TMT CIKKEK
https://doi.org/10.3311/tmt.13242

Webarchiválás a nemzeti könyvtárban

Helyzetkép hat év után

Drótos László

A cikk a digitálisan születő kultúra, azon belül a webtartalom megőrzésének kérdéseivel foglalkozik. A szerző az Országos Széchényi Könyvtár (OSZK) webarchívumának elmúlt hat évi tapasztalatai alapján mutatja be, hogy milyen döntéseket kell hozni a gyűjtőkör kialakítása során, miért fontos a „born digital" dokumentumok gyűjtése, milyen feladatai vannak ezen a téren a különböző szervezeteknek és intézményeknek, elsősorban a közgyűjteményeknek. Az írás második fele az OSZK-ban folyó webarchiválás jelenlegi gyakorlatát ismerteti, végül pedig kitér a lehetséges jövőbeli feladatokra is. A cikk a székesfehérvári Vörösmarty Mihály Könyvtár „Távcső a történelemre" című konferenciáján 2023. április 24-én elhangzott előadás szerkesztett változata.

digitális kultúra, webarchiválás, hosszú távú megőrzés, nemzeti könyvtár

Mottó:
Internet örök,
Mély méltósággal őrzi
Adatok árnya.

Bevezetés

A mottóként használt haikut a ChatGPT mögött álló mesterséges intelligencia írta a digitálisan születő kultúra megőrzésének fontosságáról. Ha megpróbáljuk megérteni, hogy mire gondolt a „költő", akkor két fontos gondolatot fedezhetünk fel ebben a három rövid sorban. Az első, hogy az internetes tartalom archiválása egyfajta örökkévalóságot biztosít az elsősorban valós idejű információmegosztásra tervezett és használt világhálónak. A vers második fele viszont arra hívja fel a figyelmet, hogy ez a médium valójában megőrizhetetlen, csak halvány vetületét tudjuk eltenni annak a hatalmas adattömegnek, ami az interneten folyamatosan keletkezik, megváltozik és eltűnik.

Az Országos Széchényi Könyvtár 2017-ben kezdett el foglalkozni a webarchiválással és ennek a cikknek az írásakor éppen az első hat év során szerzett tapasztalatok összegzése, a kialakult infrastruktúra és munkafolyamatok újratervezése zajlik. Az alábbiakban ezekről lesz szó, kiegészítve egy rövid kitekintéssel a többi *born digital* műfajra is.

Mit?

A digitális információhordozókon, legyen szó az olyan kézbe fogható eszközökről, mint az optikai-, a mágneses- és a flash-elvű tárolók, vagy pedig az internet felhőjében levő szerverekről, nagyon sokféle műfaj található – ez az egyik fő nehézség a digitális megőrzésben. Már a csoportosítás sem egyszerű, mert egy dokumentumon belül is keveredhet mondjuk szöveg, videó és programkód, gondoljunk csak egy Prezi-ből exportált prezentációra. Az itt következő listában a legjellemzőbb tulajdonság szerint vannak rendezve az ismertebb típusok, de természetesen mindig lehetnek kivételek, mert például a blog eredetileg egy naplószerű szöveges műfaj, de vannak már kizárólag képeket közlő blogok is. És persze mindegyik kategória folytatható további – esetleg még meg sem született – műfajokkal, ezt jelzi a három pont a sorok végén.

- Írott kommunikáció: **elektronikus levél**, azonnali üzenet, **internetes fórum**, **tweet** …
- Írott dokumentum: ***szövegfájl***, ***e-könyv***, ***e-folyóirat***, ***blog***, **wiki** …

Drótos, L. *Webarchiválás a nemzeti könyvtárban. Helyzetkép hat év után*, Tudományos és Műszaki Tájékoztatás, 70(2), p. 190–200, 2023.
https://doi.org/10.3311/tmt.13242

- Adathalmaz: táblázat, adatbázis, strukturált szövegfájl …
- Kép: ***digitális fotó***, számítógépes grafika és animáció, ***infografika***, mém, térkép …
- Videó: **rögzített felvétel** (pl. vlog), élő közvetítés, videokonferencia …
- Hang: **rögzített felvétel** (pl. ***podcast***), élő közvetítés (pl. online rádió) …
- Térhatású: 3D modell, VR játék, metaverzum …
- Program: operációs rendszer, felhasználói szoftver, mobil applikáció, számítógépes játék, forráskód …
- Vegyes: ***honlap***, ***hírportál***, ***virtuális kiállítás***, **közösségi média**, digitális műalkotás, ***prezentáció***, e-learning tananyag …

A **kivastagított** műfajok digitális megőrzésével már van valami tapasztalatunk az OSZK-ban, de csak a ***dőlt betűsek*** esetében mondhatjuk el, hogy több éve foglalkozunk a gyűjtésükkel, feldolgozásukkal és szolgáltatásukkal a Magyar Elektronikus Könyvtár, az Elektronikus Periodika Archívum és Adatbázis, a Digitális Képarchívum, valamint az OSZK Webarchívum keretében. (Az első háromnál fontos megjegyezni, hogy az eleve digitális formában készült dokumentumok mellett sok digitalizált anyagot is tartalmaznak, így ezek hibrid gyűjtemények.)

A fejezetcímben szereplő „mit archiváljunk?" kérdésre a gyűjtőkör definiálásával adhatunk választ, melyben a gyűjteni szándékozott műfaj vagy műfajok mellett számos egyéb szempontot is rögzíteni kell. Ilyen lehet például a téma, a nyelv, a formátum, a minőség, a nyilvánossági és veszélyeztetettségi szint, vagy akár a megcélzott felhasználói kör. Megfogalmazhatunk kizárási szabályokat is, például, hogy vírusos fájlokat vagy spamet nem archiválunk (bár akár ezek is érdekes kutatási források lehetnek). Dönteni kell továbbá arról, hogy egy digitálisan született mű esetében megelégszünk-e csak a tartalom eltárolásával, vagy szeretnénk a formát, az eredeti külalakot is megőrizni, sőt akár az interaktív funkciókat is működőképes állapotban tartani, és esetleg az eredeti dokumentum időbeli változását is lekövetni valahogyan? Webes tartalmak esetében a környezet megőrzése is eldöntendő kérdés. Mit kezdjünk az oldalakba beágyazott vagy belinkelt elemekkel, az automatikusan ajánlott további forrásokkal, a reklámokkal, a kommentekkel és lájkokkal? Minél teljesebb és hűségesebb másolatokra törekszünk, annál nagyobb lesz a feladat technikailag és sokszor élőmunka szempontból is.

