# 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**

# 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

## TARTALOMJEGYZÉK

# ELŐSZÓ

Tisztelt Olvasó!

A 35. Networkshop konferenciakötete harminc tanulmányt tartalmaz. A kötet szerzőinek többsége nem először publikál ebben a kiadványsorozatban – mint ahogy a Networkshop résztvevői közül is sokan visszatérő előadók vagy hallgatók. Nagy öröm a szerkesztőknek, hogy új szerzőket is üdvözölhetnek. 2026-ban a Debreceni Egyetem volt a műhelytalálkozó házigazdája, nem először, remélhetőleg nem is utoljára.

A tartalomjegyzéket böngészve elgondolkodhatunk azon, hogy kerül egy kötetbe jogi, informatikai, információtudományi, pedagógiai tanulmány? Az összekötő szál üvegből van: a számítógépes hálózat. A hazai felsőoktatási, kutatási, közgyűjteményi hálózat működtetői és a hálózaton átvitt információ előállítói, gondozói osztják meg eredményeiket ezeken az oldalakon.

Budapest, 2026. augusztus 10.

A szerkesztők

DOI: 10.31915/NWS.2026.1



# AZ ARK (ARCHIVAL RESOURCE KEY) AZONOSÍTÓ BEVEZETÉSE ÉS KÖZZÉTÉTELE A MAGYAR NEMZETI LEVÉLTÁR "ADATBÁZISOK ONLINE" FELÜLETÉN

*Debreczeni László, Magyar Nemzeti Levéltár Digitális Szolgáltatásfejlesztési Osztály*
*debreczeni.laszlo@mnl.gov.hu*

*Simon András, Magyar Nemzeti Levéltár Digitális Szolgáltatásfejlesztési Osztály*
*simon.andras@mnl.gov.hu*

## Köszönetnyilvánítás

A szerzők szeretnék köszönetüket kifejezni a „Digitális Irodalmi Akadémia köteteinek és hierarchikus egységeinek ARK azonosítóval való ellátása" munkaanyag szerzőinek a dokumentumban közölt értékes elemzésért és információkért.[1]

## Absztrakt

A Magyar Nemzeti Levéltár nyilvános webes felületein egyre több digitalizált levéltári irat elérhetőségét biztosítja, emellett egyre nagyobb mennyiségben tesz megtekinthetővé földrajzi- és személynévtér rekordokat is. Ezek közvetlen, felhasználóbarát módon történő hivatkozhatóságát biztosítani kell. A hivatkozás módjának a világban az egyik elterjedt módja az ARK (Archival Resource Key) azonosító használata, melyet ma már számos levéltár, mindenekelőtt a Francia Nemzeti Levéltár (ANF – Archives Nationales de France) is használ. Az azonosító a nemzeti levéltárban való bevezetésével kapcsolatban számos kérdést kellett eldönteni (saját adatbázisból származó vagy generált rekordazonosítók használatba vétele, „okos" vagy „buta" azonosítók alkalmazása, a csatolt digitális tartalmakkal nem rendelkező metaadatrekordok hivatkozásának a felvállalása, hivatkozás megoldása a teljes dokumentumok mellett azok részeire is, feloldás saját vagy külső eszközzel, ARK azonosító módosíthatósága és törölhetősége). Az előadás az intézmény által meghozott döntések ismertetése mellett bepillantást enged a döntési folyamatba és bemutatja a megszületett konkrét megoldásokat is.

**Kulcsszavak:** Permalink, ARK azonosító, levéltári nyilvántartás, webes szolgáltatások

1 Készítők: Horváth Ádám, Mohay Anikó, Radics Péter, Bánki Zsolt, Szabó Mihály, Csáki Zoltán, Virág Gabriella

## Abstract

The Hungarian National Archives provides a great amount of digitized archival documents on its public web interfaces and also makes geographical and personal namespace records available in increasing quantities. Their direct, user-friendly references must be ensured. One of the possibilities is to use the ARK (Archival Resource Key) identifier used by many archives, especially the French National Archives (ANF – Archives Nationales de France). Several issues had to be decided in connection with the introduction of the identifier in the national archives (the use of record identifiers from its own database or generated ones, the use of opaque or non-opaque identifiers, the acceptance of referencing archival catalogue records without attached digital content, the resolution of references to parts of documents in addition to entire documents, resolution with its own or external tool, the modifiability and deletion of the ARK identifier). In addition to describing the decisions made by the institution, the paper provides an insight into the decision-making process and presents the specific solutions that were created.

**Keywords:** Permalink, ARK identifier, archival catalogues, web services

## Permalink a levéltári nyilvános felületeken

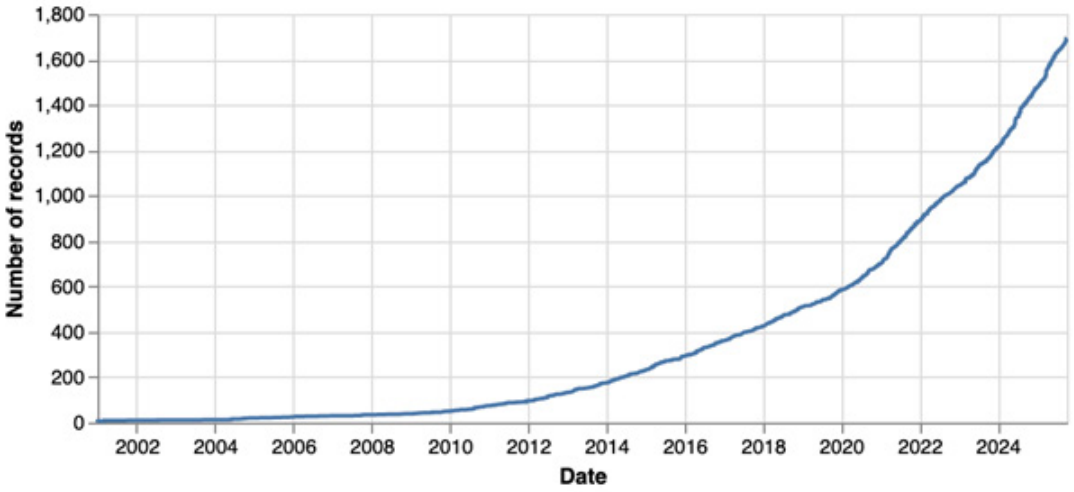
Egy URL link átlagos élettartama nem nagy. Nem kell sok időnek eltelnie, hogy egy korábban elmentett linkre kattintva a "404 Not Found" üzenet fogadjon bennünket. A perzisztens linkek és az ezekről a gyűjteményi tartalomra mutató, a link feloldását biztosító rezolverek alkalmazásával közelebb kerülhetünk ahhoz, hogy felhasználóink tartósan és közvetlenül elérjék és hivatkozhassák az általunk nyilvánosságra hozott dokumentumokat, akkor is, ha az intézményünk gyűjteményi struktúrája vagy nyilvános felülete lényegesen megváltozott.[2]

Ezek alkalmazása a levéltárakban kiemelt fontosságú, mert gyűjteményeik nagy többségben nem publikált dokumentumokat tartalmaznak, ezért azok nem rendelkeznek DOI-val sem. A levéltári iratokat nehezebb is azonosítani, mint egy könyvet, folyóiratcikkeket vagy akár a képeket. Sok esetben címmel és szerzővel sem rendelkeznek. Másrészt, vélelmezhetően tartósan hozzáférhetők lesznek az iratőrző adatbázisában, és mint primer források sohasem avulnak el. A dokumentum hozzáférhetőségének biztosítását a Magyar Nemzeti Levéltár – mint roppant stabil intézmény – biztosítani tudja.

2 https://arks.org/about/ark-features/

## Az ARK azonosító

A lehetséges permalinkmegoldások közül a Magyar Nemzeti Levéltár Digitális Szolgáltatásfejlesztési Osztályán figyelmünk az ARK azonosítóra (Archival Resource Key) irányult. Az ARK ugyanis egyre szélesebb körben elterjedt, mindenekelőtt a Francia Nemzeti Levéltár (ANF – Archives Nationales de France) is használja.




*1. ábra: NAAN azonosítóval ellátott intézmények száma*[3]

Alkalmazása nem jelent többletköltséget a levéltár számára. Egyszerű, logikus szerkezet, emellett hatékony hierarchia és verziókezelés jellemzi, és belesimul az intézményi DNS-rendszerbe.

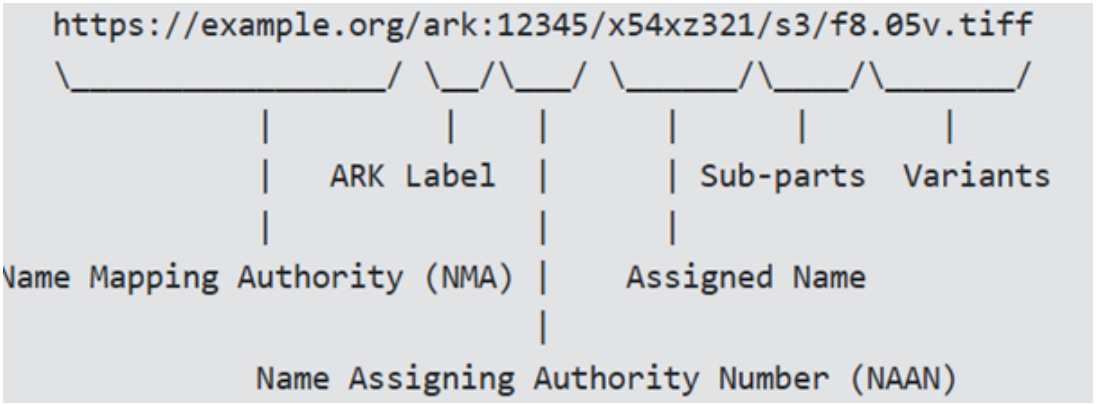
Az azonosító bevezetéséhez első lépésként Name Assigning Authority Number (NAAN) azonosítót kellett igényelnünk az https://arks.org weboldalon.

A bevezetéskor döntést kellett hozni néhány alapvető kérdésben:

- Mely domainnévhez akarjuk hozzárendelni a NAAN azonosítót?
- Mely gyűjteményi egységeket és mely adatbázisrekordokat kívánjuk ezzel az azonosítóval ellátni?
- Milyen legyen az általunk generált ARK azonosító szerkezete és hogyan legyenek ezek előállítva?
- Hogyan biztosítjuk az állandóságot, és mi legyen a sorsa a mégis megszűnt azonosítóknak?
- Támasztunk-e elvárásokat a hivatkozott objektumok metaadataira vonatkozóan?

3 https://arks.org/community/

A döntések meghozatalához az ARK szerkezetével kellett mindenekelőtt megismerkednünk.




*2. ábra: az ARK azonosító szerkezete*[4]

Name Mapping Authority: Az intézmény neve az URL-címben.
ARK Label: Ez mindig „ark:”
Name Assigning Authority Number (NAAN) azonosító
Assigned Name: Az objektumhoz rendelt egyedi azonosító.
Sub-parts: a dokumentum alábontásának követése az azonosítóban.
Variants: lehetséges változatok megjelenítése.

### Az azonosító képzése a nemzeti levéltárban

Az azonosító bevezetése során első körben az AOL-felületen (Adatbázisok Online) szolgáltatott dokumentumokkal és azon belül két alapkategóriával foglalkozunk:

- Metaadatrekord, amely online felületen elérhető digitális objektumra mutat.
- Metaadatrekord, amely a nyilvános felületen el nem érhető digitális objektumra mutat.
- Metaadatrekord, amely analóg dokumentumra mutat.
- Metaadatrekord, amely névtérelem (személy, testület, földrajzi) azonosítójára mutat.

Az ARK azonosítók előállítása során az adatbázisban már meglévő azonosítók (belső rekordazonosítók, raktári jelzetek, leltári számok) is felhasználásra kerülhetnének, de ez az azonosító által tartalmazott stringben a felhasználható jelek kötött száma miatt nem túl praktikus.

4 https://arks.org/

Számjegyek: 0123456789

Betűk, de kizárólag a mássalhangzók (kis és nagybetűk különbsége számít), leszámítva az „l” mássalhangzót.

Az alábbi központozási jelek: = ~ * + @ _ $ . /

A magánhangzók azért vannak mellőzve, hogy elkerülhető legyen, hogy az azonosító értelmes szavakat adjon ki, az l betű az i-hez való hasonlósága miatt marad ki.

Mi a felhasználásra tervezett karakterekből elhagytuk a központozási jeleket és a nagybetűket, így az ARK azonosítónkban az alábbi karakterek fordulnak majd elő:

0123456789

bcdfghjkmnpqrstvwxz

A dokumentumok egyedi azonosítója 7 karaktert foglal majd el, a kiadható, összesen 29 féle karakter így több mint 14 milliárd (huszonkilenc a hetediken) lehetőséget jelent.

A későbbiekben tervezzük a – karaktert, melyet az ARK-ban vizuális tagolásra használnak.

Osztályunkon megszületett az a döntés is, hogy – rendelkezvén a szükséges infrastruktúrával – helyi feloldóprogramot (local resolver) fogunk alkalmazni, a webszerver rewriting rules használatával, vagy más, saját fejlesztésű feloldószoftverrel. A táblázatban minden dokumentum egy sor, az ARK azonosítóhoz hozzákeresi az UUid-t, bármilyen változás van a rendszerben, a táblázatot kell módosítani.

Emellett az azonosítók tartalmazni fognak „shoulder-t”, mely az objektumhoz rendelt egyedi azonosító első tagja lehet. Ez a string a hivatkozott rekord jellegére, illetve a szolgáltató tagintézményre nézve fog tartalmazni utalást. Ez az egyediség biztosítását is segíteni fogja. A megbeszéltek szerint tehát a shoulder 4 vagy 5 karakterből áll, és a nagyobb egységeket, tagintézményeket, és azon belül adatbázisokat és adatféleségeket jelöli majd. Az azonosító a shoulderrel együtt egy Oracle adatbázisba lesz generálva, mindig az egyes objektumokhoz egy-egy levéltári adatbázis publikálásakor. A generáláskor lesz ellenőrizve, hogy az azonosító nem lett-e már korábban esetleg kiosztva az adatbázison belül.[5]

Az ARK azonosító „Sub-parts” részét képezhetik a „qualifierek”. Ezek legkézenfekvőbb szerepe egy adott dokumentumon belül az azok részeire (pl. egy több oldalas irat egyes oldalaira) való utalás lehet.

5 A https://mnl.gov.hu domain-hez kapott azonosítónk: 52333 lett. Az ehhez tartozó alárendelt domainek shoulderekkel kerülnek majd azonosításra.

Pl: ark:/12345/**x5**wf6789/c2/s4.pdf

12345 – NAAN

x5 - shoulder

wf6789 - azonosító

c2/s4.pdf - qualifier[6]

A qualifierek alkalmazását – azaz a hivatkozás megoldását a teljes dokumentumok mellett azok részeire is – akkorra halasztottuk, amikor a dokumentumok hivatkozása már széles körben megoldódott, és az egyéb meghozott döntéseket a gyakorlat igazolta.

Az azonosító kialakításakor választani kellett az „okos" non-opaque ARK és a „buta" opaque ARK között. Ez előbbi tartalmazza az azonosítót és egyéb információkat a dokumentum jellegéről, értelmes a felhasználó számára, információt ad számára anélkül is, hogy rákattintana. Mi az utóbbit (opaque) választottuk. Ezt egyszerűbb generálni, rövidebb, a lehetőségek száma nagyobb, az illetéktelen adatgyűjtő programok számára nehezebben kezelhető, és ha a dokumentum bármilyen tekintetben megváltozik, nem kell hozzányúlni. Az egyes azonosítók egyébként – éppen az illetéktelen, automatizált adatgyűjtés elleni védekezésül – nem a stringek ABC-, illetve számsorrendjében lesznek generálva.

## A bevezetés mérföldkövei

1. Levéltári iratrekordok hivatkozása (akár van már hozzá digitalizált teljes tartalom, akár nincs)
2. Levéltári dokumentumok egyes fejezeteinek, részleteinek azonosítása
3. Névtérrekordok (földrajzi, személy és testületi) hivatkozása

Az azonosítók generálása megkezdődött, előállításuk folyamatban van.

