Drótos László – Németh Márton

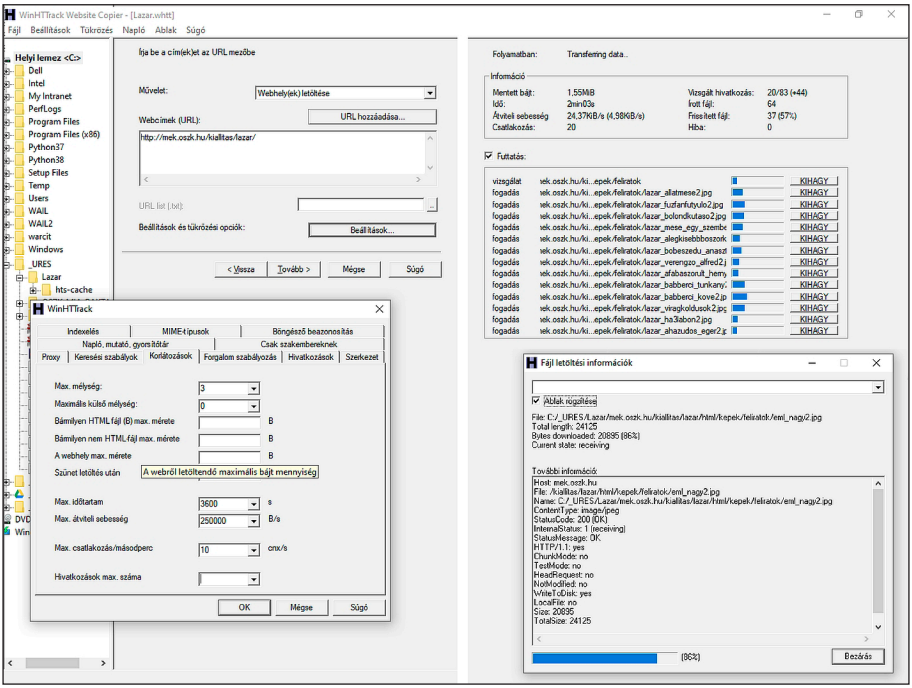
Egyedi mentésekre szolgáló webarchiváló szoftverek

A nemzeti vagy nemzetközi szintű webarchívumok többnyire olyan, nagy tömegű tartalom letöltésére alkalmas, automatikus működésű szoftverrobotokat használnak, mint amilyen az Internet Archive által kifejlesztett *Heritrix*. Viszont az elsősorban még a hagyományos webre optimalizált Heritrix nem igazán jó a modern, dinamikusan generált, esetleg a felhasználó közreműködését is igénylő honlapok vagy a közösségi média archiválására. Továbbá sokszor nem a tömeges webtartalom letöltése a cél, hanem csak egy-egy webhelyet vagy adott weboldalakat szeretnénk lementeni. Cikkünkben ilyen feladatokra alkalmas ingyenes szoftvereket mutatunk be, amelyek személyes vagy kisebb intézményi archívumokhoz egyaránt hasznosak lehetnek.

HTTrack

A teljes nevén *HTTrack Website Copier* a Heritrixnél is régebbi, nyílt forráskódú program, 1998 óta egészen 2017 áprilisáig fejlesztgette egy francia informatikus, Xavier Roche. Windows, macOS, Linux és Android rendszerekre egyaránt telepíthető, a felülete pedig magyarra is átállítható. A robot működését sokféle paraméterrel szabályozhatjuk, egy szövegfájlban akár több kiinduló URL-címet is megadhatunk neki, a letöltési folyamatot is figyelemmel kísérhetjük, és akár közbe is avatkozhatunk.

