A TUDOMÁNY, AZ OKTATÁS ÉS A KÖZGYŰJTEMÉNYI KISZOLGÁLÁS ÚJ INFORMATIKAI SZINERGIÁI

NETWORKSHOP 2026
35. Országos Informatikai Konferencia

2026. március 31–április 2.
Debreceni Egyetem, Debrecen

Szerkesztette: Tick József, Kokas Károly, Holl András

HUNGARNET Egyesület
Budapest, 2026

HUN-REN
Magyar Kutatási Hálózat

NETWORKSHOP

Szerkesztette: Tick József, Kokas Károly, Holl András

Tipográfia és tördelés: Vas Viktória
Korrektúra: Danyi Melinda
Angol nyelvi lektor: Lukács Katalin

Networkshop 2026 konferencia előadásainak közleményei
Debreceni Egyetem, Debrecen
2026. március 31–április 2.

ISBN 978-615-6792-29-7
DOI: https://doi.org/10.31915/NWS.2026

Kiadja a HUNGARNET Egyesület
az MTA Könyvtár és Információs Központ közreműködésével
Budapest
2026

Borítókép: freepik.com

DOI: 10.31915/NWS.2026.3

A MAGYAR WEBARCHÍVUM ÚJ NYILVÁNTARTÓ ADATBÁZISA

Kalcsó Gyula
Magyar Nemzeti Múzeum Közgyűjteményi Központ Országos Széchényi Könyvtár
Digitális Bölcsészeti Központ, Digitális Filológiai és Webarchiválási Osztály
kalcso.gyula@oszk.hu

Absztrakt

A tömeges webarchiválás egyik visszatérő problémája, hogy miként lehet rögzíteni a célzott tartalmat és a kapcsolódó URL-ek időbeli változásait. Ez a kérdés összefügg a seedlisták karbantartásával is, mivel ki kell zárni azokat a webhelyeket, amelyek korábban mentésre kerültek, de már nem működnek, vagyis egy adott URL mögött már nincs tartalom, vagy az már nem tartozik az adott webhelyhez. A cikk egy rugalmas koncepciót mutat be, amely felhasználható a különböző struktúrájú URL-ek (http vagy https protokollal vagy anélkül, www-vel vagy anélkül) közötti kapcsolatok, azok időbeli változásai és a webhelyhez mint entitáshoz való kapcsolódásuk kezelésére. A megoldás lényege egy entitásalapú SQL-adatbázis, amely képes az időbeli változásokat redundancia nélkül rögzíteni a 3. normálforma biztosításával. Az adatbázisban tárolt fő entitások, mint például az archiválásra kijelölt webhely és az URL, összekapcsolódnak egymással, önmagukkal és az őket tartalmazó táblákkal kapcsolótáblák segítségével. Ez a megoldás biztosítja a skálázhatóságot, azaz az egyes entitásokról tárolt információk tetszőlegesen bővíthetők, és a kapcsolótáblák „date_from" és „date_to" mezői felhasználhatók az adott kapcsolatok érvényességi idejének rögzítésére. Az entitástáblák egymáshoz való kapcsolásával például alternatív URL-eket kapcsolhatunk össze időben. Az egyes entitásokról tárolt információk komplex lekérdezéseket tesznek lehetővé. Például az archiválandó tartalom esetében a típus (webhely, weboldal, fájl stb.), vagy az URL-ek esetében a státuszkód külön táblában van tárolva. A kapcsolótáblák biztosítják azt is, hogy az időbeli változások rögzítésre kerüljenek, így például lehetséges lekérdezni, hogy egy adott időszakban melyik URL tartozott egy adott entitáshoz (pl. egy weboldalon található fájlhoz). Mindez nagyban hozzájárul a fenntarthatósághoz, mivel sokkal gazdaságosabb, könnyebben használható és rugalmasabb lekérdezési megoldást kínál, mint a korábbi adattárolási módszerek, például a Google-táblázatok.

