For more information about this spreadsheet, please visit https://github.com/archivers-space/research/tree/master/web_archiving

	General information					Users & developer features						"Out of the box" website crawling capabilities										Advanced data harvesting capabilities				Archive management features			Other notes		Date/version checked	
Name (link to home page)	*Open source?*	*Source repo*	*Operating system(s)*	*Primary dev. language*	*Target audience*	*CLI*	*GUI*	*WUI*	*Library API*	*Network API*	*Extensibility framework*	*Parallel crawling*	*Scheduled crawling*	*Crawl storage format(s)*	*Capture raw responses*	*Follow links*	*URL filtering*	*Advanced filtering*	*Extract links from JavaScript*	*Run JavaScript*	*Handle React*	*Extract links from Flash*	*Run Flash*	*Targeted scraping*	*Manual form interaction*	*Auto form interaction*	*Extract file data*	*Browse*	*Playback*	*Full-text search*	*Notable users*	*Notes and comments*	*Evaluation date*	*Version examined*
Whole site archiving systems																																	
Archive-It	½	n/a	web app	Java	enterprise	✖	✖	✔	✖	✔	✖	✔	✔	ARC, WARC		✔	✔		✔	✔	✖	✔	✖		✖	✖		✔	✔	✔	Run by IA. Used by 100's of institutions.	Paid service. Core software is Heritrix, with some extensions.	2017-05-21	
Brozzler	Apache	GitHub	Lin, Win, macOS	Python	user	✔	✖	✔	✖	✔	✖	✔	✖	WARC	✔	✔	✖	✖	✔	✔	✔	✖	✖	✖	✖	✖	✖	✔	✔	✖	IA	Uses Chromium to fetch pages. Uses warcprox and pywb.	2017-07-14	1.b11
Crawler	✔	GitHub	Lin, Win, macOS	PHP	user	✖	✖	✔	✖	✖	✖	✖	✖	MySQL	✖	✔	✖		✖	✖	✖	✖	✖	✖	✖	✖	✖	✖	✖	✖	Made by the FCC.	Bare bones crawler. License not stipulate. Purpose unclear.	2017-06-01	2012-06-04
Crawler4j	Apache	GitHub	Lin, Win, macOS	Java	user	✖	✖	✖	✔	✖	Java classes	✔	✖	files on disk	✖	✔	✖	✖	✖	✖	✖	✖	✖	✖	✖	✖	✖	✖	✖			2017-05-23	
Crawljax	Apache	GitHub	Lin, Win, macOS	Java	user	✔	✖	✔	✔	✖	Plug-ins	✔	limited	log file; plug-ins can do more	✖	✔	✖	✔	✔	✔	✔	✖	✖	✔	✔	✖		✔	✖	✖		Several papers written about the implementation.	2017-06-01	3.5
grab-site	MIT	GitHub	Lin, Win, macOS	Python	user	✔	✖	✔	✖	RPC	Python	✖	✖	WARC		✔	✔		✔	✖	✖	✖	✖		✖	✖	✖	✖	✖	✖		Uses wpull internally.	2017-05-15	
gecco	MIT	GitHub	Lin, Win, macOS	Java	user	✖	✖	✖	✔	✖	Java classes	✔	✖		✖	✔	✖		✔	✔	✔	✖	✖	✔	✔	✖	✖	✖	✖	✖		Code comments are in Chinese.	2017-06-01	
Heritrix	Apache+	GitHub	Lin	Java	user	✔	✖	✔	✖	JMX	Java classes	✔	✖	ARC, WARC	✔	✔	✔		✔	✖	✖	✔	✖	✖	✖	✖		✖	✖	✖	Used by IA.		2017-05-15	
HTTrack	GPL	GitHub	Lin, Win	C	user	✔	✔	✖	✔	✖	C callbacks	✔	✖	files on disk	✖	✔	✔	✔	✔	✖	✖	✖	✖	✖	✖	✖	✖	✖	✖	✖			2017-07-03	
ItSucks	GPL	SF	Lin, Win, macOS	Java	user	✔	✔	✖	✔	✖	Java classes	✔	✖	files on disk	✖	✔	✔	✔	✖	✖	✖	✖	✖		✖	✖		✖	✖	✖			2017-05-15	
NetarchiveSuite	LGPL	GitHub	Lin	Java	user, enterprise	✔	✖	✔	✖		Java classes	✔	✔	ARC, WARC	✔	✔	✔		✔	✖	✖	✖	✖	✖	✖	✖		✔		✔	Netarkivet at The Royal Library of Denmark	Uses Heritrix for crawling.	2017-07-14	
Nutch	Apache	Apache	Lin, Win, macOS	Java	user, enterprise	✔	✖	✖	✔	✔	Plug-ins	✔	✔	several db options	✖	✔	✔	✔	✔	✔	✔	✖	✖	✔	✔	✖	✔	✖	✖	✖			2017-06-01	
Octoparse	✖	✖	Win	.NET	user, enterprise	✖	✔	✖	✖	REST	✖	✔	✔	database, CSV, Excel, files on disk	✔	✔	✔	✔	✔	✔	✔	✖	✖	✔	✔		✔	✔				Seems to run in the cloud, but there's a downloadable console or something.	2017-09-03	6.4.3
PageFreezer	✖	✖	web app	n/a	user, enterprise	✖	✖	✔	✖	✔	✖	✔	✔	web pages	✖	✔						✖	✖		✖	✖	✖	✔	✔		EDGI web monitors use it		2017-10-04	
