

Megkezdődhet a magyar internet-tartalmak módszeres mentése – évekkel a hasonló külföldi kezdeményezések után, és csak akkor, ha elhárítanak számos technikai és jogi akadályt

Az első szoftver megtervezi és elindítja, a második pedig elvégzi a mentést. A harmadik elmenti azt is, amire a második nem volt képes, a negyedik segítségével pedig a felhasználók megnézhetik, mi került az archívumba. Már ha került egyáltalán valami. Mert igen gyakran előfordul, hogy hiba történik, és csak részlegesen vagy sehogy sem tudják menteni egyik-másik oldalt. Valahogy így működik most az internetarchiválás fáradságos folyamata, amellyel mostanában az *Országos Széchényi Könyvtár* (OSZK) szakemberei is küszködnek.

Egy évtizednyi tervezgetés után egy munkacsoport ezekben a hónapokban készíti elő, hogy (remélhetőleg) jövő évtől módszeresen tudják menteni a magyar interneten fellelhető tartalmakat. Erre annál is nagyobb szükség van, mert bár inkább azt szokás emlegetni, hogy „az internet nem felejt”, egy-egy weboldal átlagos élettartama mindössze egy-két hónap, aztán eltűnik vagy megváltozik. A hírportálok esetében pedig ez az idő sokszor csak egy-két napnyi. Az elmúlt két évtizedben folyamatosan nőtt a kizárólag az interneten megjelenő (majd onnan eltűnő) tartalmak aránya.

Korlátozott kereshetőség

A világ legnagyobb internetes archívuma, a San Franciscó-i székhelyű Internet Archive (IA) nonprofit szervezet 1996 óta rendszeresen készít ugyan mentéseket, de az sem elég, hogy már több mint 300 milliárd weboldalt gyűjtött össze. Az információk mentése szórványos (ami például egy bizonyos újságcikk megtalálását igen valószínűtlenné teszi), és a tartalom nincs indexálva. Ezért nem működik az olyan címszavas kutakodás, mint a megszokott internetes keresőgépeknél, hanem csupán az ta-

lálható meg, amiről a kíváncsiskodó már tud valamit: dátum, internetcím (URL) és – tavaly óta – az oldal neve alapján lehet keresgélni.

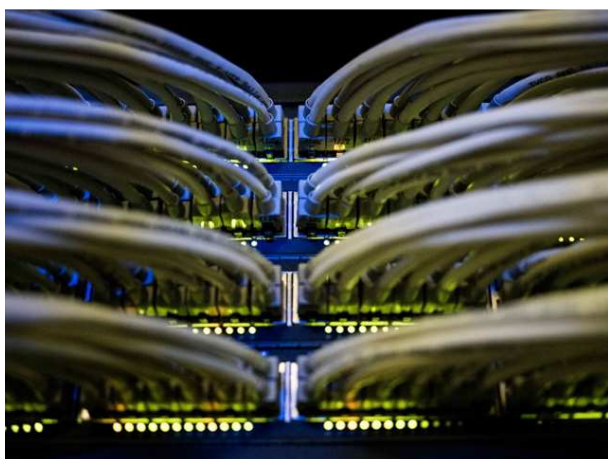
A teljes szövegű keresés a tervezett reprezentatív (vagyis lehetőleg teljes körű) magyar internetmentések esetében sem fog működni, hiszen azokat hasonló módszerrel és szoftverrel végzik majd, mint az IA gyarapítását – magyarázza az OSZK-s munkacsoport vezetője, *Drótos László* főkönyvtáros. A szakmában webaratásnak nevezett művelet kulcsszereplője, a Heritrix nevű szoftver válogatás nélkül „dobálja” egy fájlba a talált adatokat, aztán lezárja ezt a „tartályt”, amikor elér egy megadott méretet, mondjuk egy gigabyte-ot. Az ebben lévő tartalom indexálva lesz ugyan, és készíthetők róla statisztikai kimutatások, sőt akár a benne található linkek kapcsolati hálója is összerakható, de a benne lévő szavakra keresni már nem lehet. Ráadásul, mivel az aratások párhuzamosan rengeteg honlapon zajlanak, egy weboldal részei akár több konténerbe is kerülhetnek.

