

Crawlerek és scraperek

Webes tartalmak mentésére szolgáló programok

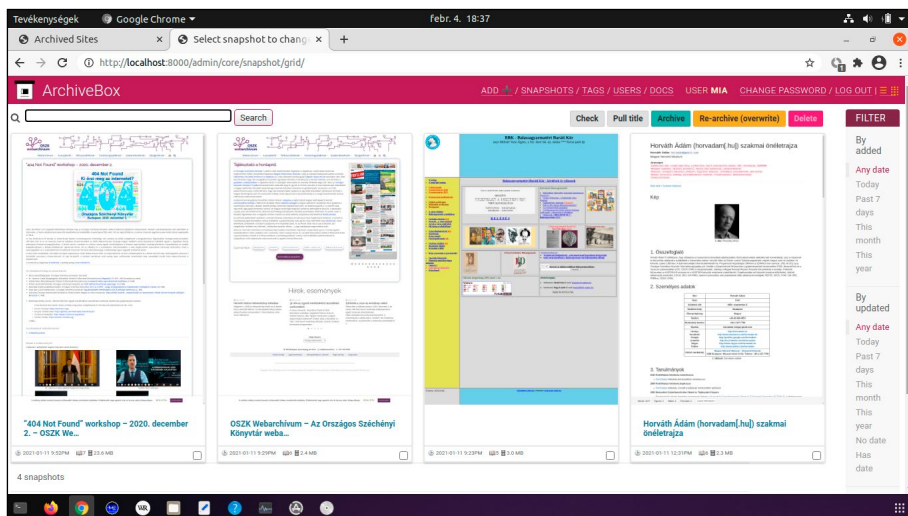
A *Könyv, Könyvtár, Könyvtáros* 2020. decemberi számában¹ Németh Márton szerzőtársammal bemutattunk néhány olyan programot, amelyeket weboldalak vagy teljes webhelyek mentésére fejlesztettek ki, elsősorban magánhasználatra, nem tömeges archiváláshoz. Jelen írás ennek a cikknek a folytatása, és ezúttal a *MLA Wiki*², vagyis a Magyar Internet Archívumhoz készülő tudásbázis legfrissebb szócikkeiből válogattam ki néhány érdekes alkalmazást. A wiki 2021 nyarán 136 új bejegyzéssel bővült, s mivel ezek többsége valamilyen program rövid ismertetése, így a *Szoftverek* kategória 2021 szeptemberére elérte a 372 tételt. A paletta igen széles, az egyetlen részfeladatra alkalmas egyszerű szkriptektől a komplett archiváló keretrendszerreig terjed. A két legnépesebb csoportot a *crawlerek* és a *scraperek* alkotják, vagyis a linkeket követő és a weboldalakat letöltő „aratógépek”, illetve a weboldalokról adatokat kigyűjtő „gereblyék”, melyek vagy a tartalomszolgáltató szerver programból lekérdezhető felületéhez (API) kapcsolódnak, vagy maguk is egy crawlert futtatnak, és az azzal összeszedett fájlokat elemzik ki.

ArchiveBox³

Az *ArchiveBox* egy fiatal szoftvermérnök, *Nick Sweeting* által Python nyelven fejlesztett nyílt forráskódú archiváló szoftver. A kerékpár- és zenemániás Nick jelenleg az Egyesült Államokban és Kanadában működtet egy tanácsadással foglalkozó startupcéget, de gyerekkorát a „Nagy Tűzfal” mögötti Kínában töltötte, így mindennapos élménye volt, ahogy eltűntek vagy elérhetetlenné váltak a számára érdekes weboldalak, vagy hirtelen az egész hálózat elment az internet, ha

cenzurázott szavakra próbált rákeresni. Ezért kezdett el először saját maga és barátai számára a webarchiválással kísérletezni.