simplecrawler	BSD	GitHub	Lin, Win, macOS	Node.js	user	✔	✖	✖	✔	✖	Node modules, ES6+	✔	✖	files on disk	✔	✔	✔	✔	✔	✖	✖	✖	✖	✖	✔	✖	✖	✖	✖	✖			2017-09-03	1.1.5
Squidwarc	GPLv3	Github	Lin, macOS	Node.js	user, enterprise	✔	✖	✖	✔	✖	Node modules, ES6+	✔	✖	WARC	✔	✔	✖	✖	✔	✔	✔	✖	✖	✖	✖	✖	✖	✖	✖	✖		A high fidelity archival crawler that uses Chrome or Chrome Headless	2017-07-21	d4ca0b8
StormCrawler	Apache	GitHub	Lin, Win, macOS	Java	user, enterprise	✔	✖	✖	✔	✖	Java classes	✔	limited	several db options	✖	✔	✔	✔	✔	✔	✔	✖	✖	✔	✔	✖	✔	✖	✖	✖	Several companies, apparently.	https://github.com/DigitalPebble/storm-crawler/wiki/Presentations	2017-06-01	
WAIL (Electron)	GPLv3	GitHub	Lin, Win, macOS	Node.js (Electron)	user	✖	✔	✖	✔	✖	✖	✔	✔	WARC	✔	✔	✔	✖	✔	✔	✔	✖	✖	✖	✖	✖	✔	✔	✖		Uses Chrome Browser and Heritrix for crawling. Pywb, Twitter Monitoring And Automatic Archival	2018-12-13	1.2.0-Beta2
WAIL (py)	MIT	GitHub	Win, macOS	Python	user	✖	✔	-	✖	✖	✖	✔	✖	WARC	✔	✔	✖	✖	✖	✖	✖	✖	✖	✖	✖	✖	✔	✔	✖		OpenWayback and Heritrix, cf. WAIL (Electron)	2018-12-13	v0.2016.07.09
WebMagic	Apache	GitHub	Lin, Win, macOS	Java	user	✖	✖	✖	✔	✖	Java		✖	files on disk		✔	✔	✔	✔	✔	✖	✖	✖	✔	✔	✖	✔	✖	✖	✖		Intended as a programming framework, not end-user app. Comments are in Chinese	2017-10-04	0.7.3
WebRecorder.io	Apache	GitHub	Lin, Win, macOS	Python	user	✔	✔	✔	✔	✖	✖	✖	✖	WARC	✔		✖	✖	✔	✔	✖	✖	✖	✖	✔	✖	✖	✔	✔		Interactive, high-fidelity web archiving tool	2017-07-03	eccea96
wget	GPL	Savannah	Lin, Win, macOS	C	user	✔	✖	✖	✖	✖	✖	✖	✖	WARC, files on disk	✔	✔	✔	limited	✖	✖	✖	✖	✖	✖	✖	✖	✖	✖	✖	✖		Use option --save-headers to save HTTP headers	2017-05-17	1.19
wpull	GPL	GitHub	Lin, Win, macOS	Python	user	✔	✖	✖	✔	✔	plug-ins, scripts	✖	✖	WARC	✔	✔	✔		✔	✖	✖	✖	✖	✖	✖	✖	✖	✖	✖	✖			2017-05-20	2.01
Single page snapshot/archiving systems																																	
Archive.is	✖	n/a	web app	n/a	user	✖	✖	✔	✖	REST	✖	✔	✖	web page	✖	✖	✖	✖	✖	✔		✖	✔	✖	limited	✖	✖	✖	✔	✖		Very good quality page captures.	2017-05-15	
curl	MIT	GitHub	Lin, Win, macOS	C	user	✔	✖	✖	✔	✖	✖	✖	✖	files on disk	✔	✖	✔	✖	✖	✖	✖	✖	✖	✖	limited	✖	✖	✖	✖	✖	Standard with many operating systems		2017-05-21	
FreezePage	✖	n/a	web app	n/a	user	✖	✖	✔	✖	✖	✖	✖	✖	web page	✖	✖	✖	✖	✖	✔	✖	✖	✖	✖	✖	✖	✖	✔	✔	✖		Seems to be free for use, but not open source.	2017-06-25	
Paparazzi	✖	n/a	macOS	n/a	user	✔	✔	✖	✖	✖	✖	✖	✖	PDF, PNG, JPG, TIFF	✖	✖	✖	✖	✖	✔	✖	✖	✔	✖	limited	✖	✖	✖	✖	✖	mhucka uses this all the time	Very good quality full-page captures.	2017-05-24	
Perma.cc	MIT + GPL	GitHub	web app	Python	user, enterprise	✖	✖	✔	✖	REST	Django	✖	✖	WARC, PDF, PNG	✖	✖	✖	✖	✖	✔		✖	✖	✖	✖	✖	✖	✔	limited	✖			2017-05-24	
WARCreate	MIT	GitHub	Chrome extension	JavaScript	user	✖	✔	✖	✖	✖	✖	✖	✖	WARC	✖	✖	✖	✖	✖	✔	✖	✖	✖	✖	✖	✖	✖	✖	✖	✖			2017-07-14	
webkit2png	MIT	GitHub	Lin, Win, macOS	Python	user	✔	✖	✖	✖	✖	✖	✖	✖	PNG	✖	✖	✖	✖	✖	✔	✖	✖	✔	✖	✖	✖	✖	✖	✖	✖			2017-05-27	
Data scraping systems																																	
Bixo	Apache	GitHub	Lin, Win, macOS, EMR	Java	user, enterprise	✔	✖	✖	✔	✖	Java classes	✔	✖	files	✖	✔	✔		✖	✖	✖	✖	✖	✖		✖	✖	✖	✖	✖		Author seems to be creator of Krugle. Uses Apache Nutch, Hadoop, Tika, others.	2017-06-25	
import.io	✖	n/a	web app	n/a	user, enterprise	✖	✖	✔	✖	REST	✖	✔	✔	JSON, CSV, Gdocs, Tableau	✖	✔	✔		✔	limited		✖	✖	✔	✔	✖		✖	✖	✖		Offers edu & charity discounts. Has a pretty active user forum.	2017-05-24	
iRobotSoft.com	✖	✖	Win	n/a	user, enterprise	✔	✔	✖	✖	✖	✖	✔	✔	files on disk, user db	✖	✔	✔	✖	✔	✔		✖	✖	✔	✔	✖	✔	✖	✖			2017-10-04	2.8.2