A HTTrack folytatni tudja a megszakadt letöltéseket, illetve frissíteni is lehet vele egy korábbi mentést, de időgépfunkció nincs benne, vagyis a letöltött weboldalak korábbi állapotait nem őrzi meg automatikusan. A fájlok egy map-



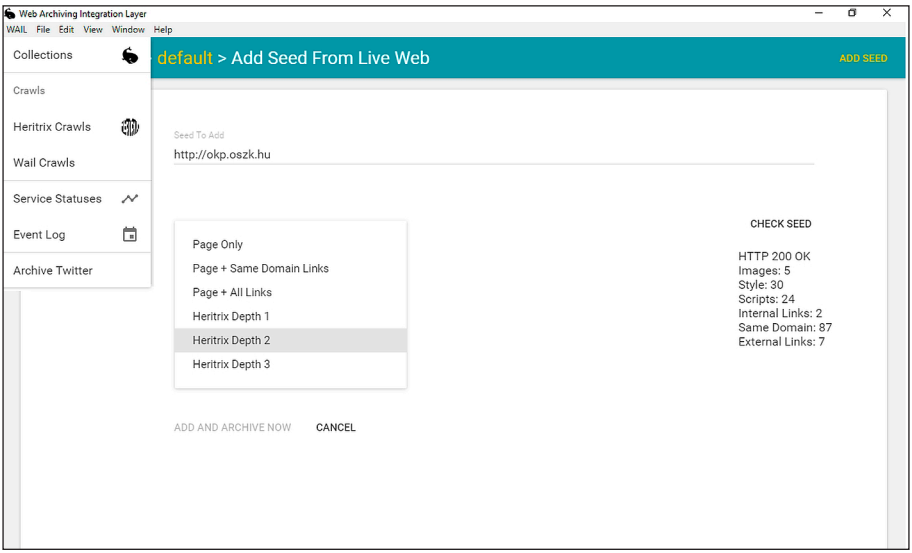
A HTTrack Windows alatti grafikus felületének két részlete: a paraméterező és a letöltést monitorozó ablakok

pastruktúrába kerülnek, a weboldalakban és a scriptekben talált belső hivatkozásokat relatív, helyi linkekre alakítja át, így a lementett tartalom internetkapcsolat nélkül is böngészhető. Szolgáltatófelületet nem tartalmaz ugyan, de ha a program által létrehozott előtét oldalt és fájlrendszert feltöltjük egy webszerverre, akkor egy helyi vagy nyilvános szolgáltatást is kialakíthatunk. A *Warcit* nevű segédprogrammal pedig WARC-formátumra konvertálhatjuk az archivált anyagot, ami már egy hosszú távon is megőrizhető, szabványos állomány. Érdemes még megemlíteni, hogy az 1996 óta működő ausztrál nemzeti webarchívum saját fejlesztésű PANDAS rendszerébe (PANDORA Digital Archiving System) a HTTrack programot építették be annak idején letöltő eszközként.

WAIL

A *Web Archiving Integration Layer* egy több eszközt integráló, Windows, macOS és Linux alatt is futtatható grafikus felület egyes weboldalak vagy webhelyek kis mélységű letöltéséhez, gyűjteményekbe szervezéséhez és visszanézéséhez. Az amerikai Old Dominion University webtudománnyal és digitális könyvtárakkal foglalkozó kutatócsoportja egyik projektjének keretében fejlesztették ki 2015 és 2017 között. Létezik egy *újabb verziója* is, amelyet 2019-ben frissítettek utoljára, de ebben kevesebb funkció van. Az eredeti WAIL a Heritrix robotját és a *PyWb*

megjelenítőt kombinálta össze, de utólag belekerült a Chrome böngészőn keresztül való letöltés lehetősége is („Page” opció), amivel a komplexebb, sok JavaScriptet használó weboldalak jobb minőségben archiválhatók.