Az ArchiveBox legelső verzióját 2017-ben – talán nem véletlenül – a függetlenség napján, július 4-én tette fel a GitHub kódmegeosztó oldalra.⁴ A cikk írásakor még mindig csak a 0.6.2-es változatnál tart, de már most is jónéhány hasznos funkciót tartalmaz ez a rendszer, amely a *wget* nevű linuxos crawlert, a *Google Chrome headless* módban futtatható motorját és a *youtube-dl* videóletöltőt integrálja. Sokféle mentési lehetőséget választhatunk, akár egyszerre többet is (szabványos WARC-formátum, fájlrendszer, egyetlen HTML-állomány, JSON-szöveg, PNG-oldalkép, PDF-dokumentum), és külön is letölthetők a megadott weboldal egyes elemei (title, favicon, response header, médiafájlok), illetve az oldal automatikusan generált technikai metaadatai, sőt az Internet Archive szerverére is elmenthetünk egy másolatot. Fontos korlátozás, hogy a linkeket legfeljebb egy szintig követi a robot, vagyis teljes webhelyek aratására már nem használható. A letöltendő URL-címek (*seed*) megadhatók egy webes űrlapon vagy egy szövegfájlban, de könyvjelzőkből vagy bookmark-szolgáltatásokból (pl. Pocket, Pinboard), böngészési előzményekből és RSS feedekből is ki tudja gyűjteni őket az ArchiveBox. Továbbá képes olyan szoftverrepositoriumok egyes részeinek klónozására is, mint a GitHub, a Bitbucket vagy a GitLab.



Lementett weboldalak az ArchiveBox adminisztrátori felületén

A távlati fejlesztési tervek között szerepel a *PyWb* programmal való archiválás és megjelenítés beépítése, valamint a felhasználó által definiálható szkriptek futtatása a böngészőben, ami az emberi aktivitást igénylő, webkettes platformokról való automatikus mentésekhez lenne nagyon hasznos. Az ArchiveBox összes

funkciója csak parancsmódban érhető el, de létezik hozzá egy webes adminisztrátori felület is, amivel paraméterezhetők és elindíthatók egyszerűbb feladatok, valamint visszanezethetők a korábbi mentések. Érdekes még megemlíteni, hogy az ArchiveBox wikijében van egy kiváló linkgyűjtemény⁵ a különféle archiváló szervezetekről, projektekről, szoftvekről, illetve a témával foglalkozó cikkekről, blogokról és fórumokról.

Warrior⁶

A *Warrior* egy civilekből és szakemberekből álló laza szerveződés, az *Archive Team*⁷ számára összeállított virtuális gép, melynek segítségével önkéntesek is be tudnak kapcsolódni az archiváló projektekbe a saját számítógépükkel. A csoport tagjai elsősorban az egyéni szintű archiváláshoz nyújtanak segítséget információkkal és technológiákkal, továbbá figyelik a veszélyeztetett, bezárásra készülő webhelyeket vagy online platformokat, és ilyenkor mentési akciókat szerveznek. A mozgalmat az osztrák közszolgálati média által a digitális archiválás frontemberének nevezett különc, *Jason Scott* indította el 2009-ben, aki egyéb tevékenységei (technológiatörténész, filmkészítő, színész és előadó) mellett a régi szövegfájlokból és ASCII-karakteres rajzokból álló <http://textfiles.com/> honlap létrehozója, és annak a *Sockington* nevű macskának az egyik gazdája, „akinek” több mint egymillió

The screenshot shows the Archive Team Warrior web interface. At the top, there's a navigation bar with 'sites.google.com Website Leaderboard'. Below it, a yellow status bar shows progress for tasks like 'CheckIP', 'GetItemFromTracker', 'PrepareDirectories', 'WigetDownload', 'PrepareStatsOfTracker', 'MoveFiles', and 'Upload'. The main content area is divided into sections for 'Current project' and 'Available projects'. Two projects are listed: 'Item site:tutoronlineforme' and 'Item site:cirurgiaplasticafacebookcom'. Each project has a list of tasks with progress indicators (checkmarks) and a detailed log of operations and file transfers. A network speed monitor in the bottom left shows 6.2 MB/s and 129.7 Kbit/s.

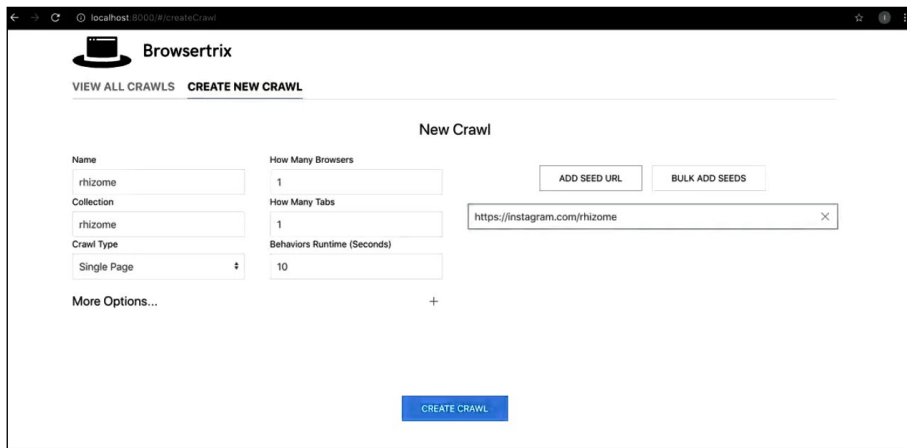
Két futó munkafolyamat az Archive Team Warrior nevű virtuális gépén

