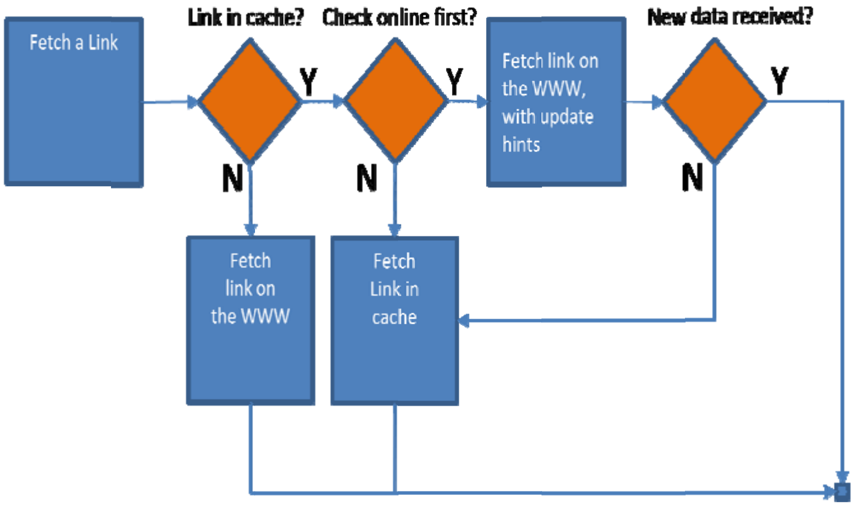
i. To get up-to-date versions

ii. To regularly store a copy that would allow retrieving the site on a specific moment

b) Repeating exactly the same operations is very inefficient, especially with big websites containing large media files. One solution is to mimic browsers by handling a cache used by the crawler to check the freshness of already downloaded data.

i. Use the remote document date to ensure that the resource is always up-to-date.

ii. Use an opaque string aimed to identify specific resource content.



**Caching mechanism**
**(Masanes, J., Web Archiving, Fig 4.4., page 110)**

## VI. Mining Web Collections

### A. Materials for Web Archives

a) Web pages

The first and most important materials for Web archives obviously are the actual Web pages.

Two main features about Web pages are

i. Hypertext

ii. Semi-structured nature

b) Metadata

Metadata – chiefly defined as data about data – also play a pivotal role on the Web and consequently also in Web archives. These background data conveying a wealth of information in addition to the obvious content of the Web objects can be gleaned from the Web or they are produced through the application of mining techniques.

There are two main metadata types of interest with regard to Web mining:

i. Metadata about the object

ii. Technical metadata in context, obtained with the transmission of the object via the Internet

c) Usage Data

The primary source of usage data is server logs. Every time a user sends a request to a Web server, the Web server protocols record that, together with some additional data about the request and where it came from.

d) Infrastructure Data

When data is routed through the Internet it is passed from one local network – also referred to as an autonomous system – to the neighboring unit until it reaches the destination. This data that reflects the Internet infrastructure is the routing tables.

### B. User cases

a) Analyzing Web Content and Web Technology

One important concept in this kind of research is Data Warehouse (DWH).

b) Exploring Web Communities

One research topic is identifying Web communities, with two famous techniques, HITS (hyperlink-induced topic search) and Page rank.

c) Screening Web Users

This kind of research tries to answer questions like what parts of a website a user visited in a single session, and where she spent most time.

d) Researching Networks

Current research on the Internet includes self-organization and fractal growth, graph theory, as well as game-theoretic analyses (Czumaj et al. 2002).

e) Planning Network Infrastructure

Monitoring and analyzing global network traffic to ultimately make networks more robust and efficient

## 10.Resources

**Required Readings:**

Lyman, P. School of Information Management and Systems University of California, Berkeley (2002). Archiving the World Wide Web. Building a National Strategy for Preservation: Issues in Digital Media Archiving. Council on Library and Information Resources Washington, D.C. and Library of Congress, Page 38-51. http://www.clir.org/PUBS/reports/pub106/pub106.pdf#page=42

Masanes, J. (2005). Web Archiving Methods and Approaches: A Comparative Study. Library Trends, Vol. 54, No. 1, Summer 2005
http://muse.jhu.edu/journals/library_trends/v054/54.1masanas.pdf

Masanes, J. (2006). Web archiving: issues and methods. In J. Masanes (Ed.), Web archiving., Berlin Heidelberg New York: Springer, page 1-46
http://www.springer.com/computer/database+management+%26+information+retrieval/book/978-3-540-23338-1

**Recommended readings:**

Bergmark, D., Lagoze, C., & Sbityakov, A. (2002). Focused crawls, tunneling, and digital libraries. Paper presented at the 6th European Conference on Research and Advanced Technology for Digital Libraries, Roma, Italy
http://mercator.comm.nsdlib.org/CollectionBuilding/ECDLpaper.pdf

Biblarz, D., Tarin, M.-J., Vickery, J., & Bakker, T. (2001). Guidelines for a collection development policy using the conspectus model. International Federation of Library Associations and Institutions, Section on Acquisition and Collection Development
http://www.ifla.org/VII/s14/nd1/gcdp-e.pdf

Crowston, K. & Williams, M., "The Effects of Linking on Genres of Web Documents," hicss,pp.2006, Thirty-Second Annual Hawaii International Conference on System Sciences-Volume 2, 1999
http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.42.3030

Masanes, J. (2002). Towards continuous Web archiving: First results and an agenda for the future. D-Lib Magazine, 8(12)
http://www.dlib.org/dlib/december02/masanes/12masanes.html

Masanes, J. (2006). Selection for Web Archives. In J. Masanes (Ed.), Web archiving., Berlin Heidelberg New York: Springer, page 71-90

National Library of Australia. (2005). Online Australian publications: Selection guidelines for archiving and preservation by the National Library of Australia
http://pandora.nla.gov.au/selectionguidelines.html

Roche, X. (2006). Copying websites. In J. Masanes (Ed.), Web archiving. Berlin Heidelberg New York: Springer, page 93-112

## 11. Concept map

## 12. Exercises/Learning activities

a) Discussion activity (15 minutes): Why do we need to archive web pages and what should be collected?

b) In class, break students into groups of 3~4. Have them discuss Web archiving regarding. (15 minutes):
   a. Personal perspective
   b. Commercial perspective
   c. Academic perspective
   d. Social impacts
   e. Technological impacts

c) Web Archiving faces five main problems (see the body of knowledge); each group should discuss these problems and give some suggestions (15 minutes).

d) Return to a full-class discussion and ask the students how they answered the above questions. Each group prepares a Word document or PowerPoint slides (5 minutes) and presents it to the class (10 minutes).

## 13. Evaluation of learning achievement

In their answers to the discussion questions, students demonstrate an understanding of

- Different Web Archiving perspectives
- Different Web Archiving problems and limitations
- Why Web Archiving is needed

## 14. Glossary

**Entry Point (EP) -** the first node from where path to other documents will be found in a crawling process.

**Extensive Archiving -** when horizontal completeness is preferred to vertical completeness.

**Horizontal completeness -** the number of relevant entry points (site home pages) found within the designated perimeter.

**Intensive Archiving -** when vertical completeness is preferred to horizontal completeness.

**Vertical completeness -** the number of relevant linked nodes found from this entry point.

## 15. Additional useful links

Internet Archive: www.archive.org
Dublin Core: www.dublincore.org
World Wide Web Consortium: www.w3c.org

## 16. Contributors

| | |
|---|---|
| Authors: | Spencer Lee, |
| | Tarek Kan’an, |
| | Jian Jiao |
| | |
| Evaluators: | Dr. Fox |