Sok technikai problémát kell még megoldani a rendszeres archiválás elindításához. Gyakran megghiúsulnak a mentési próbálkozások, amikor a felkeresett site-on valami megakasztja az aratóprogramot: például a nem logikus felépítés, a következetlen kódolás vagy egy olyan speciális formátum, amelyet csak valamilyen böngészőkiegészítővel lehet megtekinteni. Ezeket a típushibákat is gyűjtik az OSZK-ban, ahol az év végéig futó tesztidőszak során több tucat hazai honlapot próbálnak rendszeresen menteni (a demóváltozat néhány héten belül nyilvános lehet), és a tapasztalatok alapján egyrészt egyeztetnek a szolgáltatókkal, másrészt általános ajánlást készítenek, miként lehet aratórobot-baráttá tenni egy weboldalt.

Jogilag sem minden tiszta

Fel kell készülni arra is, hogy az oldalak jó részén különféle csapdák leselkednek a szép magyar néven szüretelőrobotokként emlegetett archiváló-szoftverekre. Becsapósak például a beépített naplárak, amelyek napjait és bejegyzéseit a végtelen-

ségig mentetheti egy rájuk találó robot – kivéve, ha ezt nem próbálja kivédeni a program elindítója vagy az oldal üzemeltetője. Mivel Heritrixet és társait az 1990-es évek derekán még az akkori weboldalakra fejlesztették, egyelőre leküzdhetetlen kihívást jelentenek számukra a mára elterjedt, dinamikusan változó tartalmú, a felhasználók által folyamatosan alakított honlapok (amilyen például a Facebook). Ezeken a site-okon a korábbi bejegyzések csak görgetésre töltődnek le a szerverről, ezért olyan speciális aratóprogramokkal kísérleteznek, amelyek ezt is képesek szimulálni – mondja Drótos.

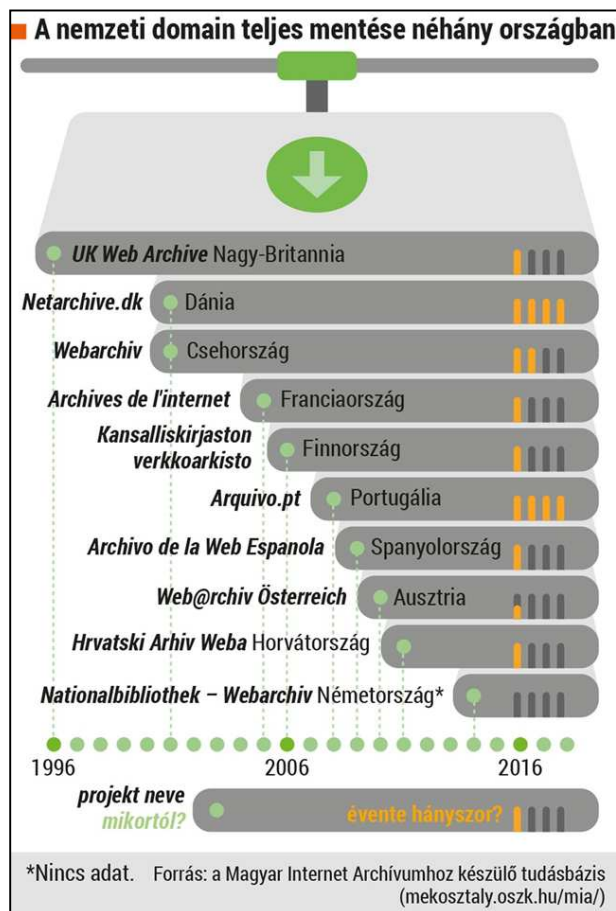


© Stiller Ákos

Még komolyabb akadály, hogy a honlapok döntő többségén nem látják szívesen a robotokat, ezzel pedig a szüretelőket is kitiltják. Be lehet ugyan állítani úgy a szoftvert, hogy megkerülje a tiltást, ám a nemzeti könyvtár archiválói nem erre készülnek, hanem arra, hogy akár a szolgáltatókkal egyénileg megállapodva, akár jogszabály segítségével jussanak be a most még elzárt helyekre.

