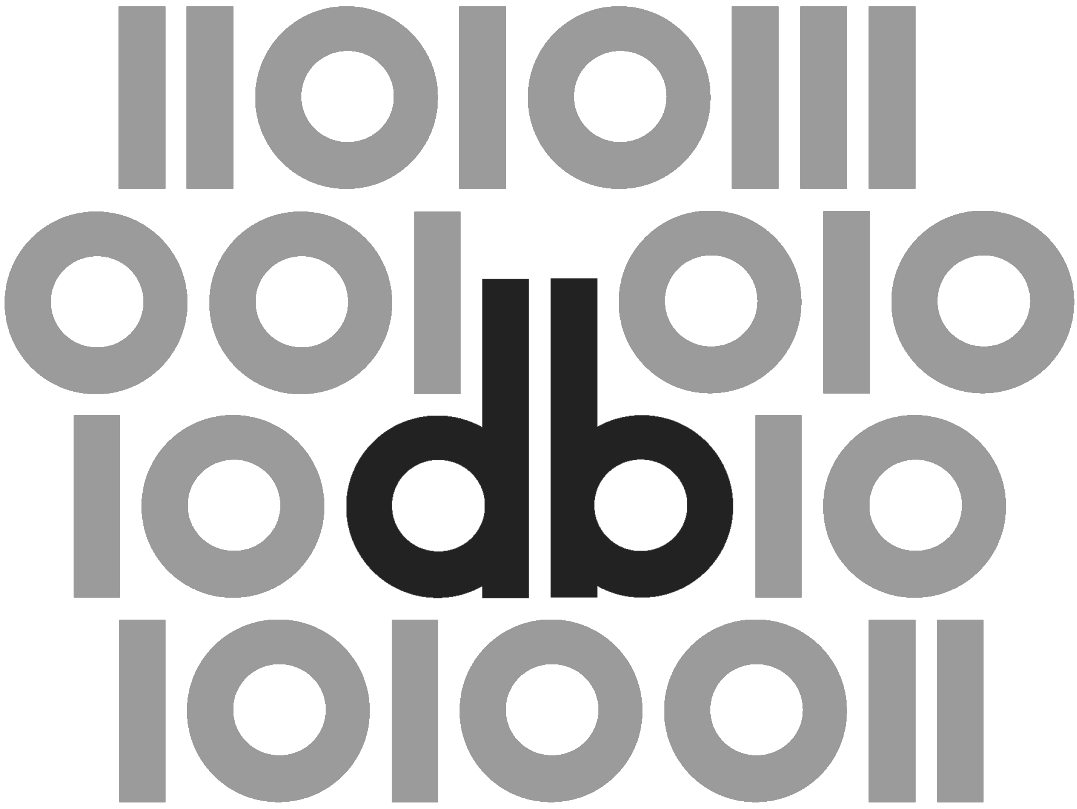
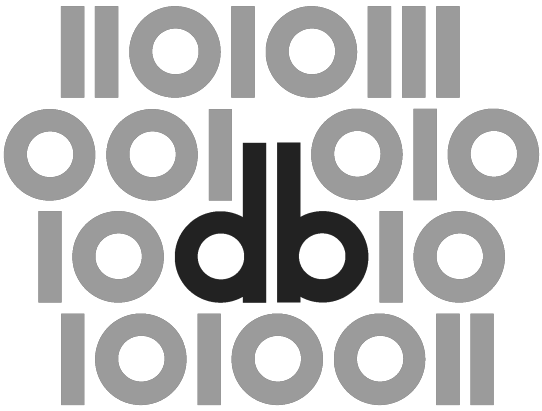
2018.01. <DIGITÁLIS BÖLCSÉSZET>



2018.01. </DIGITÁLIS BÖLCSÉSZET>

Digitális Bölcsészet
2018., első szám

<DIGITÁLIS BÖLCSÉSZET>



2018.01.

ISSN 2630-9696
DOI 10.31400/dh-hun.2018.1

Kiadja az ELTE BTK Régi Magyar Irodalom Tanszéke, 1088 Budapest, Múzeum krt. 4/A.
Felelős kiadó az ELTE BTK Régi Magyar Irodalom Tanszék vezetője.

Megjelenik az Open Journal Systems (OJS) v. 3. platformon, melynek működtetését az ELTE Egyetemi Könyvtár- és Leváltár biztosítja.

Honlap: `http://ojs.elte.hu/index.php/digitalisbolcseszet`
Email cím: dbfolyoirat@gmail.com

Tördelés: Hegedüs Béla
Grafika: Hegyi Gábor

Tartalom

Drótos László

Országos Széchényi Könyvtár

mekdl@iif.hu

Kokas Károly

SZTE Klebelsberg Könyvtár

kokas@ek.szte.hu

Webarchiválás és a történeti kutatások

A digitálisan születő tartalom sokkal részletesebb és teljesebb leképezése a jelennek, mint ami régebbi korokban a hagyományos információhordozó eszközökkel rögzíthető volt. A tanulmány első része arról ad áttekintést, hogy milyen próbálkozások és technológiák léteznek ennek a digitális jelennek a megőrzésére, illetve milyen korlátai vannak a már működő webarchívumoknak. A dolgozat második része azt vizsgálja, hogy a történeti szempontú kutatásoknak hogyan lehet hasznára mindez, s hogyan lesz elsősorban a közelmúlt történetének is elsőrangú forrása. A szerzők arra is rámutatnak, hogy a webaratások következtében előálló hatalmas adatsilók egészen új típusú forráskezelést és módszertant kívánnak majd meg, miközben azzal kecsegtetnek, hogy egészen új típusú eredményeket is fel lehet majd mutatni segítségükkel.

Kulcsszavak:
webarchiválás, digitális megőrzés, digitális bölcsészet, webhistoriográfia

A történelmet már nagyrészt online írják.[1]

Olyan mértékben függünk tőle, mint az elektromos hálózattól, de amíg azon csak energia folyik, az interneten információ, mégpedig az élet minden területéről, az intim magánügyektől a globális közügyekig. A világhálón áramló elektromágneses impulzusok minden pillanatban minden korábbinál részletesebb lenyomatát adják civilizációnk jelenének. Ezeket a lenyomatokat valahogyan meg kellene őrizni ahhoz, hogy a múlttá váló jelenünk értelmezhető legyen a jövőből visszanézve. Két-három évtizedet már elvesztettünk… bár azért nem teljesen.

[1] „De geschiedenis van vandaag wordt vooral online geschreven." Peter de Bode, René Voorburg, „Webarchivering," hozzáférés: 2018.05.22, `https://www.kb.nl/organisatie/onderzoek-expertise/e-depot-duurzame-opslag/webarchivering`.

1. Ki őrzi meg a netet?

A digitális univerzum elképesztő tágulási üteme elfedi azt a tényt, hogy ez a világ rendkívüli tempóban pusztul is. A tudományos publikációkban hivatkozott internetes források esetében öt és tíz év közötti felezési értéket mutattak ki a különböző vizsgálatok, vagyis ennyi idő alatt a linkek fele eltörik, eltűnik mögülük az eredeti tartalom. A *webkettes* helyeken még gyorsabb az erózió, egyes Facebook-posztok, Twitter-üzenetek, YouTube-videók élettartama csak napokban mérhető. A 404-es hibaüzenet a legtöbbször megtekintett internetes tartalom.[2]

Az 1970-es és 1980-as évtizedek hálózatainak: a BBS-eknek, a CompuServe-nek, a Minitelnek, az EARN/BITNET-nek, a nálunk is elterjedt X.25-nek, vagy a korai Internetnek a bitjei és bájtjai már nagyrészt eltűntek. Az 1980-ban indult Usenet fórumainak üzeneteit 1995-től elkezdte egy Deja News[3] nevű vállalkozás gyűjteni és megőrizni, majd 2001-ben a Google megvette az akkorra már több mint félmilliárdnyi üzenetből álló gyűjteményt, kiegészítette más forrásokból 1981 májusáig visszamenőleg, és beleolvasztotta a saját Google Groups szolgáltatásába. A kilencvenes évek első felének legnépszerűbb internetes tartalomszolgáltató felülete, a Gopher is majdnem teljesen eltűnt a süllyesztőben, ha 2007 júniusában egy John Goerzen nevű programozó le nem menti a még működő szervereket, és a körülbelül 780 ezer dokumentumot tartalmazó 15 gigabájtos tömörített csomagot oda nem adja az Internet Archive-nak.[4] Sajnos ez egy megkésett akció volt, mert az internettörténeti szempontból legérdekesebb Gophereket akkorra már régen leállították.

A web esetében szerencsére kisebb volt a késés. Öt évvel az első weboldal megszületése után, 1996-ban San Franciscóban létrejött az Internet Archive nevű nonprofit szervezet és archívum, amely az egyéb digitális média (könyv, kép, hang, videó, szoftver) mellett gyűjti a webhelyek tartalmát is. Ezeket részben a jelenleg már az Amazon cégcsoportjába tartozó és főként az internetes oldalak forgalmának mérésével és rangsorolásával foglalkozó Alexa Internet cégtől kapja. 2017 elején 279 milliárd weboldal volt a gyűjteményben, melyek különböző időpontokbeli mentései visszanézhetők a Wayback Machine[5] nevű szolgáltatással. De ez a hatalmas szám is csak töredéke a teljes webtérnek, mert sok webhely bejárhatatlan és lementhetetlen automatikus módszerekkel: vagy mert olyan technológiát használ, vagy mert jelszóval védett, vagy csak egyszerűen ki vannak róla tiltva a robotok.