morph.io	Affero	GitHub	Lin, Win, macOS, Docker, web app	Ruby	user, enterprise	✔	✖	✔	✔	✔	Plug-ins (supports many langs)	✔	limited	SQLite	✖	✔	✔		✔	✔		✖	✖	✔	✔	✖		✖	✖	✖		Offers cloud-based scraping. Active user forum.	2017-05-29	
Portia	BSD	GitHub	Lin, Win, macOS, Docker	Python	user	✖	✖	✔	✔	✔	Python		✔	files, MySQL, git	✖	✔	✔		✔	✔	✔	✖	✔	✔	✔	✖		✔	✖	✖		Has visual scraping definition editor.	2017-05-29	
WebScraper.io (fork)	LGPLv3	GitHub	Chrome extension	JavaScript	user	✖	✖	✔	✖	✖	✖	✖	✖	CSV, CouchDB	✖	✔	✔	✖	✖			✖	✖	✔		✖	✔	✖	✖	✖	A version of this is used by WebScraper.io (commercial)	This is a fork; the original (by Martins Balodis) has not changed since 2014.	2017-10-04	0.3.1
Web Scraper Plus+	✖	n/a	Win	n/a	user	✔	✔	✖	✖	✖	✖	✖	✖	files, user db	✖	✔	✔	✔	✔	✖	✖	✖	✖	✔	✔	✖	✔	✖	✖	✖		Can extract data & store in database. May no longer be supported.	2017-05-23	
– not yet categorized –																																	

For more information about this spreadsheet, please visit https://github.com/archivers-space/research/tree/master/web_archiving

	General information					Users & developer features						"Out of the box" website crawling capabilities											Advanced data harvesting capabilities				Archive management features			Other notes		Date/version checked	
Name (link to home page)	*Open source?*	*Source repo*	*Operating system(s)*	*Primary dev. language*	*Target audience*	*CLI*	*GUI*	*WUI*	*Library API*	*Network API*	*Extensibility framework*	*Parallel crawling*	*Scheduled crawling*	*Crawl storage format(s)*	*Capture raw responses*	*Follow links*	*URL filtering*	*Advanced filtering*	*Extract links from JavaScript*	*Run JavaScript*	*Handle React*	*Extract links from Flash*	*Run Flash*	*Targeted scraping*	*Manual form interaction*	*Auto form interaction*	*Extract file data*	*Browse*	*Playback*	*Full-text search*	*Notable users*	*Notes and comments*	*Evaluation date*	*Version examined*
80legs.com																																
Abot																																
AbotX																																
Andjing																																
Anemone																																
Aperture																																
Apifier																																
Arachnid (Java)																																
Arachnid (PHP)																																
arachniweb																																
Arale																																
ArchiveBot																																
ARCOMEM																																
ASPseek																																
Bingol																																
blekko	✖	–																														
CCBot																																
cl-web-crawler																																
CrawlBot	✖	–	web app							✔																						
crawler.js																																
crawwwler																																
Datapark Search																																
DeepArc																																
DeepVaccum	✖		macOS																													
Django Dynamic Scraper																																
dryscrape																																
EIS Archiver																																
Ex-Crawler																																
F(b)arc																																
Gungho																																
Hounder																																
html-snapshots																																
html2warc																																
HyperSpider (JS)																																
icrawler																																