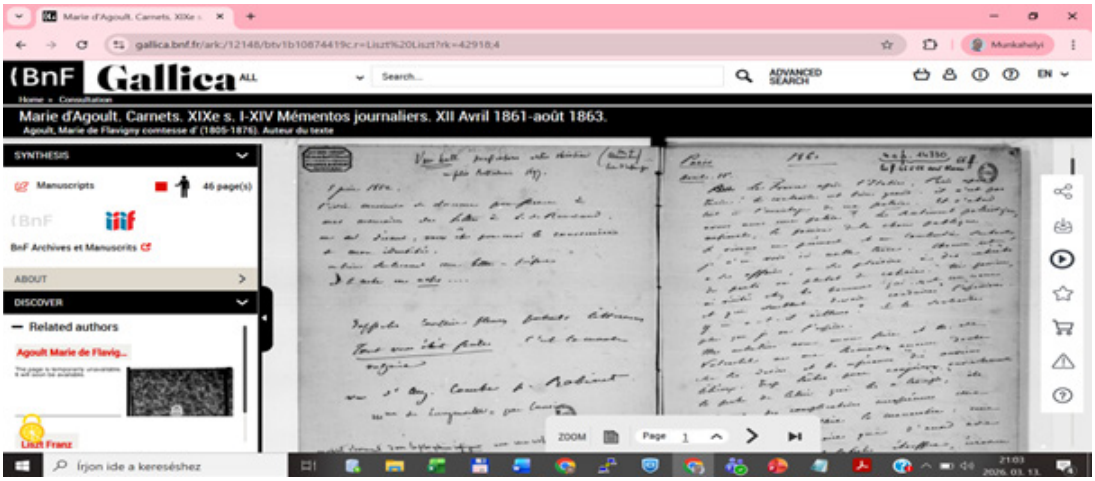
## Teendők egy ARK azonosító esetleges megszűnése esetére

Módosításra, törlésre fel kell készülni. Lehet törölni egy ARK azonosítót, és persze – főként a resolver által használt táblában lehet módosítani a dokumentum elérési útját. Mindenképpen törekedni fogunk rá, hogy a kiadott ARK azonosítók mögött ne legyenek törött linkek. Ha megszűnik egy dokumentum vagy egy állomány, és töröljük a rá hivatkozó rekordot, akkor egy standard oldalra visz majd az ARK azonosító. Ha olyan strukturális változás következik be, hogy az ARK azonosító módosul, mert a shoulderben a 2–3 vagy a 4. pozíción valami módosul, akkor erre a linkre való kattintáskor tájékoztató üzenetet küldünk. Ha az adatbázisban történik változás, akkor az eggyel fentebbi levéltári nyilvántartási szintre irányítjuk az azonosítót, és erről tájékoztató üzenetet küldünk a felhasználónak.

6 https://arks.org/

Példák a világban:

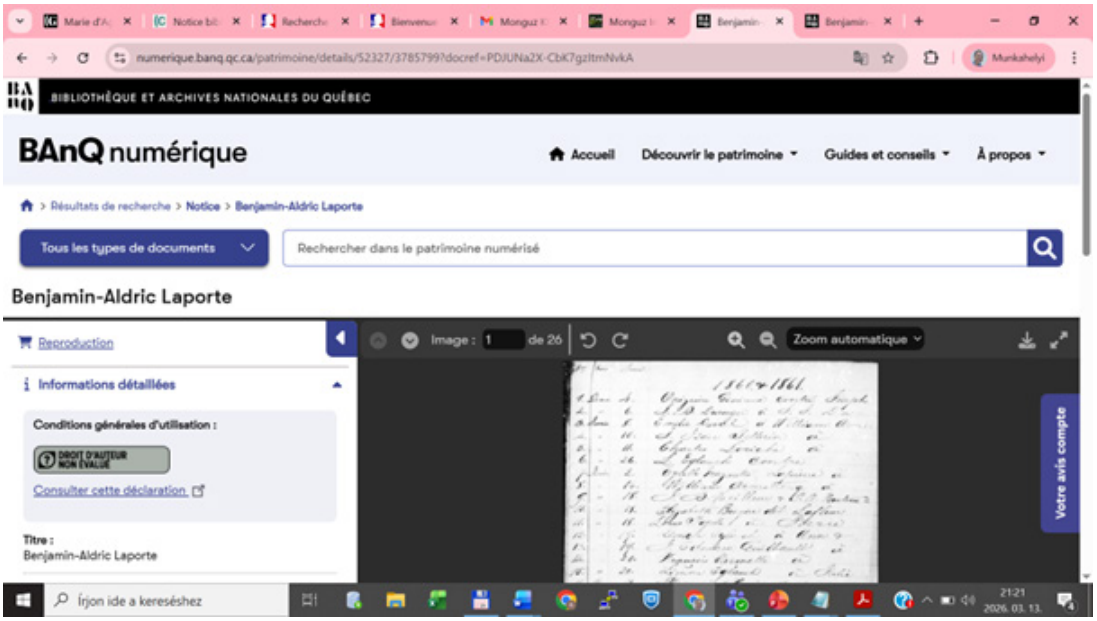

*3. ábra: ARK azonosítóval hivatkozott rekord a Francia Nemzeti Levéltár weboldalán*[7]




*4. ábra: ARK azonosítóval hivatkozott rekord a Gallica weboldalán*[8]

7 https://catalogue.bnf.fr/ark:/12148/cb43509416d

8 https://gallica.bnf.fr/ark:/12148/btv1b10874419c.r=Liszt%20Liszt?rk=42918;4




*5. ábra: ARK azonosítóval hivatkozott rekord a Québec állam Nemzeti Levéltárának weboldalán*[9]

### Saját megoldásunk bemutatása

Jelenlegi permalink a saját weboldalunkon teszt verzióban:

https://test-zala.natarch.hu/aol/ark:52333/0416rbbd4nws

A permalink feloldott állapotban:

https://test-zala.natarch.hu/aol/adatbazis/a-masodik-vilaghaboru-vesztesegi-adatbazisa/adatlap/6FA5D5ED58C44E412DD92202AB2B65C4

9 https://collections.banq.qc.ca/ark:/52327/3785799


*6. ábra: ARK azonosítóval hivatkozott rekord a nemzeti levéltár weboldalán*

A tervezett permalink struktúrája éles felületen így néz majd ki:

| *shoulder* | *domain* |
|---|---|
| ark:52333/s1 | https://aol.mnl.gov.hu/ark:${tartalom} |
| ark:52333/s2 | https://nevter.mnl.gov.hu/ark:${tartalom} |

Bár számos intézmény igényelt már mostanra NAAN azonosítót, valójában még kevés felületen alkalmazzák. Magyarországon tudomásunk szerint elsőként fogunk tömegesen ARK azonosítókat előállítani, és saját felületünkön szolgáltatni. Reméljük ezzel a módszertani újítással szolgáltatásunk elnyeri a szakma tetszését, és hozzájárulhatunk az interneten való tömeges, hatékony és hosszú távon megbízható tartalomszolgáltatáshoz.

DOI: 10.31915/NWS.2026.2



# HAZAI GYÉMÁNT OPEN ACCESS INFRASTRUKTÚRÁK – AMI VAN, ÉS AMIRE SZÜKSÉG LENNE

*Holl András*
*MTA KIK*

**Absztrakt**

Az utóbbi években, a nyílt hozzáférés üzleti kiadói megvalósításának („gold Open Access") ellentmondásos térnyerése után előtérbe került a gyémánt (nem üzleti alapon működő) infrastruktúrák szükségessége. Áttekintjük a hazai helyzetet, a létező infrastruktúra-elemeket és a működési problémákat. Javaslatot teszünk az átfogó hazai infrastruktúra létrehozására, az MTMT3 tervezésével párhuzamosan. Mind az országos hatókör, mind az MTMT-vel való kapcsolat szükséges a megfelelő működéshez.

**Kulcsszavak:** nyílt hozzáférés, közösségi fenntartású rendszerek, MTMT, infrastruktúra

**Diamond OA infrastructures in Hungary**

**Abstract**

The importance of diamond OA infrastructures is evident now. We discuss the Hungarian diamond OA landscape, and the existing problems. We propose a national infrastructure connected to the national CRIS system, the MTMT. As independent/small publishers have difficulties mastering and maintaining the necessary technology, a national supporting infrastructure is needed to provide help in order to increase the visibility and harvestability of green / independent gold OA.

**Keywords:** Diamond Open Access, infrastructures

A Budapesten indult nyílt hozzáférés kezdeményezés a tudományos publikációkhoz való akadálytalan hozzáférést tűzte célul, mivel a folyóiratok előfizetéses rendszere a felhasználók egy részénél akadályt jelentett a közlemények elérése terén, és hatalmas költségeket jelentett a globális tudományos közösség szintjén. Két út kínálkozott a nyílt hozzáférés elérésére: a zöld (repozitóriumi) és az arany (kiadói). Bár a 2000-es évek elején is létezett néhány kutatóintézeti, tudományos társasági, egyetemi kiadásban megjelenő szabadon elérhető folyóirat, ezeket ma már nem arany, hanem gyémánt típusúnak nevezzük, nonprofit jellegük miatt. Az üzleti alapon működő kiadók csak a repozitóriumok használa-

tának elterjedése által rájuk gyakorolt nyomás hatására, és a hasznukat biztosító üzleti modell (a közlési díjak általános bevezetése) megteremtése után álltak át a nyílt hozzáférésre. Mára a nyílt hozzáférés arany útja széles körben elterjedt a tudományos publikálásban, azonban ez a látványosnak tűnő siker valójában óriási kudarc. A hozzáférés elől elhárultak ugyan az akadályok, viszont a közlés előtt újabb, sokak számára leküzdhetetlen korlátok emelkedtek. A folyóiratokon keresztül történő tudományos kommunikáció rendszere pedig többe kerül, mint valaha.

Az Európai Unió az elmúlt években számos projektet indított a közösségi alapú, non-profit kiadás (a nyílt hozzáférés gyémánt útja, Diamond Open Access)[1,2] lehetőségeinek vizsgálatára és megteremtésére - ilyen volt például a DIAMAS projekt, és az European Diamond Capacity Hub (EDCH) kezdeményezés.[3] A gyémánt útról Taubert (2026) ad részletes tájékoztatást, szakirodalmi áttekintéssel, irodalomjegyzékkel együtt. Bár akkor még nemhogy a „gyémánt" megjelölés, de a nyílt hozzáférés elnevezés sem volt használatban, egy – ma így címkézhető folyóirat - már az 1990-es évek közepén megjelent Magyarországon: az Information Bulletin on Variable Stars (Holl, 2012).

Feltehetjük a kérdést: mi a helyzet itthon a nyílt hozzáférés gyémánt útját illetően, és milyen teendőink vannak e területen?

## Helyzetkép

A hazai tudományos kiadás egy része nemzetközi mércével mérve kis kiadók kezében van, mint az Akadémiai Kiadó, a l'Harmattan, és számos további kisebb, üzleti kiadó. Ezek kereskedelmi vállalkozások, nem felelnek meg a gyémánt kritériumoknak (ugyanakkor üzleti hasznuk össze sem hasonlítható a nagy multinacionális cégekével, nyereségük bizonytalan). A hazai kereskedelmi kiadók repozitóriumokban archivált anyagai, és kiadó által nyílt hozzáféréssel megjelentetett kötetei viszont bekerülhetnek a szabadon hozzáférhető tudományos irodalom körforgásába. A hazai és nemzetközi láthatóság növelése viszont fontos lenne.

A gyémánt körbe tartoznak az egyetemi, kutatóintézeti kiadók és a tudományos társaságok által megjelentetett folyóiratok és könyvek.

1 https://scienceeurope.org/our-priorities/open-science/diamond-open-access/

2 https://www.unesco.org/en/diamond-open-access

3 https://diamas.org/services

| ÉV | 2016 | 2017 | 2018 | 2019 | 2020 | 2021 | 2022 | 2023 | 2024 | 2025 |
|---|---|---|---|---|---|---|---|---|---|---|
| folyóirat | 1265 | 1221 | 1245 | 1218 | 1240 | 1253 | 1231 | 1203 | 1145 | 1098 |
| WoS indexált folyóirat | 64 | 64 | 65 | 62 | 65 | 63 | 61 | 62 | 64 | 59 |
| SJR indexált folyóirat | 125 | 124 | 123 | 123 | 122 | 121 | 120 | 118 | 122 | 113 |
| WoS/SJR indexált folyóirat | 134 | 131 | 132 | 130 | 129 | 128 | 126 | 125 | 130 | 121 |
| Teljes tudományos könyv | 1828 | 1849 | 1967 | 1683 | 1570 | 1573 | 1443 | 1264 | 1208 | 1102 |
| Teljes tudományos könyv DOI-val | 36 | 32 | 50 | 44 | 67 | 110 | 113 | 161 | 144 | 142 |

*1. táblázat. MTMT hazai kiadói statisztika – magyar kiadású folyóiratok és könyvek*

| Év | TuDoKK | MTMT |
|---|---|---|
| 2024 | 30939 | 50655 |
| 2025 | 22701 | 45931 |

*2. táblázat. Publikációk a TuDoKK-ban[4] és teljes tudományos, nyilvános forrásközlemények az MTMT-ben*

A meglévő hazai gyémánt infrastruktúra az aratható, informatikai rendszerek közötti kommunikációra képes publikáló platformokból (OJS: Holl és Bilicsi, 2019; OMP: Fülöp, Molnár és Hoczopán, 2022), a másodlagos közleményelhelyezést megvalósító repozitóriumokból, az előbbiek aggregátorából (TuDoKK) és a gyémánt OA számára többféle potenciális szolgáltatásra képes MTMT-ből áll. (Cikkünk írásakor a TuDoKK 21 minősített repozitóriumot és 28 minősített, OJS-alapú folyóiratot aggregált.) Természetesen léteznek hazai közösségi fenntartású folyóiratok egyedi tartalomkezelő rendszerekben, azonban ezek kezelése, aggregálása, láthatóvá tétele problematikus, cikkünkben ezekkel nem foglalkozunk.

A modern tudományos kommunikáció esszenciális összetevői az állandó egyedi azonosítók. A fentebb felsorolt infrastrukturális elemek az egyedi azonosítók használatára épülnek. Ezen túl az egyedi azonosítókhoz kötődő hazai informatikai infrastruktúra sem áll rendelkezésre, bár szükség lenne rá.

4 Tudományos Dokumentumok Közös Keresője – https://tudokk.mtak.hu

Az említett infrastruktúrák „kemény” informatikai rendszerek, szolgáltatások, amelyeknek magja valamilyen szoftver. Azonban az infrastruktúráknak lehet egy „puha” része is: olyan szervezetek, amelyek minősítik, népszerűsítik, támogatják a szoftverek/szolgáltatások megfelelő használatát. Ilyen jelenleg az MTMT Repozitóriumminősítő Szakbizottsága[5], ilyen a hazai repozitóriumokat és általában nyílt tudományt támogató szervezet: a HUNOR[6] és az openscience.hu weboldal.

### MTMT, MTMT3

A modern tudományos infrastruktúrákat használó hazai kiadóknál megjelent publikációk jelentős mértékben hozzájárulnak az MTMT3 céljaihoz: az automatikus adatátvétel szerepének növeléséhez, a szerzői feltöltés arányának csökkentéséhez, az adatok megbízhatóságának, minőségének javításához. Ugyanakkor az MTMT mint általánosan használt országos szoftver, valamint a hazai kutatók és publikációk teljes és egyértelműsített nyilvántartása – mintegy névtere – jelentősen hozzájárulhat az egyedi azonosítók megfelelő használatának támogatásához.

### Teendők

Mindeddig a hazai repozitórium-minősítés nem foglalkozott a fenntartók közösségi, nonprofit mivoltával, az erre vonatkozó információt a minősítési kérdőívbe be kell illeszteni, az eddig minősített folyóiratoknál ellenőrizni kell. A HUNOR keretében elindult a gyémánt folyóiratok részletes felmérése.

Mint arról a TuDoKK használói meggyőződhetnek, a repozitóriumok és OJS-alapú folyóiratok használati (konfigurálási, információ-szervezési) gyakorlata nem egységes. Többféle publikációs és repozitóriumi szoftver számos verziója van használatban, de még az azonos szoftverek használata sem egységes – az intézményi repozitóriumok fenntartói kreatívan alakítják az adatkitöltés szabályait és a folyóirat megjelenését befolyásoló beállításokat, és néha magát a szoftvert is (plugin-ek, bővítmények, eredeti kód módosítása). A helyzet igen hasonló az OJS rendszereknél is. OMP rendszer kevés működik, ezeknek aggregációja sem megoldott. Sok esetben a fenntartók OMP helyett az OJS rendszerekben helyezik el az általuk kiadott könyveket is – ez az aratásnál okoz problémát. A sokféle szoftver önmagában előnyös is lehet, különböző esetekben más-más szoftver alkalmazása lehet optimális. A verziók és konfigurációs beállítások, felhasználási sémák egységesítése azonban szükséges, legalábbis addig a szintig, hogy a hazai és külföldi aggregátorok jól tudják kezelni az adatforrásokat.