A közösségi média különösen nehezen archiválható, pedig az internet legfontosabb szegmense jelenleg. Ennek az előbb említett okok mellett a felhasználók által generált tartalom puszta mennyisége a magyarázata. Magukon a szolgáltatókon kívül senkinek

[2] Dion Hoe-Lian Goh and Peng Kin Ng., „Link Decay in Leading Information Science Journals," *Journal of the Association for Information Science and Technology* 58, 1. sz. (2007): 15–24, `https://onlinelibrary.wiley.com/doi/full/10.1002/asi.20513`; Frank McCown, Sheffan Chan, Michael L. Nelson and Johan Bollen, „The Availability and Persistence of Web References in D-Lib Magazine," in *5th International Web Archiving Workshop and Digital Preservation* (2005), `http://arxiv.org/ftp/cs/papers/0511/0511077.pdf`.

[3] *Wikipedia,* „Google Groups – Deja News," hozzáférés: 2018.05.22, `https://en.wikipedia.org/wiki/Google_Groups%2523Deja_News`.

[4] John Goerzen, „2007 Gopherspace Mirror," film, hozzáférés: 2018.05.22, `http://archive.org/details/2007-gopher-mirror`.

[5] Wayback Machine, Internet Archive, hozzáférés: 2018.05.22, `http://web.archive.org/`.

 DOI 10.31400/dh-hun.2018.1.129

nincsen elég pénze és megfelelő technikája erre, ők viszont csak addig érdekeltek a megőrzésben, amíg az élő szolgáltatást nyereségesen tudják üzemeltetni. Így szűnt meg például a Microsoftnak a 2006-ban még 120 millió regisztrált taggal rendelkező Windows Live Spaces[6] nevű blogplatformja 2011-ben, a Hyves[7] nevű, 10 milliós létszámú holland közösségi oldal 2013-ban, valamint a magyar iWiW[8] is 2014 június végén.

Hogy mekkora problémát jelent hosszú távú és kutatható archívummá alakítani a webkettes tartalmakat, arra jó példa az amerikai Library of Congress [Kongresszusi Könyvtár] esete a Twitterrel.[9] 2010-ben a Library of Congress 2006 márciusáig, vagyis a Twitter indulásáig visszamenőleg megkapta az összes, mintegy 170 milliárd nyilvános *tweet*et, valamint az *élő folyam*ot is, ami akkor napi 50 millió üzenet volt, ám ez a szám 2014-re már megtízszereződött, és azóta is folyamatosan nő. Mivel nemcsak a legfeljebb 140 karakterből álló szövegeket, hanem az azokhoz tartozó több mint százféle metaadatot is tárolni és indexelni kell, ezért a könyvtár – amúgy nem gyenge – számítógépes infrastruktúráján egy egész napig tartott volna egyetlen keresőkérdés lefuttatása. Magáncégek bevonásával 2014 közepére ígértek egy kísérleti szolgáltatást, de az azóta sem készült el.

1.1. Archívumfajták

A legtöbb internetarchívum jelenleg webarchívum, vagyis weboldalak vagy webhelyek valamilyen rendszeres vagy rendszertelen időközönként megismételt mentései. Egyre több a második generációs rendszer közöttük, amelyeket egy néhány éves üzemeltetés után alapjaiktól újraterveztek – és az első verzió honlapja jó esetben még megtalálható valamelyik webarchívumban. Céljuk és létrehozójuk alapján a főbb típusok a következők.

1.1.1. Magánarchívumok A cél lehet valamilyen érdeklődési körhöz vagy kutatási munkához való anyaggyűjtés, a hosszú távú hivatkozhatóság biztosítása, esetleg bizonyítékként való felhasználás. A magáncélú archiváláshoz felhasználóbarát célszoftverek (pl. HTTrack),[10] böngészőkiegészítők (pl. Fireshot, ZipTabs), illetve ingyenes vagy fizetős online szolgáltatások és felhőtárhelyek (pl. Save Page Now, PageFreezer)[11] állnak rendelkezésre. De persze egy-egy weboldalt el lehet menteni magukkal a böngészőkkel is vagy a Zotero[12] nevű hivatkozáskezelő programmal, melyhez olyan modul (Hiberlink plugin for Zotero)[13] is létezik, amely rögtön valamelyik nagy web-

[6] *Wikipedia,* „Windows Live Spaces," hozzáférés: 2018.05.22, `https://en.wikipedia.org/wiki/Windows_Live_Spaces`.

[7] *Wikipedia,* „Hyves," hozzáférés: 2018.05.22, `https://en.wikipedia.org/wiki/Hyves`.

[8] *Wikipedia,* „iWiW," hozzáférés: 2018.05.22, `https://hu.wikipedia.org/wiki/IWiW`.

[9] Drótos László, „Michael Zimmer: A Kongresszusi Könyvtár Twitter archívuma," recenzió (Michael Zimmer, *The Twitter Archive At the Library of Congress: Challenges for Information Practice and Information Policy* (2015)) *Tudományos és Műszaki Tájékoztatás* 62, 11–12. sz. (2015): 445–447, `https://tmt.omikk.bme.hu/tmt/article/download/610/581`.

[10] *HTTrack Website Copier*, hozzáférés: 2018.05.22, `https://www.httrack.com/`.

[11] Wayback Machine, „Save Page Now," hozzáférés: 2018.05.22, `http://web.archive.org/`; PageFreezer, hozzáférés: 2018.05.22, `https://www.pagefreezer.com/`.

[12] Zotero, hozzáférés: 2018.05.22, `https://www.zotero.org/`.

[13] Hiberlink Zotero plugin, hozzáférés: 2018.05.22, `http://hiberlink.org/zotero.html`.

archívumba menti a megőrizni kívánt oldalt, és az onnan visszakapott archív URL-t is felveszi a Zotero adatbázisába.

1.1.2. Céges archívumok Az üzleti szférában egyre jellemzőbb az internetes tartalmak mentése. A motiváció a vállalat történetének megőrzése vagy egyszerűen az a törvényi előírás, hogy archiválniuk kell minden hivatalos kommunikációt az ügyfelekkel – beleértve a honlapjukon és a különböző webkettes csatornáikon közzétett tartalmaikat is. Előbbire jó példa a Coca-Cola,[14] amelynek az archívuma hatmillió weboldalt őriz a cég különböző internetes felületeiről az 1995-ös első honlapig visszamenően. A saját anyagok mentése mellett a versenytársak vagy az adott üzletág online tartalmait is szokták gyűjteni statisztikai, adatbányászati, piac- és trendkutatási célokból. Több kulcsrakész, professzionális rendszer (pl. Q-Suite, Presurf)[15] is kapható ma már, amelyekkel nemcsak webes dokumentumok, hanem videók, Skype-beszélgetések, mobiltelefonos üzenetváltások, üzleti tranzakciók (pl. a cég webshopjából) egyaránt rögzíthetők, időbélyeggel és digitális aláírással hitelesítve, hogy egy jogi vita esetén a bíróság is elfogadja őket bizonyítékként. Ilyen rendszereket SaaS (Software-as-a-Service) formában is lehet bérelni (pl. NetTrack, Cloud Preservation, Scrapinghub),[16] ahol a megrendelőnek nem kell semmit telepítenie és tárolnia, hanem egy adminisztrációs felületen keresztül tudja ütemezni az archiválási feladatokat, és a lementett tartalom valamilyen felhőtárhelyen kerül megőrzésre.

1.1.3. Intézményi archívumok Közgyűjtemények (könyvtárak, levéltárak, múzeumok), egyetemek és kutatóintézetek, tudományos és civil szervezetek, kormányzati szervek egyaránt építenek alkalmi jelleggel vagy hosszabb távon webarchívumokat. Az egyik ok ezeknél is az intézmény történetének dokumentálása, vagyis a saját honlap és egyéb internetes felületek szisztematikus mentése. Az állami szervek esetében pedig több országban jogszabály írja elő, hogy elérhetőknek kell maradniuk a korábbi, esetleg már érvényüket vesztett, ezért az élő honlapról lekerült dokumentumoknak is. Emellett tematikus gyűjteményeket is építenek egyre több helyen : a városi könyvtárak például helyismereti, helytörténeti tartalmakat mentenek, a tudományos intézmények a kutatási területüknek megfelelő forrásokat, a civil szervezetek pedig az általuk képviselt ügy internetes lenyomatait. Az intézményi webarchívumok is készülhetnek a céges archívumoknál említett professzionális rendszerekkel, illetve fizetős felhőszolgáltatásokkal, de gyakoribb a nyílt forráskódú

[14] Ted Ryan, „1s and 0s : The History of The Coca-Cola Company's Website," *Coca-Cola Company,* 2012. nov. 08., `http://www.coca-colacompany.com/stories/1s-and-0s-the-history-of-the-coca-cola-companys-website`.

[15] Q-Suite, hozzáférés : 2018.05.22, `https://www.qumram.com/products`; Presurf, hozzáférés : 2018. 05.22, `http://www.capsis.nl/en/websitearchiving/presurf/introduction/`.

[16] NetTrack, hozzáférés : 2018.05.22, `http://www.capsis.nl/en/websitearchiving/nettrack/introduction/`; Cloud Preservation, hozzáférés : 2018.05.22, `http://www.nextpoint.com/`; Scrapinghub, hozzáférés : 2018.05.22, `https://scrapinghub.com/`.

 DOI 10.31400/dh-hun.2018.1.129

szoftverekből összerakott saját megoldás (pl. NetarchiveSuite),[17] valamint a nonprofit archiváló szolgáltatások (pl. Archive-It, archive.is, ArchivetheNet)[18] használata.