követője van a Twitteren. Az ArchiveTeam magyar honlapja⁸ 2021 elején indult, és saját webarchívumot is szolgáltat *Régi Szép Idők* (RSZI) néven. A Warrior Windows, macOS és Linux alatt egyaránt futtatható VirtualBox vagy VMware Player virtualizációs környezetben, de Docker-formában is telepíthető. A *wget* crawlert használja, az egyes jobokat pedig a központi szerveren futó *Tracker-rendszer* vezérli. A felhasználó kiválaszthatja, hogy hány ilyen munkafolyamatot akar futtatni párhuzamosan, illetve hogy az Archive Team aktuális projektjei közül melyekben szeretne részt venni. Ezután a Warrior a Trackertől kapott paraméterek alapján elindítja a jobokat, majd ha valamelyik webhely letöltése befejeződött, akkor a WARC-fájlokat úgynevezett *rync-kapcsolaton* keresztül feltölti előbb az Archive Team tárhelyére, onnan pedig bekerülnek az Internet Archive gyűjteményébe is. A Warrior paraméterezése és működésének követése böngészőből történik, egy nagyon egyszerű, logikus felületen a *localhost:8001* címet megnyitva. Regisztrációt nem igényel, csak egy felhasználói nevet kell választanunk, amely az „önkéntes bedolgozók” aktuális listáján fog megjelenni a Tracker weboldalan.

Browsertrix⁹

A *Browsertrix* egy böngészőalapú archiváló szoftver, mellyel a weblapok HTML-kódját elemző „buta” robotoknál jobb minőségben – igaz, jóval lassabban – lehet lementeni a mai bonyolult, gyakran szkriptekkel generált weboldalaikat. A *Browsertrix* mögött az a San Franciscóban élő *Ilya Kramer* áll, aki a már említett tavalyi 3K-cikkben ismertetett *Webrecorder* és *ReplayWeb.page*-t is írta, több más nagyszerű archiváló és megjelenítő eszközzel együtt. Korábban az Internet Archive-nál dolgozott, majd átment a *Rhizome* nevű, digitálisművészet- és digitálismédia-megőrzéssel foglalkozó nonprofit szervezethez. A *Rhizome* működteti a *Webrecorder* technológiáját használó *conifer.rhizome.org* archiváló szolgáltatást és az *oldweb.today* oldalt, ahol emulált környezetben futó régi böngészőkkel nézhető meg a régi webhelyek. A *Browsertrix* a nevét a „browser” és a „Heritrix” szavak összevonásából kapta (utóbbi az Internet Archive crawlere). Fejlesztése a 2010-es évek közepére nyúlik vissza, és három fázisra tagolható. A legelső változat a weboldalak bárki által annotálhatóvá tételével foglalkozó *Hypothes.is* projekt Annotator szoftvere számára készült 2015-ben, de mivel már az is nyílt forráskódú volt, így más rendszerekbe is be lehetett építeni.¹⁰ A *Selenium* nevű böngészőautomatizáló eszközt használta a *Chrome* vagy a *Firefox* vezérlésére is, és a betöltött weboldalaikat elküldte a *webrecorder.io* vagy a *Save Page Now* szolgáltatásnak. (Előbbi a Conifer elődje, utóbbi pedig az Internet Archive *archive-on-demand* funkciója.) A technikai metaadatok – beleértve az eredeti és az archív URL-címet is – pedig egy JSON-naplófájlba kerültek. Az IIPC (International Internet Preservation Consortium) 2019-es konferenciáján *Ilya Kramer* bemutatott egy továbbfejlesztett változatot, amely egy Docker-konténerben futtatható önálló archiváló eszköztár.¹¹ Ehhez

