netpreserve.org
international internet preservation consortium

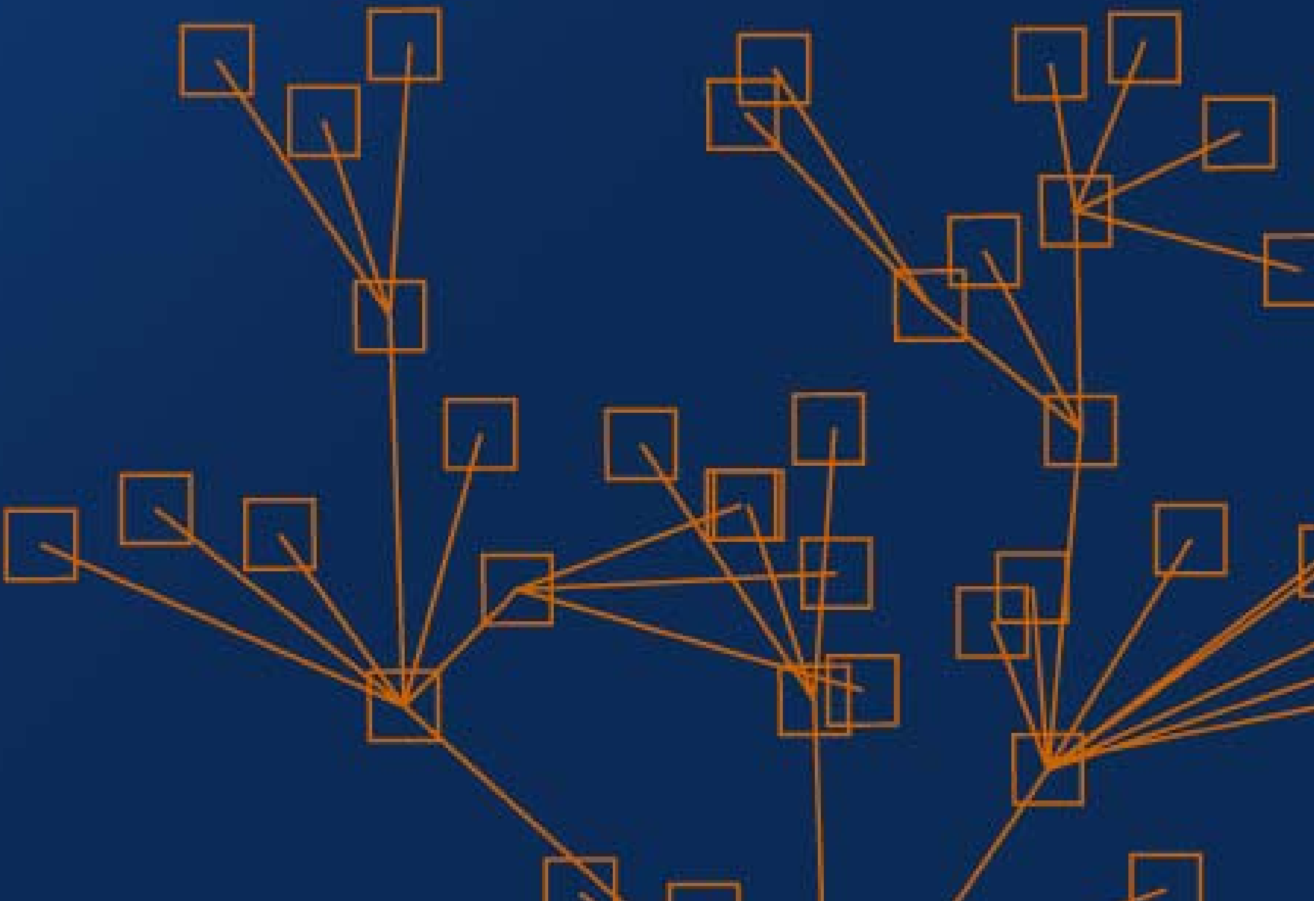
Web Harvesting Survey

Jennifer Marill

Andrew Boyko

Michael Ashenfelder

Version 1



Jennfer Marill
Andrew Boyko
Michael Ashenfelder
Martha Anderson
Gina Jones

The Library of Congress
101 Independence Ave, SE
Washington, DC 20540
USA

Web Harvesting Survey

This survey is an attempt to identify and classify many of the conditions found on Web sites that influence the harvesting of content and the quality of an archival crawl. This table is based on Ketil Albertsen's report, "A taxonomy for the 'the deep web'," and on discussions of the Library of Congress Web harvesting team (LCWHT). The LCWHT departs from the notion of identifying conditions as Deep vs. Surface Web; rather, we see these conditions in a continuum. Further, the harvestability of these conditions will shift in this continuum as the crawling tools mature.

There are three distinct phases involved in harvesting content:

Acquisition - Retrieving content from its home server and storing it–and its metadata–locally

Parsing - Correctly interpreting the format of acquired files and extracting any links to other retrievable content

Presentation - Rendering harvested content to be viewed within the archive in its original form

Each phase is a distinct process with its own issues, and the degree to which a given type of content may have problems in one phase does not necessarily affect its outcome in another phase (though content that cannot be acquired also cannot be parsed or presented).