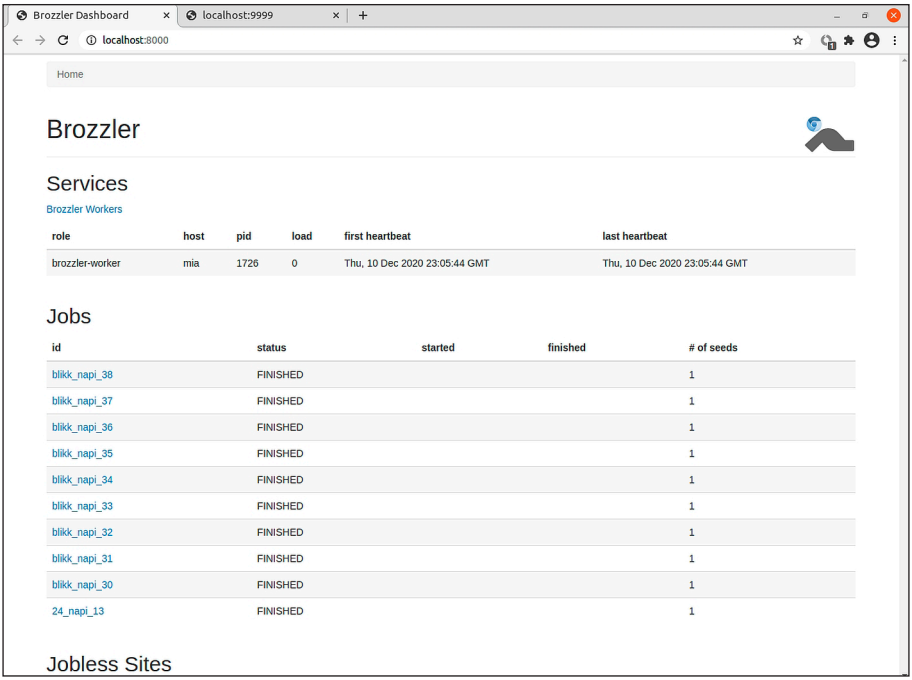
A WAIL főmenüje és az archiválás indítására szolgáló aloldala

Érdekesség, hogy van benne egy Twitter-modul is. Ezzel az általunk megadott csatornákon megjelenő híreket menthetjük folyamatosan, maximum 10 órán keresztül. Az eltárolandó tweetek szavak vagy hashtagek szerint szűrhetők is. Mindegyik beépített archiváló eszköz WARC-fájlokba ment, a mentések részgyűjteményekbe rendezhetők, és visszanézhetők egy böngészővel. A program használata nagyon egyszerű, de egyben ez az egyik hátránya is, mert túl kevés paraméterezési és metaadatolási lehetőséget biztosít. A másik hibája, hogy néha többször is el kell indítani, mire minden komponens rendesen elkezd futni a háttérben. A WAIL által létrehozott WARC-fájlokat egy webszerverre feltöltve, és a PyWb-vel leindexelve egy kisebb helyi szolgáltatást is ki lehet alakítani.

Brozzler

A *Brozzler* a *browser* és a *crawler* szavak összevonásából kapta a nevét, mely egyben a működését is kifejezi: a Chrome böngésző motorját kombinálták össze egy olyan robottal, ami követni tudja a weboldalakban talált linkeket. Beleépítették továbbá a *youtube-dl* videóletöltőt is, így kiválóan alkalmas például hírportálok vagy webkettes szolgáltatások archiválására, mert nemcsak felépíti és végiggörgeti az oldalakat, hanem a beágyazott videókat is megpróbálja elindítani és lementeni. A Brozzlert 2016-ban mutatta be az Internet Archive az International Internet Preservation Consortium (IIPC) éves konferenciáján, mint egy újgenerációs

webarchiváló eszközt. Azóta beépítették az előfizetős *Archive-It* szolgáltatásba, és ezt használja az Internet Archive *Save Page Now* funkciója is, amellyel bárki archiválhat egy weboldalt. A program Linux alatt használható, és már a telepítése sem egyszerű. Az aratások paraméterezéséhez nincsen grafikus felület, csak az éppen futó, illetve már lezárult mentések nézhetők meg egy weblapon.



Lezárult aratások listája a Brozzler „műszerfalán”

A robot viselkedését egy YAML-formátumú szövegfájlban tudjuk konfigurálni, majd ezt a fájlt kell megadnunk a Brozzler parancsmódú indításakor paraméterként. Mivel minden oldal előbb betöltődik a böngészőbe, majd egy script végiglapozza, hogy azok az elemek is letöltődjenek, amelyeket csak görgetéskor küld el a webszerver, ezért a Brozzler jóval lassabb, mint a Heritrix vagy a HTTrack robotja, viszont az eredmény sokkal jobb tud lenni. Ez is WARC-formátumban ment, de megjelenítő modul nincs beleépítve.