már készült egy egyszerű, grafikus felület is, de bonyolultabb aratásokat csak parancsmódban vagy YAML-formátumú fájlokon keresztül lehet konfigurálni. A headless-módban futtatott Chrome-hoz külön definiálhatók „viselkedések” (*behaviors*), ezekkel vezérelve például a Twitter- vagy a Facebook-oldalak letöltését, a Webrecorder *autopilot* funkciójához hasonlóan. Opcionálisan oldalképek is készíthetők vele, a weboldalak szövege pedig *Sohral* indexelhető. A bejelentkezést igénylő webhelyekhez külön böngészőprofilok definiálhatók, így elkerülhető (bár nem százszázalékos biztonsággal), hogy a bejelentkezési adatok bekerüljenek a WARC-fájlba. A mentéseket PyWb-vel lehet visszaneézni. Két évvel később a projekt ismét új irányt vett: egyetlen integrált rendszer helyett modulokra szedte szét a fejlesztője. 2021 őszén még csak a crawler-komponens volt letölthető (szintén Dockerben), amihez még nem készült el a grafikus felület.¹² A Browsertrix Crawler a *Puppeteer* nevű eszköz segítségével vezérli a böngészőt, a mentést pedig a PyWb végzi *capturing*-üzemmódban. Van egy *screencasting*-opciója is, amivel a böngészőben figyelhetjük, ahogy a robot letölti a weboldalakat. A kiinduló URL-lista fájlban is megadható, és seedenként külön konfigurációs szabályok határozhatók meg. Több szálon futtatható és ehhez is vannak *behavior*-szkriptek (pl. autoscroll, video autoplay, valamint webhelyspecifikus viselkedések). A WARC- mellett WACZ-formátumba is tud menteni, ami a ReplayWeb.page számára szükséges indexeket és technikai adatokat tartalmazza, de természetesen PyWb-vel is visszaneézhető az archivált tartalom.



Instagram-oldal mentésének beállítása a Browsertrix programban

TAGS¹³

A *TAGS*, ami a Twitter Archiving Google Sheet rövidítése és persze a Twitteren használt *hashtagek*re utal, egy régi és sajnos már nem frissülő esz-

köz, de két szempontból is érdemes bemutatni. Egyrészt, mert egy egyszerű példa arra, hogy hogyan lehet adatokat „összegereblyézni” a webről, másrészt a története jól illusztrálja, hogy az archiváló technológiák milyen gyorsan elavulnak az internetes szolgáltatások gyakori változásai miatt. *Martin Hawkesey*, aki most a skóciai *Edinburgh Futures Institute* tanulótervezési és oktatástechnológiai vezetője 2010-ben hobbiprojektként kezdett el foglalkozni azzal, hogy hogyan lehetne a közösségi médiából gyűjtött adatokat felhasználni az oktatásban. Ekkor írta meg a *TAGS scraper*, majd egy évvel később a *TAGSExplorer* modult, amivel elemezhető, vizualizálható és kereshető a begyűjtött információk. A scraperek más megközelítést alkalmaznak, mint az előzőekben bemutatott crawlerek, és jellemzően más célokra is használjuk őket. Ahelyett, hogy az eredeti weboldalakról próbálnánk meg minél hűségesebb másolatokat készíteni, azok lényegi tartalmát és/vagy metaadatait gyűjtjük be, például tudományos vagy piackutatási elemzésekhez. Ez technikailag egyszerűbb feladat, mint komplex webhelyek lementése és megtekinthető állapotban tartása, viszont elvész az eredeti külalak és kontextus. A TAGS egy sablon és egy szkript

The screenshot shows the TAGS v6.1.9.1 interface, which is a Google Spreadsheet. It contains several sections:

- With this spreadsheet you can:** Loading...
- Instructions:**
 - If you've never run TAGS > Setup: Twitter: Access is so now (this should only need to be done once for all your TAGS sheets)
 - Enter term: `#webarchive OR #webarchiveview #RT -from:stillio` (without quotes)
 - you can use search operators like AND OR as well as from: and to: eg #JobNow AND from:BarackObama
- Note:** Make a one off collection with TAGS > Run now! or set a trigger to collect every hour TAGS > Update archive every hour. To change the frequency open Tools > Script: Editor then Triggers > Current script's triggers... and adjust
- Advanced Settings:**
 - Period: default
 - Follower count filter: 0 (if search term is being spammed you can set the minimum followers a person must have to be included in archive)
 - Number of tweets: 5000 (maximum varies based on the type of archive you are collecting)
 - Type: search/tweets (use a search term in step 3 above to get results from last 7 days)
- Stats:**

Number of Tweets	447
Unique tweets	373
First Tweet	29/05/2021 10:04:00
Last Tweet	08/10/2021 15:55:22
- Make interactive:** Turn your archive into an interactive online resource using TAGSExplorer
- Note:** Share > Anyone with link to use these views
- Buttons:** TAGSExplorer (conversation explorer), TAGS Archive (searchable archive)
- Top Tweeters Table:**