Given the current state of harvesting tools, we have rated the harvestability of a given Web site according to the following values:

Easy – Tools can harvest content now

Difficult – Tools may or may not be able to harvest content now. Tools will require additional analysis and modification.

Future – Tools cannot harvest content now. Tools require significant additional development that may not be worthwhile.

NOTE: We have not supplied examples for every listed condition, only examples that illustrate the difficulty or uniqueness of certain conditions.

Classification	Condition / Issue	Reference / Notes	Example	Harvesting:
Static HTML documents	Stand alone files: HTML, GIF, JPEG			Acquisition: Easy Parsing: Easy Presentation: Easy
Alternate content types	Flash, PDF, XML (RSS, RDF), MS Office formats, Java applets	The files can be harvested, but embedded navigation links and other functionality may not be. May require plug-ins.		Acquisition: Easy Parsing: Difficult Presentation: Future
Forms	Drop-down navigation form box	There are cases when this information is not properly harvested, and results in cgi errors on the harvested site	Select a state from the "Find Your Senators" drop-down menu at http://web.archive.org/web/20030603195808/www.senate.gov/, Appropriate relative links to senators got harvested. Select a representative from the drop-down menu at http://web.archive.org/web/20030601130533/http://www.house.gov/. Links lead to live sites.	Acquisition: Easy Parsing: Difficult Presentation: Difficult

Classification	Condition / Issue	Reference / Notes	Example	Harvesting:
Forms	Content acquired from search or query box	Though the Web page that contains the form is harvestable, data obtained through the form is not. In a search or query operation, user dialogs with web server and database by entering data into form fields. It is not possible for a harvester to programmatically determine all possible values to input into form fields. In a harvested form disconnected from the source database, users will get an error message when searching or querying. Occasionally, the archived form may continue to link to a live server, which may give the appearance of displaying resulting data from within archives. Complete harvesting of all data may require: 1) Source data-base 2) Server applica-tion 3) Possible emula-tion Acquisition of a large dataset may require a capture strategy of "push" or "pull." NOTE: If the same information retrieved via the query boxes is also presented via hard-coded links, the information can be harvested.	Entering "finance" in the search field at the top of http://web.archive.org/web/20030601233155/www.house.gov/house/CommitteeWWW.html generates results from the live site. A sample Library of Congress hard-coded database link is on http://memory.loc.gov/cgi-bin/ampage?collId=mgw2&fileName=gwpage018.db&recNum=123 For other hard-coded database links, In the right column, on http://web.archive.org/web/20030610084037/appropriations.senate.gov/, click items listed under Latest News.	Acquisition: Future Parsing: Future Presentation: Future

Classification	Condition / Issue	Reference / Notes	Example	Harvesting:
JavaScript	Drop-down navigation menus	Resulting content may be from the live site and not the archived site, depending on the JavaScript implementation.	Harvested drop-down menus work incorrectly at http://web.archive.org/web/20030202033054/bayh.senate.gov/index1.html	Acquisition: Difficult Parsing: Difficult Presentation: Difficult
JavaScript	Content opening in new browser window	Resulting content may be from the live site and not the archived site, depending on the JavaScript implementation.	In the column on the right side of Web page at http://web.archive.org/web/20030513062256/www.senate.gov/pagelayout/reference/b_three_sections_with_teasers/reference_home.htm, click on "Biographical Directory." Content displays from live site.	Acquisition: Difficult Parsing: Difficult Presentation: Difficult
JavaScript	Current date or other live information displaying	There may be files with executable code displaying the current date. This code needs to be modified to reflect the circumstances under which the harvesting was done.	Archived page at http://web.archive.org/web/20030621143030/http://www.billnelson.senate.gov/index.html displays current date, but page was archived in 2003. JavaScript is reading system clock.	Acquisition: Future Parsing: Easy Presentation: Future
JavaScript	Voice application	Proprietary voice software that requires no plug-in.	"Talking Web" at http://nihseniorhealth.gov/. Click "Turn Speech On" at top of page.	Acquisition: Future Parsing: Future Presentation: Future

Classification	Condition / Issue	Reference / Notes	Example	Harvesting:
JavaScript – client-side	URLs generated by dynamic mechanisms	Scripts are activated by manual user actions – e.g. buttons or menu selections – and URL is assembled from smaller pieces. URLs may also be hidden in data elements handled by plug-ins, such as Flash movies. A variant (JavaScript applets) is when text is retrieved as a database object, and links to other objects in the database are embedded in the text as dynamically generated, temporary URLs. 1) It is not currently possible for the harvester to determine all possible URLs 2) Harvester cannot determine if an object is already harvested or not, as the URL will be different in each session 3) Requires plug-ins (e.g., where URLs are encoded in Flash)		Acquisition: NA Parsing: Future Presentation: Future
Non-streaming Media	Direct URL link to audio or video file	May require plug-ins	Click an mp3 link on http://web.archive.org/web/20040220080416/http://billycoopersmusic.com/cd.htm. The next displayed page is http://web.archive.org/web/20040220082113/http://billycoopersmusic.com/au/cd/healinghands.mp3	Acquisition: Easy Parsing: Easy Presentation: Future