Kulcsszavak: webarchiválás, adatbázis-építés, born digital archiválás

The New Registry Database of the Hungarian Web Archive

Abstract

One recurring challenge in large-scale web archiving is how to capture changes over time in the target content and its associated URLs. This issue is also related to the maintenance of seed lists, as it is necessary to exclude websites that were previously archived but are no longer operational—that is, where a given URL no longer contains content or no longer belongs to that website. The article introduces a flexible concept that can be used to manage connections among URLs of different structures (with or without the http or https protocol, with or without www), their changes over time, and their association with the website as an entity. The core of the solution is an entity-based SQL database capable of recording temporal changes without redundancy by ensuring third normal form (3NF). The main entities stored in the database, such as the website designated for archiving and the URL, are linked to each other, to themselves, and to the tables containing them via junction tables. This solution ensures scalability, meaning that the information stored about each entity can be expanded as needed, and the "date_from" and "date_to" fields in the junction tables can be used to record the validity period of the given relationships. By linking entity tables to one another, for example, we can correlate alternative URLs over time. The information stored about individual entities enables complex queries. For instance, in the case of content to be archived, the type (website, web page, file, etc.) or, in the case of URLs, the status code is stored in a separate table. The junction tables also ensure that changes over time are recorded, making it possible, for example, to query which URL belonged to a given entity (e.g., a file on a webpage) during a specific period. All of this greatly contributes to sustainability, as it offers a much more cost-effective, user-friendly, and flexible query solution than previous data storage methods, such as Google Sheets.

Keywords: web archiving, database building, born digital archiving

A magyar webarchívum és nyilvántartott adatai

Az Országos Széchényi Könyvtárban 2017 óta működik a webarchiváló osztály, majd csoport, jelenleg a Digitális Bölcsészeti Központ Digitális Filológiai és Webarchiválási Osztályának részeként[1]. A dokumentumok gyűjtése háromféle módon történik: válogatva a legfontosabb magyar webhelyekről, kiemelt eseményekhez kötődve a főbb hírforrásokból, illetve általános jelleggel a magyar webtérről. Szelektíven kerül gyűjtésre a tudományos, kulturális, oktatási, közéleti jellegű tartalmak meghatározott köre. Az általános gyűjtés a .hu domén alatt regisztrált vagy egyéb doménhez tartozó, de magyar közönséget megcélzó nyilvános webhelyekre terjed ki. A webaratás csupán azon szervereket érinti, ahonnan technikailag biztosítható a nyilvánosan elérhető tartalom automatikus lementése.

1 Az előadás időpontjában még a megadott szervezeti keretben, a cikk megjelenésének az időpontjában viszont már a Digitális Megőrzési és Webarchiválási Osztály részeként.

2020. május 19-én megszavazta az országgyűlés a kulturális törvényt módosító javaslatokat, köztük azt is, amely a nemzeti könyvtár feladatává teszi a webtartalom megőrzését (2020. évi XXXII. törvény, 7. A muzeális intézményekről, a nyilvános könyvtári ellátásról és a közművelődésről szóló 1997. évi CXL. törvény módosítása). Az 1997. évi CXL. törvénybe bekerültek a webarchiválásról szóló részek (elsősorban az 59/A §): „(1) A webtartalom archiválás keretében, webaratással történő begyűjtésének, feldolgozásának, másolásának, hosszú távú megőrzésének és webarchívumba rendezésének, továbbá felhasználásának feladatát (a továbbiakban együtt: webarchiválás) a nemzeti könyvtár látja el. A könyvtárak együttműködnek a nemzeti könyvtárral az archivált webtartalom hozzáférhetővé tételében" (Kult. tv.). A Kormány 626/2020. (XII. 22.) számú rendeletben kiadta a webarchiválás részletes szabályait, amelynek értelmében a nemzeti könyvtár évente kétszer köteles webtérszintű aratást végezni.

A magyar webarchívum az alábbi fő részekből áll: demóarchívum; téma, műfaj vagy földrajzi hely szerinti részgyűjtemények; eseményelapú gyűjtemények; webtéraratások (a .hu doméntartomány mentése); speciális hibrid gyűjtemények (főként eseményekhez köthető, a webes tartalmon kívül a nemzeti könyvtár más digitális gyűjteményeiből származó dokumentumokkal). A demóarchívum, és a speciális hibrid gyűjtemények egy része nyilvánosan is megtekinthető a honlapunkon, a többi részgyűjtemény jogi okokból csak a nemzeti könyvtár olvasótermében elhelyezett terminálokon.