Webrecorder és Conifer

A korábbiakban már szó esett arról, hogy aratórobotok segítségével lehet igazán hatékonyan a statikus webtartalmat begyűjteni. A számítógépes világháló fejlődésével azonban a honlapok szerkezete komplexebbé vált, egyre több programkomponens összjátéka szükséges a tartalom és a dizájn helyes megjelenítéséhez. Ezt a bonyolultsági szintet számos esetben a robotok konfigurálásával már nem

lehet lekövetni. Arról nem is beszélve, hogy sok olyan közösségimédia-platform és egyéb online ökoszisztéma jelent meg, melyek aratásával a „buta" robotok egyáltalán nem tudnak megbirkózni. Erre a kihívásra is kínálkoznak azonban megoldások. Az alapelgondolás az, hogy ha úgyis böngészőprogramot használunk a webes tartalmak megtekintésére, vessük azt be archiválási célra is. A Google által fejlesztett Chrome felszíne alatt egy *Chromiumnak* nevezett böngészőmotor található, melyre számos egyéb felületet is ráépítettek már, például a Microsoft Edge böngészőt. A Chromiumra nemcsak egy hagyományos böngészőfelületet illeszthetünk rá, hanem olyan programokat is, amelyek az általunk megnézett tartalmak archiválására használhatók. Sőt arra is van mód, hogy ne is legyen egyáltalán grafikus felület, csak egy parancssorból vezérelhető szolgáltatás, mely az archiválást a böngészőmotornak küldött parancsokkal hajtja végre. Persze az archivált tartalom megjelenítéséről is gondoskodni kell, tehát olyan szoftverekre is szükség van, amelyek kiveszik a többnyire tömörített WARC-formátumban tárolt archívumi egységekből a megfelelő fájlokat és összeállítják belőlük a weboldalakat. A továbbiakban két olyan archiváló programot mutatunk be, amelyek egy böngészőmotorra épülnek, és van bennük visszanéző funkció is.

A *Webrecorder* eredetileg egy ingyenesen és akár regisztrálás nélkül is használható online szolgáltatás volt a webrecorder.io címen, melyet a Rhizome nevű non-profit szervezet hozott létre 2016-ban. Minden regisztrált felhasználó kapott egy bizonyos mennyiségű tárhelyet és napi letöltési korlátot. Az online felületen kiválaszthatta, hogy milyen típusú böngésző funkcióit szeretné emulálni, majd megadta az archiválandó webhely URL-címét, betöltődött a kezdőlap, s elindult a kiválasztott böngésző segítségével a végigkattintott oldalak és médiafájlok mentése. Leginkább ahhoz hasonlít ez, mint ahogyan régen televízióadásokat rögzítettünk videomagnóval. Korábban tartozott hozzá egy PC-kre feltelepíthető program is *Webrecorder Player* néven, amellyel a felhőben levő tárhelyéről letöltött WARC-fájlokat a saját gépünkön is vissza lehet nézni.

Tulajdonképpen a rendszer működése napjainkban is az előbb felvázolt elveken alapul. Időközben azonban a videomagnó-funkcióhoz kifejlesztésre került egy Windows, macOS és Linux alatt is futtatható változat *Webrecorder Desktop* néven. Az online és az asztali verzió fejlesztése elvált egymástól: a Webrecorder nevet ez utóbbi viszi tovább, a felhőszolgáltatást a New York-i Metropolitan Múzeum által vezetett nonprofit múzeumi konzorcium működteti *Conifer* néven, a Webrecorder Playert pedig az asztali verziót továbbfejlesztő csapat teljesen újraírta, és átnevezte ReplayWeb.Page-re.