5 https://www.mtmt.hu/repozitoriumminosito_szakbizottsag/

6 https://openscience.hu/hunor/

A másik nagy probléma a megfelelő támogatás hiánya. Sokszor a költségvetés is szűkös, gyakorta informatikai szakemberek sem állnak rendelkezésre. Mivel mind a repozitóriumok, mind a publikációs platformok fontos szerepet játszanak a hazai kutatás dokumentumellátásában (hazai szerzők, hazai kiadók, magyar nyelvű publikációk esetében leginkább), továbbá olcsó információforrást jelentenek a kereskedelmi kiadókhoz képest. Érdemes lenne csekély ráfordítás ellenében a minőségen javítani. A kis intézmények esetében megoldás lehet egy központi, országos szolgáltatási platform igénybevétele. A REAL-PhD árva repozitórium – más intézmények anyagait fogadja be, helyi repozitórium hiányában. Franciaországban a HAL nemzeti kutatási repozitórium, Horvátországban az OJS-Hrcak szolgáltatás segíti az egyetemi kiadókat. Megvalósítható lehet központilag hosztolt szerveren az egyes kis intézmények repozitóriumait külön-külön virtuális gépeken futtatni, központi hardver és szoftver támogatással, de intézményi kiadói, könyvtári üzemeltetésben. Segíthetne az erre a célra központi forrásból szerezhető anyagi támogatás is – a jó szakemberekkel rendelkező intézmények ne csak szívességi alapon tudjanak segíteni kevésbé szerencsés partnereiknek.

Sok lehetőséget rejt magában az MTMT-re való támaszkodás. Már ma is léteznek olyan esetek, ahol az MTMT jótékony hatással lehet a hazai gyémánt open access fejlődésére, más esetekben várni kell az MTMT fejlesztésére, az MTMT3 elkészültére.

Leginkább az egyedi azonosítók használatában nyújthat az MTMT támogatást. Az MTMT az egyetlen olyan hazai adatbázis, amelyik egyértelműen azonosít gyakorlatilag minden hazai kutatót és publikációt, és mindezekről bőséges adatokkal is rendelkezik. Amennyiben a hazai ORCID konzorcium – esetleg ORCID HUB – az MTMT-vel összekapcsolva működik, elérhető, hogy a hazai kutatók többségének legyen ORCID azonosítója, de csak egy, és abban publikációs (és esetleg munkahelyi, végzettségi) adatok is szerepeljenek. A DOI azonosítók esetében az ezeken alapuló MTMT-be való feltöltés biztosíthatja, hogy a publikáció után rögvest ellenőrzésre kerüljön, működik-e az azonosító? Ugyancsak az MTMT jelenthet megoldást a hazai publikációk periodikus DOI működési tesztelésére is, sőt, az elérhetetlen DOI-k esetében az URL átirányítás kiváltójaként is működhet. (Ugyanakkor a DOI-k használata az MTMT szempontjából is sok hasznot jelent.)

A hazai nemzeti ORCID konzorcium megalakítása lehetővé tenné, hogy a kutatást végző intézmények magasabb szinten férjenek hozzá az ORCID API-khoz, és az általuk foglalkoztatott kutatók profiljába affiliációs információt is elhelyezhessenek.

A hazai kiadású folyóiratok és könyvek esetében egy kívánatos eljárás lehet az, hogy a publikáció OJS vagy OMP platformon történjen, minden publikáció DOI azonosítót kapjon, és minden szerző ORCID azonosítóval rendelkezzen, mely azonosítók a folyóiratcikkben is szerepeljenek és a DOI regisztrációs ügynökségnek bejelentésre kerüljenek. A közlemény adatai a publikációkor kerüljenek be a DOI regisztrátor adatbázisába, az irodalomjegyzékkel együtt. Az MTMT értesüljön az újonnan élesített DOI-król, majd a leíró bibliográfi-

ai adatokat és a szerzői hozzárendelést a DOI alapján beemelje, végül a publikációk az MTMT-ből SWORD-del feltöltve archiválásra kerüljenek egy repozitóriumban. Másrészről a közvetlen OJS-MTMT kommunikáció olyan információk átvitelét is lehetővé tehetné, amelyek a DOI regisztrátor adatbázisába nem kerülnek be. Minden hazai publikációs platform (és repozitórium) szerezzen minősítést, és legyen kereshető a TuDoKK-ban is.

A hivatkozások akár a DOI-val történő importtal, akár a CrossRef adatbázis bányászatával jussanak el az MTMT-be. Végül a hazai publikációk adatai kerüljenek be a hazai (és persze, nemzetközi) szakkönyvtárak keresőinek (discovery) látókörébe, (például a TuDoKK-on keresztül). Hozzátehetjük még, hogy nem csupán növelni kell a láthatóságot, de a használat dokumentálása is fontos. A DOI és a CrossRef Cited-by szolgáltatás is egy lépés ebbe az irányba, a COUNTER-statisztikák elterjesztése pedig még messzebbre vihet. Az esetleg DOI-t nem tartalmazó publikációkból a hivatkozásokat a REAL-ban elhelyezett kópiák nagy nyelvmodellek segítségével történő bányászata útján kerülhetnek be az MTMT-be.

## Irodalom


Fülöp, Tiffany és Molnár, Tamás és Hoczopán, Szabolcs (2022) *Open Monograph Press e-könyvplatform a Szegedi Tudományegyetemen.* In: Valós térben – Az online térért : Networkshop 31: országos konferencia. 2022. április 20–22. Debreceni Egyetem. Kiadja a HUNGARNET Egyesület az MTA Könyvtár és Információs Központ közreműködésével, Budapest, pp. 227–234. https://real.mtak.hu/155510/

Holl, András (2012) *Information Bulletin on Variable Stars – Rich Content and Novel Services for an Enhanced Publication.* D-LIB MAGAZINE, 18 (5/6). 05holl. https://real.mtak.hu/69988/

Holl, András és Bilicsi, Erika (2019) Nyílt publikálási szoftverek és platformok. In: Networkshop 2019. HUNGARNET Egyesület, Budapest, pp. 54–60. https://real.mtak.hu/104942/

Taubert, N. (2026). Research on Diamond Open Access in the Long Shadow of Science Policy. *Publications*, *14*(1), 20. https://doi.org/10.3390/publications14010020

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

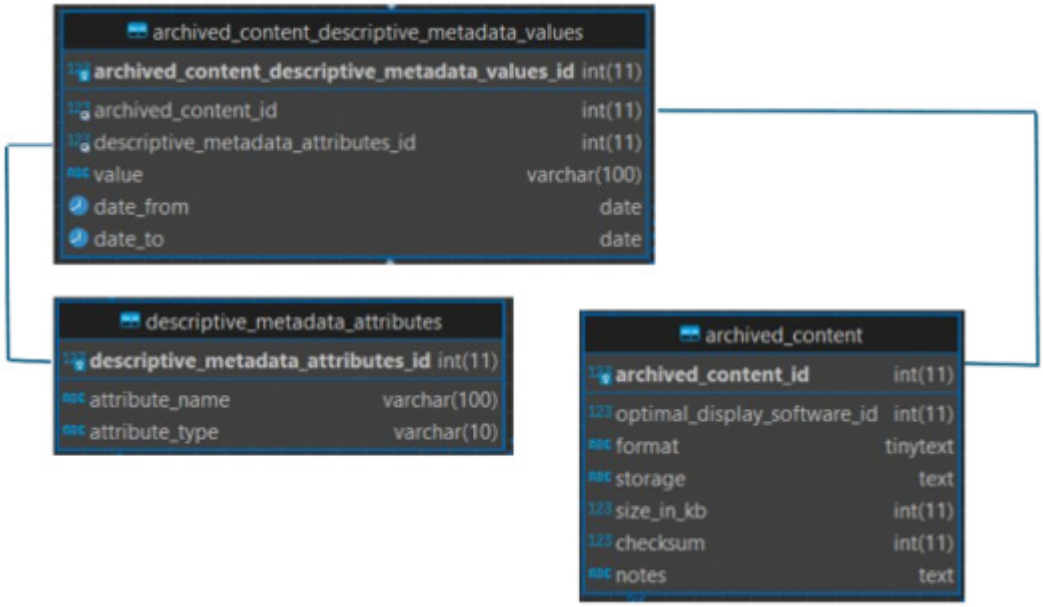
keretei alapvetően meghatározzák az archiválás menetét, az archivált tartalom tárolását és szolgáltatását, ezért az együttműködő partnerek adatainak és a jogi információknak a nyilvántartása fontos komponenst jelent. Az automatizációban és a monitoringban nagy szerepet játszik a szoftvereknek és konfigurációs, 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

DOI: 10.31915/NWS.2026.4



# NEMZETKÖZI ADATBÁZISOK REKORDJAINAK INTEGRÁLÁSA AZ SZTE DISCOVERY RENDSZERÉBE

# INTEGRATION OF INTERNATIONAL DATABASE RECORDS INTO THE SZTE DISCOVERY SYSTEM

*Hoczopán Szabolcs*
szabolcs.hoczopan@ek.szte.hu
*ORCID:* 0000-0002-7892-9974
*SZTE Klebelsberg Könyvtár és Levéltár*

*Nagy Róbert*
*Monguz Információtechnológiai Kft.*

*Rózsa Patrik*
*Monguz Információtechnológiai Kft.*

## Absztrakt

Az SZTE Klebelsberg Könyvtár és Levéltár 2026 elején indítja el a Monguz Információtechnológiai Kft.-vel közösen fejlesztett discovery rendszerét. A fejlesztés eredeti célja a könyvtár által gondozott rendszerek – például a katalógusok, repozitóriumok, valamint az OJS és OMP platformok – adatainak integrálása és közös kereshetőségének biztosítása egyetlen keresőfelületen.

A fejlesztés következő lépése egy teljes körű discovery rendszer kialakítása az előfizetett nemzetközi adatbázisok bibliográfiai rekordjainak és teljes szövegű hozzáféréseinek integrálásával.

Az így létrejövő keresőrendszer képes a felhasználói információkeresési igények széles körű kiszolgálására, valamint a különböző adatforrások egységes elérésének biztosítására, jelentősen megkönnyítve a tájékozódást a hagyományos könyvtári állományok és az online dokumentumok között.

## Abstract

The Klebelsberg Library and Archives of the University of Szeged will launch its discovery system, developed in collaboration with Monguz Information Technology Ltd., in early 2026. The original goal of the project is to integrate the data of the systems maintained by the library – including catalogs, repositories, and the OJS and OMP platforms – and to provide unified access to them through a single search interface.

The next phase of the development is the implementation of a comprehensive discovery system through the integration of bibliographic records and full-text access links from subscribed international databases.

The resulting search system is intended to support a wide range of user information needs while providing unified access to heterogeneous data sources, significantly facilitating navigation between traditional library collections and online documents.

## Kitekintés

A digitális információs környezet rohamos fejlődése alapjaiban alakította át a könyvtárak szerepét és a felhasználók információkeresési szokásait. Az online tartalmak mennyiségének elsöprő növekedése, a különböző adatbázisok, repozitóriumok és elektronikus források sokfélesége, egyre összetettebbé teszi az információkhoz való hozzáférést. Ebben a környezetben a hagyományos könyvtári katalógusok már nem képesek önmagukban kielégíteni a felhasználói igényeket, hiszen jellemzően csak a könyvtár saját állományára korlátozódnak, és nem biztosítanak egységes hozzáférést a különböző típusú digitális tartalmakhoz.

A felhasználók elvárásait nagymértékben formálták a nagy, kereskedelmi keresőmotorok használata, amelyek egyszerű, gyors és átfogó keresési élményt kínálnak. Ennek következtében a könyvtárakkal szemben is megjelent az igény arra, hogy hasonlóan intuitív és hatékony eszközöket biztosítsanak az információkereséshez. A discovery jellegű keresőrendszerek erre a kihívásra adnak választ. Olyan integrált felületek, amelyek képesek egyetlen keresőmezőből elérhetővé tenni a könyvtár különböző forrásait, beleértve a nyomtatott dokumentumokat, digitális tartalmakat, elektronikus folyóiratokat, e-könyveket és külső adatbázisokat is.

E rendszerek egyik legnagyobb előnye az egységes keresési élmény megteremtése. A felhasználóknak nem szükséges külön-külön adatbázisokban keresniük, hanem egyetlen felületen, egyszerre férhetnek hozzá a releváns információk széles köréhez. Emellett a discovery rendszerek fejlett keresési algoritmusokat, relevanciaalapú rangsorolást és széles körű szűkítési lehetőségeket kínálnak, amelyek jelentősen javítják a keresés hatékonyságát.

A discovery rendszerek bevezetése nem csupán technológiai fejlesztést jelent, hanem szemléletváltást is a könyvtári keresőszolgáltatások terén, ahol a hangsúly a felhasználóközpontúság felé tolódik, és a cél a lehető leggyorsabb és legkényelmesebb (“Google-szerű”) hozzáférés biztosítása a releváns, főleg tudományos információkhoz.

## SZTE Discovery előzmények

Az SZTE Klebelsberg Könyvtár és Levéltár hosszabb ideje törekszik egy minden könyvtári forrás keresésére alkalmas, valódi discovery rendszer kialakítására. A nemzetközi és hazai trendekkel összhangban célunk, hogy az olvasóknak a szakirodalom keresése során minél kevesebb különálló keresőfelületet kelljen használniuk. Manapság már alapvető felhasználói elvárás, hogy egyetlen keresőrendszer biztosítsa a könyvtári tartalmak minél teljesebb körű elérését.

A fejlesztési folyamat kezdetén ugyanakkor szükséges volt egy fontos korlátozás megfogalmazása. Nyilvánvalóvá vált, hogy a discovery rendszerekben elsősorban a klasszikus könyvtári dokumentumtípusok, például könyvek, e-könyvek és tudományos publikációk integrált keresése valósítható meg hatékonyan. Korábbi kísérleteink során a teljesség igényével olyan adatbázisokat is integráltunk, amelyek szócikk-alapú rekordstruktúrával rendelkeztek (például EBM-adatbázisok[1]). Az ilyen rendszerekből származó, rendkívül nagy számú, külön tételként kezelt rekord azonban nagy mértékben rontotta a találati listák áttekinthetőségét, gyakorlatilag használhatatlanná téve a szűretlen eredményhalmazt. Ennek tapasztalatai alapján a továbbiakban eltekintettünk az ilyen típusú források integrálásától.

A discovery rendszer fejlesztése során több részeredményt is sikerült elérni, azonban az eddigi megoldások mindegyike tartalmazott hiányosságokat: vagy az állomány nem volt teljeskörűen lefedve, vagy a rendszer nem biztosította a szükséges funkcionalitást. Elkészült például a repozitóriumok VuFind-alapú közös keresője[2], amely azonban nem tartalmazta sem a könyvtári katalógust, sem az előfizetett online forrásokat. A katalógusba integrált KBART-metaadatok révén ugyan megjelentek az előfizetett források[3], azonban ezek nem cikk- vagy könyvfejezet-szinten voltak kereshetők.

Egy későbbi megoldás során az előfizetett discovery rendszerbe a katalógus rekordjai is integrálásra kerültek (és tervben volt a repozitóriumok bevonása is), azonban ez a rendszer nem támogatta a személyre szabott funkciókat, például a kölcsönzési műveleteket. További kísérletként az akkori Bodza-alapú webkatalógusba integráltuk az EBSCO Discovery Service API-t. Elméletileg ez a megoldás már közel állt egy egységes keresőfelület megvalósításához, mivel a katalógus és az előfizetett források egy felületen jelentek meg. A gyakorlatban azonban a katalógus nem tudta megfelelően kezelni az EDS által szolgálta-

1 Evidence Based Medicine

2 Digitalizált tartalmaink közös keresője VuFind alapokon az SZTE Klebelsberg Könyvtárban
Farkas, Richárd és Sándor, Ákos (2020) Digitalizált tartalmaink közös keresője VuFind alapokon az SZTE Klebelsberg Könyvtárban. In: Networkshop 2020. Országos Online Konferencia. 2020. szeptember 2–4. HUNGARNET Egyesület, Budapest, pp. 22–32. ISBN 9786150103761. https://doi.org/10.31915/NWS.2020.3

3 Hoczopán Szabolcs: Online források a hagyományos katalógusokban. In: TUDOMÁNYOS ÉS MŰSZAKI TÁJÉKOZTATÁS 66 : 9 pp. 464–466., 3 p. (2019)
URL: https://journals.bme.hu/tmt/article/view/35069

tott, heterogén minőségű rekordok nagy mennyiségét, így a rendszer nem bizonyult alkalmasnak szélesebb körű alkalmazásra.

## Qulto Discovery

2025-ben, RRF pályázati keretben indult el egy új discovery rendszer fejlesztése a Monguz Információtechnológiai Kft. bevonásával. Az épülő rendszerrel szemben alapvető követelmény volt, hogy képes legyen adatcserére a könyvtár valamennyi meglévő keretrendszerével. Ennek megfelelően integrációt kellett biztosítania a Corvina könyvtári katalógussal, az EPrints-alapú teljes szövegű repozitóriumokkal, az Omeka-alapú multimédia repozitóriummal, az Invenio adatrepozitóriummal, valamint az Open Journal Systems és az Open Monograph Press platformokkal.

A belső rendszerek integrációján túl elengedhetetlen volt az előfizetett adatbázisokból származó, folyamatosan bővülő tartalmak bevonása is, amely API-alapú kapcsolaton keresztül valósult meg. A különböző adatforrások eltérő metaadatsémáinak összehangolása jelentős kihívást jelentett, amelynek részleteit kollégáim külön publikációja tárgyalja[4].