A 626/2020. (XII. 22.) Korm. Rendelet a webarchiválás részletes szabályairól a következőt írja elő: „1. § (1) A nemzeti könyvtár a muzeális intézményekről, a nyilvános könyvtári ellátásról és a közművelődésről szóló 1997. évi CXL. törvény (a továbbiakban: Kultv.) 61. § (4) bekezdés p) pontja szerinti feladatkörében a webtartalom begyűjtése érdekében a Kultv. 59/A. § (3) bekezdése szerinti webtartalomról jegyzéket […] vezet." Az aratásokhoz szükséges, valamint leíró, technikai és adminisztratív metaadatokat tartunk nyilván Google-táblázatokban, XML-metaadatfájlokban, valamint az aratások kimenetei között megtalálható report- és logfájlokban. Fontos szerepet játszanak az ún. seedlisták, amelyek a crawlerek számára az aratások kiinduló URL-jeit jelentik. Ezek karbantartása (a megszűnt webhelyek eltávolítása, az alternatív, átirányított URL-ek kezelése stb.), valamint bővítése ugyancsak fontos nyilvántartási feladat (l. Kalcsó 2025).

A megoldandó problémák és a megoldási javaslat

A webarchívumok metaadatolását világszerte különböző módszerekkel oldják meg. Az Online Computer Library Center (OCLC) Research Library Partnership Web Archiving Metadata Working Group 2018-ban kiadott egy ajánlást a webarchívumok metaadatolására vonatkozóan (Dooley & Bowers 2018), amelyben alig több mint egy tucat metaadatelemre tesznek javaslatot Dublin Core, EAD, MARC 21, és MODS formátumokban. A metaadat-tárolás a Google-táblázatoktól az XML-en át egészen az RDF-ig terjed.

A magyar webarchívumban kialakított leíró metaadatolási módszerről l. Drótos & Németh 2019.

Sajnos azonban a részletes leíró metaadatolás eddig csupán a demóarchívum esetében valósulhatott meg, továbbá a leíró metaadatokon kívül számos technikai és adminisztratív metaadatot is kezelnünk kell (pl. a korábbi aratások adatait nagy tömegben). Az adatok több helyen, különféle formátumokban, sok esetben egymással össze nem kapcsolható módon vannak tárolva. A részgyűjtemények nyilvántartásai inkonzisztensek és redundánsak: különböző adatmezőket használnak ugyanarra az adatra, ugyanazt az adatot több helyen is nyilvántartjuk. A nyilvántartásban meg kellene valósítani az entitásalapúságot: külön kellene kezelni az élő webre, valamint a mentett tartalmakra vonatkozó adatokat. Bizonyos esetekben szükséges lenne az adatok időbeli változását követni (l. alább, *Az időbeli változások követése* c. részt).

Fontos követelmény lenne, hogy biztosítható legyen az adataink összekapcsolhatósága más rendszerekkel (szolgáltatófelületekkel, katalógusokkal stb.). A webarchiválás informatikai rendszerében az adatok felhasználásával jelentősen növelni lehetne az automatizációt, valamint segíthetnének a monitoringfolyamatok fejlesztésében is. Mindezek miatt célszerűnek látszik a sok helyen tárolt, különböző típusú adatok egyetlen adatbázisba szervezése.

Az adatbázis-tervezés alapelvei

Mivel adataink integritása és konzisztenciája fontos szempont, valamint jelenleg is óriási a mennyiségük (és a jövőben várhatóan ütemesen növekedni fog), továbbá fontos szempont az adatbevitelkor a validálás és a sémakényszerítés, ezért egy SQL-adatbázis építése mellett döntöttünk. Jelenleg egy MariaDB épül, InnoDB motorral. Az érvényesített modern adatbázis-tervezési szempontok közül kiemelendő, hogy 3. normálformára törekszünk, ezáltal is próbáljuk a redundanciát a minimálisra csökkenteni (Salzberg 1986), valamint a több a többhöz kapcsolatokat kapcsolótáblákkal valósítjuk meg (Date 2003). Törekszünk a skálázhatóságra: EAV (entitiy-attribute-value) modellt (Kleppmann 2017) alkalmazunk, azaz külön táblákban tároljuk az entitás adatait, az attribútumokat és az értékeket.