Top Tweeters	No.	@%	% RT	Twitter Activity
keesone	04	3	#N/A	[Line Graph]
archivelborg	40	15	#N/A	[Line Graph]
shawmijones	37	2	#N/A	[Line Graph]
UKWebArchive	20	8	#N/A	[Line Graph]
stillio	17	16	#N/A	[Line Graph]
InternetHistor	15	20	#N/A	[Line Graph]
yeamsurer	14	#N/A	#N/A	[Line Graph]
Respadon Proj	11	2	#N/A	[Line Graph]
internetarchive	10	27	#N/A	[Line Graph]
sophiewackles	10	#N/A	#N/A	[Line Graph]
NetPreserve	9	27	#N/A	[Line Graph]
unleasharchives	7	5	#N/A	[Line Graph]
- Summary Table:**

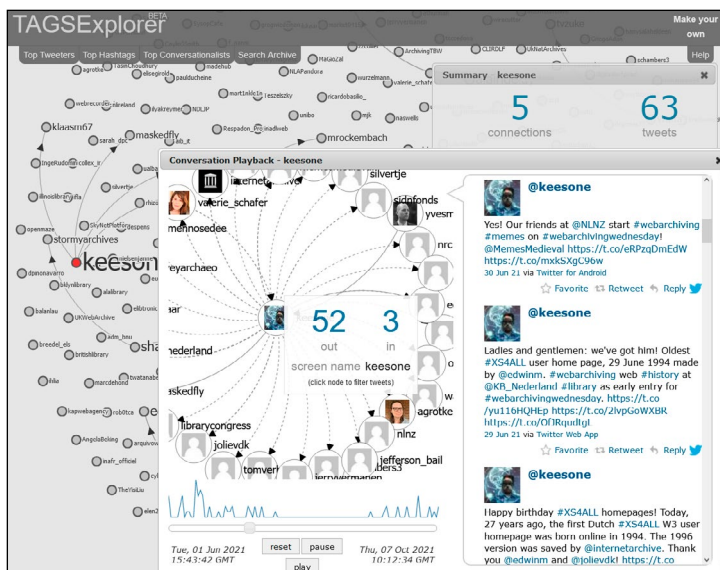
Number of links	408
Number of RTs	0
Number of Tweets	447
Unique tweets	373
First Tweet in Archive	29/05/2021 10:04:00
Last Tweet in Archive	08/10/2021 15:55:22
In Reply ids	76
In Reply @s	19
Tweet rate (w/min)	0.11

A TAGS paraméterezése a webarchiválással kapcsolatos tweetek gyűjtéséhez és az összesített adatokat tartalmazó munkalap egy részlete

a Google Dokumentumok táblázatkezelőjéhez, amiről egy másolatot készíthetünk a saját tárterületünkre, majd engedélyeznünk kell, hogy hozzáférjen a Google- és a Twitter-fiókunkhoz. (Ez a biztonsági szigorítások miatt manapság

egy elég körülményes művelet sor.) A táblázat *Readme/Settings* munkalapján tudjuk paraméterezni a szkriptet. Az archiválható tweetek szűrhetők hashtag és felhasználó szerinti keresőkérdéssel, illetve a kedvencek listája alapján, továbbá megadható a maximális számuk és a frissítés gyakorisága is. A scraper a TAGS menü *Run Now!* menüpontjára való kattintás után a Twitter API-ján keresztül elkezdheti letölteni a tweetek szövegét és tizenhétféle metaadatát (pl. felhasználói név, profilkép-URL, időpont, földrajzi hely, nyelv), melyek az *Archive* munkalapon jelennek meg. Visszamenőleg legfeljebb az utolsó hét nap üzeneteit tudjuk leszedni a nyilvános API korlátai miatt. A menüből két további fület (*Summary* és *Dashboard*) is hozzáadhatunk a táblázathoz, ahol összesítéseket és grafikonokat látunk a begyűjtött adatokról.

A TAGSExplorer linkre kattintva pedig egy külön böngészőablakban megnézhető az emberek közötti kapcsolati háló, és visszajátszható egy kiválasztott felhasználó tevékenysége harminc másodpercre sűrítve. A Google és a Twitter szigorításai miatt Martinnak többször is módosítania kellett a szkriptet, míg végül 2018-ban – úgy tűnik – végleg abbahagyta a karbantartását, így ez az egyszerű és sokak által használt eszköz lassan működésképtelen lesz. A TAGSExplorer és a TAGS Archive-linkek például már nem jók az eredeti sablonban, és az egyik fórumtag által 2020 szeptemberében javasolt módosítás¹⁴ is csak az előbbit javítja meg, e cikk írásakor az archívum keresője már nem volt elérhető.