Néhány érdekesebb projekt: York University Web Archives[19] (az egyetem saját webhelyei, metaadatokkal együtt letölthető csomagokban is), Web Archive of Cacak[20] (a szerb Čačak város könyvtárának kis helyismereti gyűjteménye), CyberCemetery[21] (az Egyesült Államok megszűnő kormányzati honlapjainak utolsó állapota), Human Rights Web Archive[22] (a Columbia University Libraries által mentett emberi jogi témájú webhelyek), Contemporary Composers Web Archive[23] (54 modern zeneszerző – köztük Ligeti György – honlapjai az amerikai zenei könyvtárak szakembereinek válogatásában), Digital Archive for Chinese Studies[24] (az Universität Heidelberg Institut für Sinologie 2001 óta épített gyűjteménye), Latin American Web Archiving Project[25] (a University of Texas LANIC központjának gyűjtése latin-amerikai politikai pártokról és választásokról).

1.1.4. Nemzeti archívumok Rendszerint a nemzeti, állami könyvtár vagy az általa vezetett intézményi konzorcium tartja fenn ezeket, és a helyi kötelespéldányra vonatkozó törvény szabályozza a működésüket. A cél az adott nemzet digitálisan születő kultúrájának megőrzése a jövő számára. Ez történhet a nemzeti webtér időnkénti (éves vagy féléves) aratásával, illetve egy kellően reprezentatív, néhány ezer vagy néhány tízezer webhelyet tartalmazó részhalmaz gyakoribb (havi, heti vagy akár napi) mentésével. A legtöbb országban mindkettőt alkalmazzák, mert jól kiegészítik egymást: a teljes körű aratás egy átfogó pillanatképet rögzít, de ritkábban, a szelektív gyűjtéssel pedig gyakrabban és jobb minőségben lehet az érdekesebb, értékesebb tartalmakat archiválni, akár az országdoménen kívüli szerverekről is – és ezek kisebb számossága még azt is megengedi, hogy részletesebben metaadatolják, katalogizálják őket, ami a visszakeresést nagyban megkönnyíti. E mellett szokás még eseményalapú mentéseket is csinálni néhány napig vagy hétig, például valamilyen világraszóló rendezvény, választási kampány vagy természeti katasztrófa esetén, hogy a sajtóban és a közösségi fórumokon megjelenő információkból és reakciókból minél többet tudjanak rögzíteni. A szellemi tulajdont és személyiségi jogokat védő szabályok miatt a legtöbb országban csak helyben, a könyvtáron vagy a zárt könyvtári hálózaton belül

[17] NetarchivSuite, hozzáférés: 2018.05.22, `https://sbforge.org/display/NAS/NetarchiveSuite`.

[18] Arcive-It, hozzáférés: 2018.05.22, `http://archive-it.org`; archive.is, hozzáférés: 2018.05.22, `http://archive.is`; ArchivetheNet, hozzáférés: 2018.05.22, `http://archivethe.net/en`.

[19] York University Web Archives, hozzáférés: 2018.05.22, `https://digital.library.yorku.ca/yul-232039/web-archives`.

[20] Web Archive of Cacak, hozzáférés: 2018.05.22, `http://cacak-dis.rs/digital/english/web--archive-of-cacak/`.

[21] CyberCemetery, hozzáférés: 2018.05.22, `https://govinfo.library.unt.edu`.

[22] Human Rights Web Archive, hozzáférés: 2018.05.22, `https://hrwa.cul.columbia.edu`.

[23] Contemporary Composers Web Archive, hozzáférés: 2018.05.22, `https://library.columbia.edu/bts/web_resources_collection/contemporary_composers_web_archive.html`.

[24] Digital Archive for Chinese Studies, hozzáférés: 2018.05.22, `http://www.zo.uni-heidelberg.de/boa/digital_resources/dachs/index_en.html`.

[25] Latin American Web Archiving Project, University of Texas, hozzáférés: 2018.05.22, `http://lanic.utexas.edu/project/archives/`.

lehet hozzáférni a webarchívumhoz, dedikált gépekről, másolási lehetőség nélkül. De a szelektíven mentett webhelyekből – amelyeknél erre az eredeti tartalomgazda engedélyt adott – szoktak azért egy távolról is elérhető, böngészhető szolgáltatást is csinálni (ilyen pl. a szlovén webarchívum),[26] illetve a metaadatok általában a teljes gyűjtemény esetében nyilvánosak, és vagy egy külön adatbázisban (pl. a Library of Congress webarchívumának kereső- és böngészőfelülete),[27] vagy a könyvtár katalógusában visszakereshetők (pl. egy archivált webhely rekordja a svájci nemzeti könyvtár katalógusában).[28] A nemzeti könyvtárak többsége saját, nyílt forráskódú szoftverekből álló rendszert működtet a webarchiváláshoz (a katalán webarchívumhoz például ezeket a szoftvereket használják: Heritrix, Wayback, NutchWax, Wera, Web Curator Tool, Hadoop), gyakran egy IT-partner segítségével (pl. a szlovák nemzeti könyvtár a Tempest céggel),[29] de arra is van példa, hogy kiszervezték a feladatot, és egy professzionális webarchiváló céget bíztak meg vele (pl. az írországi web mentését és annak szolgáltatását a nemzeti könyvtár számára az *Internet Memory Foundation* végzi).[30] Az élő webről való aratás mellett létezik olyan megoldás is, hogy maga a tartalomszolgáltató küldi be valamilyen szabványos adatcsere-csatornán át a webhelyén megjelent új tartalmakat a könyvtári archívumba. A folyamatosan változó és robotokkal amúgy is nehezen bejárható, dinamikusan generált weboldalakból álló hírportálok esetében ez a legjobb megoldás.

Jelenleg körülbelül 40 projekt sorolható a nemzeti szintű archívum kategóriájába, de ez csak valamivel több mint 30 országot jelent, mert egyes tartományoknak, tagállamoknak vagy nagy nemzetiségeknek külön archívuma van. Néhány példa:

- PANDORA: ausztrál könyvtári konzorcium keretében működik 1996 óta, szelektíven mentenek, valamint katalogizálnak már közel 50 ezer webcímet, és az Internet Archive segítségével időnként a teljes .au domént is learatják.[31]
- LCWA: a Library of Congress 2000-ben – akkor még MINERVA néven – indított projektje, melynek keretében több mint 11 ezer webhelyet archiválnak, és eseményekről is csinálnak részgyűjteményeket, például a szeptember 11-i terrortámadás, a 2002-es téli olimpia, az iraki háború.[32]
- UKWA: 2004-ben egy könyvtári együttműködés keretében létrejött brit webarchívum, amelynek három gyűjteménye van: egy több mint 15 ezer oldalból álló válogatott állomány, az Internet Archive-tól átvett 1996–2013 közötti mentés

[26] Spletni arhiv, Narodne in univerzitetne knjižnice, hozzáférés: 2018.05.22, `http://arhiv.nuk.uni-lj.si`.

[27] Library of Congress, hozzáférés: 2018.05.22, `https://www.loc.gov/websites/`.

[28] Helveticat, Schweizerische Nationalbibliothek (NB), hozzáférés: 2018.05.22, `http://www.helveticat.ch/lib/item?id=chamo:1745898`.

[29] Central Archiving Platform, hozzáférés: 2018.05.22, `https://www.tempest.sk/products-and-services/central-archiving-platform-2d5.html`.

[30] Web Archive, National Library of Ireland, hozzáférés: 2018.05.22, `https://www.nli.ie/en/web_archive.aspx`.

[31] PANDORA, Australia's Web Archive, hozzáférés: 2018.05.22, `http://pandora.nla.gov.au`.

[32] Archived Web Sites, Library of Congress, hozzáférés: 2018.05.22, `https://www.loc.gov/websites/`.

 DOI 10.31400/dh-hun.2018.1.129

az Egyesült Királyság webhelyeiről és az .uk címtartomány 2013 utáni saját mentései.[33]
- WebArchiv: a cseh nemzeti könyvtár 2000-ben indult projektje, melyben szelektív, eseményalapú és a teljes cseh webtérre kiterjedő archiválást végeznek. Eddig 5129 tartalomszolgáltatóval kötöttek szerződést.[34]
- WARP: a japán National Diet Library 2002-től fejlesztett, már harmadik generációs webarchiváló rendszere, mellyel 2015-ben közel 11 ezer webhelyet mentettek, és ezekből 280 ezer fontosabb dokumentumot kigyűjtve önállóan is katalogizáltak.[35]

1.1.5. Globális archívumok A korábban már említett, messze a legnagyobb Internet Archive mellett van még egy-két projekt, amelyek nem nemzet, földrajzi hely vagy téma alapján fókuszáltak. Ilyen például a Common Crawl nevű, kaliforniai székhelyű webarchiváló kezdeményezés,[36] mely 2011 óta ingyenesen letölthető és kutatható halmazokat gyűjt a nyilvános webről, jelenleg évi négyszeri aratással. 2015 végén már 1.82 milliárd weboldalt tettek ily módon elérhetővé az Amazon felhőtárhelyéről. Ide sorolható még egy 2011–2013 közötti kísérleti EU-s projekt, a BlogForever.[37] Ennek keretében közel 210 ezer blogot mentettek és elemeztek ki azzal a céllal, hogy kidolgozzák ennek a műfajnak az archiválási technológiáját. Itt érdemes megemlíteni a 2009 óta létező, főként a veszélyeztetett internetes szolgáltatások megőrzésére szerveződött, Archive Team nevű – civilekből és szakemberekből álló – „akciócsoportot",[38] illetve annak WikiTeam részét is,[39] amely eddig már 27 ezer önálló *wiki*t és több *wikifarm*ot mentett le az Internet Archive-ba.[40]

A különböző nemzetközi, nemzeti és helyi internetarchiválási kezdeményezéseket egy 2003-ban a francia nemzeti könyvtár és 12 partnerintézmény által alapított konzorcium, az IIPC (International Internet Preservation Consortium)[41] fogja össze, jelenleg 54 tagja van. A szervezet céljai: az internet megőrzésével foglalkozók közötti tapasztalatcsere, az ehhez szükséges technológiák közös fejlesztése, a szabványosítás. Külön munkacsoportok foglalkoznak a begyűjtés, a megőrzés, a hozzáférés, a ráépülő szolgáltatások és az oktatás témáival. Éves konferenciákat rendez, közös projekteket koordinál, szoftvereket fejleszt.