A Conifer és a Webrecorder felülete és funkcionalitása tehát nagyon hasonlít egymáshoz, de tapasztalataink szerint a különféle közösségimédia-platformok nem ugyanolyan hatékonysággal és minőségben archiválhatók a kétféle eszközzel, s a sebességben is jelentős különbségek vannak. A Coniferben, mint már korábban említettük, beállítható, hogy milyen típusú és képességű böngészőt

Egy blogoldal archiválása a Conifer automatikus görgetőfunkciójával

szeretnénk az archiváláshoz használni. A Webrecorderben erre nincs mód, mert csak egyféle Chromium-alapú böngészőmotor áll a rendelkezésünkre. Viszont ennél tudunk mobil böngészőt is emulálni, tehát a kifejezetten mobileszközökre fejlesztett, illetve azokra optimalizált webtartalmak mentésére is lehetőségünk van. A Webrecorderben a „Preview" gombbal a böngészési folyamatot még az archiválás megkezdése előtt el lehet indítani, így például előzetesen be tudunk jelentkezni felhasználónevet és jelszót igénylő szolgáltatásokba úgy, hogy az azonosító adatok nem lesznek a WARC-fájlban eltárolva.

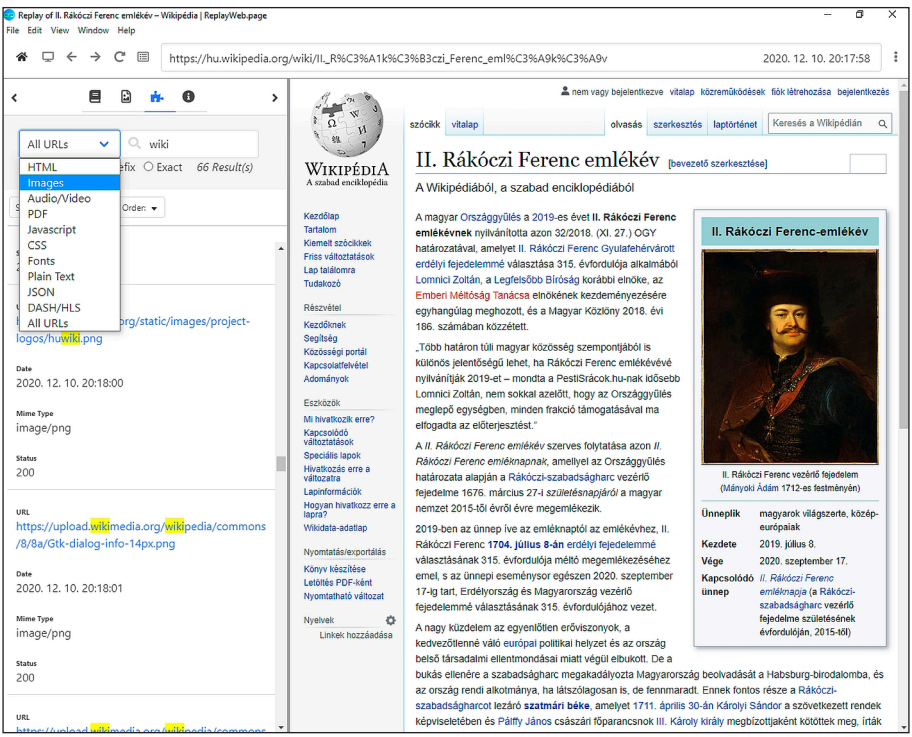
Egyre több olyan webhely van, amelyeknél a tartalom folyamatosan töltődik le a szerverről, ahogyan a felhasználó lefelé görgeti az oldalt (például egyes blogszolgáltatások, hírportálok és közösségi platformok). Ezt az emberi interaktivitást próbálja szimulálni az „Autopilot" nevű funkció, melyet mindkét rendszerbe beépítettek. Az automatikus görgetés mellett az Instagram-, a Twitter-, a Facebook-, a YouTube-, a SlideShare- és a SoundCloud-oldalak esetében további műveleteket is tud ez a „robotpilóta", például megnyitja az egyes posztokat, betölti az összes kommentet és az azokra írt válaszokat, elindítja a videókat stb. A gyakorlatban sajnos – különösen a Facebook esetében – ez a megoldás csak korlátozott eredménnyel jár. A nagy platformokat állandóan továbbfejlesztik, módosí-