A fejlesztési folyamat kezdetén nemzetközi példák elemzésére is sor került. A vizsgálat során elsősorban nem a technikai megvalósításra, hanem a felhasználói élményre, a funkciók elrendezésére és a felület kialakítására helyeztük a hangsúlyt. A leginkább követendő példának az Oxfordi Egyetem SOLO[5] nevű, Primo-alapú discovery rendszere bizonyult. A rendszer letisztult felülete, valamint a különböző forrásokból származó találatok magas szintű integrációja olyan felhasználói élményt biztosít, amelyben az eltérő adatforrásokból érkező rekordok közötti különbségek csak szakértői szemmel érzékelhetők.

## Az új relevancia szerinti rendezés

Az új discovery rendszer fejlesztése során kiemelt feladat volt a relevancia szerinti rendezés újragondolása és finomhangolása. A rendszer alapját képező Qulto-katalógus ezen a téren korábban nem nyújtott megfelelő teljesítményt, ugyanakkor egy több adatforrást integráló keresőrendszer esetében a használhatóság alapvetően a keresőmotor minőségén múlik.

Az új relevancia szerinti rendezés egy pontszámítási algoritmuson alapul. Az alapértelmezett pontszámot több tényező határozza meg. Egyrészt minél többször fordul elő a keresőkifejezés egy adott rekordban, annál magasabb pontszámot kap. Másrészt az is befolyásolja az értéket, hogy a találat milyen mezőben jelenik meg: a rövidebb karakterszámú mezőkben (például cím) történő előfordulás nagyobb súlyt kap, mint a hosszabb mezők-

4 Zeller Rozália, Farkas Richárd: Repozitóriumok integrációja, metaadat konverziók az SZTE Discovery rendszerébe. In: Networkshop 2026 Országos Informatikai Konferencia.....

5 https://solo.bodleian.ox.ac.uk/discovery/search?vid=44OXF_INST:SOLO (2026.05.07)

ben való egyezés. Például az „Ady" keresőkifejezés esetén az a rekord, amelynek a címe „Ady", magasabb pontszámot kap, mint az, amelynek a címe „Ady Endre szerelmei".

Az alapértelmezett pontszámítás egy további súlyozási mechanizmussal került kiegészítésre. Ennek segítségével lehetőség van arra, hogy a keresőkifejezés bizonyos, előre meghatározott indexekben való előfordulása többletsúlyt kapjon. A konfiguráció során ki kellett jelölni azokat az indexeket, amelyeket kiemelten fontosnak tartunk, és ezekhez relatív súlyértékeket kellett rendelni. Például a „szerző" indexhez 100-as, a „cím" indexhez 50-es súly rendelhető. Ennek eredményeként, ha a keresőkifejezés a szerző mezőben is szerepel, a rekord sokkal előrébb kerül a találati listában, mint azok, ahol az egyezés csak kevésbé releváns mezőkben található.

A relevancia befolyásolásának további lehetősége, hogy nemcsak a felhasználó által megadott keresőkifejezésekhez, hanem előre definiált, fix feltételekhez is rendelhetők súlyok. Így például meghatározható, hogy azok a rekordok, amelyek a „lelőhely" mezőben tartalmazzák a „Klebelsberg Könyvtár" értéket, többletsúlyt kapjanak. Ezzel biztosítható, hogy a helyi állomány példányai a találati lista elején jelenjenek meg. Hasonló módon kiemelhetők például a nyomtatott dokumentumok vagy a távolról elérhető elektronikus tartalmak is.

A végső pontszám az alapértelmezett relevanciaérték és a konfiguráció során megadott súlyok kombinációjából adódik, amely meghatározza a találati lista sorrendjét.

### Olvasói műszerfal

A discovery rendszer nemcsak az adatforrások integrált keresését biztosítja, hanem a személyre szabott felhasználói funkciókat is egységes felületen teszi elérhetővé. Bejelentkezést követően az olvasók megtekinthetik személyes adataikat, kölcsönzött dokumentumaikat, valamint egy felületen elérhetővé válik valamennyi SZTE-s könyvtári tagságuk és kölcsönzésük.

A felhasználói felület csempékre épülő struktúrát alkalmaz. Két fő csempetípus különíthető el: az úgynevezett „okos csempék", valamint az "ugrópont" csempék.

Az okos csempék egy-egy konkrét funkció köré szervezett modulok. Ide tartozik például az elektronikus olvasójegy, amely vonalkódos és arcképes azonosítást tesz lehetővé, és alkalmas a könyvtári azonosítás kiváltására. Az olvasójegyen keresztül nyomon követhetők a kölcsönzések, valamint a folyamatban lévő és átvehető raktári kérések a különböző fiókkönyvtárakban.

A „könyvespolc" nevű okos csempe lehetőséget biztosít a találati listában megjelent dokumentumok mentésére és rendszerezésére. A felhasználók a kiválasztott rekordokat egyéni mappákba (listákba) rendezhetik, amelyeket a találati lista és a rekordnézet felől egyaránt elérhetnek. A mappák láthatósága különböző módokon szabályozható: lehetnek teljesen privátak, jelszóval megosztottak vagy nyilvánosan elérhetők. Ez a funkció lehetősé-

get teremt arra, hogy például az oktatók a kötelező irodalmat dinamikusan, a hozzáférési információkkal együtt osszák meg hallgatóikkal.

A harmadik kiemelt okos csempe a „munkafüzet", amely a saját tartalmak – például repozitóriumi dokumentumok vagy elektronikus kiadványok – aktív feldolgozását támogatja. A rendszerben ezek a dokumentumok nem csupán metaadat-szinten érhetők el, hanem teljes szövegükben indexelve vannak, így lehetőség nyílik könyvjelzők elhelyezésére, jegyzetelésre és szövegkiemelésre. A munkafüzet csempében a felhasználók áttekinthetik saját jegyzeteiket és módosított dokumentumaikat.

A webes tartalommegjelenítés az IIIF[6] (International Image Interoperability Framework) nyílt szabványcsaládra épül, amely lehetővé teszi nagy felbontású digitális képek egységes megjelenítését és kezelését.

További személyre szabott elemek közé tartozik a „korábban megtekintett dokumentumok" listája, valamint a korábbi keresések alapján generált ajánlórendszer. Ezek dinamikusan, a felhasználói aktivitás függvényében alakulnak.

Az ugrópont csempék ezzel szemben nem konkrét funkciókat valósítanak meg, hanem tartalmi és navigációs elemeket jelenítenek meg. Ezek segítségével a könyvtár különböző információkat, híreket és linkeket oszthat meg a felhasználókkal. Az ugrópont csempék egyik fontos sajátossága, hogy tartalmuk felhasználói csoportonként differenciálható. A könyvtár különböző célcsoportjai – például kutatók, hallgatók vagy külső olvasók – eltérő, számukra releváns tartalmakat kaphatnak. A jelenlegi tervek szerint a felhasználók három fő csoportba kerülnek besorolásra: kutatók, hallgatók és egyéb felhasználók. Ez a felosztás még kezelhető adminisztrációs terhelést jelent, mégis biztosítja a tartalmak megfelelő szintű testreszabását.

### A Summon API integrációja

Az előfizetett források rekordjainak integrációja a ProQuest Summon rendszerén keresztül, API-alapú kapcsolattal valósult meg. A Summon API integrálása egy, a miénkhez hasonló egyedi fejlesztésű könyvtári discovery rendszerbe jelentős technológiai kihívást jelentett. Bár a szükséges API-kulcs és a dokumentáció rendelkezésre állt, a külső rendszer működésének beillesztése nem pusztán adatlekérdezési feladat volt, hanem komplex architekturális tervezést igényelt.

Az alábbiakban a fejlesztés során felmerült legfontosabb problémákat és az ezekre adott megoldásokat ismertetjük.

6 https://iiif.io/ (2026.05.07)

### Autentikáció és lekérdezés

A Summon API használatához az érvényes API-kulcs mellett egy speciális digitális aláírás (digest) előállítása is szükséges, amely a lekérdezési paraméterekből képződik. A HTTP-fejlécek összeállítása szigorúan meghatározott formátumot követel meg: tartalmaznia kell az API URL-t, a pontosan formázott aktuális dátumot, a generált digitális aláírást, valamint a JSON formátum megadását.

A keresési paraméterek kezelése szintén számos sajátosságot mutatott. Az olyan paraméterek, mint például az *s.q* (kulcsszavas keresés), az *s.ff* (facetták) vagy az *s.sort* (rendezés), több esetben nem egyértelmű működést eredményeztek. Külön problémát jelentett, hogy ha egy paraméter explicit módon megadásra került, akkor a hozzá tartozó összes kapcsolódó paramétert is kötelező volt definiálni, még abban az esetben is, ha egyébként rendelkezik alapértelmezett értékkel. Ennek elmulasztása hibás vagy hiányos eredményhalmazhoz vezetett.

A lapozás implementálásakor további korlátozással kellett számolni: a rendszer legfeljebb 1000 találat lekérését teszi lehetővé. Ez azt jelenti, hogy hiába alkalmazható például 50-es oldalméret és 50 oldal, a két érték szorzata már meghaladja a lekérdezési limitet, ami hibát eredményez.

A facetták kezelése szintén összetett feladatnak bizonyult. Elkülönítve kellett kezelni a facetták megjelenítési nevét és technikai azonosítóját, továbbá külön figyelmet igényelt a több feltétel egyidejű alkalmazása során az ÉS/VAGY logikai kapcsolatok helyes felépítése.

A szerveroldali válaszként kapott, nagyméretű JSON-adathalmazból célzottan kellett kiválasztani azokat az adatokat, amelyek a felhasználói megjelenítés szempontjából relevánsak (például találatszám, facettaértékek, oldalszám, nyelvi adatok).

### A rendszerek architekturális összehangolása

A legnagyobb kihívást az jelentette, hogy egy eltérő struktúrájú külső rendszer találatai úgy jelenjenek meg, hogy azok illeszkedjenek a már megszokott felhasználói élményhez. Ez kétirányú megközelítést igényelt.

Egyrészt biztosítani kellett, hogy a felhasználó által megadott keresési feltételek megfelelően transzformálhatók legyenek a Summon API által elvárt paraméterekre. Másrészt a visszakapott találatokat úgy kellett megjeleníteni, hogy azok illeszkedjenek a meglévő discovery felület működéséhez.

Az integráció során arra a következtetésre jutottunk, hogy a Summon API-ból érkező találatokat célszerű a helyi adatforrásokból származó találatoktól elkülönítve kezelni. Ez a megoldás biztosítja a szűrés, a rendezés és a lapozás kiszámítható és stabil működését.

A gyakorlatban ez azt jelenti, hogy a felhasználó által indított keresés egy kiterjesztéssel kerül továbbításra az API felé, és az eredmények egy külön találati listában jelennek meg, ugyanazon keresési kontextuson belül.

Az így kialakított megoldás egy aggregátoralapú architektúrát eredményezett, amely a jövőben több külső forrás egyidejű lekérdezését is lehetővé teszi. A fejlesztés egyik legfontosabb technológiai hozadéka az új, SPA[7] (single-page application)-alapú technológia bevezetése, amely hosszabb távon a teljes discovery rendszer modernizációjának irányát jelöli ki.

További kihívást jelentett a klasszikus értelemben vett részletes rekordnézet hiánya. A Summon rendszerben nem létezik hagyományos rekordoldal; ehelyett egy permalinken keresztül elérhető, egyelemű találati lista szolgál a részletes megjelenítés alapjául. Ennek kiváltására egy saját fejlesztésű „Quick Look" (gyorsnézet) funkció került bevezetésre, amely lehetővé teszi, hogy a felhasználók a találati lista elhagyása nélkül férjenek hozzá a bővebb metaadatokhoz, például tárgyszavakhoz, DOI-azonosítóhoz, ISBN-adatokhoz vagy a befoglaló folyóirat adataihoz.

A megjelenítés tervezése során figyelembe kellett venni azt is, hogy bizonyos mezők, különösen az absztraktok, tárgyszóláncok jelentős terjedelműek lehetnek. Ezen adatok teljes megjelenítése a találati listában rontaná az áttekinthetőséget. Jelenleg mindkét mező a gyorsnézet funkción belül, alapértelmezetten összecsukott formában jelenik meg, ugyanakkor a felhasználók igény szerint lenyithatják és teljes terjedelmükben megtekinthetik azokat.

### A Summon API tartalmi finomhangolása

A bevezetés során további kihívást jelentett az API által szolgáltatott tartalom megfelelő szűrése. Kezdetben feltételeztük, hogy a Summon natív felületén beállított konfigurációk automatikusan érvényesülnek az API-n keresztül is. A gyakorlatban azonban ez nem bizonyult helytállónak.

Az első tesztek során nagy számban jelentek meg olyan rekordok, amelyekhez az intézmény nem rendelkezett hozzáféréssel. Mivel ezekhez nem tartozott tényleges teljes szövegű elérés, a felhasználói élmény jelentősen romlott. További nehézséget jelentett, hogy a rekordok egységesen az Ex Libris 360 Link linkfeloldó URL-jét tartalmazták, ami megnehezítette az egyes adatforrások azonosítását.

A probléma megoldására az API oldalon egy úgynevezett „inHolding" szűrő került bevezetésre, amely már az intézményi előfizetésekhez igazította a visszaadott rekordok körét. Ez azonban újabb problémát vetett fel: számos előfizetett adatbázis – például a Scopus, a MathSciNet vagy a Web of Science – csak bibliográfiai rekordokat tartalmaz, teljes szövegű hozzáférés nélkül.

7 https://en.wikipedia.org/wiki/Single-page_application (2026.05.07)

Ezeknek a rekordoknak a megjelenítése technikailag indokolt, azonban felhasználói szempontból félrevezető lehet, mivel a rekordhoz nem tartozik közvetlen hozzáférés. A helyzetet tovább bonyolítja, hogy a linkfeloldó ezekhez az esetekhez is teljes szövegre mutató hivatkozást generál, amely nem minden esetben vezet ténylegesen elérhető tartalomhoz.

Ennek kezelésére egy további szűrési lépés került bevezetésre, amely kizárólag azokat a rekordokat engedi megjelenni, amelyekhez a Summon tudásbázis szerint is rendelkezésre áll teljes szövegű hozzáférés. Bár ez a megoldás a találatok teljességének bizonyos mértékű csökkenésével jár, számottevően javítja a felhasználói élményt, mivel elkerülhetővé teszi a hozzáférhetetlen tartalmakból adódó félreértéseket.

A relevancia szerinti rendezés tekintetében nem volt szükség további beavatkozásra, mivel a Summon API saját, fejlett rangsorolási mechanizmussal rendelkezik, amely a natív felülettel azonos találati listát biztosít.

Fontos tapasztalat volt továbbá, hogy a korábbi API-megoldásokkal szemben a Summon rendszer már eleve válogatott, jó minőségű rekordokat szolgáltat. A beépített duplumszűrés, valamint a további szűrési mechanizmusok révén a kezelendő rekordmennyiség is megfelelő keretek között maradt.

### Fejlesztési tervek

Az API találati lista jelenleg dokumentumtípus, tárgyszó, tudományterület és forrásadatbázis szerinti dedikált facettákkal szűrhető. A közeljövőben a szűrési lehetőségeket egy „Lektorált" és egy „Open Access" facettával kívánjuk bővíteni.

Az Open Access szűrő lehetővé tenné, hogy az API-találatok közül az egyetemi hálózaton kívülről is egyszerűen azonosíthatók legyenek az autentikáció nélkül elérhető dokumentumok. Emellett ez a funkció ma már gyakorlatilag alapelvárásnak tekinthető a korszerű tudományos keresőrendszerekben.

A „Lektorált" facetta ezzel szemben a szakmailag ellenőrzött tartalmak gyors szűrését támogatná, segítve a felhasználókat a releváns tudományos publikációk azonosításában.

Rekordszinten szintén több fejlesztést tervezünk. A jelenlegi implementációban például a publikációk évfolyamadata nem minden esetben jelenik meg a folyóiratszám mellett, továbbá az API-ból érkező oldalszámadatok esetében szeretnénk a kezdő és záró oldalszámokat is megjeleníteni a találati listában.

A hivatkozási adatok jelenleg kizárólag a Web of Science API-n keresztül érhetők el. Ezek lehetővé teszik, hogy a felhasználók gyors áttekintést kapjanak arról, hogy az adott publikáció milyen mértékben jelent meg a tudományos diskurzusban. A jövőben a Web of Science adatait Scopus idézettségi adatokkal is ki kívánjuk egészíteni.

Emellett célunk egy alternatív metrikai rendszer integrálása is, amely nem kizárólag a klasszikus idézettségi mutatókat, hanem az online tudományos jelenlétet és társadalmi

visszhangot is képes mérni. Jelenleg a Dimensions Badge és az Altmetric megoldásait vizsgáljuk abból a szempontból, hogy melyik integrálható hatékonyabban a discovery rendszer rekordszintű eszközkészletébe.