*A Holland Királyi Könyvtár digitális gyűjteményében dolgozó,
magyar származású történész, Kees Tszelszky webarchiválási témájú
tweetjeinek visszajátszása a TAGSExplorerrel*

Octoparse¹⁵

Az *Octoparse* az egyik legjobb webscraper-rendszer, ami Windows vagy macOS rendszerű asztali gépen futtatható változatban és előfizetéses felhőszolgáltatásként is elérhető. A kaliforniai központú *Octopus Data Inc.* céget 2015 szeptemberében hozta létre a kínai *Shenzhen World Information Technology* és a következő év márciusában mutatták be az eredetileg *Octopus Collector* nevű szoftverük Octoparse-ra átkeresztelt, angol nyelvű változatát. (A szóösszetétel feltehetően arra utal, ahogy a program polipszerű karjaival kinyúl az internetre, majd kielemez a begyűjtött weboldalakat.) Sok más hasonló célú versenytársától eltérően az Octoparse használata nagyon könnyen megtanulható a rámutatással és kattintással működő konfiguráló felületnek, az intelligens elemző algoritmusnak és a helyzetérzékeny sűgónak köszönhetően. Miután megadtuk a kiinduló weboldal URL-címét, a program betölti azt és felismeri a főbb HTML-elemeket, majd táblázatos formába rendezi a bennük talált tartalmakat, melyek lehetnek adatok, szövegek és linkek. A táblázat oszlopait átnevezhetjük, átrendezhetjük vagy törölhetjük, illetve felvehetünk továbbiakat is úgy, hogy az egérrel kijelölünk elemeket a weblapon.

#	Title	Title_URL	Author	Download_URL
1	Web scraping w...	https://books.google...	R Mitchell - 201...	
2	Web scraping us...	https://journals.sagep...	A Bradley, RJE Ja...	https://journals.sagepub.co...
3	Practical Web Sc...	https://books.google...	S vanden firou...	https://www.academia.edu/...
4	Social media we...	https://www.scienced...	LC Dewi, A Chan...	https://www.sciencedirect...
5	Web Scraping	https://arxiv.org/abs/...	A Hermawan, Sri...	https://arxiv.org/abs/1808.0...

Web scraping témájú publikációk adatainak begyűjtése a Google Tudós találati listájából az Octoparse oldalelemző funkciójának segítségével

Ha van lapozási vagy további tartalombetöltési lehetőség az oldalon, akkor azt is megtaníthatjuk a scrapernek – amennyiben nem ismeri fel automatikusan. A főbb közösségi, üzleti, utazási és egyéb szolgáltatásokhoz előre elkészí-

tett sablonok állnak rendelkezésre, így azokkal még könnyebb ez a munkafázis. A *Run* gombra kattintva a beépített böngészőben elindul a weboldalak betöltése és azokból az adatok begyűjtése, melyek azután XLS-, CSV-, JSON-, HTML-fájlokba vagy SQL-adatbázisokba exportálhatók.

Az egyes *taskok* beállításánál több hasznos lehetőség is van: kiválasztható, hogy a program milyen böngészőként azonosítsa magát, blokkolja-e a reklámokat, letöltse-e a képeket, leálljon-e, ha úgy tűnik, hogy végtelen ciklusba került, cserélgesse-e az IP-címeket a kitiltás elkerülése érdekében stb. Az aratások természetesen ütemezhetők is, és amennyiben a saját gépünk helyett a felhőben futtatjuk őket, akkor kérhetünk e-mail-értesítést, amikor véget értek, és ilyenkor API-n keresztül is letölthetők az adatok. Az ingyenes változatnál tíz scraper definiálható, legfeljebb két feladat futhat egy időben és csak a saját gépünkön, maximum tízezer adatrekord exportálható ki egyszerre, viszont a bejárható oldalak száma nem korlátozott.

The screenshot shows the Octoparse web scraper interface. The main window displays search results for 'web scraping' on Google. The results list several articles, including 'Web scraping: state-of-the-art and areas of application' and 'An overview on web scraping techniques and tools'. An 'Export local data' dialog box is open, showing options to export data as Excel (xlsx), JSON, CSV, or HTML. The 'Export to database' section includes options for SqServer and MySQL. At the bottom, a table of extracted data is visible, with columns for #, Title, Title_URL, Author, Download_URL, Author_URL, Text, Cited_URL, and Citation_Number. The table contains 10 rows of data. Below the table, there are buttons for 'Cloud Backup', 'Export Data', and 'Run'.