tanak a felületen, ezért időről-időre hozzá kellene illeszteni az Autopilot scriptjeit ezekhez a változtatásokhoz. E cikk írásakor például a Coniferben a Facebook-oldalakon csak az automatikus görgetés választható, de az is szinte azonnal lefagy. Egy másik probléma, hogy a nagy méretű weblapoknál a böngésző egy idő után kifut a memóriából, illetve az archiválás közben használt gyorsítótárhelyből, ez is okozhatja az automatizált görgetés leállását.

Az egyes *session*ök, vagyis munkamentek gyűjteményekbe rendezhetők, megjegyzésekkel láthatók el, a fontosabb oldalak kiemelhetők listákba, a Conifer esetében nyilvánossá is tehetők, továbbá letölthetők WARC-formátumban, és persze törölhetők is, akár egyenként, akár gyűjtemény szinten. Mindkét rendszerbe beleintegrálták a visszanéző funkciót is és arra is van lehetőségünk, hogy más programokkal készített WARC-fájlokat feltöltsünk visszanézés céljából az online Conifer-fiókunkba, vagy megnyissunk a Webrecorder programmal.

ReplayWeb.Page

Az archiváló rendszerekbe beépített megjelenítőknél funkciógazdagabb és esetenként jobb eredményt produkál a Webrecorder Player utóda, a kifejezetten a WARC-fájlok nézegetése céljából kifejlesztett *ReplayWeb.Page* program, amely online szolgáltatásként és telepíthető alkalmazásként is használható. Ami