A hosszabb időtávú – akár csupán pár évtizednyi – megőrzésnél felmerülhet az elavuló formátumok problémája is. Néhány ilyennel már mi is találkoztunk az elektronikus könyvtár és a webarchívum története során (*DOC*, *LIT*, *DjVu*, *Flash Video*, *SilverLight*, *Java applet*). Ilyen esetben háromféle megoldás jöhet szóba: vagy időállóbb formátumra kell konvertálni a fájlokat, vagy meg kell őrizni az eredeti környezetet (akár az operációs rendszer, sőt az azt futtató hardver szintjéig), vagy emulálni kell azt az eszközt, amivel a dokumentum annak idején megjeleníthető volt. Utóbbi a webarchívumok esetében elsősorban a böngészőt jelenti, és vannak is ilyen rendszerek (pl. a felhőszolgáltatásként és saját gépre telepíthető formában is létező Conifer).

A digitális dokumentumok és esetleg a technikai környezet megőrzése mellett a metaadatok előállítása és biztonságos tárolása is nagyon fontos, hogy értelmezhető legyen az archívumban levő sok giga- vagy terabájt.

A nemzeti könyvtárban a 2021. január 1-én életbe lépett 626/2020 számú kormányrendelet alapján folyik a webhelyek archiválása. 2022 nyarán az OSZK megújuló Gyűjtőköri Szabályzatához írt szövegjavaslatban definiáltuk a „born digital hungarikum" fogalmát és azon belül a gyűjtendő műfajok körét, beleértve a webarchívum feladatait is. Utóbbihoz készítettünk egy önálló, 10 oldalas szabályzatot is, amiben részletesebben meghatároztuk az egyes tartalomtípusok és részgyűjtemények válogatási elveit, valamint a velük kapcsolatos munkafolyamatokat.

> „Az OSZK webarchívumának gyűjtőkörébe tartozik a magyar webtérben létező vagy valaha létezett, nyilvánosan közzétett digitális tartalmak összessége, beleértve tehát azokat is, amelyek már az élő weben nem elérhetők, de valahol még megőrződtek.
>
> A magyar webtér – nyelvtől és tulajdonostól függetlenül – kiterjed a .hu tartomány alá bejegyzett doméneken és aldoméneken lévő webhelyekre, valamint a külföldi és nemzetközi legfelső szintű domének alatt magyar természetes vagy jogi személyek által létrehozott webes tartalmakra,

továbbá minden olyan oldalra a weben, amelyet magyar célközönségnek (is) szánnak, vagyis magyar nyelvű vagy magyar vonatkozása van.
Nyilvánosan közzétettnek minősül minden olyan digitális tartalom, melyet az előállítója bárki számára online elérhetővé tett, beleértve a regisztrációt vagy előfizetést igénylő szolgáltatásokat is. Nem tartozik viszont ebbe a körbe a magáncélra, vagy csak adott kör (pl. csoport, szervezet, intézmény) tagjai számára az internetre vagy intranetre feltöltött tartalom. Ilyeneket a webarchívum aktívan nem gyűjt, de beadás esetén ezeket is befogadja."[1]

Miért?

A számítógépekkel és mobil eszközökkel ellátott társadalmakban az internet mindig és mindenütt jelenlevőnek tűnik, ami egyrészt azt a hamis illúziót kelti, hogy ez mindig így is marad, másrészt, hogy ami oda felkerül, az többé nem tüntethető el. Ezért sokan nem értik, hogy miért van szükség a webarchiválásra, vagy ha vannak is mellette érvek, akkor miért nem elég erre a világ legrégibb és legnagyobb gyűjteménye, az amerikai Internet Archive? Mi értelme van a nemzeti szintű webarchívumoknak és a még azoknál is jóval kisebb intézményi vagy magángyűjteményeknek?

Az interneten a fizikai világunk folyamatos leképezése zajlik, legyen szó globális eseményekről vagy jelentéktelen magánügyekről. Mivel számos műfaj csak digitális formában létezik, a világháló elterjedése óta eltelt három és a következő ki tudja még hány évtized dokumentálása szempontjából fontos lenne minél többet megőrizni ezekből a különleges információforrásokból, hogy a jelen és a jövő generációi értelmezni tudják azokat a dolgokat, melyek civilizációnknak ebben a fejlődési szakaszában történtek és történnek.

A digitális állományok sérülékenyek, könnyen törölhetők vagy megváltoztathatók szándékosan vagy véletlenül. Ha egy born digital dokumentum elveszett vagy megjeleníthetetlenné vált, nem lehet újradigitalizálni, ellentétben az analóg hordozókról digitalizált anyagokkal, melyek újra előállíthatók – a technika előrehaladtával akár a korábbinál jobb minőségben is – mindaddig, amíg az eredetiből létezik legalább egy példány. Így süllyedtek el a múltnak kútjában gyakorlatilag nyom nélkül az 1970–1980-as évek és az 1990-es évtized első felének online világai: a telefonvonalas BBS-ek és a francia Minitel, a nálunk is népszerű X.25 és Bitnet hálózatok szolgáltatásai, valamint a korai internet FTP, Gopher és WAIS szerverei.

Egy másik, már a jelenben is súlyos probléma, hogy az internet egy olyan „világkönyvtár", amiben nincs katalógus és nincs raktári jelzet, sem stabil lelőhely, így az online dokumentumok nem azonosíthatók be és nem hivatkozhatók biztonságosan. Ha egy vagy több *mementót* készítünk róluk, vagyis adott időpont(ok)ban lementjük őket, majd ezeket a mentéseket egyedi azonosítóval és néhány további metaadattal látjuk el, akkor már jobban hasonlítanak valódi könyvtári objektumokra. Az archivált tartalmak az élő webnél sokkal inkább alkalmasak tudományos publikációkban és tananyagokban való hivatkozáshoz, vagy mondjuk bizonyítéknak vitás ügyekben.

Mivel az internetes tartalom eleve digitális, ezért számítógéppel közvetlenül feldolgozható, kutatható, rendkívül értékes „nyersanyag". Ez a hatalmas információtömeg képezi az alapját a napjainkban zajló mesterséges intelligenciarobbanás mögött álló gépi tanuló rendszereknek. Nem véletlen, hogy e cikk írásakor a világ tíz legértékesebb vállalata közül hét digitális technológiák fejlesztésével (is) foglalkozik és többen élenjárók az adatok gyűjtése és feldolgozása terén is (Apple, Microsoft, Google, Amazon, Meta/Facebook). Önmagában ez a tény, vagyis hogy a digitális információ már az olajnál is értékesebb, indokolja a kisebb-nagyobb webarchívumok szükségességét.

Ki?