#	Title	Title_URL	Author	Download_URL	Author_URL	Text	Cited_URL	Citation_Number
1	Web scraping: state-of-	https://eexplore.ies...	R Drou, EN Sari, O S...	https://scholar.google...	https://scholar.google...	Main objective of Web Sc...	https://scholar.google...	Létezők száma: 23
2	An overview on web scra...	http://www.ijfrcn...	AV Saurkar, KS Patna...	http://www.ijfrcn...	http://scholar.google...	From the evolution of W...	https://scholar.google...	Létezők száma: 52
3	Personalized content ext...	https://www.ijglo...	T Karthikeyan, K Sek...	https://ericbrasil...	https://scholar.google...	Web scraping is a techn...	https://scholar.google...	Létezők száma: 70
4	Legality and ethics of w...	https://www.resear...	V Krotov, L Silva - 20...	https://www.resea...	https://scholar.google...	Automatic retrieval of da...	https://scholar.google...	Létezők száma: 37
5	Web scraping and Naive...	https://opscience...	C Slamet, R Andrian, ...	https://opscience...	https://scholar.google...	Many organisations (gov...	https://scholar.google...	Létezők száma: 72
6	Scraping the demos. Digi...	https://www.tandfon...	L Ulbricht - Democra...	https://www.tandof...	https://scholar.google...	Scientific, political and b...	https://scholar.google...	Létezők száma: 17
7	Data analysis by web scr...	https://eexplore.ies...	DM Thomas, S Math...	https://scholar.google...	https://scholar.google...	The standard informatio...	https://scholar.google...	Létezők száma: 28
8	A novel web scraping ap...	https://eexplore.ies...	E Uzun - IEEE Access...	https://eexplore.ies...	https://scholar.google...	Web scraping is a proces...	https://scholar.google...	Létezők száma: 13
9	Cloud based web scrapin...	https://eexplore.ies...	RS Chaulagain, S Pan...	https://www.resea...	https://scholar.google...	With the penetration of...	https://scholar.google...	Létezők száma: 38
10	Conceptual approach for...	https://www.umwib...	P Miley Economic un...	https://www.umwib...	https://scholar.google...	The paper focuses on the...	https://scholar.google...	Létezők száma: 23

Az Octoparse által begyűjtött adatok exportálása Excel-táblázatba

Web Scraper¹⁶

A szimplán csak *Web Scrapper*ek elnevezett ingyenes böngészőkiegészítő, és mellé a fizetős, felhőalapú, webes adatgyűjtő szolgáltatást egy 2017-ben alapított cég, a Lettországból bejegyzett *Web Graph* fejlesztette ki. Előbbinek az alapötlete nagyon egyszerű: magát a Chrome vagy Firefox böngészőt használjuk scraperként, így a JavaScripttel generált, Ajax-kéréseket küldő vagy bejelentkezést igénylő weboldallal sincs gond. A bővítmény telepítése után az F12 gomb

megnyomásával elérhető fejlesztői eszközök között megjelenik egy további fül három fő menüponttal. Új feladatot a *Create new sitemap* menüben hozhatunk létre, és ugyanitt arra is lehetőségünk van, hogy egyszerűen bemásoljunk egy kész utasítássort JSON-formátumban. (Ilyeneket a honlapon levő oktatóvideók alatt és a fórumban találhatunk.) Új sitemap esetén adni kell neki egy nevet, valamint beírni egy vagy több URL-címet, ahonnan majd a scraper elindul. A már létrehozott feladatok a *Sitemap* alatt listázhatók ki és szerkeszthetők vagy törölhetőek. Szerkesztés módban a Sitemap menü *Selectors* alpontja nyílik meg, itt tudjuk megadni a *Select* gomb megnyomásával, majd az egérrel való rámutatással, hogy a weboldal mely elemét válassza ki a program, és hogy az milyen típusú. A legegyszerűbb a *text* típus, melynél csak az adott oldalelem szövege (pl. egy név vagy e-mail-cím) kerül begyűjtésre. Ha több ilyen elem is van az oldalon, akkor a shift billentyű lenyomása mellett érdemes még néhányat kiválasztani, hogy megtanítsuk a Web Scrapert ezek felismerésére, és ilyenkor a *Multiple* opciót is ki kell pipálni a selector űrlapján. A *text*en kívül további tizenkét elemtípus közül választhatunk, amikkel különböző műveleteket végeztethetünk el, például linkek követése, oldal görgetése, lapozás. Ezek a selectorok