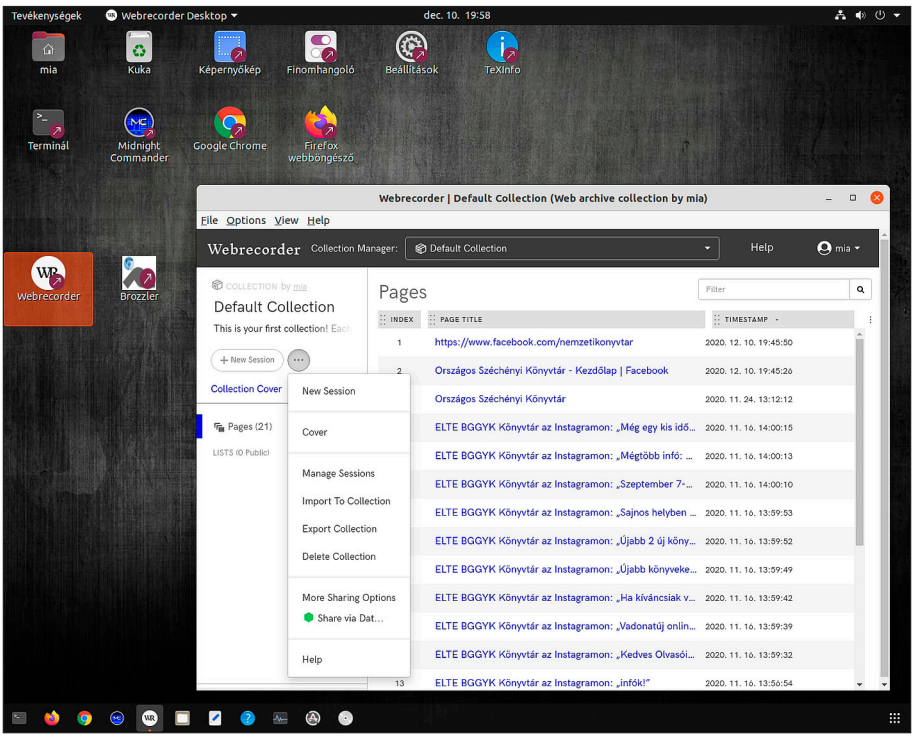


Egy Wikipédia-szócikket tartalmazó WARC-fájl visszanézése a ReplayWeb.Page-ben

az igazi hozzáadott értékét adja a programnak, az az archivált tartalom böngészésén túl az, hogy a WARC-konténerben található fájlokat többféle szempontból elemezni is lehet vele, és azok külön-külön is megnézhetők. Tud például fájltípus és domainnév szerint is szűrni, s ily módon meg lehet vizsgálni, hogy milyen tartalmi elemeket sikerült lementeni. Hátránya, hogy egyszerre csak egy WARC-állományt tud leindexelni majd megnyitni, így egy teljes webarchívum szolgáltatófelületeként nem használható, ellentétben a már korábban említett PyWb-vel.

Virtuális gép

2020 nyarán tesztelési és oktatási célokra összeállítottunk egy virtuális gépet, amivel Windows alatt lehet Ubuntu Linux rendszert futtatni, s így olyan archiváló szoftvereket is ki lehet próbálni, amelyek Windows alatt nem, vagy nem elég hatékonyan működnek. Ennek a megoldásnak további előnye, hogy nem kell két operációs rendszer telepítésével bajlódni ugyanazon a PC-n, illetve a számítógépünkben levő erőforrásokat az általunk kívánt arányban megosztva, párhuzamosan használható a két rendszer. Mi az Oracle *VirtualBox* szoftverét választottuk ki erre a célra, ami szintén egy ingyenes program. A Linux alaprendszerre tele-



A Webrecorder Desktop futtatása az Ubuntu rendszerű virtuális gépen

pítettük a cikkünkben ismertetett Brozzlert és a Webrecorder Desktop alkalmazást (ami Linuxon sokkal gyorsabb, mint a Windows alatti változata), továbbá a PyWb megjelenítőt, a PyWb és a Brozzler által használt Python környezetet, a RethinkDB adatbáziskezelőt, valamint a Chrome és a Firefox böngészőket, a Midnight Commander fájlkezelőt, és néhány más hasznos segédprogramot is kitettünk az asztalra.

A virtuális gép egyetlen 40 GB-os fájlban tölthető le egy felhőtárhelyről, két kis segédfájllal együtt, így egy komplett, előzetesen beállított tesztkörnyezetet tudunk az érdeklődők részére rendelkezésre bocsátani, akiknek elég csak a VirtualBoxot feltelepíteni, majd hozzáadni ahhoz a letöltött virtuális gépet. Szintén egy előre telepített virtuális gép formájában lehet tesztelni a *Web Curator Tool* nevű keretrendszert, ami a Heritrix robotjának vezérlésére szolgál és akár egy nagyobb intézményi webarchívumot is lehet vele menedzselni. A Web Curator Toolt 2006-ban a National Library of New Zealand és a British Library kezdte el fejleszteni, és az elmúlt évben a projekt új lendületet kapott az IIPC támogatásának köszönhetően.

Akit érdekelnek ezek a virtuális gépek, az írjon a *mia@mek.oszk.hu* e-mail-címre, és megadjuk a letöltési helyet és a rendszergazdai jelszót, illetve segítünk a beüzemelésben. Nagy örömünkre szolgált, hogy az idei webarchiválási konferencia után már néhány könyvtárból érkezett is ilyen megkeresés, ami azt jelzi, hogy van törekvés a digitálisan születő kultúra megőrzésére más közgyűjteményekben is.

Ez a cikk a 2020. december 2-án videókonferencia formájában megtartott „404 Not Found" – Ki őrzi meg az Internetet? című rendezvény délutáni workshopján bemutatott szoftvereket ismertette. A workshophoz tartozó linkgyűjtemény megtalálható a konferencia weboldalán: webarchivum.oszk.hu/404-workshop-2020-december-2/. További információk és internetes források az Országos Széchényi Könyvtár Webarchívumának honlapján a „Szakembereknek" menüpont alatt elérhető tananyagban és wikiben vannak. A gyakorlati ismereteket pedig a Könyvtári Intézet által félévente meghirdetett – legközelebb 2021 februárjában online elvégezhető – tanfolyam keretében sajátíthatják el az érdeklődők.