Az *Internet Archive*[2] 2023. május végén több mint 800 milliárd mementót, vagyis egyszer vagy többször lementett webes dokumentumot tartalmazott, de még ez is csupán kis töredéke a világháló nyilvános részének, ami pedig szintén csak a jéghegy csúcsa, mert a deep és dark webre egyáltalán nem jutnak el az archiváló robotok. A webarchívum csak egy része a San Francisco-ban székelő nonprofit szervezet által gondozott hatalmas gyűjteménynek, ami milliószámra tartalmaz digitalizált és digitálisan született könyveket, képeket, hang- és videofelvételeket, valamint szoftvereket is. Az Internet Archive méretei miatt a legtöbb mun-

kafolyamat teljesen automatizált vagy külső partnerek és a támogató közösség felé kiszervezett, aminek számos előnye mellett megvannak a hátrányai is. Például nincsenek egyértelmű válogatási és gyűjteményépítési elvek, hiányoznak vagy nem egységesek a metaadatok, a nyilvános szolgáltatásba kikerülő tartalom jogszerűségét nem vizsgálják, csak az értesítés utáni eltávolítás elvét követik. A webarchívum méretéből fakadó informatikai kihívás mellett személyiségi és szerzői jogi okai is lehetnek annak, hogy a *Wayback Machine*[3] nevű szolgáltató felületen nincs teljes szövegű keresési lehetőség, az URL-en kívül csak az oldalra mutató linkek szövegében előforduló szavak alapján lehet megtalálni egy weblapot.

Digitalizált dokumentumoknál is indokolt lehet esetenként a más eszközzel vagy más beállításokkal történő újraszkennelés, a komplex és gyakran változó webes tartalmak esetében viszont sokkal inkább igaz az, hogy minél több mentés készül róluk eltérő technológiákkal és időpontokban, annál nagyobb eséllyel marad meg belőlük valami az utókornak. Ezért is jó, hogy az Internet Archive mellett a világban számos helyen épülnek webarchívumok.

A legfontosabb szereplők természetesen a közgyűjtemények, melyeknél jó esetben megvan a szükséges szakértelem, valamint a jogi és informatikai környezet, és intézményileg stabilabbak, mint egy adományokból és szponzori támogatásokból élő civil szervezet. De fontos a non-profit szféra aktivitása is, akárcsak az üzleti alapon működő cégeké, mert előbbieknél az elkötelezettség és a nyitottság, utóbbiaknál a technikai tudás és a pénzügyi háttér jelenthet előnyt. Az egyetemeken és kutatóintézetekben működő tudományos műhelyeknek szintén lehet feladatuk a digitális megőrzésben. Ezek a szakterületükhöz kapcsolódóan építhetnek kisebb és többnyire csak rövid távra szánt gyűjteményeket vagy adatbázisokat a webről származó tartalomból. Végül említsük meg jó példaként azokat a tartalomgazdákat és -szolgáltatókat is, akik vagy maguk oldják meg azt, hogy a lecserélt vagy már nem fejlesztett régi webhelyeik azért továbbra is elérhetők maradjanak egy másik szerveren, vagy beküldik őket a helyileg illetékes webarchívumba, vagy értesítik annak kezelőit, hogy a leállás előtt még legyen lehetőségük lementeni az utolsó állapotot.

A közgyűjteménytípusok között is kialakítható egy munkamegosztás a born digital anyagok megőrzésében. Van néhány ország, ahol a nemzeti audiovizuális archívum gyűjti a rögzített és a sugárzott hang- és videotartalmakat. Arra is van példa, hogy az állami, kormányzati webhelyek archiválása a nemzeti levéltár feladata. A múzeumok esetében pedig a kortárs alkotók online felületeinek, elektronikus levelezésének, illetve digitális alkotásainak gyűjtése és megőrzése az egyik jellemző tevékenység, de hagyományos gyűjteményeik kiegészítéseként is foglalkoznak webarchiválással, például események dokumentálása céljából.

A könyvtári szférán belül szintén tovább osztható a feladat. Míg a nemzeti könyvtár a nemzeti webtérnek minél szélesebb részéről próbál lenyomatokat készíteni, addig a szak- és felsőoktatási könyvtárak a saját gyűjtőkörüknek megfelelő webes tartalmakból építenek kisebb, de jobban megválogatott és feldolgozott archívumokat, a regionális hatókörű közkönyvtárak pedig értelemszerűen elsősorban a helyi vonatkozású online források összegyűjtésében és megőrzésében érdekeltek.

Az OSZK-ban jelenleg 3 munkatárs foglalkozik közvetlenül a webarchiválással a 2022-ben létrejött Digitális Bölcsészeti Központ (DBK) keretén belül. Mivel ez a létszám korábban sem volt sokkal nagyobb, ezért már a projekt kezdetekor egy széles együttműködési kör kialakításán gondolkoztunk. A webarchívum honlapján[4] a „Szakembereknek” menüpont alatt található egy felhívás ezzel kapcsolatban és ugyancsak itt helyeztünk el önképzéshez használható forrásokat: egy magyar nyelvű wikit, egy nemzetközi szakirodalmi bibliográfiát és annak „Az internet archiválása mint közgyűjteményi feladat” című tanfolyamnak az anyagát, melyet a Könyvtári Intézet minden tavasszal és ősszel meghirdet. Az első komolyabb együttműködésekre a 2019-es KDS pályázaton nyert könyvtárakkal került sor, melyek besegítettek a megyéjükhöz kapcsolódó webhelyek címeinek gyűjtésébe. Azóta is folyamatosan keressük a kapcsolatot elsősorban a közkönyvtárakkal, az egyetemi könyvtárakkal és tanszékekkel. Arra is van már példa, hogy helyi archívumok kialakításához tudtunk szakmai segítséget adni (Aranybulla-webarchívum[5], *Karikó Katalin* virtuális kiállítás[6]).

A nemzetközi kapcsolatok a tudásmegosztás és a jó gyakorlatok megismerése miatt nagyon hasznosak. Az OSZK 2018 óta tagja a webarchiválással foglalkozó szervezeteket és intézményeket tömörítő *International Internet Preservation Consortium*-nak[7], részt veszünk az IIPC munkacsoportjaiban és rendezvényein. Ugyancsak aktív tagjai voltunk a 2023-ban záruló *WARCnet*[8] projektnek, ami a webarchívumok kutatási célú hasznosításával foglalkozott. Kapcsolatban vagyunk az Internet Archive-val, amely egy teljes szövegű keresőt alakított ki a .hu doménről archivált anyagához. A dán királyi könyvtárban fejlesztett *SolrWayback* szoftver első tesztelői között voltunk és az OSZK nyilvános gyűjteményével demózzák a rendszert a készítői. Most formálódik a kapcsolat a luxemburgi nemzeti könyvtárral, ahol egy már üzemszerűen működő rendszert alakítottak ki a fizetőfal (*paywall*) mögött levő híroldalak archiválására, ami számunkra még egy megoldandó feladat.