The screenshot shows the Open Web Scraper interface overlaid on a browser window. The browser is displaying the Magyar Elektronikus Könyvtár (MEK) search results page. The Open Web Scraper overlay includes a 'Learn how to use' section with links to video tutorials, documentation, test sites, and a forum. It also features an 'Automate Web Scraper in Cloud' section with options for scheduling data extraction, managing via API, and exporting to Dropbox. A 'Go premium' button is also present. The browser's developer tools are open, showing the 'Web Scraper' extension interface with a table of selected items.

web-scraper-o	Selector	konvy_link	konvy_link.href	szamlalo
1633859138-1f	Scrape	/keresek/keres.phtml?tip=uj	https://mek.oszk.hu/22300/22368	SZÁMLÁLÓ: 310
1633859142-1f	Browse	/keresek/keres.phtml?tip=uj	https://mek.oszk.hu/22300/22370	SZÁMLÁLÓ: 454
1633859140-1f	Export Sitemap	/keresek/keres.phtml?tip=uj	https://mek.oszk.hu/22300/22369	SZÁMLÁLÓ: 357
1633859145-1f	Export data as CSV	/keresek/keres.phtml?tip=uj	https://mek.oszk.hu/22300/22345	SZÁMLÁLÓ: 220
1633859135-1f	Export data as CSV	/keresek/keres.phtml?tip=uj	https://mek.oszk.hu/22200/22276	SZÁMLÁLÓ: 226

*Megtekintésszámok begyűjtése a Magyar Elektronikus Könyvtárból
az újdonságok listája alapján Web Scraperral*

szülő-gyerek kapcsolatban lehetnek egymással, így egy munkafolyamattá fűzhetők össze, mely gráfformában is megnézhető. Az adatgyűjtés a Sitemap menü *Scrape* alpontjával indítható, az eredmény pedig a *Browse* menüpont alatt jelenik meg táblázatként, ami azután CSV-fájlba exportálható.

A program működési logikájának megtanulása időigényes, de szerencsére jó dokumentáció, videós oktatóanyag, blog, fórum és többféle tesztoldal is van hozzá, így aki ezekbe beleássa magát, az egészen bonyolult feladatokat is meg tud oldani a böngészőjével. A cég felhőjében futó, *Web Scraper Cloud* elnevezésű szolgáltatás többféle csomagban is előfizethető. A legdrágábbnál gyakorlatilag korlátlan a robottal bejárható oldalak száma, ötnél több párhuzamos job is futtatható, megadható külső proxy, az aratások ütemezhetőek, az eredmények hatvan napig maradnak a felhőtárhelyen, és a CSV- mellett van XLSX-, JSON-, Dropbox- és Google Sheets-export, valamint API-n keresztüli elérés is.

Jegyzetek

1. Drótos László – Németh Márton: *Egyedi mentésekre szolgáló webarchiváló szoftverek*. = Könyv, Könyvtár, Könyvtáros, 29. évf. 2020. 12. sz. 3–11. p. https://epa.oszk.hu/01300/01367/00334/pdf/EPA01367_3K_2020_12_003-011.pdf (2021.09.20.)
 2. <https://webarchivum.oszk.hu/mediawiki/> (2021.09.20.)
 3. <https://archivebox.io/> (2021.09.20.)
 4. <https://github.com/ArchiveBox/ArchiveBox> (2021.09.20.)
 5. <https://github.com/ArchiveBox/ArchiveBox/wiki/Web-Archiving-Community> (2021.09.20.)
 6. https://wiki.archiveteam.org/index.php/ArchiveTeam_Warrior (2021.09.20.)
 7. https://en.wikipedia.org/wiki/Archive_Team (2021.09.20.)
 8. <http://archiveteam.hu> (2021.09.20.)
 9. <https://github.com/ikreymer/browsertrix> (2021.09.20.)
 10. <https://github.com/ikreymer/browsertrix> (2021.09.20.)
 11. <https://github.com/webrecorder/browsertrix> (2021.09.20.)
 12. <https://github.com/webrecorder/browsertrix-crawler> (2021.09.20.)
 13. <https://tags.hawksey.info> (2021.09.20.)
 14. <https://tags.hawksey.info/forums/topic/explorer-archive-hyperlinks-broken-in-6-1-9-1/> (2021.09.20.)
 15. <https://www.octoparse.com> (2021.09.20.)
 16. <https://webscraper.io> (2021.09.20.)
-
-