Természetesen a teljes discovery rendszer finomhangolása a bevezetést követően is folyamatos marad. A jelenlegi webkatalógus leváltását követően az olvasói visszajelzések várhatóan kiemelt szerepet játszanak majd a rendszer további fejlesztési irányainak meghatározásában.

## Összegzés

A bemutatott fejlesztési folyamat jól mutatja, hogy egy valóban integrált discovery rendszer kialakítása nem pusztán technológiai feladat, hanem folyamatos kompromisszumok és ismétlődő döntések sorozata. A különböző adatforrások eltérő struktúrája, a metaadatok minőségének változatossága, valamint a külső rendszerek korlátai mind olyan tényezők, amelyek komolyan befolyásolják a végeredményt.

A fejlesztés során szerzett tapasztalatok alapján egyértelművé vált, hogy a „minden egyben" típusú keresés csak bizonyos határok között valósítható meg hatékonyan. A túlzottan heterogén tartalom integrálása könnyen a felhasználói élmény rovására mehet, ezért kulcsfontosságú a tudatos szelekció és a megfelelő szűrési mechanizmusok kialakítása.

A Summon API integrációja rávilágított arra is, hogy a külső szolgáltatások beépítése nem csupán technikai kérdés, hanem jelentős mértékben befolyásolja a felhasználói percepciót is. A hozzáférhető és nem hozzáférhető tartalmak megkülönböztetése, valamint a találati listák érthetősége kritikus tényezők a rendszer elfogadottsága szempontjából.

A relevancia szerinti rendezés finomhangolása, valamint a személyre szabott funkciók bevezetése egyértelműen hozzájárult a rendszer használhatóságának javításához. A tapasztalatok azt mutatják, hogy a felhasználók számára nem feltétlenül a források teljes körű lefedettsége a legfontosabb, hanem az, hogy gyorsan és megbízhatóan jussanak el a számukra valóban releváns tartalmakhoz.

A kialakított architektúra, különösen az aggregátor jellegű megközelítés és az SPA-alapú fejlesztési irány, megfelelő alapot biztosít a rendszer további bővítéséhez. A jövőben lehetőség nyílik további külső források integrálására, valamint a felhasználói felület és a keresési funkciók további finomítására[8].

Összességében elmondható, hogy a discovery rendszer fejlesztése nem tekinthető lezárt projektnek, hanem egy folyamatosan fejlődő szolgáltatás, amely a felhasználói igények, valamint a technológiai környezet változásaihoz igazodva alakul tovább.

8 https://discovery.bibl.u-szeged.hu/ (2026.05.07.)

DOI: 10.31915/NWS.2026.5



# A RAG A VÉGFELHASZNÁLÓI SZINTEN

# RAG AT THE END-USER LEVEL

*Ungváry Rudolf*
*Országos Széchényi Könyvtár*
*ungvaryr@gmail.com*

## Absztrakt

A RAG (Retrieval-Augmented Generation)– a *forráselőírt információkeresés* – elfogadott magyar meghatározása még nem létezik. Ami van, azt lényegében az informatika szaknyelvén fogalmazzák meg. (Pl. beágyazott visszakereséssel végzett generálás, lekéréses kibővített generáció, **a generatív MI képességeit külső információforrásokkal ötvöző adatlekérdezés,** visszakereséssel bővített generálás, adatlekérésre alapozott generálás/ mesterséges intelligenciák.) A természetes nyelvet használó, nem informatikus végfelhasználó számára nehezen derül ki, hogy pontosan, az ő gondolkodása, az ő teendői szempontjából miről van szó. Egyszerűen fogalmazva meg kell adnia általa kiválasztott forrásokat, beleértve külső adatbázisokat is az MI rendszerének, (mint amilyen pl. a ChatGPT), hogy azokat is „beágyazva" a rendszer elvégezze a megadott tárgyú információ keresését. A többnyire speciális, olykor az interneten hiányzó forrásokat megadva a felhasználó pontosabb, szakszerűbb válaszokat kap, adott esetben a saját szakterületén használatos különleges szakkifejezésekkel megfogalmazva. Más szóval a generatív MI nem csak saját „tudására", hanem konkrét, külső forrásokra is támaszkodik a válaszadáskor. Kétségtelen, hogy ezáltal a felhasználó jobb eredményeket kaphat, mintha e források nélkül kérdezne (fogalmazná meg a promtot). A 2025. évi konferencián már ismertettünk ilyen külső, a felhasználó által megadott forrást alkalmazó eljárást a MARC21 magyarra fordításához, de magával az MI-vel végzett, forráselőírt információkereséssel nem foglalkoztunk. Most ismertetjük, hogyan használhatja a végfelhasználó a számára rendelkezésre álló eljárást, és összehasonlítjuk a RAG ama fejlesztői, ipari, professzionális változatával, melynek kialakításához informatikai szakismeret szükséges – noha használatához, ha jó a kialakított rendszer, a felhasználónak szintén nincs szüksége beható informatikai szakismeretekre. Konklúziók: 1. Az LLM nem tudástár. 2. „Tudásához" meg kell adni a forrásokat. 3. A legspecifikusabb információk nincsenek a hálón 4 Nem a legfrissebb információ van a hálón. 5. Az igazán összetett, mélységi forráselőírt kereséshez nem elég a végfelhasználói ismeret. 6. A jövő a specializálódott RAG szövegalapú párbeszédrendszereké („chatbot").

## Abstract

The accepted Hungarian definition of *RAG* – source-prescribed information retrieval – does not yet exist. What does exist is essentially formulated in the technical language of computer science (e.g., generation with embedded retrieval, retrieval-augmented generation, data querying that combines the capabilities of generative AI with external information sources, generation enhanced with retrieval, generation based on data retrieval/artificial intelligences). For an end user who is not an IT specialist but uses natural language, it is difficult to understand exactly what this means from the perspective of their own thinking and tasks.

Simply put, the user must provide their AI system (such as ChatGPT) with selected sources – including external databases – so that the system can incorporate ("embed") them and perform information retrieval on the specified topic. By supplying sources that are often specialized and sometimes absent from the internet, the user can obtain more precise and professionally formulated answers, in some cases expressed using the special terminology of their own field. In other words, generative AI relies not only on its own "knowledge" but also on specific external sources when producing answers. There is no doubt that in this way the user can obtain better results than by asking questions (formulated prompts) without such sources. At the 2025 conference we already presented a procedure using such external, user-provided sources for the translation of MARC21 into Hungarian, but we did not address source-prescribed information retrieval carried out with AI itself. We now describe how the end user can employ the method available to them, and we compare it with the developer-level, industrial, professional version of RAG, the creation of which requires IT expertise – although, if the system is well designed, its use likewise does not require deep IT knowledge from the user. Conclusions: 1. An LLM is not a knowledge repository. 2. Sources must be provided for its "knowledge.". 3. The most specific information is not on the web. 5. Truly complex, in-depth source-prescribed retrieval cannot be carried out with end-user knowledge alone. 6. The future belongs to specialized, RAG-based text dialogue systems ("chatbots")

## Bevezető

A 2025. évi NETWORKSHOP konferencián már ismertettünk egy RAG-eljárást a MARC21 magyarra fordításához (Ungváry, 2025). Ebben magával a nagy nyelvi modelleken alapuló mesterséges intelligencia rendszerrel (LLM, a továbbiakban MI-vel) végzett, angol nevén Retrieval-Augmented Generation (RAG) eljárásával, annak magyar megevezésével és alkalmazásának formáival nem foglalkoztunk. Ennek a munkának ez lesz a tárgya.

## Terminológia

A RAG elfogadott magyar meghatározása még nem létezik. Ami van, azt lényegében az informatika tolvajnyelvén fogalmazzák meg. Például „beágyazott visszakereséssel végzett generálás", „visszakereséssel bővített generálás", „lekéréses kibővített generáció", „adatlekérésre alapozott generálás/mesterséges intelligenciák", „az LLM [Large Language Modell, Nagy Nyelvi Modell, mesterséges intelligencia, MI] által generált válaszok minőségét külső tudásforrásokra alapozva javító lekérdezési technika" (IBM). Ezen megfogalmazások mögött az a szándékos vagy öntudatlan szándék húzódik meg, hogy a végfelhasználót kényszerítsék rá egy tőle idegen szaknyelv használatára. A legközelebb még az a válasz áll a természetes nyelvhez, melyet a mesterséges intelligencia (a továbbiakban MI) ad, ha rákérdeznek: *„a generatív MI képességeit külső információforrásokkal ötvöző adatlekérdezés"*.

A természetes nyelvet használó, nem informatikus végfelhasználó (a továbbiakban felhasználó) számára nehezen derül ki, hogy az ő gondolkodása, az ő teendői szempontjából miről van szó. Tehát nem az a kérdés, hogy az informatika szakterületén minek kellene nevezni a RAG-ot, noha ott is célszerű volna a lehető legegyszerűbb, egyúttal legérthetőbb név. Ahogy a „programozható elektronikus adatkezelő berendezés" helyett még a szakterületen is elég a „számítógép" kifejezés használata.

Mindezek alapján a magyar megnevezés javaslata: ***forráselőírt információkeresés* (*és válasz*)**. A továbbiakban ehhez tartjuk magunkat, de a rövidség kedvéért az angol kifejezés rövidítését – RAG – használjuk, mivel a megjelenése óta ez terjedt el magyar nyelvterületen is.

## A RAG jelentősége

A RAG segítségével a végfelhasználó a saját természetes nyelvén nemcsak tájékozódhat, hanem olyan bonyolult feladatokat is megoldhat, melyekhez eddig informatikai szaktudás volt szükséges. A természetes nyelven megfogalmazott promptok segítségével adhatja meg a megoldandó feladatokat. Ezzel teljesíti azt a legfontosabb etikai követelményt, hogy minden műszaki termék, megoldás az embert hivatott szolgálni. Ez azt is jelenti, hogy *a termék használatának nyelve* feleljen meg annak a nyelvnek, amelyen a végfelhasználó gondolkodik, nem pedig az informatika szaknyelvének. Ez a helyzet ugyan a RAG esetében még nem a legtökéletesebben valósult meg, de rendkívüli előrelépés következett be. Az alábbi tanulmányban egyrészt tárgyaljuk a RAG eme, a végfelhasználót közvetlenül szolgáló változatát. A továbbiakban ezt nevezzük végfelhasználói RAG-nak. Röviden tárgyaljuk a bonyolultabb, nagyobb forrásállományt alkalmazó, informatikai eszközökkel kialakított, de ugyancsak végfelhasználó nyelven használható változatát. A továbbiakban ezt nevezzük professzionális RAG-nak. Bemutatunk tovább egy példát a végfelhasználói RAG alkalmazására. Ebben az OSZK katalógusának az Egyetemes Tizedes Osztályozás (ETO) jelzeteinek szintaktikai helyességét ellenőriztük természetes nyelven fogalmazott

promptok segítségével. A MI elvi jelentősége éppen az, hogy lehetővé teszi a természetes nyelv széles körű használatát. Arra itt csak utalunk, hogy ugyanakkor alkalmazása lélektani veszélyekkel jár a fölkészületlen, ezért megtéveszthető felhasználó számára. Csak annyit jegyzünk meg, hogy a felhasználó minden érthető, pontos válasz ellenére nem értelmes válaszokat kap, mert nem értelemmel rendelkező lénnyel, hanem egy élettelen, mesterséges intelligenciának nevezett és annak látszó rendszerrel áll szemben, amely értelmes lényekre jellemzőnek látszó válaszokat ad. Amely ráadásul tapad, ha a végfelhasználó nem egyértelműen vet véget a munkaülésnek. Elég ez azzal kipróbálni, hogy a kérdező „udvariasan" a „Vége" közléssel zár.

A mesterséges intelligencia (a Nagy Nyelvi Modell, MI) jelentősége, hogy programozási ismeretek nélkül kommunikálni lehet egy számítástechnikai rendszerrel. Nemcsak előírt formákban, mint például a szabadon választható természetes nyelvű keresőszavakkal, hanem folyamatos, természetes nyelven. A forráselőírt információkeresés (RAG) erre még inkább alkalmas. Eredményes alkalmazásának csak a pontosan megszerkesztett, természetes nyelven megfogalmazott prompt a feltétele. Ebben a munkában ezt igyekszünk szemléltetni az említett gyakorlati példán keresztül.

## A RAG tulajdonságai

A RAG esetében az MI megadott forrásra (dokumentum, hosszabb szöveg, professzionális változatban akár nagy dokumentumtárházra) támaszkodva válaszol. A forrás külső. Azaz elvileg, valószínűleg nincs (még) az interneten, legalábbis akkor még nem volt, amikor az MI betanítása játszódott le. A betanítás folyamata összetett. Egészen nagy vonalakban a fejlesztők nem közvetlenül „az interneten" tanítják a modellt, hanem hatalmas, előre összeállított és szűrt adathalmazokon. Ezekből eltávolítják a „szemetet" (reklámokat, többszöröződéseket, hibás kódokat). Mindennek a befejezése után az MI már nincs folyamatos kapcsolatban az internettel. Amit tud, az a saját emlékezetében („belső súlyaiban") tárolt statisztikai összefüggés.

A RAG-ot alkalmazva egyszerre több forrást is megadható. A forrás lehet szöveg, kép, fájl stb. Lehet a felhasználó számítógépén tárolt állomány, vagy az internet valamelyik ugrópontjával („linkkel") azonosított forrás. Nem történik egyéb, mint hogy utasítjuk a rendszert, hogy „ezt/ezeket a forrást/forrásokat vedd figyelembe a válasz létrehozásához". A felhasználó testre szabottabb, pontosabb, szakszerűbb eredményhez jut, akár azt is megadhatja, hogy esetenként milyen különleges szakkifejezéseket használjon az MI adott tárgy esetében.

Más szóval a RAG esetében az MI nem csak saját betanított „tudására", hanem konkrét, külső, a felhasználó által megadott forrásokra is támaszkodik a válaszadáskor. Kétségtelen, hogy ezáltal az eredmények a felhasználó igényeinek jobban meg fognak felelni, mintha e források nélkül kérdezne (fogalmazná meg a promptot).

Gyakorlatilag azonban nem mindig tudható, hogy a forrás teljesen külső, vagy már be lett egykor vonva a betanításban. Ráadásul a felhasználó az interneten létező forrást is megadhat annak ugrópontjával. Ez a RAG egyik technikai paradoxonja. Nehéz éles határvonalat húzni. Lehet ugyanis adatfedés („data overlap"), mivel a forrás már szerepelt a tanítóadatok között. Viszont amikor az MI „fejből" (a kérdésben a forrást nem megadva) válaszol, akkor csak a saját emlékezetét használja (a „parametrikus memóriát"). Amikor RAG-on keresztül – tehát „csatoltan" – kap adatot, az közvetlenül kerül a promptba. Azaz – feltehetően – nem számít bele a saját emlékezetébe. Azért számít mégis külsőnek (még ha „ismerős" is), mert a rendszer számára elsőbbséget élvez (frissebb), pontosan meg van jelölve, melyik forrásból vegye az információt, és arra utasítja „magát", hogy a megadott forrást részesítse előnyben a saját belső, esetleg elavult vagy pontatlan tudásával szemben.

Valódi „külső" forrásról technikailag csak akkor beszélhetünk, ha az adat zárt hálózatból származik (pl. a magánszemély dokumentuma az **Elektronikus Egészségügyi Szolgáltatási Térből** (EESZT vagy vállalati belső PDF-ek).

A továbbiakban a ChatGPT párbeszédes rendszerből (Chat-based Generative Pre-trained Transformer) vett példákat fogunk használni.

### A végfelhasználói RAG

A felhasználó számára a kérdésnek („prompt") fenntartott mező bal szélén a „+" parancs szolgál arra, hogy megadja a forrást, melyet a rendszer automatikusan figyelembe vesz a válaszadáskor. Akár „be is húzhat" egy dokumentumot közvetlenül a párbeszédes mezőbe (ablakba). Mivel az MI operatív memóriája véges, az így megadott szöveget kisebb részekre darabolja, átalakítja a saját nyelvére („embedding"), kiválasztja a leginkább releváns részeket és válaszol („generál"). A „+" parancs tehát nem más, mint a RAG a felhasználói szinten. Felhasználóbarát RAG, amely nem igényel RAG-ra vonatkozó, mélyebb informatikai tudást, és lehetővé teszi a forrás alapú választ. A felhasználó legfeljebb annyit tud, hogy „a dokumentumot odaadtam az MI-nek, hogy abból dolgozzék".

Ha a forrás szövegét a kérdésben adnák meg (tehát nem, mint dokumentumot), az hosszabb is lehet, mint amennyit az MI egyhuzamban hatékonyan fel tudna dolgozni (nagyobb, mint a feldolgozható szövegrész-hossz, a „token limit"). Ha a szöveghossz eléri ezt a határt (ameddig hatékonyan képes dolgozni), a rendszer elkezd „felejteni". Hibaüzenetet ad vagy levágja a válasz végét. A kérdésbe foglalt („ágyazott") hosszabb szöveget, dokumentumot nem képes darabokra törve optimalizálni („skálázni") a feldolgozhatóság érdekében.