Hogyan?

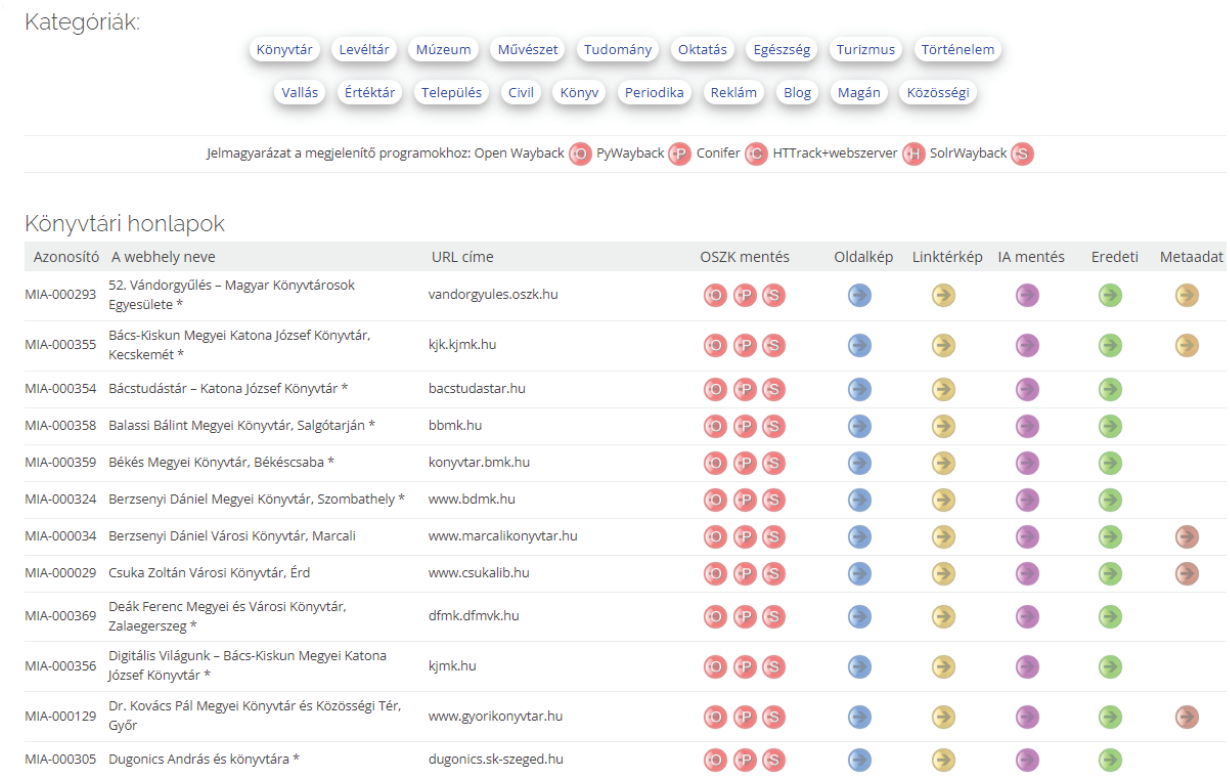
A „Hogyan őrizzük meg a webet?" kérdésre nincs egyszerű válasz. Az eredetileg statikus HTML és egyéb fájlokból álló online hipertext rendszerből mára egy dinamikusan generálódó multimédia univerzum lett. A fontos weboldalak nagy része már inkább programkód és adatállomány, mintsem dokumentum, és erősen kötődik a szolgáltató platformhoz, valamint a megjelenítést végző böngészők aktuális képességeihez. Ezért a „hogyan"-t megelőzi az első fejezetben feltett „mit" kérdés eldöntése, és attól függően lehet megválasztani a megfelelő eszközt és munkafolyamatot. Ezért van az, hogy a webarchívumok igen sokfélék és még a nemzeti szintűek között is elég lényeges különbségek vannak a használt módszerekben. Pozitív fejlemény viszont, hogy a komolyabb gyűjteményeknél már mindenhol a nemzetközi szabványnak számító WARC[9] formátumban tárolják a letöltött tartalmat, így (technikailag) megvan a lehetőség ezek összekapcsolására és a hosszú távú fennmaradásukra. További biztató jel az IIPC aktív koordináló tevékenysége a nyílt forráskódú archiváló és megjelenítő szoftverek fejlesztésében. Ez egy véget nem érő feladat, mert az állandóan változó internetes technológiák miatt mozgó célpontra kell lőni: mire kialakul egy megbízhatóan és hatékonyan működő megoldás, a világháló generációt vált és újra kell tervezni az archiváló eszközparkot.

Mivel sok közgyűjteményben és tudományos műhelyben nincs meg, vagy hosszabb távon nem tartható fenn a szükséges informatikai szakértelem és infrastruktúra, ezért még nemzeti könyvtárak esetében is van arra példa, hogy a webarchívum műszaki részét kiszervezik külső cégnek (pl. az Internet Archive fizetős *Archive-IT*[10] szolgáltatásába), ahol biztosított a megbízható működés és az archiváló technológia frissítése az aktuális fejlettségi szintre.

Az OSZK 2017-ben saját informatikai rendszer kialakítása mellett döntött, ingyenes szoftverekkel. 2022-ben már négy, a Kormányzati Informatikai Fejlesztési Ügynökség által üzemeltetett C4e felhőben futó szervert használt a webarchívum: egyet a honlap és a speciális különgyűjtemények szolgáltatására, egyet a nyilvános archívumhoz, kettőt pedig a gyűjtemény nagy részét kitevő zárt állomány bővítésére. Utóbbiak közül az egyiken csak a tömeges aratásokat végző robot fut. A zárt archívum 2023 májusában éppen egy – a korábbinál kétszer nagyobb – 300 terabájtos tárolóra költözik. A publikus szerver mögött jelenleg 10 terabájtnyi háttértár van, de ennek még csak az ötöde telt be.

Az elsősorban demonstrációs célokat szolgáló, honlapokat, blogokat és időszaki kiadványokat is tartalmazó kis gyűjteményben[11] több mint 350 archivált webhely nézhető meg nyilvánosan, melyekhez van egy teljes szövegű kereső is. (*1. ábra*) Szintén elérhető a honlapunkról az OSZK közel száz online szolgáltatásának legalább egyszeri mentése, köztük sok virtuális kiállítás.[12] Mi is készítettünk két speciális válogatást, melyekben vegyesen van webtartalom és más digitálisan született, valamint digitalizált dokumentum. Az egyik a 2019–2020-as II. Rákóczi Ferenc Emlékévben készült, ebben több mint 300 weblap található.[13] A másik pedig az OSZK alapításának 220. évfordulója alkalmából jött létre és Széchényi Ferenc életét és hagyatékát mutatja be.[14] Ennek az anyaga nagyrészt kép, de van benne 69 weboldal is, részben a könyvtáralapításról szóló cikkek.