iwebcrawler								✔																								
jedi-crawler																																
Jspider																																
JWAT																																
Knowlesys																																
LARM																																
Lassie																																
Lentil																																
LinkGrabber																																
METIS																																
Miru																																
mnoGoSearch																																
mummify.it																																
Newspaper																																
Norconex HTTP Collector																																
NutchWAX																																
OpenWayback																											✔	✔				
OpenWebSpider																																
OutbackCDX																																
PageVault																																
Panscient																																
ParseHub																																
peeep.us																																
phantom-crawler																																
Pinboard																																
Preservica																																
PromptCloud																																
Puppeteer																																

For more information about this spreadsheet, please visit https://github.com/archivers-space/research/tree/master/web_archiving

	General information					Users & developer features						"Out of the box" website crawling capabilities											Advanced data harvesting capabilities				Archive management features			Other notes		Date/version checked	
Name (link to home page)	*Open source?*	*Source repo*	*Operating system(s)*	*Primary dev. language*	*Target audience*	*CLI*	*GUI*	*WUI*	*Library API*	*Network API*	*Extensibility framework*	*Parallel crawling*	*Scheduled crawling*	*Crawl storage format(s)*	*Capture raw responses*	*Follow links*	*URL filtering*	*Advanced filtering*	*Extract links from JavaScript*	*Run JavaScript*	*Handle React*	*Extract links from Flash*	*Run Flash*	*Targeted scraping*	*Manual form interaction*	*Auto form interaction*	*Extract file data*	*Browse*	*Playback*	*Full-text search*	*Notable users*	*Notes and comments*	*Evaluation date*	*Version examined*
scrala																																	
screen-scraper																																	
ScrapBook																																	
ScrapeKit																																	
Scrapy																																	
SeimiCrawler																																	
Sentry	Affero	GitHub	Lin, Win, macOS	Go		✔						✔	✔																				
Shine																																	
Smart Cache Loader																																	
Sparkler	Apache	GitHub																															
Sphider																																	
Spiderman																																	
Spindle																																	
twarc																																	
UbiCrawler																																	
warc-discovery																																	
warcbase																																	
WARClight																																	
warcprox																																	
wayback				Java																											An older, unbranded version of IIPC OpenWayback		
Web Harvest																																	
Web::Scraper																																	
Web2Warc																																	
WebArchivePlayer																																	
WebCite																																	