A „+” paranccsal tehát forrásként megadhatók a legkülönbözőbb dokumentumok .txt, .csv, .pdf, .docx. és xlsx formákban. Itt jegyezzük meg, hogy bonyolultabb feladatok megoldásakor a legjobb, ha mindig csak az első kettővel dolgozunk. A RAG a forrásokat

- szöveggé alakítja,
- a kisebb (500–1000 karakteres) információegységekre, „chunkokra” bontja,
- az egységekből számsoros lenyomat („embeding”) készül,
- a lenyomatok a válasz létrehozásáig egy vektoradatbázisba kerülnek, utána törlődnek,
- tehát *a forrásállomány csak az adott beszélgetés keretén belül létezik*, „él”, és
- eleve nem skálázódik hatalmas mennyiségű forrásállományra.
- A válasz létrehozásakor („generálásakor”) a rendszer az így keletkezett tartalomból dolgozik.
- A releváns részeket használja föl,
- afféle „forrásként” kezeli a feldolgozott tartalmat,
- más szóval a „kérdés+dokumentum” alapján válaszol.

A lekérdező rendszer működése

- nem kulcssavakon alapszik, hanem
- jelentés-hasonlóság (pontosabban „lenyomat” hasonlóság) alapján működik. Ez afféle kereshető „jelentés-alapú index”

A keresőkérdésből („promptból”) ugyancsak lenyomatok készülnek, és ezeket hasonlítják össze a meglévő lenyomatokkal.

A felhasználói RAG esetében a feldolgozás menete csak megközelítően hasonló a professzionális RAG rendszeréhez képest. Merev, de ennek fejében

- egyszerű,
- automatikus,
- kis dokumentummennyiségre jó,
- nem testre szabható.

Mindez a „mélyben” játszódik le, a felhasználónak ehhez nem kell értenie. Az ő feladata a prompt pontos, egyértelmű megfogalmazása. Ehhez igénybe vezeti az MI-t. Megkérdezheti például, nincs-e ellentmondás a promptban, és a válasznak megfelelően módosíthat.

### A végfelhasználói RAG néhány köznapi példája

Adott nagyobb dokumentum szövege alapján fordítást kérni, lásd Ungváry (2026).

Adott betegség, egészségi állapot orvosi bizonylatai (anamnézisek, laborvizsgálatok) alapján megkérdezni ezek jelentését, a kilátásokat stb.

Adott festmény(ek) alapján megkérdezni a szerzőséget, vagy a szerző jelentőségét.

Házassági vagyonmegosztási per iratai alapján tájékozódni a kilátásokról.

Adásvételi, bérleti, vállalkozási stb. szerződés alapján megkérdezni a kockázatokat, a megfogalmazás lehetséges csapdáit stb.

### A végfelhasználói RAG néhány könyvtári példája

- Keresd azokat a fordításokat, melyeknél nem ismert a forrásnyelv vagy a fordító vagy az eredeti cím
- Keresd a XX. századi városmajori szerzők alkotásait.
- Keresd a XIX. században írt gyerekirodalmi alkotásokat.

Hasonló lehetőségek adódnak levéltárakon belül a levéltári állományok digitalizált részében a keresésre.

E példák esetén hozzá kell tudni férni az adott intézmény adatbázisához. Könyvtári területen megkönnyíti a helyzetet, hogy Magyarországon egyrészt létezik (még) a Magyar Elektronikus Könyvtár állománya. Másrészt Király (2024) analitikusan feldolgozta a nagyobb nemzeti könyvtárak katalógusait. Ezek az általa megadott ugrópontokon keresztül hozzáférhetők. Ezek a QA Catalogue állományok.

Általában szükség van a prompt pontos és igényes megfogalmazására. Ehhez olykor segítség is szükség lehet. A prompt helyességéről maga az MI is megkérdezhető, és a válaszai alapján finomítható.

### Egy konkrét könyvtári példa

A könyvtári állományok ma még túlnyomórész csak a MARC21 katalógustételei alapján használhatók fel a RAG számára. Ez egyelőre szűkíti az alkalmazhatóságot, hiszen csak a katalógus rekordok mezőtartalmai alkotják a RAG forrásállományát. Az alábbi példában az Országos Széchényi Könyvtár QA Catalogue állományát használtuk fel a lekérdezéshez: http://ddb.qa-catalogue.eu/oszk/?tab=terms&query=*:*&lang=en&facet=080a_Udc_ss

Az vizsgáltuk, hogy milyen minőségűek, milyen és mennyi formális hibát tartalmaznak az OSZK ETO-jelzetei. A QA Catalogue-ba belépve elérhető a MARC21 bibliográfiai rekord 080$a mezőjében rögzített jelzetek. A RAG első forrását tehát a 080$a állomány alkotja. Ez volt tehát a célállomány.

A referencia állományt az ETO táblázatai alkotják, mint egységesített besorolási adatok.

Választásunk oka az ETO jövőbeli jelentőségében rejlik, melyre már korán rámutattunk, lásd Ungváry (2020, 2021), miszerint az információmennyiség nagy mértékű növekedése következtében akkora lesz a kereséskor kapott zaj, hogy idővel megnő a hiteles információk jelentősége. Az ETO-val feldolgozott állományok fel fognak értékelődni.

A probléma az, hogy csak az ETO legkorábbi magyar kiadása érhető el (ETO 1958). Az 1980-as teljes magas színvonalú kiadás (ETO 1980) csak nyomtatásban létezik. Ugyancsak nyomtatásban létezik az ETO új magyar kiadása (ETO 2005), mivel csak nyomtatási célra vásárolták meg a digitalizált ETO állományát a nemzetközi konzorciumtól. Ezért ehhez a munkához nem kaphattuk meg, mivel fizetni kellett volna érte.

A második forrást ezért az ETO (1980) alkotja. Ez újabb probléma: az MI nem tudja a relatív ugróprontot megnyitni. Az abszolút ugrópontot is hiába nyomoztuk ki. A MEK oldal régi HTML-szerkezet, nem lehet jól darabolni, jól beágyazni, ezért az MI elutasítja. Kénytelenek voltunk letölteni, és magát a .pdf dokumentumot „+" paranccsal csatolni. Így az ETO formai jelzetellenőrzésére az így megadott ETO már jobban használható a RAG számára, de valójában így sincs igazán megbízható validálás, lévén, hogy a „képi" állományból a beolvasott szöveget digitálisan szerkeszthető és kereshető formátummá (OCR) kellett átalakítani. Ez a rendkívül rossz minőségű .pdf állomány miatt eleve tökéletlenül mehetett csak végbe, ezért az eredmény egyelőre ugyancsak rendkívül tökéletlen.

A tökéletes megoldás a strukturális ETO referencia. A nemzetközi Konzorcium kezelésében létezik a minden adatot tartalmazó állomány, az UDC Master Reference File (MRF), de a teljes hivatalos változat **licenchez kötött**.[1] Mindebben azonban a nem informatikus és nem intézményi felhasználó nem illetékes, kénytelen szükségmegoldással megelégednie.

További probléma, hogy amikor az OSZK cédulakatalógusát digitalizálták, a nyelvi és népi alosztásoknál a rendszer nem ismerte föl az egyenlőségjelet (=). Ennek következménye, hogy ezek a jelzetek a 080$a mezőben enélkül vannak jelen. A 700$m megjegyzés mezőben szerepel, hogy „Cédulakatalógus alapján", így ezek a jelzetek külön prompttal később majd kiválogathatók és egy másik prompttal kielemezhetők.

Még további probléma, hogy az = előzékű nyelvi és a (= előzékű népi alosztások az ETO (2005) kiadásában megváltoztak.

Az UDC nemzetközi kiadásában ráadásul megváltoztatták az összes =9... utáni jelzetet. A későbbi promptban majd azt is szabályozni kell, hogy hogyan kezelje az MI a hiányzó = jelet és a többi hiányzó jelet. Mindez csak a jéghegy csúcsa.

1 Az UDC Consortium hivatalos oldalán az UDC Master Reference File (MRF) oldalon található róla információ. Eszerint az ETO-MRF relációs adatbázisban (MySQL) *v*an tárolva és XML exportként licencelhető. Ebből például a JSON előállítható lenne. Az OSZK csak a szegényesebb középkiadást vásárolta meg a nyomdai kiadvány számára, ezért ehhez a munkához külön engedély és költségek nélkül amúgy sem lehetett felhasználni.

Ez utóbbi probléma miatt digitalizáltuk az ETO (1980) kiadásából az alosztások táblázatait, és az ETO (2005) kiadásából a 9-es főosztály táblázatait. Ezek alkotják a negyedik forrást. A több digitalizáls meghaladta a szerző erejét. Se a teljes magyar ETO (1980), se a magyar ETO (2005) nem elérhető[2]. A promptot tehát ennek hiányában kellett futtatni és annak alapján dolgozott az MI, amivel „betanították".

A többszörösen ismételt futtatások eredménye az alábbiakban látható.

1. **Ellenőrzött jelzetek száma:** 477 362
2. **Valószínű ETO-jelzet:** 332 891 (*törzs nélkül csak feltételezve*)
3. **Formailag hibás jelzet:** 125 087
4. **Nem létező főosztály / alosztály:** 27 796 (*törzs nélkül nem vizsgálható*)
5. **Hibás közös alosztás:** 52
6. **Szabálytalan összekapcsolás:** 656 (*rossz operátor pl.* `:`, +, /, `: : ;` *dupla kötőjel, hibás kombináció pl. --; rossz helyen levő kapcsolatok*)
7. **Nem kategorizálható:** 0
8. **Felülvizsgálandó:** 26

**Kiszűrt zaj 1. és 2. különbsége** (*ETO-törzs nélkül eldönthetetlen*): 130 691.

A **2. Valószínű 332 891 ETO jelzet többsége** további vizsgálatot igényel, melyről egy későbbi **CeLISR** (*Central European Library and Information Science Review = Közép-európai Könyvtár- és Információtudományi Szemle*) tanulmányban tájékoztatunk.

## A professzionális RAG

A professzionális RAG esetében a feldolgozandó forrásállomány eleve összehasonlíthatatlanul nagyobb, „ipari" méretű. Például néhány ezer szerződésből, költségtáblázatból, szabályzatból stb. álló forrásállományt kell tudni kezelni, és ebből kiindulva kellenek válaszok. A különbség tehát egyrészt a külső forrás mérete, másrészt a felhasználói felület. Ezzel természetszerűleg együtt jár a professzionális RAG jóval bonyolultabb belső szerkezete, mellyel csak vázlatszerűen foglalkozunk.

2 Pontosabban: ahhoz, hogy a 2005-ös kiadás digitalizált változata elérhető legyen, az UDC-t karbantartó konzorciumnak fizetni kell.

A professzionális RAG esetében

- biztosítva van a nagyobb feldolgozható szövegrész-hossz (a „token limit”),
- több ezer dokumentum („dokumentumkorpusz”) esetén is működik,
- gyors, mert teljesen skálázható,
- teljesen ellenőrizhető,
- folyamatosan bővíthető.

Ott alkalmazzák ahol

- ügyfélszolgálati szövegalapú párbeszédrendszerre („chatbotra”) van szükség, vagy
- meghatározott típusú forrásokból álló adatbázisból (pl. jogszabály-adatbázis) kell tudni kérdezni,
- belső dokumentumrendszert kell MI-alapúvá tenni.

A teljes RAG-architektúrában jóval összetettebb

- a lenyomat (embedding) generálás,
- a vektoradatbázis (pl. Pinecone, Weaviate, FAISS).
- Nem a teljes találati listával kerülünk szembe (mint például a hagyományos kulcsszavak adatbázis-lekérdezéskor), hanem
- csak a relevanciát mutató pontszám („score”) alapján kiválasztott legmegfelelőbb „k” darab eredményt kapjuk (ez a „top-k retrieval”).
- A megadott keresőkérdésbe automatikusan bevonódnak például a szinonimák, kijavítódnak az elírások, azaz a rendszer helyesbíti, kiegészíti és szemantikailag is némileg értelmezi a kérdést (ez a keresőkérdés „átírása”, a „query rewriting”).
- Az első lépésben kapott találati halmazból kiválasztott részhalmazt egy további, „okosabb” összehasonlító elemző megvizsgálja („cross-encoder”) **és új, finomított**/tökéletesített/javított relevancia sorrendbe rendezi, ez a „reranking”). A jövő zenéje, ha a rendszer strukturált szemantikai szótárt (pl. tezauruszt) is használ, mert akkor a keresésbe automatikusan bevonódhat a fölé- vagy alárendelt szavakkal, vagy az egészt és a részt kifejező szavakkal való keresés is („generic posting”, „partitive posting”).

A három felsorolt eljárást egyébként a webes keresők (pl. a Google) is alkalmazzák.

A professzionális RAG-rendszerek mintegy megtanulják egyensúlyba hozni a belső tudásukból (előzetes képzés) származó információkat (azaz esetünkben a ChatGPT-t) a külső lekérdezett adatokkal. Ez a folyamat javítja a rendszer azon képességét, hogy pontos és a kontextusnak megfelelő válaszokat generáljon.

A professzionális RAG-rendszerek azonban nem tartalmazhatnak egyszerre több különböző, ráadásul nagy külső forrásállományt. Más szóval nem lehet mindent egyetlen RAG-rendszerbe bepakolni.

Ezt a korlátot küszöböli ki az ügynökalapú RAG (Agentic RAG). Ez irányítja a válaszadást például a külső forrásokból történő információkeresés (RAG) eredményeire. Kereshet RAG-gal, és kereshet akár kulcsszóval hagyományos adatbázisokban is. A hagyományos RAG-gal ellentétben az Agentic RAG ügynökei képesek a kérdések elemzésére, a keresési stratégia finomítására, eszközök (pl. adatbázisok) válogatott használatára és a válaszok validálására az iteratív folyamat során.

A hagyományos professzionális RAG esetében egyszeri kérdés–keresés–válasz a folyamat. Az ügynökalapú RAG esetében a válasz dinamikus, interaktív folyamat, az ügynök értékeli az információt és szükség esetén módosítja a keresést. Az olyan válaszhoz például, hogy „van-e összefüggés a saját bank és a különböző bankok költségeinek emelkedése és a szerződésbontások között?" Használni kell hagyományos adatbázisokat és RAG-ot egyaránt.

A professzionális RAG készítésének és fenntartásának jelentős költségei vannak:

- egyszeri vagy frissítésenkénti átalakítási („embedding") költség,
- kérdésenkénti átalakítási és rendszerhívási költség.

Ezzel sok kérdés (napi 1000) esetén már számolni kell.

Mindennek az infrastruktúrája általában többe kerül, mint maga az alkalmazott MI-rendszer.

Egy professzionális RAG-rendszer:

- nem veszélyesebb, mint egy belső dokumentumkezelő rendszer, de
- ha üzletileg kritikus döntések születnek belőle, akkor már magas a kockázati szint.

Egy kis- és középvállalati RAG

- ugyan nem automatikusan sérülékeny, de
- nem is „banki szintűen" védett.

A védelem minősége nem a választott MI, hanem a tervezés függvénye.

## Első kitekintés

A felhasználói RAG-gal megadott forrást maga az MI-rendszer nem tanulja meg, és nem építi be tartósan. Ez csak az adott beszélgetés (munkaülés) összefüggésében „él". A rendszer a „+" paranccsal a munkameneten belül használja válaszhoz, de csak ideiglenesen. Lényegében a professzionális RAG esetében is ez a helyzet.

### Második kitekintés. A kognitív és a gépi kapacitáskorlát

George A. Miller szerint (1956) a rövidtávú/munkaemlékezet kapacitására az jellemző, hogy átlagosan 7±2 információegységet (szót, számot, képet – darabot, „chunk"-ot) tud egyszerre kezelni. Nevezik rövidtávú kognitív kapacitáskorlátnak is, ez az agyi információ-

feldolgozás korlátja. Ezért nincs az egyes nyelvekben például ennél több mondatrész se. Ez magyarázza, miért nehéz a túl hosszú és összetett mondatok értelmezése.

Kolmogorov (és később más kutatók, mint például Victor Yngve) vizsgálták a mondatok feldolgozásának hierarchikus mélységét („stack depth"). Megfigyelték, hogy a természetes nyelvekben a mondatokon belüli „függőségi távolság" vagy a beágyazások száma ritkán lépi át a 11-es határt. E fölött a mondat szintaktikai szerkezete már nem tartható egyben egyetlen koherens gondolati egységként. Ez a Kolgomorov-féle szám, a mondat bonyolultsági mélysége.

Ha egy mondatban túl sok a függő viszony (például egy alanyhoz túl sok jelző vagy mellékmondat kapcsolódik távolról), az agy – feltételezett – „veremmemóriája" (a „stack")[3] megtelik. Ha egy mondat túl komplex (hierarchikus mélysége túllépi ezt a 11-es bűvös határt), a megértés hatásfoka drasztikusan romlik. Az egyes nyelvek nyelvtani szerkezetei (mondatrészek típusai) valóban úgy fejlődtek ki, hogy ne kényszerítsék az agyat ezen elméleti korlát átlépésére.