[33] UK Web Archive, hozzáférés: 2018.05.22, `http://webarchive.org.uk`.
[34] Webarchiv, the Museum of Czech Web, hozzáférés: 2018.05.22, `http://www.webarchiv.cz/en`.
[35] WARP: Web Archiving Project, hozzáférés: 2018.05.22, `http://warp.da.ndl.go.jp/info/WARP_en.html`.
[36] Common Crawl, hozzáférés: 2018.05.22, `http://commoncrawl.org/`.
[37] BlogForever, hozzáférés: 2018.05.22, `http://web.archive.org/web/20160729112149/http://blogforever.eu/`.
[38] Archive Team, hozzáférés: 2018.05.22, `https://www.archiveteam.org/`.
[39] WikiTeam, hozzáférés: 2018.05.22, `http://www.archiveteam.org/index.php?title=WikiTeam`.
[40] Internet Archive: WikiTeam, hozzáférés: 2018.05.22, `https://archive.org/details/wikiteam&tab=about`.
[41] International Internet Preservation Consortium (IIPC), hozzáférés: 2018.05.22, `http://netpreserve.org/`.

1.2. A magyar helyzet

A hazai helyzet sajnos röviden összefoglalható: a kilencvenes évek második felében indított magyar webes keresők (Heuréka, Góliát, Altavizsla/Vizsla), majd a 2010 körül megjelent újgenerációs társaik (Bluu, Szörcs, Miner, PolyMeta/Johu, RichPOI) robotjai által gyűjtött magyar tartalomra nem épült webarchívum – ahogyan például a portugáloknál történt –, és azóta már el is tűntek ezek a rendszerek adatállományaikkal együtt, mert nem bírták a versenyt a Google keresőjével. Az MTA SZTAKI 2008–2013 között részt vett két európai uniós K+F-projektben (LIWA és LAWA),[42] melyek a webarchiválás technológiájának megújítását és a webarchívumok kutatási célú felhasználásához szükséges módszerek és szoftverek kidolgozását célozták. A 2010-es évek elején az ELTE Tudománytörténet és Tudományfilozófia Tanszékének tudománymetriai munkacsoportja végzett egy fókuszált webarchiválást.[43] Mintegy 400 magyar webhelyet: kutatóintézeti, valamint egyetemi és főiskolai honlapokat mentettek kéthetes periodicitással. A lementett tartalmat ki is elemezték olyan szempontból, hogy mit és mennyit kommunikálnak magukról online ezek az intézmények. A hazai könyvtári szférában 2006-ban hangzott el az első javaslat egy Magyar Internet Archívum (MIA) létrehozására.[44] A szándék az Országos Széchényi Könyvtár 2007-es munkatervébe is bekerült,[45] és bár több próbálkozás is volt a szükséges forrás megteremtésére (együttműködve a Szegedi Tudományegyetem könyvtárával, illetve az NIIF Programmal), ezek nem vezettek eredményre, így annak ellenére, hogy az OSZK-ban az egyedi internetes dokumentumok gyűjtése és feldolgozása már régóta folyik a MEK, EPA és DKA szolgáltatások[46] keretében, webhelyeket még nem archivál a nemzeti könyvtár. A 2017 elejétől 2018 végéig tartó OKR (Országos Könyvtári Rendszer) nevű projektbe viszont végre bekerült a webaratás tesztelése,[47] melyhez az infrastruktúrát a KIFÜ–NIIF[48] biztosítja. A tervek szerint néhány száz – főként kulturális és tudományos – webhely kerül többszöri lementésre, és lesz két kísérlet a .hu alá tartozó szerverek teljes körű aratására is. Az elsődleges cél egyelőre még csak a tanulás, a szükséges kutatási és fejlesztési munka elvégzése, egy üzemszerűen működő magyar webarchívum feltételeinek megteremtése.

[42] MTA SZTAKI, Living Web Archives (LiWA), hozzáférés: 2018.05.22, `https://www.sztaki.hu/en/projects/liwa`; MTA SZTAKI, Longitudinal Analytics of Web Archive Data (LAWA), hozzáférés: 2018.05.22, `https://www.sztaki.hu/projektek/lawa`.

[43] Gulyás László, „Magyar Internet Archívum pilot és elemzés," prezentáció (ELTE, 2014. április 14.), hozzáférés: 2018.07.18, https://slideplayer.hu/slide/2647111/.

[44] Drótos László, „Mi a MIA? Javaslat egy Magyar Internet Archívum létrehozására," *Tudományos és Műszaki Tájékoztatás* 53, 6. sz. (2006): 267–274, `http://tmt.omikk.bme.hu/show_news.html?id=4431&issue_id=473`.

[45] Bibliotheca Nationalis Hungariae, „Az Országos Széchényi Könyvtár programja, 2007," hozzáférés: 2018.05.22, `http://www.oszk.hu/sites/default/files/szakmai_munkaterv_2007_0.pdf`.

[46] Magyar Elektronikus Könyvtár, hozzáférés: 2018.05.22, `http://mek.oszk.hu/`; Elektronikus Periodika Archívum és Adatbázis, hozzáférés: 2018.05.22, `http://epa.oszk.hu/`; Digitális Képarchívum, hozzáférés: 2018.05.22, `http://dka.oszk.hu/`.

[47] Országos Széchényi Könyvtár, Magyar Internet Archívum, „OSZK webaratás – teszt fázis," hozzáférés: 2018.05.22, `http://mekosztaly.oszk.hu/mia/`.

[48] Kormányzati Informatikai Fejlesztési Ügynökség, hozzáférés: 2018.05.22, `http://kifu.gov.hu`; Nemzeti Információs Infrastruktúra Fejlesztési Program, hozzáférés: 2018.05.22, `https://niif.hu`.

 DOI 10.31400/dh-hun.2018.1.129

Egy kultúrának a webtérben való szereplése és láthatósága ma már a globális verseny része, ezért a döntéshozóknak óriási a felelőssége, hogy az ezen a területen keletkező hiányosságaink, mulasztásaink ne okozzanak versenyhátrányt.

1.3. Technológia

A magán és a kisebb intézményi archívumoknál a már említett letöltőalkalmazásokat használják, melyekkel az ismerős Windows-környezetben vagy akár közvetlenül a böngészőből lehet weboldalakat vagy webhelyeket lementeni. A cégek sokszor valamilyen kulcsrakész archiválórendszert vesznek meg, amely a webes tartalmak letöltése mellett API-kon (alkalmazásprogramozási felületeken át) tud menteni például levelezőrendszerekből és más kommunikációs csatornákról, webkettes platformokról vagy sugárzott médiafolyamokból, sőt akár a tranzakciós módszert is támogatja, vagyis amikor a webszerver minden olyan dokumentumból automatikusan elküld egy másolatot az archívumba, amelyet egy felhasználó lekért. A nagyméretű webarchívumok mind szoftveres robotokat, ún. keresőrobotokat *(crawler)* futtatnak. Ezek egy előre megadott URL-címlistából kiindulva derítik fel a weboldalak közötti linkeket, és az üzemeltető által definiált szabályok alapján döntik el, hogy melyeket kövessenek, melyekről töltsék le az ott található weboldalakat és a beléjük ágyazott egyéb fájlokat. A legtöbb nemzeti archívum már az Internet Archive által fejlesztett Heritrix crawlert[49] használja, amely nagyméretű, szabványos WARC-csomagokba[50] menti a megtalált digitális objektumokat. Ezekből később – az eredeti URL-címük alapján – a szintén ingyenes és nyílt forráskódú OpenWayback nevű szoftverrel[51] rekonstruálhatók és nézhetők meg az archivált webhelyek különböző időpontbeli mentései. Természetesen le is lehet indexelni az archívumban levő szöveges fájlokat, és akkor a teljes szövegű visszakeresés is lehetséges. Erre a célra többféle szoftver is szóba jöhet, mint például a kifejezetten webarchívumokhoz kialakított NutchWAX.[52] A letöltendő webhelyek nyilvántartásához, a mentések gyakoriságának, mélységének és egyéb paramétereinek beállításához, a begyűjtött anyag minőségének ellenőrzéséhez, a leíró metaadatok elkészítéséhez és az eredeti tartalomgazdáktól kapott engedélyek kezeléséhez szükség van még egy keretrendszerre. Egyes archívumok ezt maguk fejlesztik a kezdetektől fogva, de erre a célra is léteznek már nyílt forráskódú eszközök, például a Web Curator Tool.[53] Érdemes még megemlíteni a Memento Project[54] keretében kidolgozott megoldást, amely a webszerverek által használt HTTP-protokollt egészíti ki egy „Datetime" elemmel.[55] Ennek segítségével a kliens (pl. egy webböngészőt használó ember) egy

[49] Heritrix, hozzáférés: 2018.05.22, `http://crawler.archive.org/index.html`.