Az adatbázis főbb komponensei

Adatbázisunk amennyire lehetséges, entitásalapú, ezért a legfontosabb komponenseket ezek táblái jelentik: az archiválandó webhelyek (targetek), a mentett és a hosszú távú megőrzésre átadott tartalom (digitális példányok), és a hozzájuk kapcsolódó adatok (beleértve a technikai és az adminisztratív metaadatokat is). Fontos részét képzeik az archiválási események adatai, amelyek összekapcsolódnak a megfelelő entitásokkal. Hasonló elven működik a minőségbiztosítási események nyilvántartása is. Mivel tevékenységünk jogi

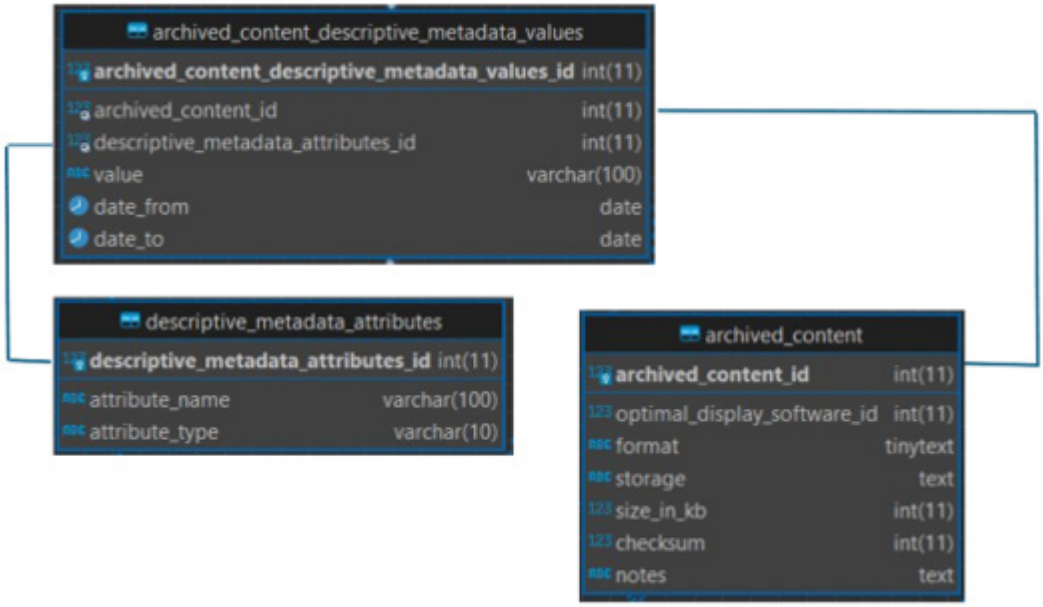
keretei alapvetően meghatározzák az archiválás menetét, az archivált tartalom tárolását és szolgáltatását, ezért az együttműködő partnerek adatainak és a jogi információknak a nyilvántartása fontos komponenst jelent. Az automatizációban és a monitoringban nagy szerepet játszik a szoftvereknek és konfigurációs, valamint technikai metaadataiknak a nyilvántartása. A rendszerintegráció érdekében tárolnunk kell a kapcsolódó rendszerek adatait is.

Az időbeli változások követése

Több olyan adatunk van, amelyek időben változhatnak. Ilyen például a webhely, és az azt azonosító URL kapcsolata. Ebben a viszonylatban sokféle változás lehetséges: egy webhely URL-je megváltozik, az adott URL-en más webhely jelenik meg, az adott URL-ről máshová van átirányítva a forgalom, megszűnik az URL. Ezek rögzítését egyfelől a különböző entitások különböző táblákban tárolásával és összekapcsolásával (URL és target, valamint URL és HTTP-státuskód), valamint az összekapcsolótáblákon a date_from és a date_to mezők használatával oldjuk meg.