A nyilvános szerveren levő mentések mind egyediek, ami azt jelenti, hogy önálló egységekként és gyakran egyedi beállításokkal lettek archi-



Kategóriák:

Könyvtár Levéltár Múzeum Művészet Tudomány Oktatás Egészség Turizmus Történelem
Vallás Értéktár Település Civil Könyv Periodika Reklám Blog Magán Közösségi

Jelmagyarázat a megjelenítő programokhoz: Open Wayback PyWayback Conifer HTTrack+webszerver SolrWayback

Könyvtári honlapok

Azonosító	A webhely neve	URL címe	OSZK mentés	Oldalkép	Linktérkép	IA mentés	Eredeti	Metaadat
MIA-000293	52. Vándorgyűlés – Magyar Könyvtárosok Egyesülete *	vandorgyules.oszk.hu						
MIA-000355	Bács-Kiskun Megyei Katona József Könyvtár, Kecskemét *	kjk.kjmk.hu						
MIA-000354	Bácstudástár – Katona József Könyvtár *	bacstudastar.hu						
MIA-000358	Balassi Bálint Megyei Könyvtár, Salgótarján *	bbmk.hu						
MIA-000359	Békés Megyei Könyvtár, Békéscsaba *	konyvtar.bmk.hu						
MIA-000324	Berzsenyi Dániel Megyei Könyvtár, Szombathely *	www.bdmk.hu						
MIA-000034	Berzsenyi Dániel Városi Könyvtár, Marcali	www.marcalikonyvtar.hu						
MIA-000029	Csuka Zoltán Városi Könyvtár, Érd	www.csukalib.hu						
MIA-000369	Deák Ferenc Megyei és Városi Könyvtár, Zalaegerszeg *	dfmk.dfmvk.hu						
MIA-000356	Digitális Világunk – Bács-Kiskun Megyei Katona József Könyvtár *	kjmk.hu						
MIA-000129	Dr. Kovács Pál Megyei Könyvtár és Közösségi Tér, Győr	www.gyorikonyvtar.hu						
MIA-000305	Dugonics András és könyvtára *	dugonics.sk-szeged.hu						

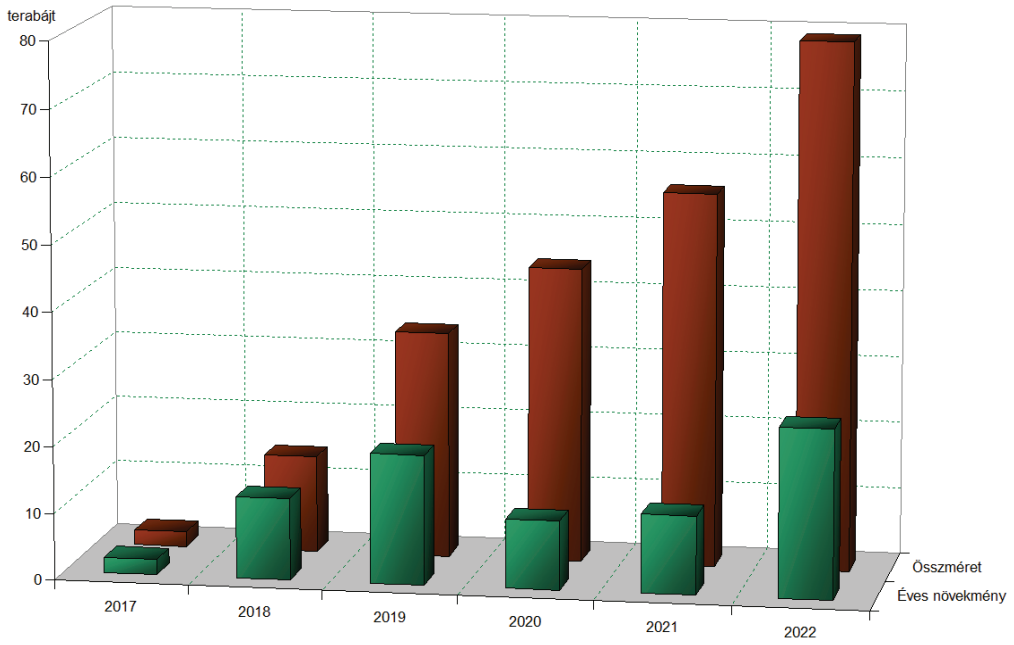
1. ábra A nyilvános webarchívum részlete

válva egy vagy több alkalommal. Többségük az Internet Archive által fejlesztett *Heritrix* robotot vezérlő *Web Curator Tool* (WCT) nevű rendszerrel készült, de használunk alternatív eszközöket is a problémás webhelyek minél jobb minőségben való megőrzése érdekében (*Conifer*, *Webrecorder*, *ArchiveWeb.page*, *PyWb*, *HTTrack*).

A csak az OSZK olvasótermében levő dedikált gépekről elérhető archívumban viszont – a közösségi médiatartalmak és a podcastok kivételével – főként tömeges aratások anyagai vannak, melyek több száz, több ezer, vagy akár több százezer webhelyre terjednek ki egy időben. Ezekhez is a Heritrixet használjuk, az aratások paraméterezése pedig egy saját fejlesztésű űrlapos felületen (*Kaptafa*) történik. Mivel a scriptekkel generált híroldalak ezzel a robottal nem, vagy csak töredékesen menthetők, ezért néhány nagyobb hírportált a *Brozzler* nevű, böngésző-alapú eszközzel is mentünk.