[50] Web ARChive (WARC) Format, hozzáférés: 2018.05.22, `https://iipc.github.io/warc-specifications/specifications/warc-format/warc-1.1/`.

[51] Wayback, hozzáférés: 2018.05.22, `http://archive-access.sourceforge.net/projects/wayback/`.

[52] NutchWAX, hozzáférés: 2018.05.22, `http://archive-access.sourceforge.net/projects/nutchwax/`.

[53] Web Curator Tool, hozzáférés: 2018.05.22, `http://dia-nz.github.io/webcurator/`.

[54] Time Travel Service, „Memento Guide – Introduction to Memento," hozzáférés: 2018.05.22, `http://www.mementoweb.org/guide/quick-intro/`.

[55] H. Van de Sompel, M. Nelson, R. Sanderson, „HTTP Framework for Time-Based Access to Resource States: Memento," *Request for Comments* (2013. dec.), `https://tools.ietf.org/html/rfc7089`.

weboldal aktuális állapota helyett le tudja kérni annak adott időpontbeli vagy ahhoz legközelebbi mentését a világ webarchívumainak valamelyikéből. Ezzel a megoldással a web olyan médium lett, amelynek múltja is van, nemcsak jelene.

1.4. Problémák

Annak ellenére, hogy az Internet Archive több mint 20 éve archiválja a globális webet, és számos országban már üzemszerűen működnek a webarchívumok, az internet megőrzésének feladata még messze nincs megoldva. Az egyik nagy probléma, hogy csak a weben van a hangsúly, és annak is leginkább a hagyományos változatait: honlapokat, blogokat, e-folyóiratokat, fórumokat stb. lehet jól aratni és visszanézhetővé tenni a jelenlegi technológiával. Az internet igazán dinamikus szegmensei, a közösségi platformok, a stream formában sugárzott rengeteg médiatartalom, az üzenő- és csevegőcsatornák, a dokumentum- és fájlmegosztó szolgáltatások, a szerverek közvetítése nélküli peer-to-peer kapcsolatokon zajló adatcsere, a számtalan, saját API-n kommunikáló mobilapplikáció és persze az egész *deep* és *dark web* kimarad ezekből a webarchívumokból. Vannak ugyan ezeken a területeken is próbálkozások, ígéretesnek tűnő szoftverek, de komoly méretű archívumot és szolgáltatást még senkinek sem sikerült ilyenekből felépítenie – és a gyors technológiai változások és magas költségek miatt nem is nagyon van rá esély.

Persze a meglevő webarchívumoknak is örülni kell, mert így is hatalmas értéket képviselnek, nélkülük teljesen elveszett volna a 20. század végi és 21. század eleji történelem internetes leképeződése. De ahhoz, hogy ne pusztán digitális *raktárak*, hanem tudományos kutatásra is alkalmas *digitális könyvtárak*, *levéltárak* és *múzeumok* legyenek, még sok mindent meg kell oldani, ki kell fejleszteni. Ha kicsit kutakodunk például az Internet Archive Wayback Machine felületén, rövid idő alatt feltűnnek a rendszer korlátai. Például csak URL-cím alapján lehet keresni. (Igaz, az idén megjelent új verzióban már az egyes webhelyek kezdőlapjára mutató linkek szövege is kereshető, de az évek óta ígért teljes szövegű kereső még mindig várat magára, ami nem is csoda, mert a feladat tulajdonképpen nagyobb, mint a Google keresőjét működtetni, mert annak „csak” az éppen létező webről kell releváns találatokat adnia.) Sok mentésnél jön hibaüzenet; vagy a linkek nem követhetők, vagy a menük és a belső keresők nem működnek – ezek mind a jelenlegi archiválótechnológia korlátait jelzik. Az automatikus módszerek mellett emberi felügyelettel működő, témára és minőségre fókuszált – ezért jóval kisebb – webarchívumoknál természetesen jobb a helyzet, de még ezeknél is bele kell törődni, hogy alapvetően töredékes, nagyon szemetes, rosszul strukturált, kevéssé metaadatolt és hatalmas bithalmazokról van szó, ráadásul tele gyorsan avuló fájlformátumokkal, amelyek megjelenítése külön problémákat fog okozni a távolabbi jövőben. Tipikus *big data* tehát egy webarchívum, amelynek a kutathatóvá, bányászhatóvá és vizualizálhatóvá tételéhez a már meglevő eszközök mellett még sok mindent ki kell fejleszteni. Hogy lesznek-e új generációs, a mainál jobb technikák az internet megőrzésére, a megőrzött tartalom elemzésére és feldolgozására, az jelentős részben attól függ, hogy a humán- és társadalomtudományok szakemberei mennyire igénylik ezeket, milyen innovatív kutatási módszereket találnak ki az internetes archívumok anyagának hasznosítására.

 DOI 10.31400/dh-hun.2018.1.129

2. Az internet és a történelem

Az internet viszonylag korán került a történeti kutatók látókörébe, természetesen először kommunikációs közegként, azután már a különféle források és irodalmak tárolásának és elosztásának módjaként is. S nem szabad megfeledkezni arról sem, hogy a nagyon korai időszakban fellelhető online könyvtárkatalógusok mennyire fontos tájékozódási pontot jelentettek. Érdekes, hogy itthon már a '90-es évek végén több kiadvány tárgyalta, illetve leírta a történész és a hálózat viszonyát,[56] a rengeteg külföldi hasonlóról nem is beszélve.[57] Az interneten fellelhető történeti források listáját és elérhetőségét ma már nem is lehet nyomtatott kiadványokban összegyűjteni.

A digitális bölcsészeti kutatások egyik legfontosabb területe a *történeti* lett, most erősen hangsúlyozva e szóhasználatnak a *történettudományin túlmutató* hatókörét. Az pedig magától értetődő, hogy a digitális bölcsészet számára a webarchívum *a jövő levéltára,* amelynek sokrétű kutatása nyomán a megismerési folyamatban új és új szempontok, módszerek, információs rétegek és persze eredmények tárulnak majd fel.[58]

2.1. A megőrzés biztonsága

Mindnyájunk tapasztalata a webes információ ilyen vagy olyan okokból való eltűnése. A webtér ingatagsága még a régi könyvek fennmaradáshoz képest is riasztó, a Vizsolyi Biblia mai fellelhetőségének valószínűsége 400 év távlatából is nagyobb, mint egy átlagos weboldalnak.[59] Ismertek példák arra is, amikor véletlenül vagy nemtörődömségből, de olykor nyilvánvaló tudatosság okán a mai politikai és gazdasági történések dokumentumai, jelenkori tudásunk forrásai tűnnek el a hálózatról.

A *The Web as History* című kötetben több példát hoznak a szerkesztők a *tudatos eltüntetésre*. 2013 végén fordult elő az a sajnálatos eset, hogy a brit Konzervatív Párt szervereiről törölték azokat a korábbi tartalmakat, amelyek a párt és David Cameron

[56] Sennyey Pongrácz, *A hálózat használata a történettudomány területén*, NIIF Információs Füzetek 1 (Budapest: NIIF, 1998). E sorozat több mint 20 füzetéből (szerk. Drótos László és Kokas Károly) több másiknak is vannak történeti referenciái, mint pl. a filozófiai, irodalmi és nyelvi vagy ókortudományi kiadványnak. A sorozat online fellelhetősége: `http://mek.oszk.hu/01200/01280/html/`); lásd még Komáromy Gábor, *Történelem az Interneten* (Budapest: Kossuth Kiadó, 1998).

[57] Ez utóbbiakról ad viszonylag korai képet a Daniel J. Cohen és Roy Rosenzweig által szerkesztett *Digital History: a Guide to Gathering, Preserving, and Presenting the Past on the Web* (Philadelphia: University of Pennsylvania Press, 2006). Az egyik legkorábbi és máig hivatkozott összefoglaló: Andrew McMichael, Roy Rosenzweig and Michael O'Malley, „Historians and the Web: A Beginner's Guide," *Perspectives on History* (1996 jan.), `https://www.historians.org/publications-and--directories/perspectives-on-history/january-1996/historians-and-the-web--a-beginners-guide`, `https://www.historians.org/publications-and-directories/perspectives-on-history/january-1996/historians-and-the-web-a-beginners-guide`.

[58] A digitális bölcsészet és a történeti kutatások általános helyzetképéről lásd Kokas Károly, „Digitális bölcsészet 2016: A bölcsészek és az informatikai megközelítés régen és most," in *MONOKgraphia: tanulmányok Monok István 60. születésnapjára,* szerk. Nyerges Judit, Verók Attila, Zvara Edina (Budapest: Kossuth Kiadó, 2016), 405–412, `http://publicatio.bibl.u-szeged.hu/10296/`.

[59] Horváth Iván is figyelmeztetett erre, lásd „A hálózat hátránya: fennmaradás helyett pusztulás?" in Horváth Iván, *Magyarok Bábelben* (Szeged–Budapest: JATEPress–Gépeskönyv, 2000), `http://magyar-irodalom.elte.hu/babel/2450.htm`.

számára kellemetlenné váltak. A politológus kutatók végül a törölt dokumentumokat a British Library webarchívumában találták meg. Az orosz-ukrán konfliktusban 2014-ben lelőtt utasszállító gép kapcsán „gondos kezek" szerették volna eltüntetni azt az orosz katonai körökből származó internetes hírt, miszerint orosz szakadárok lőtték volna le a gépet. De a bejegyzést az Internet Archive megőrizte a kutatóknak.[60]

Ezek a példák azt illusztrálják, hogy a webtér információi sorsdöntő ügyekben hiányozhatnak vagy előkerülhetnek, s ebben a kérdésben a webtér archiválásának döntő jelentősége van. Az esetek arra is figyelmeztetnek, hogy ebben a helyzetben sincs másképp, mint a hagyományos történeti kutatás anyagainál: bár úgy látszik, aminek nincs lenyomata, az nem is létezett, az előfordulhat, hogy megtalálható jobb keresési módokkal vagy más archívumokban.