webcitation.org																																	
WebEater																																	
WebLech																																	
Webrecorder Player	Apache	GitHub	Lin, Win, macOS	Python, Electron		✔	✔												✔	✔			✔										
webshot																																	
website-scraper																																	
WebSPHINX																																	
wiki-crawler																																	
WWW-Mechanize																																	
WWW::Crawler::Lite																																	
Xapian																																	
XArch																																	
xidel																																	
-- Dead or deprecated software (home pages no longer exist, etc.) --																																	
~~Combine~~																																	
~~Hyperix~~																																	
~~Nalanda~~																																	
-- Insufficient information (e.g., no explanations or descriptions, or descriptions written in an unfamiliar language) --																																	
spiderframework																																2017-07-03	
-- Too specialized for a specific purpose --																																	
ArchiveFacebook	✔	GitHub	Firefox Add-On	JavaScript	user	✖	✔	✔	✖	✖	✖	✖	✖	Original Web Resources	✖	✖	✖	✖	✖	✖	✖	✖	✖	✔	✖	✖	✔	✔	✔	✖		Firefox add-on for archiving a Facebook acct. Uses old XUL add-on format, no longer maintained	2017-07-03	
HyperSpider (Java)	Public domain	SF.net	Lin, Win, macOS	Java	user	✔	✖	✖	✖	✖	✖	✖	✖	CSV, txt, DOT, RDF, Prolog, HTML	✖	✔	✖	✖	✖	✖	✖	✖	✖	✖	✖	✖	✖	✖	✖	✖	✖		Only collects the link structure of a website	2017-09-03	2003-02-09
python-sitemap	GPLv3	GitHub	Lin, Win, macOS	Python	user	✔	✖	✖	✖	✖	✖																					Only collects the link structure of a website	2017-09-03	18683d0
pysitemap	Apache	GitHub	Lin, Win, macOS	Python	user	✔	✖	✖	✖	✖	✖																✖	✖	✖			Only collects the link structure of a website	2017-10-04	c5057f3
python-goose	Apache	GitHub	Lin, Win, macOS	Python	user	✔	✖	✖	✖	✖	✖	✖	✖	files on disk	✖							✖	✖									Specialized for extracting text of news articles	2017-10-04	1.0.25
SyncUThink	GPL	GitHub	Lin, Win, macOS	Java	user																											Crawls CiteULike accounts to download PDFs	2017-07-03	
Visual Sitemap Gener	✖	✖				✖	✖	✔	✖	✖	✖																					Does not archive or scrape. Might also be known as "DYNO Mapper".	2017-10-04	
WERA	✔			PHP								✖	✖	WARC	✖	✖	✖	✖	✖	✖	✖	✖	✖	✖	✖	✖	✖	✔	✔	✔		Tool for searching and navigating archived web document collections		

For more information about this spreadsheet, please visit https://github.com/archivers-space/research/tree/master/web_archiving

General information						Users & developer features						"Out of the box" website crawling capabilities												Advanced data harvesting capabilities				Archive management features			Other notes		Date/version checked	
Name (link to home page)	Open source?	Source repo	Operating system(s)	Primary dev. language	Target audience	CLI	GUI	WUI	Library API	Network API	Extensibility framework	Parallel crawling	Scheduled crawling	Crawl storage format(s)	Capture raw responses	Follow links	URL filtering	Advanced filtering	Extract links from JavaScript	Run JavaScript	Handle React	Extract links from Flash	Run Flash	Targeted scraping	Manual form interaction	Auto form interaction	Extract file data	Browse	Playback	Full-text search	Notable users	Notes and comments	Evaluation date	Version examined