A zárt archívum is több részre oszlik, de egyben is böngészhető és kereshető. Vannak emberi munkával válogatott tematikus és műfaji részgyűjteményeink. Előbbiekből 2023 elején 17 volt, összesen kb. 60 ezer webhellyel. Utóbbiakból pedig hatfélét hoztunk létre eddig, köztük az e-periodikák nyilvántartását több mint 7 ezer címmel, és a podcastokét, amiben mintegy 1500 csatorna adata van (ezekről 2022-ben 75 ezer adás hangfájljait töltöttük le). A robottal aratható részgyűjteményeket nagyjából negyedéves gyakorisággal mentjük. Fontosabb közéleti, sport vagy egyéb jellegű események alkalmával speciális válogatásokat hozunk létre, melyeket sűrűbben, de csak egy adott időszakban aratunk, ilyenekből már 18 készült (pl. koronavírus-járvány, olimpiák, választások). Van továbbá egy földrajzi helyhez kötött különgyűjteményünk is, mert az orosz-ukrán háború kitörése után elkezdtük menteni a Kárpátalján működő webhelyeket. A legnagyobb állomány a webtér szintű aratások során keletkezik. Ezek félévente futnak és minden eddig emberi munkával gyűjtött, vagy automatikus módszerekkel felderített, vagy más forrásból átvett URL-címre kiterjednek, amelyeken lehet magyar vonatkozású tartalom. Ez a címlista jelenleg kb. 1,37 millió domént és aldomént tartalmaz. A zárt szerveren 2017 óta összegyűjtött WARC-fájlok mérete meghaladja a 80 terabájtot. (*2. ábra*)

Mivel a webböngészők a WARC-formátumot közvetlenül nem tudják értelmezni, ezért speciális megjelenítő programokra van szükség az archivált tar-



	2017	2018	2019	2020	2021	2022
Éves növekmény	2,3	12,2	19,4	10,3	11,8	25,4
Összméret	2,3	14,5	33,9	44,2	56,0	79,1

2. ábra A zárt archívum évenkénti és összesített növekedése

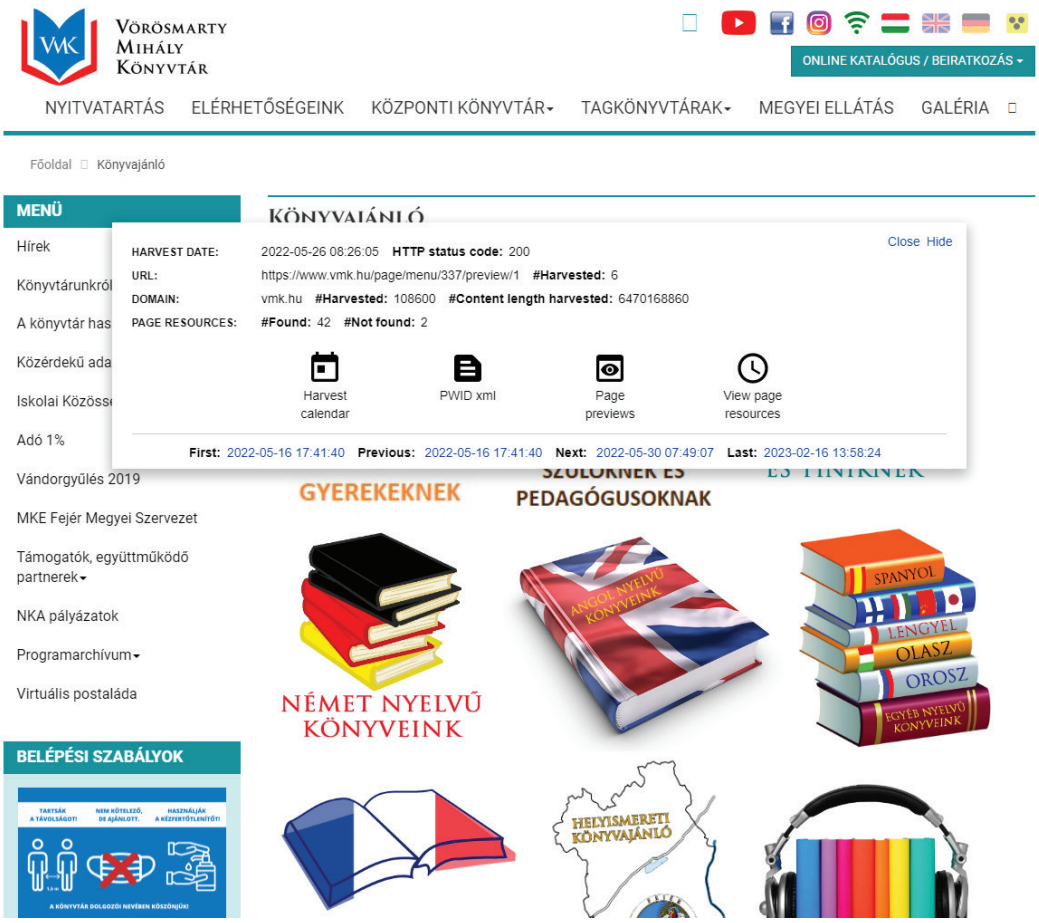
talom visszanézéséhez. Ez lehet a Windows alatt is futó *Webrecorder Player*, vagy a Chrome kiegészítőként és online is elérhető *ReplayWeb.page*, vagy az archiváláshoz is alkalmas *Conifer* felhőszolgáltatás, de ezekkel csak egy-egy WARC-fájl tartalma jeleníthető meg. Ezért mi a szervereinken a Wayback Machine-hoz hasonló *OpenWayback*-et és a nála korszerűbb, Python nyelven írt *PyWb*-t használjuk, továbbá a kép- és teljes szövegű keresőt, valamint vizualizáló funkciókat is tartalmazó *SolrWayback* rendszert. (*3. ábra*) A *HTTRack* programmal készített, fájlrendszerben tárolt mentések pedig a webszerverünkön keresztül nézhetők vissza.

A webarchívum nyilvános és a zárt szerverén különféle böngésző- és keresőfunkciók állnak a felhasználók rendelkezésére. A metaadatok és a webhelyek kezdőlapjáról készült oldalképek kicsinyített verziói minden esetben publikusak, akárcsak a tömeges aratások naplóiból gyártott statisztikák. A találati listákban az alternatív megjelenítőkre, valamint az eredeti webhelyre mutató ikonok mellett feltüntetjük az Internet Archive mentéseihez vezető linket is, mert ezek más időpontokban és más minőségben készültek, így jól kiegészítik a mi archív példányainkat.

A webarchívummal kapcsolatos főbb munkafolyamatok a következők: a robot indításához szükséges URL-címek válogatása, az aratások paraméterezése és elindítása vagy beütemezése, a keletkezett WARC-fájlok indexelése a kereséshez, az eredeti webhelyek kezdőoldalának „lefényképezése”, a nyilvános gyűjtemény esetén minőségellenőrzés és részletes metaadatolás, a zárt archívumba kerülő tömeges aratásoknál pedig statisztikakészítés. További feladat lehet az engedélykérés a publikus szolgáltatáshoz, valamint a hibás mentéseknél az aratási beállítások finomhangolása vagy más archiváló szoftver használata. A DBK létrejöttével elkezdődött a lementett tartalom elemzése is. Ennek első eredménye az orosz-ukrán háborúval kapcsolatos hírek feldolgozása, melyeket 75 hazai és határon túli magyar hírportálról mentünk 2022. február 21. óta. A gyűjtemény jogi okok miatt nem nyilvános, de keresni lehet benne a honlapunkon[15], a hírek szövegéből készített statisztikák és ábrák pedig a dHUpla nevű Digitális Bölcsészeti Platformon érhetők el.[16]

Merre?