Classification	Condition / Issue	Reference / Notes	Example	Harvesting:
Streaming Media	Indirect linkage of URLs. Plug-in specific.	Includes Real Audio, QuickTime and Windows Media. May require plug-ins	Senator Chris Dodd's archived Web page at http://web.archive.org/web/20030714230032/www.dodd.senate.gov/multimedia/welcome-jump.html contains the link/URL to a video file on a live server. Though the page itself is archived, the video link is to a live server. Similarly, Sentator Edward Kennedy's archived page http://web.archive.org/web/20030625163311/www.kennedy.senate.gov/cspan.html links to the live CSPAN feed.	Acquisition: Future Parsing: Future Presentation: Future
Streaming Media	Web cameras and radio channels displaying real-time data 24 hours a day, 365 days a year.	Archiving not realistic. Consider storing samples taken at intervals.	http://web.archive.org/web/20030602174158/www.wwoz.org/live_broadcast_stream_wm.html is an archived Web page calling up a live radio broadcast.	Acquisition: Future Parsing: Future Presentation: Future
Password required	Always same password required	It is not currently possible for the harvester to determine all possible user ID and password combinations	User ID and password at http://web.archive.org/web/20020827114707/www.bankofamerica.com/index.cfm	Acquisition: Difficult Parsing: Difficult Presentation: Difficult

Classification	Condition / Issue	Reference / Notes	Example	Harvesting:
Password required	Password randomly generated, encrypted, and/or displayed as a graphic	It is not currently possible for the harvester to determine all possible user password combina-tions, decrypt encrypted passwords, or read and implement passwords embedded and displayed within graphics.	During the process of purchasing a ticket at http://www.ticketmaster.com a randomly selected graphic password is generated for the user.	Acquisition: Future Parsing: Future Presentation: Future
Encoding	Sites with non-Western character sets	Current crawlers cannot yet read all possible character sets (charset) and therefore fail to detect – and follow – links in those encoded documents.		Acquisition: Easy Parsing: Difficult Presentation: Easy
Server-side scripts	Dynamically generated pages	Several separate files may be combined and assembled by server, and displayed as a single document. Includes PHP, ASP, and Cold Fusion.	On the right side of the Web page at http://web.archive.org/web/20030621143030/http://www.billnelson.senate.gov/index.html, click “Perspective” or any similar title in that section. Subsequent Cold Fusion page may be assembled from separate parts and displayed as a whole.	Acquisition: Easy Parsing: Easy Presentation: Easy
Server-side scripts	Dynamically displayed information	Date and time, “Page hit” counters	Visitor counter on lower right corner of http://web.archive.org/web/20030618155415/http://lancearmstrong.com/	Acquisition: Easy Parsing: Easy Presentation: Easy

Classification	Condition / Issue	Reference / Notes	Example	Harvesting:
Proprietary software	Documents displaying a window into a larger document	Includes geographical maps, enabling user to drag around and zoom-in on sections. Map view is generated on the fly from a database. Requires all source files, plug-ins, underlying model, appropriate software, and possible emulation.	US National Map Viewer at http://nmviewogc.cr.usgs.gov/viewer.htm cannot be easily archived, as shown by going to the archived version of the page at http://web.archive.org/web/20030204083341/nationalmap.usgs.gov/nmjump.html and clicking "Go to the National Map Viewer" button.	Acquisition: Future Parsing: Future Presentation: Future
Cookies	Sites that provide conditional content based on user information	Cookies gather and track user information. Harvester itself doesn't supply any useful user information. Therefore some aspects of the site that are based on user input might not display and be harvested.	A programmatic harvest of http://www.amazon.com cannot add personalized features, via a cookie, as can a user browsing the site for specific things.	Acquisition: Difficult Parsing: NA Presentation: NA
Cookies	Sites that provide conditional content based on link followed	Site can provide different content for the same URL with different cookie set		Acquisition: Difficult Parsing: NA Presentation: NA

Classification	Condition / Issue	Reference / Notes	Example	Harvesting:
Database requests	Certain database data unavailable to harvester	A portion of company information may be displayed on their Web pages, based on a handful of attributes in a few tables. Other tables, unrelated to the published data, may exist in the same database but not be displayed. Requires: 1) Database 2) Application making selection and preserving data 3) Filter tables and applications not required to run applications 4) Knowledge of database schema 5) Possible emulation		Acquisition: Future Parsing: Future Presentation: Future

Other Elements to Note:

Classification	Condition(s) / Issue	Reference / Notes	Example
"No Archiving" mechanism	- robots.txt files on web servers - X-No-Archive for NNTP entries and SMTP messages - META ROBOTS in HTML files	Harvest is policy dependent.	
URL links	Harvester redirection and 404s	A large number of URLs may direct a crawler to a single target page, increasing a publisher's search engine ranking.	
External		Verifies that user has a	

Classification	Condition(s) / Issue	Reference / Notes	Example
authentication server		right to access a document. Requires user password and deposit of database or set of files. 1) Data may be integrated with authentication server 2) May require authentication server and emulation	