Akadály az is, hogy a legutóbb 1998-ban módosított kötelezpéldány-rendelet még nem foglalkozott a honlapokkal. Az OSZK illetékesei úgy vélik, ideje lenne azokat is bevonni e körbe, hogy könnyebb legyen az internetes archiválás. A külföldi példák többnyire azt mutatják, hogy a sokfelé már az elmúlt másfél évtizedben elindult webaratást (lásd infografikánkat) a jogalkotás kissé késve követi. Vagy sehogyan sem: Hollandiában egy évtizede végez már archiválást a nemzeti könyvtár, ám ehhez máig egyesével kérnek engedélyt a tartalomszolgáltatóktól. Teljes mentést azért sem készítenek a holland webről, mert több mint 8 millió honlap tartozik bele, ezért csak a legfontosabbakat szüretelik, jelenleg 16 ezret – említette az OSZK-ban e témában rendezett októberi műhelybeszélge-

tésen Kees Teszelszky, a Holland Nemzeti Könyvtár munkatársa.



© HVG

Az olvasni tudó szoftvereket is meg kell őrizni

Magyarország ennél könnyebb helyzetben van: a magyar internet „csupán” nagyjából 700 ezer honlapból áll. A mostani tervek szerint évente kétszer készítenének róla nagy, reprezentatív mentést a nemzeti könyvtárban. A .hu domain alatti oldalak mellett learatnák a határon túli, a magyarsághoz köthető site-okat is. Becslések szerint mindez alkalmanként néhány száz terabyte-tól legfeljebb egy-két petabyte-nyi (1000–2000 terabyte) adatot jelentene – biztonsági másolattal, indexálással együtt. A mentések során csak 2–3 szintig „menének le” a robotok a honlapokon, videókat pedig nem töltenének le.

Ez a reprezentatív adatállomány nem is lenne nyilvános, nem úgy, mint a válogatott weblapok archívuma, amelyet sokkal rendszeresebben (egy-két havonta, de szükség szerint akár gyakrabban is) és nagyobb mélységben frissítenének, valamint

teljes szövegűen kereshetővé tennének. Elsősorban kulturális és tudományos intézetek (könyvtárak, múzeumok, levéltárak, kutatóközpontok, egyetemek), kormányzati intézmények és sajtótermékek oldalait gyűjtenék itt, de a szakemberek szerint akár az internetező nagyközönség is ajánlhatna az utókor figyelmére méltó tartalmakat. Arról azonban még nincs döntés, hogy pontosan milyen archív anyagot ki hol és milyen feltételekkel tanulmányozhat. Erre a világszerte mostanáig felépíteni kezdett mintegy negyven nemzeti archívumban sincs egységes szabályozás.

Az utókor csak akkor lesz képes tanulmányozni a múltat a webarchívumokon keresztül, ha a tartalom olvasására alkalmas szoftvereket is megőrzik. Kérdés persze, mire lehet majd használni évtize-

dek múlva a most gyűjtött információtömeget, amelyben az értékek mellett nyilvánvalóan rengeteg a szemét is. Még akkor is, ha a – többségében külföldi szerverekről működtetett – magyar torrent- vagy pornóoldalak tartalmait a nemzeti könyvtár nem fogja szűrni. Drótos László válasza:

„A régészek számára a szemétdomb az egyik legfontosabb információforrás. Sokkal többet elmond a mindennapi élet apróságairól, mint a szépen megrajzolt kódexek.”

Forrás:

http://hvg.hu/tudomany/201803_a_magyar_internet_archivalasa_aratoszoftverek_digitalis_szemetdomb_mentes_ugyanugy#rss

Válogatta: Fonyó Istvánné