Miller és Chomsky (1963) kapcsolták össze először a mondatstruktúrát és a memória-korlátot, és fogalmazták meg, hogy a nyelvhasználók úgy működnek, mint a véges memóriájú automata jellegű rendszerek. Ezzel klasszikus kapcsolatot teremtettek a kognitív kapacitás és a formális nyelvi modellek között.

Az MI-rendszereknek ugyancsak vannak kontextus- és memória-korlátai („context window", „token limit", „lost in the middle"). Feldolgozás közben a verem mélysége nő, és ezzel párhuzamosan nehezebb a feldolgozás.

Egy MI-rendszernél ez azt jelenti, hogy mekkora szöveget képes egyszerre a memóriájában tartani és feldolgozni egy adott munkamenetben. A méret a kontextus-ablak. Ennek kezelése érintőlegesen tekinthető a modern informatika válaszának ugyanerre a „Miller-Kolmogorov" problémára. Ahogy az emberi agy túlterhelődik 7–11 egység után, az MI-rendszereknél is megfigyelhető az „elveszés a középen" („lost in the middle" jelenség: hiába hatalmas a feldolgozható szövegrész-hossz („token-limit"), ha túl sok irreleváns információt (zajt) zsúfolunk a promptba, a modell figyelmi mechanizmusa „elfárad", és a lényeg elsikkad. Romlik az MI „megértő képessége".

Tehát a természetes nyelvekhez hasonlónak látszó probléma jelent meg. Van egy optimális nagyságú információs egység, amely fölött az MI esetében is egyre nehezebb az „értelmezés". Az optimum érdekében a professzionális RAG-nál alkalmazzák a korábban tárgyalt újrarendezést („reranking"), amely éppen azt csinálja, amit az emberi figyelem: megpróbálja a beérkező információt arra a „bűvös" néhány egységre redukálni, amit a modell még hallucináció nélkül, egyetlen logikai egységként képes átlátni.

3 A RAM egy dedikált, gyors területe, amely a programok futása során a lokális változókat, függvényhívásokat és visszatérési címeket kezeli

## Összegzés

1. Az LLM nem tudástár. 2. Azt tudja, amit a betanítás időpontjában adtak meg neki. 3. „Tudásához" meg kell adni a forrásokat. 4. A legspecifikusabb információk nincsenek a hálón. 5. Nem a legfrissebb információ van a hálón. 6. Az igazán összetett, mélységi forráselőírt kereséshez nem elég a végfelhasználói ismeret. 7. A jövő a specializálódott RAG szövegalapú párbeszédrendszereké („chatbotoké"), még inkább az ügynökalapú RAG.

## A szakirodalomról

A RAG-ot tárgyaló, a végfelhasználó számára is érthető szakkönyv, vagy akár tanulmány jelenleg még nincs. Ennek az előadásnak/tanulmánynak a gondolatait a néhány, 2024 óta már megjelent angol és német szakkönyvből, számtalan tanfolyami és internetforrásból, szakértők személyes közléseiből – mondhatni – információs forgácsokból állítottuk össze.

Ezért nem lehet egyetlen vagy néhány mérvadó forrást megadni. Csak példaként utalunk Anand Vemula (2024) rövid kézikönyvére, melyhez hasonlókat a szerző az MI területéről sorozatban publikált a Kindle kiadónál.

A kognitív kapacitáskorlátra vonatkozóan csak a legfontosabb forrásokat említjük meg. A „klasszikusokat", mint Miller (1956), Yngve (1960), Miller, Chomsky (1963).

A kognitív szintaxis és a neurális nyelvmodellekről Ming, Vitányi (2008).

A felhasználói RAG-gal először a 2025. évi Networkshop konferenián tartott előadásunkban és a konferencia kötetbe fölvett tanulmányban foglalkoztunk (Ungváry (2025).

A QR Catalogue-ról szóló számos publikáció közül lásd Király (2024).

Külön köszönet illeti tanácsaiért Radnai Miklóst, a ServeusAI vezetőjét.

## Irodalomjegyzék

*ETO* (1958). Egyetemes Tizedes Osztályozás. A nemzetközi táblázatok hivatalos magyar kivonata. Szerk. Vekerdy Gyula. Budapest, Bibliotheca Kiadó. – Az Országos Széchényi Könyvtár Kiadványai 40. https://mek.oszk.hu/19600/19690/ (2026. 04.01.)

*ETO* (1980). Egyetemes Tizedes Osztályozás. Teljes kiadás. Magyar Szabványügyi Hivatal. MSZ 16500–80 (FID Publ. No. 390). Budapest, Szabványkiadó.

ETO (2005) Egyetemes Tizedes Osztályozás. UDC Publ. No. P057. Országos Széchényi Könyvtár. Könyvtári Intézet., Budapest, 2005. 1. kötet. Táblázatok. 1. rész: Segédtáblázatok és főtáblázati számok 0/5999.89. 2. rész: Főtáblázati számok 6/94(94).05.

Király P. (2024). QA Catalogue – A Quality Assessment Tool for Library Catalogues. GWDG Nachrichten 04–05. https://gwdg.de/about-us/gwdg-news/2024/GN_04-05-2024_www.pdf#page=19 (2026. 04.01.)

Miller G. A., N. Chomsky 1963. Finitary models of language users. In: Luce et al. (1963, 419–491). Magyarul: A nyelvhasználók véges modelljei. In: Pléh Cs. (szerk.) Szöveggyűjtemény a pszicholingvisztika tanulmányozásához. Budapest: Tankönyvkiadó. (57–110).

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychological Review. **63** (2): 81–97.

https://labs.la.utexas.edu/gilden/files/2016/04/MagicNumberSeven-Miller1956.pdf (2026. 04.01.)

Ming L., Vitányi P. (2008). *An Introduction to Kolmogorov Complexity and Its Applications.* 3. Auflage. Springer-Verlag, New York 2008, ISBN 978-0-387-33998-6, DOI: https://doi.org/10.1007/978-0-387-49820-1 (2026. 04.01.)

Ungváry, Rudolf (2025). *A ChatGPT felhasználása a MARC21 fordítására = Using ChatGPT for MARC21 Translation.* In: Oktatási, kutatási és közgyűjteményi infrastruktúrák és tartalmak: digitális transzformáció felsőfokon : NETWORKSHOP 2025 : 34. Országos Informatikai Konferencia : 2025. május 13–15. Széchenyi István Egyetem, Győr. Hungarnet Egyesület, Budapest, pp. 222-231. ISBN 978-615-6792-15-0 https://real.mtak.hu/229661/1/223_fejezet_NETWORKSHOP_2025.pdf

https://videosquare.eu/hu/recordings/details/18790, A_MARC21_eddigi_magyar_forditasai_es_ChatGDP_forditasa._Osszehasonlito_vizsgalat (2026. 04.01.)

Ungváry, Rudolf (2021. *Bemerkungen zu der Qualitätsbewertung von MARC-21-Datensätzen. Qualität in der Inhaltserschliessung*. In: Qualität in der Inhaltserschliessung. De Gruyter, Saur, 2021. (Bibliotheks- und Informationspraxis, Band 70.), pp. 177–229.

Ungváry, Rudolf (2020). Ismeretszervező-könyvtári rendszerek tartalmi feltárásának összehasonlító vizsgálata MARC21 környezetben. In: Tudományos és Műszaki Tájékoztatás. 2020 67. évf. 11. sz., pp. 655–680. https://journals.bme.hu/tmt/article/view/35274 (2026. 04. 01.)

Vemula A. (2024) Mastering the RAG: A Practical Guide to Deploying AI-Powered Data Retrieval and Generation in Your Enterprise -ERP, SAP, SFDC. Kindle Edition.

Yngve (1960). A Model and Hypothesis for Language Structure. 369. kötet/Technical report / Research Laboratory of Electronics, Massachusetts Institute of Technology, Massachusetts Institute of Technology Research Laboratory of Electronics. 466 p. Band 70. https://aclanthology.org/www.mt-archive.info/50/ProcAmPhilSoc-1960-Yngve.pdf (2026. 04. 01.)

DOI: 10.31915/NWS.2026.6



# MARC-BOTOK HARCA: A GENERATÍV MESTERSÉGES INTELLIGENCIA MODELLEK LEHETŐSÉGEI ÉS KORLÁTAI A KÖNYVTÁRI KATALOGIZÁLÁSBAN

# THE BATTLE OF THE MARC BOTS: THE POTENTIAL AND LIMITATIONS OF GENERATIVE ARTIFICIAL INTELLIGENCE MODELS IN LIBRARY CATALOGING

*Takács Margit*
*Debreceni Egyetem BTK, Könyvtár- és információtudományi Tanszék, egyetemi adjunktus*
*takacs.margit@arts.unideb.hu*
*ORCID: 0009-0004-9206-733X*

*Borbély Mária*
*Debreceni Egyetem BTK, Könyvtár- és információtudományi Tanszék, egyetemi adjunktus*
*borbely.maria@arts.unideb.hu*
*ORCID: 0000-0002-1161-1419*

## Absztrakt

A könyvtári katalogizálás technológiai fejlődése során a hagyományos katalóguscédulákat mára felváltották az online rendszerek, amelyek feldolgozómunkáját legújabban már a mesterséges intelligencia is támogatja. A generatív GPT-modellek képesek a bibliográfiai metaadatok és MARC-rekordok automatikus előállítására, ami új távlatokat nyit, ugyanakkor komoly minőségbiztosítási kérdéseket is felvet. A tanulmány a kifejezetten könyvtári feldolgozásra fejlesztett generatív GPT-modellek által létrehozott MARC-rekordok elemzésére fókuszál. Az összehasonlító vizsgálat olyan szempontokat érvényesít, mint a szintaktikai megfelelőség, a tartalmi validitás, a feldolgozás hatékonysága és a konzisztencia. Az elemzés megmutatja az MI jelenlegi és jövőbeli helyét a katalogizálási folyamatokban.

**Kulcsszavak:** Mesterséges intelligencia (MI), katalogizálás, MARC21

## Abstract

As library cataloging technology has evolved, traditional catalog cards have now been replaced by online systems, whose processing tasks are now supported by artificial intelligence. Generative GPT models are capable of automatically generating bibliographic metadata and MARC records, which opens up new possibilities while also raising serious quality assurance issues. This presentation focuses on the analysis of MARC records generated by generative GPT models specifically developed for library cataloging. The com-

parative study evaluates aspects such as syntactic correctness, content validity, processing efficiency, and consistency. The analysis highlights the current and future role of AI in cataloging processes.

**Keywords**: Artificial Intelligence (AI), cataloging, MARC21

## Bevezetés

A mesterséges intelligencia (MI, Artificial Intelligence, AI), különösen a nagy nyelvi modellek (Large Language Models, LLM) megjelenése nemcsak az információszerzésre van hatással, hanem a könyvtári munka különböző területeit is befolyásolja, köztük a feldolgozómunkát. A generatív MI képességei ugyanis a szabadszöveges tartalomgyártáson túlmenően a strukturált metaadatok generálása terén is megújulási lehetőséget hordoznak. A könyvtári szakemberek számára a jelenlegi kihívást már nem annak eldöntése jelenti, hogy indokolt-e az MI-alapú eszközök alkalmazása a feldolgozási munkafolyamatokban, hanem sokkal inkább a katalogizálásra fejlesztett megoldások hatékonyságának és megbízhatóságának kiértékelése.

E kihívásra reagálva a tanulmány négy, kifejezetten könyvtári feldolgozásra létrehozott generatív MI-eszköz (CatalogerGPT, CATMELK, Cataloging Assistant, MARC-AI Cataloger) teljesítményét hasonlítja össze. A kutatás bemutatja az alkalmazott módszertant és az eszközöket, majd egy ember által készített referenciarekordhoz viszonyítva elemzi az algoritmusok kimeneteit, hogy meghatározza a mesterséges intelligencia jelenlegi és jövőbeli helyét a katalogizálási folyamatokban.

## Generatív AI-alkalmazások

A generatív mesterséges intelligencia (GAI), például a GPT-technológia, betanult adatok alapján képes új tartalmakat generálni és valós idejű interakciót biztosítani. Bár az általános célú nyelvi modellek (pl. ChatGPT) felismerik a bibliográfiai struktúrát, kutatások (Taniguchi, 2024; Niveditha és mtsai., 2024) igazolják, hogy a MARC 21-rekordok készítése során gyakran pontatlanok, strukturális hibákat vétenek és hajlamosak fiktív adatok generálására (hallucinációra).

Ezen hiányosságok kiküszöbölésére a könyvtári informatikai fejlesztések két fő irányát figyelhetjük meg. Az egyik megközelítést a kereskedelmi integrált könyvtári rendszerekbe (ILS) beépített MI-alapú katalogizálástámogató eszközök jelentik, mint például az Alma AI metaadat-asszisztense vagy az OCLC új mesterséges intelligencia alapú katalogizáló eszközei. Ezek a munkafolyamatokba integrált intelligens asszisztensek szöveg vagy kép alapján generált MARC 21-rekordvázlatokkal és automatikus osztályozási szám (DDC, LCC), valamint tárgyszó (LCSH) javaslatok generálásával támogatják a feldolgozást, anélkül, hogy kivennék a döntést a könyvtáros kezéből (OCLC, 2025., Ex Libris, 2025.)

A másik fejlesztési irányt az önálló, szakterület-specifikus katalogizáló asszisztensek jelentik, melyeket szintén kifejezetten a MARC 21 formátumhoz és a különböző osztályozási rendszerekhez igazítottak. A jelen kutatás négy ilyen, konkrét intézményi szabályoktól és kereskedelmi rendszerektől független eszközt elemez: a CatalogerGPT, a CATMELK, a Cataloging Assistant és a MARC-AI Cataloger alkalmazásokat.

### CatalogerGPT

A CatalogerGPT egy egyéni fejlesztés, amelyet Glen Greenly, több mint 30 éves tapasztalattal rendelkező kanadai könyvtári szakember hozott létre. Az alkalmazás kiemelkedő képessége, hogy a könyvről készült képek alapján azonosítja a bibliográfiai adatokat, kitölti a megfelelő MARC-mezőket, és képes a meglévő rekordok hibaellenőrzésére is. Továbbá tárgyszavakat, osztályozási jelzeteket (Library of Congress Classification=LCC, Dewey Decimal Classification=DDC) javasol (Toolify.ai., 2026; Greenly, é.n.).

### CATMELK

A CATMELK (Cataloging using Machine Learning Knowledge) egy kutatási projekt keretében jött létre, amelyet két könyvtári szakember, Ruwan Gamage és Priyanwada Wanigasooriya fejlesztett ki Srí Lankán. A projekt fő erőssége a vizuális adatkinyerés. OCR-technológiával (optikai karakterfelismeréssel), fényképek alapján állít elő RDA[1]-kompatibilis MARC 21-rekordokat. Automatikusan pótolja a hiányzó szabványos mezőket (pl. 3XX), és validálja a bibliográfiai leírásokat. Emellett a CATMELK javaslatokat tesz a megfelelő osztályozási jelzetekre (DDC[2], LCC[3]) és tárgyszavakra is (Gamage & Wanigasooriya, 2024; Gamage & Wanigasooriya, é.n.).

### Cataloging Assistant alkalmazások

A ChatGPT-ben három olyan közösségi fejlesztésű Cataloging Assistant modell érhető el, amelyet a feldolgozómunka felgyorsítására hoztak létre. Ezek a megadott adatokból RDA-kompatibilis MARC-rekordokat generálnak, emellett – a választott eszköztől függően – automatikusan meghatározzák az adott könyvtári osztályozási jelzeteket (LCC vagy

1 Az RDA (Resource Description and Access = Forrásleírás és -hozzáférés) egy korszerű bibliográfiai keretrendszer, amelyet a könyvtári tartalmak leírására fejlesztettek ki. Célja olyan adatok létrehozása, amelyek a felhasználóközpontú, kapcsoltadat-alapú alkalmazások fejlesztése érdekében kialakított nemzetközi modellek szerint is jól formáltnak tekinthetők.

2 A DDC (Dewey Decimal Classification vagy Dewey-féle tizedes osztályozás) egy könyvtári osztályozási rendszer, amelynek célja a könyvek témák szerinti rendszerezése.

3 Az LCC (Library of Congress Classification) egy betűkből és számokból álló jelzeteket alkalmazó osztályozási rendszer, amely a dokumentumok tartalmi feltárására és visszakereshetőségének biztosítására szolgál.

DDC), valamint a polcrendi elrendezéshez szükséges Cutter-számokat (LC vagy Sanborn). (Cataloging Assistant LCC and LC Cutter, é.n.; Cataloging Assistant DDC and LC Cutter, é.n.; Cataloging Assistant LCC and Sanborn Cutter, é.n.)