A 2017–2019-es tanuló fázis után sikerült összerakni egy üzemszerűen működő, tömeges archivá-



3. ábra Egy archivált könyvtári weboldal a SolrWayback megjelenítőben

lásra és speciális mentésekre is képes infrastruktúrát, valamint kialakítani a válogatás, a minőségellenőrzés, a metaadatolás és statisztikakészítés munkafolyamatait. 2023-ra viszont elértünk olyan határokat, melyek miatt újra kell gondolnunk és át kell alakítanunk a jelenlegi rendszer egyes részeit. Vannak továbbá olyan területek is, amelyekkel eddig még egyáltalán nem foglalkoztunk, de érdemes volna ezekben az irányokban is elkezdeni legalább a kísérletezést.

Újratervezés:

– Az archívum mérete miatt a jelenlegi gépeken már teljesítménygondok vannak, amiknek az egyik fő oka, hogy az arató szerver és a WARC-fájlok indexelését és szolgáltatását végző szerver ugyanazt a tárterületet használja és a sok lemezművelet lassítja a folyamatokat. Megoldásként felmerült egy puffer terület kialakítása vagy a műveletek szétválasztása, a nagy méretű webtér aratásoknál pedig egy további szerver beállítása.

– Tervezés alatt van az OSZK hosszú távú digitális megőrzésre alkalmas rendszere, melybe a digitalizált dokumentumok mellett a webarchívum anyagát is fel kellene majd tölteni. Utóbbinál viszont nem egyértelmű, hogy mi legyen az a raktári egység, amit a nyílt archívumi információs rendszer (OAIS) szabványnak megfelelően be lehet majd küldeni ebbe a raktárba, hiszen a tömeges aratások során az egy webhelyhez tartozó fájlok különböző WARC-konténerekbe kerülnek, összekeveredve a robot által ugyanakkor más webszerverekről lekért állományokkal. Ráadásul helytakarékosságból mindig csak az új vagy az előző mentés óta megváltozott fájlok kerülnek eltárolásra, és ez a deduplikáció tovább bonyolítja azt, hogy hol is tárolódik egy adott webhely anyaga. Ötletként felmerült, hogy a jelenlegi tematikus és műfaji részgyűjtemények tömeges aratása helyett át kellene térni az egyedi mentésekre, ahogyan ez jelenleg a nyilvános gyűjteménynél történik. Erre a célra vagy a WCT-t vagy a Kaptafa egy módosított változatát lehetne használni.

– Mivel a régi típusú honlapokra kidolgozott Heritrix egyre kevésbé alkalmas a modern weboldalak megfelelő minőségű mentésére, ezért jobb volna böngészőn keresztül való mentést használni legalább a híroldalakon és a közösségi felületeken. Ez ugyan lassabb, de jobb eredményeket ad, mert a böngészővel kombinált robot lefuttatja a HTML-fájlokba ágyazott vagy belinkelt programkódokat, végiggörgeti az oldalakat és egyéb műveleteket is el tud végezni, mintha csak egy ember ülne a gép előtt. A „Hogyan?" fejezetben említett *Brozzler* már egy próbálkozás volt ebben az irányban 2020-ban, de azóta megjelent – bár még mindig csak béta állapotban – egy fejlettebb eszköz *Browsertrix Crawler* néven, amihez van egy felhasználóbarát kezelőfelület is. A közeljövőben mindkét komponenst szeretnénk beüzemelni és először a nagyobb hírportálok archiválását megoldani vele, beleértve az előfizetést igénylő részeket is.
– A Browsertrix a *Kubernetes* nevű, konténer alapú alkalmazáskezelő szoftver alá telepíthető, így a szerveren levő többi programtól elkülönülten, egy saját virtuális környezetben fut. Mivel már most is elég sok alkalmazás van a szervereinken, melyek különböző rendszerkörnyezetet igényelnek, ezért az ütközések elkerülése, valamint az alkalmazások frissítése és skálázása könnyebb és automatizálhatóbb lenne, ha mindent a Kubernetes alá szerveznénk át.
– A jelenleg Google táblákban és XML-fájlokban tárolt metaadatok nyilvántartására egy egységes, rugalmasan bővíthető és az adatokban történt változtatásokat visszakövethető módon tároló megoldás kellene. Ez lehet egy saját fejlesztésű adatbázis, ami össze van kötve a WCT-vel és esetleg a Kaptafával is, de van egy olyan javaslat is, hogy mindent XML-formátumra kellene konvertálni és a *Git* verziókezelő szoftverre bízni a változások nyilvántartását. Fontos lenne a technikai és az adminisztratív jellegű metaadatok minél nagyobb részét automatizált módon kezelni, mert már olyan méretű a gyűjtemény, hogy ezekre a feladatokra nincsen emberi erőforrás.
– Meg kell oldani az úgynevezett „közpénzes", vagyis állami vagy önkormányzati finanszírozású webhelyek archivált változatainak nyilvános szolgáltatását, melyekhez a korábban említett kormányrendelet értelmében nem szükséges egyedi szerződéseket kötnie az OSZK-nak a tartalomgazdákkal. Erre a célra a PyWb biztosít egy *access control* nevű funkciót, amivel szabályozható, hogy milyen URL-címek nézhetők meg bárhonnan és melyek azok, amelyek csak a könyvtáron belül érhetők el. Ezért a PyWb megjelenítőt össze kell kapcsolni a megújítandó metaadat-nyilvántartással, a SolrWayback rendszert pedig úgy módosítani, hogy a teljes szövegű keresésnél a találatokat a PyWb-ben nyissa meg. Ha ezt sikerül megoldani, akkor a jelenlegi önálló publikus szerverre nem lesz szükség, így elkerülhetők a párhuzamosságok.

Mindezeknek az új megoldásoknak a kipróbálására egy tesztszervert hozott létre az OSZK rendszergazdája 2023 májusában. A különféle szoftverek feltelepítése és beállítása után elkezdődhet majd a tesztelés és a tapasztalatok gyűjtése az éles rendszerek újratervezéséhez.