Más szempontból különösen fontos e területen, hogy a *nyílt adat* (open data) mentalitás és gyakorlat uralkodóvá váljon. Ez az alapelv meghatározó módon elvárja az adatok *elérhetőség*ének és *hozzáférhetőség*ének a megvalósítását, az *újrafelhasználás* és *továbbterjeszthetőség* feltételeinek rendezettségét és a feldolgozásban az univerzális részvétel lehetőségét, amely *a lehető legkevesebb korlát* felállítását engedi csak meg. Mindennek egyfajta következménye a nagyon magas fokú *interoperabilitás,* amely a különböző adathalmazok együttműködésének vagy vegyíthetőségének lehetőségét is jelenti, hisz ez teszi lehetővé a *különböző komponensek együttműködését.* Mindez nem csupán az adatbiztonsággal, az elérés és kutathatóság demokratizálásával függ össze. Ez a gondolkodás maximálisan támogatja a kutatások, az adatbányászati módszerek legteljesebb használatának lehetőségét is.[61]

2.2. A webtér mint történeti forrás[62]

Az a gondolat, hogy maga a teljes internet forrása a történeti kutatásoknak, valamivel később, illetve párhuzamosan kerül elő a webarchiválási projektek értelmének és szükségességének indoklásaiban.[63] Úgy tűnik, a webtér és a történész viszonyának legalább négy vonatkozása van:

- a kifejezetten történeti kutatás számára készült repertóriumok, adatbázisok és más szolgáltatások;
- ezek a webarchiválás részeként mentésre kerülnek, így másodlagos elérhetőséget is ad az adott webarchívum, illetve gondoskodik arról, hogy az eredeti anyag ne tűnjön el a webtérből;

[60] Niels Brügger and Ralph Schroeder, „Introduction: the Web as History," in *The Web as History: Using Web Archives to Understand the Past and the Present,* eds. Niels Brügger and Ralph Schroeder, (London: UCL Press, 2017), 1–2.

[61] Vö. *Open Data Handbook*, hozzáférés: 2018.05.22, `http://opendatahandbook.org/guide/hu/what-is-open-data/`.

[62] A kérdés legújabb és legátfogóbb vizsgálata a fent idézett *The Web as History* kötetben. Ez a szerkesztett kötet az első monografikus igényű kiadvány, amely arra összpontosít, hogy miként lehet a múlt archivált webtartalmait felhasználni a társadalom fejlődésének széleskörű kutatásához.

[63] Minderről általában és összefoglalóan: Niels Brügger, „Web History and the Web as a Historical Source," *Zeithistorische Forschungen/Studies in Contemporary History, Online-Ausgabe* 9, 2. sz. (2012): 316–325, `http://www.zeithistorische-forschungen.de/2-2012/id=4426`.

 DOI 10.31400/dh-hun.2018.1.129

- mindennek egy külön vonatkozása, hogy a periodikusan lementéseket tartalmazó webarchívum rétegeiben *az online történelmi kutatások eszköztárainak historiográfiája* is kutathatóvá válik;
- a *webtér archiválása,* amikor az internet maga válik történeti forrássá.[64]

Dolgozatunk szempontjából természetesen most ez az utóbbi mód és út a lényeges. A web negyedszázada kezdte el átfogni, befogadni és befolyásolni az életünket. Ez a folyamat olyan gyorsan zajlott le, és a trend növekedése olyan intenzív volt, hogy már jelen állapotában szinte teljes körűnek mondható. Nyilvánvaló, hogy aki az elmúlt 25 év történetével – bármilyen aspektusú történetével – kíván foglalkozni, az nem kerülheti el a webtér vizsgálatát.

2.2.1. A szakirodalom, a módszertan és a forrástípusok A webarchívumok történeti kutatásokban való felhasználásának a tapasztalata értelemszerűen nem túl nagy. Bár számos példa van már arra, hogy történeti, irodalomtörténeti stb. tanulmányok hivatkozásában találunk webarchívumos adatot (lásd az alábbi fejezetekben), az archivált webtér mint forrástípus módszertani megközelítésének irodalma és a téma elméleti feldolgozása még kezdeti állapotban van. Ha rákeresünk a problémát legjobban megragadó *web historiography* kulcsszóra, ma még alig találunk többet, mint pár tucat cikket és néhány könyvet. Kis túlzással a hivatkozott szakirodalom felét-harmadát a téma legaktívabb kutatója, Niels Brügger (aki talán a *webhistoriográfia* szakkifejezés első alkalmazója is lehet) írta, szerkesztette vagy inicializálta, aki a dán Centre for Internet Studies vezetője és az Aarhus Egyetem professzora.

A szakirodalmi feldolgozás hiányosságainak – azon túl, hogy a webaratások és projektek kiteljesedése alig több mint egy évtizedes múltra tekint vissza – az is oka, hogy ennek a forrástípusnak az elemzése és jellegzetességeinek a feltárása valószínűleg a legkomplexebb forrásismereti kihívás a kutatóknak.

Ennek okai – többek között – a következők lehetnek:

- a vizsgálatok óriási terjedelmű adatokra vonatkozhatnak *(big data)*;
- rengeteg más, korábban már létező, de egymással föltétlen össze nem kapcsolt egyéb forrást kívánunk egyben kezelni;
- minden létező, elektronikusan egyáltalán reprodukálható és megragadható médiatípus része lehet a fájlcsomagnak;
- az adattömeg belső rendje és metaadatolása nagyon különféle lehet, s annak metódusához még nincs kialakult gyakorlat;
- a megjelenő és a kívánatos módszerek a történészi körökben még kevéssé ismert és naponta újdonságokat felmutató mesterséges intelligencia (MI, angol rövidítés AI) határmezsgyéjén mozognak;

[64] Érdemes ehhez referálni a Web Archives for Historical Research (WAHR) csoport weblapját, ahol így fogalmaznak: „This project is among the first attempts to harness data in ways that will enable present and future historians to usefully access, interpret, and curate the masses of born-digital primary sources that document our recent past.“ WAHR, hozzáférés: 2018.05.22, `https://uwaterloo.ca/web-archive-group/about`.

- ha a *webhistoriográfia* a történettudomány segédtudományává válik majd, különös jellemzője lesz, hogy több tudományterületet maga is segédtudományaként használ (az informatika, könyvtártudomány, matematika, szociológia, statisztika és szociálpszichológia biztosan előkerül ebben a kontextusban);
- különlegessége a forrásnak, hogy látványosan a jövőnek készül, hiszen úgy gondolják a készítők, hogy a puszta elmentése a webtérnek a benne foglalt primer információkon túl később a fejlettebb módszerekkel egyre inkább és sokkal hatékonyabban kutatható lesz.

A fentieket és a hozzájuk hasonló szempontokat, megfigyeléseket most érdemes felvetni. Ezek felderítéséhez és igazi megoldásához azonban az apró részleteken keresztül vezet az út, vagyis részletes esettanulmányok sokaságának elkészítésével.[65]

2.2.2. **A teljesség és a rész** A webarchiválás kezdeteitől világos tehát, hogy mindennek a teljes kultúra történetére és annak kutatására vonatkozó aspektusa is van. Különösen felerősödött ez a gondolat akkor, amikor később nyilvánvalóvá vált, hogy a politikai történések jelentős része a weben zajlik, és az összes többi médium is azt referálja. A szubkultúrák szinte teljesen az internetre költöztek, nem beszélve arról, hogy a fiatalok majdnem kizárólagos globális kommunikációs környezete a 21. század elejétől maga a webtér lett (YouTube, Facebook, Twitter stb.), hogy aztán az évszázad második évtizedével már a fejlett világ általános kommunikációs közművévé váljon.

A jövőben az archivált web várhatóan fontos szerepet kap majd a média- és a kommunikációtörténetben, de azon történészek, politológusok stb. számára is nélkülözhetetlen lesz, akik a közelmúlt folyamatainak mélyrétegeit vizsgálják. Mindebből következik továbbá, hogy érdemes vizsgálni az archivált webtér történeti forrásként való kezelésének sajátosságait, valamint azt, hogy milyen kihívások elé állítja a történészt az új médium, amelyet forrásként kíván kutatni.[66]

A mentett webtömeg (akár óriás projekteket, akár tematikus vagy nemzeti programokat nézünk) méreténél és átfogó természeténél fogva speciális terepet ad a kutatásnak. Maga a puszta adatmennyiség és az a tény, hogy az egészében is indexelt (ellentétben például azzal, hogy a levéltárak és azok együttese nem kutatható egészében digitálisan), különös lehetőségeket ad, de meg is rettenti a felhasználót. A webtérnek is megvannak a valósághoz képest a maga torzításai, amelyek hamis illúziót kelthetnek, hiszen az, hogy mi, mikor és milyen mértékben kerül a webre, nem föltétlen függ össze a dolog súlyával és fontosságával. Például a globális webtérben föltétlen vizsgálandó és figyelembe veendő az, hogy itt mindig is egyértelmű volt az angolszász kultúra, s különösen az angol nyelv dominanciája. Ezen belül a folyamat elején – az internet amerikai eredete miatt – az amerikai kulturális hatás egészében befolyásolta a webtermést

[65] Vö. Josh Cowls, „Cultures of the UK Web," in *The Web as History: Using Web Archives to Understand the Past and the Present,* eds. Niels Brügger and Ralph Schroeder (London: UCL Press, 2017), 220–237. Itt a .uk tartomány alá eső mintegy 65 terabájtos mentés módszertani elemzéséről van szó, hogy az a lehető legjobban kutathatóvá váljon. A British Library, a University of London Történeti Kutatóintézetének és a University of Oxford Internet Intézete által vezetett Big UK Domain Data for Arts and Humanities (BUDDAH) projekt eredményeit mutatja be ez a dolgozat.