Az entitásalapú leíró metaadatok kezelése

A fentebb már említett entitásalapúság nem csupán adatbázis-tervezési szempontból fontos. A metaadatolás, azon belül főként a leíró metaadatok kezelése összetett problémát jelent. Egyfelelől biztosítani szeretnénk, hogy adatbázisunk megfeleljen a legfontosabb könyvtári konceptuális modelleknek (pl. LRM), másfelől az is követelmény, hogy más könyvtári rendszerekkel összekapcsolható legyen, azaz metaadatszabvány szempontjából skálázható legyen: ne legyen „bevésett" metaadatséma, hanem tetszőlegesen alkalmassá tehető legyen lényegében bármely szabványnak megfelelésre. Ennek érdekében alkalmaztuk az EAV-modellt (Kleppmann 2017), amelynek jegyében az entitásokhoz tartozó attribútumokat (a leíró metaadatelemeket), valamint a hozzájuk tartozó értékeket külön táblákban tároljuk. Ily módon az attribútumok száma tetszőlegesen növelhető, pl. a Dublin Core mellé bevezethetőek bármely metaadatszabvány (pl. MARC21) mezői. Ez a módszer lehetővé teszi az RDA irányába történő elmozdulást is.



1. ábra: Az EAV-modell alkalmazása az adatbázisban: a mentett tartalom mint entitás, az azt leíró attbribútumok, és a hozzájuk tartozó értékek külön táblában tárolása (a szerző szerkesztése)

Összegzés, kitekintés

Az előadás a magyar webarchívum új nyilvántartó adatbázisának tervezési folyamatát mutatta be. Szólt a betöltendő adatok sokféleségéről, az entitásalapú, 3. normálformára hozott, skálázható SQL-adatbázis tervezéséről, továbbá az adatbázis főbb komponenseiről. Bemutatott két példát: az időbeli változások adatbázisban rögzítésének a megoldásáról, valamint a a leíró metaadatkezelésnek az EAV-modell szerinti entitásalapú megvalósításáról. A Google-táblázatokban és XML-fájlokban tárolt adataink betöltése folyamatban van, előttünk áll még a report- és logfájlok adatainak a retrospektív betöltése, valamint egy rugalmasan testreszabható adatbeviteli és lekérdezőfelület kialakítása, továbbá a rendszerintegráció.

Irodalom

2020. évi XXXII. törvény a kulturális intézményekben foglalkoztatottak közalkalmazotti jogviszonyának átalakulásáról, valamint egyes kulturális tárgyú törvények módosításáról. https://net.jogtar.hu/jogszabaly?docid=a2000032.tv Letöltve: 2026.03.30.

Date, C. J. (2003). *An introduction to database systems* (8. kiadás). Addison Wesley.

Dooley, J. & Bowers, K. (2018). *Descriptive Metadata for Web Archiving: Recommendations of the OCLC Research Library Partnership Web Archiving Metadata Working Group*. https://doi.org/10.25333/C3005C

Drótos, L. & Németh, M. (2019). Metadata Management and Future Plans to Generate Linked Open Data in the Hungarian Web Archiving Pilot Project. *ITLIB*, *2019*(2). https://itlib.cvtisr.sk/buxus/docs/38-metadata.pdf Letöltve: 2026.03.30.

Kalcsó, Gy. (2025). A magyar webtér aratásával kapcsolatos kurátori feladatok. In: Tick, József; Kokas, Károly; Holl, András (szerk.) *Oktatási, kutatási és közgyűjteményi infrastruktúrák és tartalmak: digitális transzformáció felsőfokon : NETWORKSHOP 2025* : 34. Országos Informatikai Konferencia : 2025. május 13–15. Széchenyi István Egyetem, Győr. Budapest, Magyarország : Hungarnet Egyesület (2025) 238 p. pp. 194–201. https://doi.org/10.31915/NWS.2025.21

Kleppmann, M. (2017). *Designing Data-Intensive Applications: The Big Ideas Behind Reliable, Scalable, and Maintainable Systems*. O'Reilly Media.

Korm. rend. 626/2020. (XII. 22.) Korm. rendelet a webarchiválás részletes szabályairól. https://njt.hu/jogszabaly/2020-626-20-22 Letöltve: 2026.03.30.

Kultv. 1997. évi CXL. törvény a muzeális intézményekről, a nyilvános könyvtári ellátásról és a közművelődésről. https://net.jogtar.hu/jogszabaly?docid99700140.tv Letöltve: 2026.03.30.

Salzberg, B. (1986). Third normal form made easy. *ACM SIGMOD Record 15*(4), 2–18. https://doi.org/10.1145/16301.16302