Új irányok:

– A tesztszerver megfelelő „játszótér" lenne újabb archiváló technikák kipróbálására is. Ilyen az alkalmazásprogramozási felületen (*API*-n) keresztül való tartalomletöltés a népszerű közösségi platformokról, vagy a weboldalakon levő szövegek, adatok és fájlok célzott „gereblyézése" (*web-scraping*). Utóbbival lehetne például képeket vagy podcast adásokat gyűjteni metaadatokkal együtt. A visszanézési problémákon segítene, ha a régi böngészőverziókat is futtatni tudó Conifert és az ArchiveWeb.page programmal készített felvételeket leginkább viszszajátszani képes ReplayWeb.page szoftvert a saját szerverünkön is tudnánk működtetni.
– A Digitális Képarchívummal való együttműködés egyik formája a képmegosztó oldalak anyagának célzott begyűjtése lenne, majd pedig ezek visszakereshetővé tétele automatikus képfelismeréssel, amihez jó tanító korpusz a DKA-ban levő több mint 124 ezer, emberi munkával tárgyszavazott képi dokumentum.
– A gépi tanulás és a mesterséges intelligencia a webarchívum anyagának feltárásában is nagy segítség volna. A webhelyek automatikus tárgyszavazása mellett szóba jöhet az entitások (pl. földrajzi és tulajdonnevek) azonosítása és összekapcsolása a Nemzeti Névtérrel[17], illetve a Wikidata adatbázissal[18]; a hangfájlok kereshető szöveggé alakítása; az e-könyv és a PDF-formátumú fájlok osztályozása műfaj és téma szerint stb.

– Az OSZK webarchívumában levő sok terabájtnyi tartalom különféle kutatásokhoz és fejlesztésekhez szolgálhat „nyersanyagként". Ilyenre már most is van érdeklődés, de ahhoz, hogy megfelelő adathalmazokat tudjunk kigyűjteni, illetve egy helyben használható kutatási környezetet kialakítani, különféle adatbányász és adatvizualizáló eszközöket is be kell építeni a szolgáltatási palettába.

Lehetne még sorolni a továbbfejlesztési irányokat, de már ennyiből is sejthető, hogy a következő hat év sem lesz unalmas a „webkönyvtárosok" számára. Reméljük, hogy más közgyűjteményekben és tudományos műhelyekben is egyre többen kezdenek el foglalkozni ezzel a területtel, mert a digitális kultúra megőrzése az információs társadalmak egyik legnagyobb kihívása.

Irodalom

Drótos, L. *Webarchiválás a nemzeti könyvtárban. „Távcső a történelemre"*, Aranybulla-webarchívum bemutató és konferencia, Vörösmarty Mihály Könyvtár, Székesfehérvár, 2023. április 24. Elérhető: https://webarchivum.oszk.hu/wp-content/uploads/2023/04/OSZK_webarchivum_DL.pptx (Utolsó elérés: 2023. május 30.)

Drótos, L. *Digitálisan született tartalom megőrzése a nemzeti könyvtárban*, OSZK belsős intézményi workshop, Országos Széchényi Könyvtár, Budapest, 2022. december 1. Elérhető: https://webarchivum.oszk.hu/wp-content/uploads/2022/11/Born_digital_DL.pptx (Utolsó elérés: 2023. május 30.)

Drótos, L. *Az OSZK Webarchívum 2022. évi eredményei*, „404 Not Found – Ki őrzi meg az internetet?", konferencia és workshop, Országos Széchényi Könyvtár, Budapest, 2022. december 8. Elérhető: https://webarchivum.oszk.hu/wp-content/uploads/2022/12/Drotos_Laszlo_Az_OSZK_Webarchivum_2022.pptx (Utolsó elérés: 2023. május 30.)

Drótos, L. *Az idő fogságában. Ki őrzi meg a közösségi médiát?*,In: Tudományos és Műszaki Tájékoztatás, 68(7), p. 428–439, 2021. Elérhető: https://tmt.omikk.bme.hu/tmt/article/view/13062 (Utolsó elérés: 2023. május 30.)

Kalcsó, Gy. *Archivált webes tartalom kutatási célú hasznosítása*, „Távcső a történelemre" – Aranybulla-webarchívum bemutató és konferencia, Vörösmarty Mihály Könyvtár, Székesfehérvár, 2023. április 24. Elérhető: https://webarchivum.oszk.hu/wp-content/uploads/2023/04/Archivalt_webes_tartalom_kutatasi_celu_hasznositasa_KGY.pptx (Utolsó elérés: 2023. május 30.)

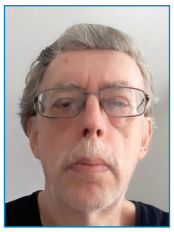
Visky, Á. L. *Rákóczi Emlékév Archívum – egy digitális gyűjtemény mintaalkalmazás*, „Távcső a történelemre" – Aranybulla-webarchívum bemutató és konferencia, Vörösmarty Mihály Könyvtár, Székesfehérvár, 2023. április 24. Elérhető: https://webarchivum.oszk.hu/wp-content/uploads/2023/04/Rakoczi_archivum_mintaalkalmazas_VMK2023_VAL.pptx (Utolsó elérés: 2023. május 30.)

Visky, Á. L. *Együttműködési lehetőségek a webarchiválás területén*, In: Könyvtári Figyelő, 67(1), p. 39–45, 2021. Elérhető: http://ojs.elte.hu/kf/article/view/2297 (Utolsó elérés: 2023. május 30.)

Hivatkozások

1 Webarchiválási munkafolyamatok szabályzata. Készítette: OSZK Digitális Bölcsészeti Központ Digitális Filológiai és Webarchiválási Csoport. Utoljára módosítva: 2022-08-04 (belső anyag)
2 https://archive.org
3 https://web.archive.org
4 https://webarchivum.oszk.hu
5 https://webarchivum.vmk.hu
6 https://mediateka.ek.szte.hu/exhibits/show/kariko_katalin_szte/
7 https://netpreserve.org
8 https://cc.au.dk/en/warcnet/
9 https://www.iso.org/obp/ui/#iso:std:iso:28500:ed-2:v1:en
10 https://archive-it.org
11 https://webarchivum.oszk.hu/demo-kezdolap/
12 https://webarchivum.oszk.hu/oszk-s-archivum-kezdolap/
13 https://rakoczi2019.webarchivum.oszk.hu
14 Megjelenés alatt.
15 https://ukrajnapublic.webharvest.oszk.hu/solrwayback/
16 https://dhupla.hu/page/kreativ/
17 https://magyarnemzetinevter.hu
18 https://www.wikidata.org

Beérkezett: 2023. június 2.



Drótos László
könyvtáros
Országos Széchényi Könyvtár, Digitális Bölcsészeti Központ
Digitális Filológiai és Webarchiválási Osztály
E-mail: drotos.laszlo@oszk.hu