[66] Brügger, „Web History," 316–325.

 DOI 10.31400/dh-hun.2018.1.129

(például a franciák és a francia nyelvű anyag később érkezett és más prioritásokkal), míg a jelenben éppen a kínai webkultúra erőteljes térfoglalása zajlik, ami az egész webtér arányait is megváltoztatja. De ez nem csak a nyelvi kultúrákra igaz: vannak jellemzően hamar webre kerülő témák, intézmények, és vannak olyanok, amelyek ezen a téren sokkal visszafogottabbak (hasonlítsuk össze például a filmkultúra és az egyházi tartalmak webfoglalásának tér- és időszerkezetét). Ezeknek a szempontoknak a jelenkor-történeti webkutatásoknál való felvetése és bevitele külön lehetőségeket és persze veszélyeket rejt.

Ismert mondás, hogy a jövő generációi számára az a létező, amit a webtérben fellelhetünk. De a történésznek máshogy illik gondolkodnia, mert nem a webtartalmak történetét kutatja valójában, hanem a valóság lenyomatait keresi. Így számolnia kell azzal is, ami nincs, vagy ami abból a korszakból még nem hagyott lenyomatot a hálózaton. El kell különíteni az illúziókat a valóságos arányoktól.

E belső aránytalanságokat növeli és növelheti, hogy a már elkészült és működő webarchívumok kutathatósága sem egyforma: különféle személyiségi és szerzői jogi megfontolások korlátozhatják. A hozzáférés puszta kényelmének is van kutatást torzító hatása.[67]

2.3. Forrástípusok és esetek

Nem képzelhető el a lementett webtér jó történeti hasznosítása anélkül, hogy magának az anyagnak a természetéről ne alkotnánk pontos képet. Ez is különleges tudást követel, nyugodtan betehetjük hát a webtérarchiválást a kronológia, a pecséttan, az oklevéltan mint történeti segédtudományok közé. A mai kutatások ezen a téren jószerivel csak példákat produkálnak, néha sporadikusan, de a kísérletezés veleje, hogy megtaláljuk és kialakítsuk a megfelelő módszertant, meglegyenek azok a szempontok, amelyek a kutatás e speciális szegmensében érvényesíthetők.

Természetesen a történeti forráskezelésben feltett kérdések sorra következhetnek: mi pontosan a forrás? Milyen típusú „objektum" az? Annak egészéről vagy részéről beszélünk? Ki hozta azt létre? Köthető-e pontos időponthoz? Mi annak a hitelessége? stb. Látható, hogy a szokásos gondok a webtér történeti kutatásában sokszor egészen másképp vetődnek fel, olykor még az is előfordulhat, hogy a századokig természetesnek tartott forráskezelési szempont nem érvényes vagy egyenesen értelmezhetetlen. Az a problémák egészen különleges vonatkozása e kontextusban, hogy számtalan hivatkozást kapunk és kaphatunk számunkra elérhetetlen vagy éppen elveszett forrásokra, gondoljunk csak az anyagban beágyazott linkekre, amelyek *nem* részei egyik archívumnak sem, illetve a rengeteg, akár elvileg a webtérben lévő, de magánjellegű szövegre (elsősorban a levelezésekre).

A másik szokatlannak tűnő szempont, hogy az analóg világ forrásaival ellentétben itt sokszor *elvileg* rendelkezhetünk az adott webdokumentum mások által való felhasználásának adataival is, tehát azzal, hogy számszerűsíthetően az adott forrást hányan és mikor használták. Mekkora hatást gyakorolhatott? A gondolat, történés recepciótörténete így a jövőben egészen új aspektust kaphat, mert a részben vagy

[67] Vö. Niels Brügger and Ralph Schroeder, „Introduction: the Web as History," 1–17.

egészben alátámasztott vélelmek, feltételezések helyett nagyon konkrét hivatkozásokkal számolhat. Mindennek persze elsődleges feltétele az, hogy a webarchiválás és annak metainformációs rendszere és struktúrája alkalmas legyen ilyen vizsgálatokra.

Magáról a webes információ természetéről is érdemes megfontolásokat tenni. Minden történész tudja, mi a különbség – egy adott témában – a levéltárban feltárt jegyzőkönyv, egy egykori magánlevél, egy újságban megjelent vezércikk vagy riport, illetve egy politikusi nyilatkozat között. Ismereteink vannak ezekről a forrástípusokról, tudjuk őket azonosítani, s értékük és jellegük szerint kezeljük őket. Vajon megvane ez a biztonsága a történésznek a hálózati források esetében? Akkor, amikor azok nem az analóg kultúra puszta digitális lenyomatai, hanem egészen új típusú források, olyanok, amelyek már a hálózat világában születtek, mint például a blogbejegyzések vagy egy cikk kommentjei.

S természetesen külön módszertani probléma az, hogy a webarchiválás konténerszerű, magában foglaló természete révén a legkülönfélébb digitális objektumok szerepelhetnek az éppen használt archív anyagban, amelyeknek mind megvannak a tartalmi és formai sajátosságaik, sokszor nemcsak értelmezési, de műszaki probléma elé is állítva a kutatókat. (Az magában külön téma lehetne, hogy ezen médiumtípusok némelyike igen kicsiny múltra tekint vissza, ezért az elemzés is gyerekcipőben jár. Gondoljunk például a mobileszközök révén elszaporodott videódokumentumokra vagy a digitális tudás szélesülésével elkövetett – vidám vagy komoly – hálózati hoaxokra, továbbá az ún. mémkultúrára, amely magában is a politika s később a történelem nagyon különleges *karikatúramozzanata* lesz.)

2.3.1. A keresés mint a kutatás maga A szakirodalomban már kirajzolódik, hogy ezen új típusú és iszonytató tömegű forrás kapcsán átértelmeződik a *keresés* fogalma is, hiszen a jól keresni tudás a magas szintű kutatásnak egyik alapfeltétele lesz.[68] A holland kutatók (egy egyetem, egy kutatóintézet és a nemzeti könyvtár együttműködésében) a *search as research* gondolathoz metodológiát és a különféle lehetséges forgatókönyvekhez példákat is publikáltak. Külön kiemelik, hogy vizsgálatuk célja elsősorban a keresőmotorok algoritmusainak tanulmányozása, melyek a felhasználók számára valójában rejtett módon rendezik és rangsorolják a találatokat (például a látogatottság, vagy az IP-címtartomány földrajzi közelsége szerint). Mindezért az ún. lekérdezés-tervezés, illetve a keresési eredmények összehasonlító elemzése azt a célt szolgálja, hogy a különféle keresőmotorok algoritmusát (a webtörténész) a nyilvánossághoz közelebb hozza, és azokat átláthatóvá tegye. A kereshető webarchívumok esetében maga az egész archívum is egy ún. analitikai egység lehet, amelyet globális *big data* metódusokkal is lehet értelmezni, ahol a szövegbányászati algoritmusok (mint például az egyre gyakrabban emlegetett *n-gram* feldolgozás), a mesterséges intelligencia bevonása (statisztikai alapú elemzések), és persze a grafikonmegjelenítés és -elemzés is fontos eszköz lehet.

Nem tárgya most dolgozatunknak, de felvethető, hogy a webarchívumon kívül a valóságos webkeresésben ugyanezek a szempontok, a keresőmotorok által előállított

[68] Anat Ben-David and Hugo Huurdeman, „Web Archive Search as Research: Methodological and Theoretical Implications," *Alexandria* 25, 1–2. sz., 93–111.

 DOI 10.31400/dh-hun.2018.1.129

eredmények milyen viszonyban vannak a teljes valósággal? És mindez vajon milyen mennyiségben és minőségben befolyásol mindent, amit az internetes keresési eredmények következtében teszünk és visszatermelünk a webtérbe, hogy aztán az, legalábbis részben, kvázi manipulált webtörténelemmé váljon? (E ponttól – a szövegvilágok befogadásában is – eljuthatunk akár Heisenberg nevezetes határozatlansági relációjának *bölcsész applikációjához*, a Luhmann alkotta tézishez, ti. a megfigyelő és megfigyelés torzító hatásmechanizmusának leíráshoz, hiszen ebben az esetben és gondolatmenetben lényegében ugyanarról van szó.)[69]

2.3.2. Esettanulmányok és példák A történeti webkutatás fejlődésének jelen szakaszában a fenti és a fentihez hasonló elméleti és módszertani megfontolások mellett nagyon nagy szükség van jól kidolgozott és alaposan dokumentált esettanulmányokra. Messze vagyunk még attól, hogy mindez rutinná váljon, és ameddig a témán belüli specializáció le nem zajlik, addig fontos, hogy minden érintettet (majdnem) minden kutatás érdekeljen. Ennek oka, hogy ezen kutatásoknak a megismerése intuitív lehet: ötleteket és módszertani tanácsot adhat, új területekre, friss gondolatokra hívhatja fel a figyelmet, s az utánzásnak és az analógiáknak nagy jelentősége lehet. Ilyen szempontból is óriási haszna van az említett összefoglaló kötet esettanulmányainak, illetve a weben egyre szaporodó, tárgyunkba vágó iniciatíváknak.

Az ismert és tekintélyes webarchívumok anyagának hasznosításával tehát sorra születnek az elemzések, és ezek – madártávlatból nézve – legfőbb szembetűnő jellegzetessége a fantasztikus változatosság. Már most meglepő, hogy hányféle módszer és megközelítés lehetséges, s mennyi különleges metszete rajzolható ki az anyagnak. Vannak már nagyobb és átfogóbb, hagyományos történészi megközelítésben gondolkodó feldolgozások is, de talán érdekesebbek azok, amelyek különleges és friss szempontokat vetnek fel.

Egy egészen friss tanulmány a webarchívumok fontosságáról a humán tudományokban azt boncolgatja, hogy szinte minden archivált információ fontos lehet történelmi szempontból. Sőt, a kutatási szempontok mindig újabb és újabb vetületét mutathatják az anyagnak, mindez csak attól függ, hogy milyen kontextusban folyik éppen a kutatás. Például egy ingyenes hirdetés, amelyben egy használt gyerekkerékpár képe szerepel, első látásra bagatellnek tűnhet, és legföljebb az általános megőrzés szempontjából fontos. Azonban húsz év múlva kiderülhet, hogy ez a híres kerékpárbajnok első kerékpárja, vagy érdekes lehet – hasonlókkal együtt – egy olyan kutató számára, aki éppen a kerékpártervezés technológiai fejlődését kutatja. Ez a példa jól mutatja, hogy milyen nehéz egy webarchivátornak eldönteni, mi a fontos és mi nem a jövő szempontjából.[70]

Az idősorokban és a földrajzi vonatkozásokban *(georeferenciák)* is nagy potenciál van. Az említett tanulmány hangsúlyozza ezt, és példaként bemutatja, hogy a 2003-ban archivált portugál választási eredmények állami kezelésű honlapjai hogyan alkotnak az archívumban változatosan vizsgálható idősort egészen 1997 óta. Mindennek

[69] Vö. Szajbély Mihály, *A nemzeti narratíva szerepe a magyar irodalmi kánon alakulásában Világos után* (Budapest: Universitas Kiadó, 2005). A könyv az első magyar nyelvű kísérlet a Niklas Luhmann által kidolgozott rendszerelmélet adaptációjára a kifejezett bölcsészettudományok területén.

[70] Daniel Gomes and Miguel Costa, „The Importance of Web Archives for Humanities," *Journal of Humanities & Arts Computing* 8, 1. sz. (2014), 113–114.

következtében a 2002-es választások eredményeinek vizsgálata és webprezentációs története is más megvilágításba került.[71]

De nem csupán a történelmi tények tekintetében fontos a webtér kutatása. Egy ausztrál vizsgálatban a hosszan elnyúló abortuszvitát illetően például érdekesebbnek tűnik maguknál az elsődleges eredményeknél az a szempont, hogy a *hálózaton való megtalálhatósága* a különféle álláspontokat közvetítő weblapoknak (az ún. *pro-choice* és *pro-life* érvelésnek) mennyire befolyásolhatta az álláspontok és vélemények alakulását.[72] Így azután a keresési metódusok, a különféle keresőszolgáltatások fontosabb szereplőként tűntek fel, mint azt korábban gondolták. (Kicsit hasonló ez ahhoz, hogy az adott hír milyen nézettségű televíziós csatornán vagy hogy melyik idősávban látható.)

A tartalomelemzés, a szövegbányászat és annak teljes módszertana egyrészt jóval régebbre nyúlik vissza, mint a webarchiválás, másrészt még maga is fejlődésének kezdeteinél tart, természetesen rengeteg eredménnyel és nagyon sok projektben. Nem nehéz elképzelni, hogy ezek a metódusok, a konkrét szoftverek és a hozzájuk társuló gráfos és grafikonos kimenetelű vizualizációs eljárások, a mesterséges intelligenciát felmutató programok rendre megtalálják illetve egyre jobban felfedezik a webarchívumok lehetőségeit is. Ez a folyamat még az elején tart, de kétségtelenül az egyik legnagyobb kutatási potenciállal és perspektívával rendelkező terület.[73]

3. Összefoglalás és a jövő

A nemzeti és intézményi webarchiválás minden egyes eleme fontos, de mint a Library of Congress 2001-es (!) webmegőrzésre irányuló jelentése is kimondta, ajánlatos a jövőnek való megőrzés biztonsága érdekében, hogy több nemzeti és más intézmény végezze a webtér lementését. A jövőben kiemelt fontosságú lesz, hogy ezek a projektek – akár világméretűen is – koordinálva legyenek, tervezésüket, működésüket a lehető legjobban összehangolják.[74]

Elmondhatjuk, hogy a webarchiválás a kezdeteinél tart, tudományos hasznosítása pedig sokszor még tapogatózás jellegű. Az azonban világos, hogy a jelenkortörténet kutatásának nemsokára fő eszköze lesz. Olyan lesz számukra a webhistoriográfia, mint például a középkorászoknak az oklevéltan. S valljuk be, aki az archontológiában,

[71] D. Gomes and M. Costa, „The Importance of Web Archives for Humanities," 111.

[72] Robert Ackland and Ann Evans, „Using the Web to Examine the Evolution of the Abortion Debate in Australia, 2005–2015," in *The Web as History: Using Web Archives to Understand the Past and the Present,* eds. Niels Brügger and Ralph Schroeder (London: UCL Press, 2017), 186–187.

[73] A sok közül minderre jó példa lehet az alábbi, ahol nem webarchívumokban, hanem egy tudományos folyóirat archívumában próbáltak szoftveres elemzést prezentálni, de nyilvánvaló, hogy hasonlókat webaratásokban is el lehet képzelni. J. Comins and L. Leydesdorff, „RPYS i/o: Software Demonstration of a Web-based Tool for the Historiography and Visualization of Citation Classics, Sleeping Beauties and Research Rronts," *Scientometrics* 3 (2016), `https://arxiv.org/ftp/arxiv/papers/1602/1602.01950.pdf`.

[74] „Many organizations and individuals will collect and preserve materials independently of any relationship to the Library of Congress. Indeed this independence is important for preservation. If different organizations, in different countries with different cultures, carry out different preservation programs the materials that they collect are vulnerable in different ways. This diminishes the risk of everything being lost through a single disaster or mistakes of technology and organization." William Y. Arms, „A Report to the Library of Congress: Web Preservation Project Final Report," Library of Congress, 2001. jún., `https://www.loc.gov/webarchiving/include/webpresf2.pdf`.

 DOI 10.31400/dh-hun.2018.1.129

diplomatikában, heraldikában és társaiban nem jártas valamennyire, nem is lehet igazi középkorkutató. A webarchívum történeti hasznosítása azonban magára a történeti kutatásra is visszahat, hiszen olyasmiket tudhatunk általa – rejtett trendeket vagy egyáltalán megőrzött források sokaságát, a nagy időszeletek kutatásának lehetőségét stb. –, amelyek megsokszorozzák a történeti anyag mennyiségét, reflektálhatóságát és egészen új szempontokat vetnek fel.

A szerzőknek, akik könyvtárosok, engedtessék meg még egy szempont: a struktúra és a tételek feltárása a webarchívumban. Köztudott, hogy a hagyományos levéltárak, kézirattárak és könyvtárak tudományos (pl. történészi) használatának színvonala mennyire függ az anyag rendezettségétől és feltártságától. Ez ránk, könyvtárosokra (levéltárosokra, archivátorokra) nagyon nagy felelősséget ró, hiszen a metaadatolás kulcskérdés lesz.[75] E folyamatokban jól látható, hogy az igazi digitális bölcsészeti kihívás az, hogy a kutatói (történészi), technikai és könyvtári készségek, motivációk össze kell, hogy kapcsolódjanak. Ez nem csupán az ismeretek kibővítését jelenti, hanem az együttműködések egy magasabb szintjét is.

Igaza van Niels Brüggernek, hogy a kutatóknak nagyon sok segítségre van szüksége, hiszen ma még a történészek nem ismerik a webes archiválásban felmerülő, folyamatosan változó technikai kihívásokat, talán a webmegőrzés hosszú távú biztosításának jelentőségét sem.[76] Ezért a szakembereknek nem csupán a tulajdonképpeni feladatra, hanem arra is kell gondolniuk, hogy proaktív módon artikulálják az archiválás szükségességét. Mindezt annak érdekében, hogy ösztönözzék a meglévő kulturális örökségi intézményeket helyi, tematikus vagy nemzeti webarchívumok létrehozására, és megmutassák, szemléltessék e nagyszerű új forrásegyüttes felhasználásának lehetőségeit.

Web Archiving and Historical Research

Born digital content offers a more detailed and complete record of the present than what traditional sources provided of the past. The first section of this paper surveys current efforts and technologies to capture the present digital universe and reflects on the limitations of current web archives. The first attempts to archive the Hungarian web were made in 2017 in the National Széchényi Library, Budapest. The second part explores how this content could be harnessed for historical research, and how it will become the principal source of our recent past. The authors point out how web archives, and the resulting scale of data, will require new strategies and methodologies to deal with born digital sources effectively. They also show that born digital sources will also make it possible to pursue new types of inquiries that yield new results.

Keywords:
web archives, digital preservation, digital humanities, web historiography

[75] Vö. a témába vágó honlap, a *Web Archives for Historians* egyik konferenciabeszámolójával: Peter Webster, „What Do We Need to Know About the Archived Web?”, hozzáférés: 2018.05.22, `https://webarchivehistorians.org/2016/11/17/what-do-we-need-to-know-about-the--archived-web/`.

[76] Brügger, „Web History,” 316–325.