**MARC-AI Cataloger**

A MARC-AI Cataloger egy egyéni fejlesztés, amelyet Asif Altaf hozott létre. Elsősorban az alapadatok (ISBN, cím, szerző) alapján generál RDA-kompatibilis rekordokat, emellett tárgyszavakat (LCSH), osztályozási jelzeteket (Library of Congress Classification=LCC, Dewey Decimal Classification=DDC) javasol. Kifejezetten hangsúlyozza az emberi ellenőrzés szükségességét a véglegesítés előtt (Altaf, é.n.).

## Módszertan

A kutatáshoz egy referenciaként szolgáló rekordot készítettünk a kiválasztott könyv digitalizált képei alapján. A referenciarekord MARC 21 formátum (Library of Congress, 2025) alapján készült figyelembe véve az RDA leírási szabályait (RDA-HU Munkacsoport, 2026). A vizsgálat során a tartalmi feltáró mezőket (osztályozás, tárgyszavak) szándékosan kizártuk, hogy az intézményi gyakorlatok és a szubjektivitás ne befolyásolja az eredményeket. Így a vizsgálat fókuszában kizárólag a formai feltárás objektíven mérhető elemei maradtak.

A MARC 21 formátumú rekordok generálására a korábban ismertetett négy specifikus MI-alapú eszközt használtunk: CatalogerGPT, CATMELK, CatalogingAssistant és MARC-AI Cataloger. A rekordokat elsősorban a dokumentumról készített képek segítségével generáltattuk. A mesterséges intelligencia által létrehozott MARC-rekordok minőségének mérésére egy sajátos értékelési keretrendszert dolgoztunk ki, amelynek elméleti alapját Bruce és Hillmann (2004) klasszikus metaadat-minőségi dimenziói adták. A szakirodalomban és a metaadat-kezelők körében végzett felmérések alapján a legfontosabb minőségbiztosítási mutatóknak a pontosság, a teljesség és a konzisztencia bizonyultak (Park & Tosaka, 2010; Zavalin & Zavalina 2025). Ugyanakkor a nagy nyelvi modellek és a generatív MI-specifikus működési sajátosságai miatt ezt a három szempontot a könyvtári katalogizálás követelményeihez igazítva öt vizsgálati területre bővítettük:

- Teljesség: Azt vizsgálja, hogy a rekord tartalmazza-e az összes releváns mezőkódot, almezőt és indikátort. Ezt az LC minimumkövetelményei, illetve az összes kinyerhető adat alapján is kiszámítottuk (a felhasznált és az összes lehetséges elem hányadosaként).
- Pontosság: A MARC-formátum szerkezeti elemeinek (kódok, indikátorok) helyes alkalmazását méri. A nem létező, fiktív kódok hallucinálása automatikusan 0 pontot vont maga után, továbbá külön értékeltük a mezők ismételhetőségére vonatkozó szabályok betartását is.

- Tartalmi validitás: Az adatok tényszerűségét és forráshűségét ellenőrzi karakteres szinten a kritikus leíró mezőkben. Ezt a kategóriát egészítette ki a hallucináció vizsgálata, amely bináris értékelést kapott: ha az AI a forrásban nem szereplő, fiktív adatot (például kitalált oldalszámot vagy sorozatcímet) generált, az érték 0 lett, ellenkező esetben 1.
- Szabványkövetés és logikai konzisztencia: A formai leírási szabályok, mint a mezőkön belüli MARC21-ben előírt írásjelek és a fix hosszúságú mezők (000, 008) karakterhelyességének betartását vizsgálja. A logikai konzisztencia a mezők közötti belső ellentmondás-mentességet biztosítja (pl. a 100-as és 245-ös mező összhangját).
- Hatékonyság és ismételhetőség: A munkafolyamatba illeszthetőséget egyrészt a generálás sebességével (időfaktor másodpercekben), másrészt annak ellenőrzésével mértük, hogy a modell képes-e azonos bemenetből többször is pontosan ugyanazt a rekordot előállítani (determinisztikus működés).

## Eredmények

A kidolgozott szempontrendszer alapján elvégeztük a négy vizsgált eszköz (CatalogerGPT, CATMELK, Cataloging Assistant, MARC-AI Cataloger) által automatikusan generált leírások kiértékelését, a kapott eredményeket pedig egy összegző táblába helyeztük (1. táblázat).

| Szempont | | CatalogerGPT | | CATMELK | | Cataloging Assistant | | MARC-AI Cataloger | |
|---|---|---|---|---|---|---|---|---|---|
| Teljesség | Mezőkódok | 0,95 | 0,86 | 0,77 | 0,6 | 0,9 | 0,93 | 0,81 | 0,8 |
| | Indikátorok | 0,84 | 1 | 0,76 | 0,8 | 0,79 | 1 | 0,92 | 1 |
| | Almezők | 0,82 | 0,85 | 0,89 | 0,8 | 0,6 | 0,9 | 0,79 | 0,85 |
| | **Összesen** | **2,61** | **2,71** | **2,42** | **2,2** | **2,29** | **2,83** | **2,52** | **2,65** |
| Pontosság | Mezőkódok | 0 | | 1 | | 1 | | 0 | |
| | Indikátorok | 0,90 | | 0,9 | | 1* | | 0,83 | |
| | Almezőkódok | 1 | | 1 | | 1 | | 1 | |
| | Ismételhetőség | 1 | | 1 | | 1 | | 1 | |
| | **Összesen** | **2,9** | | **3,9** | | **4** | | **2,83** | |
| Tartalmi validitás | Adatpontosság | 0,99 | | 0,99 | | 0,99 | | 0,99 | |
| | Hallucináció | 0 | | 1 | | 0 | | 0 | |
| | **Összesen** | **0,99** | | **1,99** | | **0,99** | | **0,99** | |

| Szempont | | CatalogerGPT | CATMELK | Cataloging Assistant | MARC-AI Cataloger |
|---|---|---|---|---|---|
| Szabványkövetés és logikai konzisztencia | MARC21 tagolás | 0,875 | 0,875 | 1 | 1 |
| | Fix hosszúságú mezők | 0,95 | 0 | 0,98 | 0,98 |
| | logikai konzisztencia | 1 | 1 | 1 | 0 |
| | **Összesen** | **2,575** | **1,625** | **2,605** | **1,605** |
| Hatékonyság és konzisztencia | Időfaktor | 36 s | 52 s | 23 s | 29 s |
| | Ismételhetőség | nem | nem | nem | nem |

*1. táblázat: A vizsgált MI-eszközök értékelésének összesített eredményei*

A modellek teljességének vizsgálata rámutatott, hogy egyik mesterséges intelligencia sem tudta maradéktalanul teljesíteni a Library of Congress által megkövetelt minimumot, ráadásul a 040-es katalogizálási forrásmező generálása szinte mindegyiknek problémát okozott. Míg a kötelező adatmezők számában a Cataloging Assistant érte el a legjobb eredményt (14 mező), teljesítménye némileg ellentmondásos, mivel inkább az opcionális mezők kitöltésében volt erős, az alapkövetelmények teljesítésében már kevésbé bizonyult megbízhatónak. Összesítve a legkiegyensúlyozottabb eredményt a CatalogerGPT nyújtotta, míg a MARC-AI Cataloger és a mindössze 9 kötelező mezőt generáló CATMELK mérsékelten elmaradtak az élmezőnytől.

Bár az összes modell hibátlanul bontotta szabványos almezőkre a szöveget és maradéktalanul betartotta az ismételhetőségi szabályokat, a CatalogerGPT és a MARC-AI Cataloger érvénytelen, nem létező MARC-kódokat is hallucinált. Ezek az érvénytelen kódok valós könyvtári rendszerekben hibát okoznak, és az így feldolgozott adatok láthatatlanok maradnak az OPAC indexelésében és a találati listákban.

A tényadatok kinyerésében – ami a kiváló OCR és entitás-felismerési (név, cím, ISBN) képességeket bizonyítja – mind a négy modell szinte hibátlanul, 0,99-es pontszámmal teljesített. A hallucinációk terén azonban jelentős különbség mutatkozott: egyedül a CATMELK tudta teljesen elkerülni a fiktív adatok generálását, míg a többi mesterségesintelligencia-alapú eszköz hajlamos volt kitalált információkkal önkényesen kitölteni a hiányzó részeket.

MARC21-tagolás szempontjából a Cataloging Assistant és a MARC-AI Cataloger maradéktalanul követte a központozási szabályokat, míg a CatalogerGPT és a CATMELK esetében az egyetlen azonosított eltérés arra korlátozódott, hogy az előírásokkal ellentétben a 300-

as mező nem ponttal zárult. A fix hosszúságú mezők területén a CATMELK teljesített a legrosszabbul, mivel ez az alkalmazás eleve nem generálta le a 008-as mezőt és a rekordfejet sem hozta létre. Bár az eszközök többsége belső ellentmondások nélkül dolgozott, a Cataloging Assistant 0 pontot kapott logikai konzisztenciára, mert a 490-es mező indikátorának beállítása ellenére nem generálta a kötelező 830-as sorozati melléktételt.

A vizsgált modellek hatékonysága (generálási ideje) eltér: a leggyorsabb a Cataloging Assistant (23 mp), míg a leglassúbb a CATMELK (52 mp) volt, ami tömeges feldolgozásnál már meghatározó különbséget jelenthet a munkafolyamatban. A legfőbb problémát azonban az ismételhetőség teljes hiánya jelenti: egyik rendszer sem volt képes azonos bemenetből kétszer pontosan ugyanazt a rekordot előállítani. Ez a kiszámíthatatlanság szigorú utólagos ellenőrzés nélkül komolyan veszélyezteti a katalógusok egységességét.

## Összegzés

A kutatás összegzése szerint a generatív mesterséges intelligencia drasztikusan felgyorsítja a katalogizálást, de még nem képes teljes mértékben kiváltani az emberi szakértelmet. Bár az adatokat kiválóan nyeri ki, a szigorú szabványok betartása, az adathallucinációk és az ismételhetőség hiánya továbbra is komoly problémát jelent. Emiatt az MI jelenleg csupán kreatív asszisztensként használható, a könyvtárosi munka fókusza pedig a manuális rögzítésről a gépi adatok hitelesítésére és validálására tevődik át. A jövőben azonban, ha az MI kiküszöböli a fiktív adatok generálását, a determinisztikus működés hiányát, valamint a specifikus könyvtári szabályrendszerek alkalmazásában mutatkozó hiányosságokat, akkor a katalogizálásban betöltött szerepe szintet léphet. Azaz az MI egy megbízható, nagymértékben automatizált technológiává válhat, aminek köszönhetően a könyvtárosok a szakértelmüket és figyelmüket az adatok végső hitelesítésére és a komplex összefüggések kezelésére fordíthatják.

## Felhasznált források


Altaf, A. (é. n.). MARC AI Cataloger [Egyedi GPT]. OpenAI.

Elérhető: https://chatgpt.com/g/g-NVFV7JHhH-marc-ai-cataloger (Utolsó elérés: 2026. 02. 26.)

Bruce, T. R., Hillmann, D. I. (2004) The Continuum of Metadata Quality: Defining, Expressing, Exploiting. ALA Editions. Elérhető: https://ecommons.cornell.edu/entities/publication/b511c75c-849f-4b52-930e-e3a9f0cba043 (Utolsó elérés: 2026. 02. 26.)

Cataloging assistant: DDC and LC Cutter [Egyedi GPT]. (é. n.). OpenAI. Elérhető: https://chatgpt.com/g/g-WtqYttitP-cataloging-assistant-ddc-and-lc-cutter (Utolsó elérés: 2026. 02. 26.)

Cataloging assistant: LCC and LC Cutter [Egyedi GPT]. (é. n.). OpenAI.

Elérhető: https://chatgpt.com/g/g-abwV39ywz-cataloging-assistant-lcc-and-lc-cutter (Utolsó elérés: 2026. 02. 26.)

Cataloging assistant: LCC and Sanborn Cutter [Egyedi GPT]. (é. n.). OpenAI. Elérehtő: https://chatgpt.com/g/g-qO3cUTaDi-cataloging-assistant-lcc-and-sanborn-cutter (Utolsó elérés: 2026. 02. 26.)

Ex Libris. (2025.). Shaping tomorrow: Introducing Alma's AI vision - Chapter 1: AI in Alma's next-generation platform & the evolving user experience. Elérehtő: https://exlibrisgroup.com/blog/shaping-tomorrow-introducing-almas-ai-vision-chapter-1-ai-in-almas-next-generation-platform-the-evolving-user-experience/

(Utolsó elérés: 2026. 06. 08.)

Gamage, R., Wanigasooriya, P. (2024) Using Generative AI for Bibliographic Description: A Study with ChatGPT 4, Journal of the University Librarians Association of Sri Lanka, 27(2), p. 257–284. https://doi.org/10.4038/jula.v27i2.8083

Gamage, R., Wanigasooriya, P. (é. n.). CATMELK [Egyedi GPT]. OpenAI. Elérhető: https://chatgpt.com/g/g-YQm6gByUe-catmelk

Greenly, G. (é. n.). CatalogerGPT [Egyedi GPT]. OpenAI. Elérhető: https://chatgpt.com/g/g-8ymg1Ftwo-catalogergpt (Utolsó elérés: 2026. 02. 26.)

Library of Congress (2025) MARC 21 format for bibliographic data. Elérhető: https://www.loc.gov/marc/bibliographic/ (Utolsó elérés: 2026. 02. 26.)

Niveditha, B., Harinarayana, N. S., Balachandran, C. (2024) Error Analysis in ChatGPT's MARC21 Records: A Study of RDA Conformity. Journal of Information and Knowledge, 61(4), p. 187–195. https://doi.org/10.17821/srels/2024/v61i4/171481

OCLC (2025.). OCLC introduces new AI tools to make cataloging faster and smarter. Elérhető: https://www.oclc.org/en/news/releases/2025/20251208-ai-recordmanager-connexion.html (Utolsó elérés: 2026. 06. 08.)

Park, J., R., & Tosaka, Y. (2010) Metadata quality control in digital repositories and collections: Criteria, semantics, and mechanisms. Cataloging & Classification Quarterly, 48(8), 696–715. Elérhető: https://cci.drexel.edu/faculty/jpark/articles/Metadata%20Quality%20Control%20in%20Digital%20Repositories%20and%20Collections.pdf (Utolsó elérés: 2026. 02. 26.)

RDA-HU Munkacsoport (2026, január 12) Javaslat az RDA magyarországi bevezetésének első szakaszára: 1. Monografikus források (nyomtatott és elektronikus könyvek). Országos Széchényi Könyvtár Könyvtári Intézet. Elérhető: https://ki.oszk.hu/sites/default/files/dokumentumtar/javaslat_az_rda_magyarorszagi_bevezetesenek_elso_szakaszara_20260112.pdf (Utolsó elérés: 2026. 02. 26.)

Taniguchi, S. (2024) Creating and Evaluating MARC 21 Bibliographic Records Using ChatGPT. Cataloging & Classification Quarterly, 62(5), p. 527–546.

https://doi.org/10.1080/01639374.2024.2394513

Toolify.ai. (2026). CatalogerGPT – Features and capabilities.

Elérhető: https://www.toolify.ai/gpts/g-8ymg1Ftwo (Utolsó elérés: 2026. 02. 26.)

Zavalin, V., Zavalina, O. L. (2025). Are we there yet? Evaluation of AI-generated metadata for online information resources. Information Research an International Electronic Journal, 30(iConf), p. 732–740. https://doi.org/10.47989/ir30iConf47215. Elérhető: https://publicera.kb.se/ir/article/view/47215/37087 (Utolsó elérés: 2026. 02. 26.)

DOI: 10.31915/NWS.2026.7



# REPOZITÓRIUMOK INTEGRÁCIÓJA, METAADAT KONVERZIÓK AZ SZTE DISCOVERY RENDSZERÉBE

# INTEGRATION OF REPOSITORIES AND METADATA CONVERSIONS INTO THE SZTE DISCOVERY SYSTEM

*Zeller Rozália*
*SZTE Klebelsberg Könyvtár és Levéltár*
*rozalia.zeller@ek.szte.hu*

*Farkas Richárd*
*SZTE Informatikai Szolgáltatási Igazgatóság,*
*Könyvtár-informatikai és Adatgazdálkodási Egység*
*richard.farkas@ek.szte.hu*

## Absztrakt

Az SZTE Klebelsberg Könyvtár és Levéltár 2025 folyamán a Monguz Kft.-vel közösen fejlesztett egy discovery típusú integrált keresőrendszert, amely 2026 elején kerül használatba. A fejlesztés eredményeképpen az olvasók egyetlen online felületen tudják keresni, böngészni a könyvtár összes adatbázisának tartalmát.

A Discovery közös kereső alá egységes, MARC21 formátumú rekordok kellett, hogy kerüljenek. A legnagyobb kihívást az jelentette, hogy az egyes kiinduló adatforrások több különböző metaadat sémájú rekordokat tartalmaznak. A közös kereshetőség biztosítása érdekében meg kellett feleltetnünk az EPrints repozitóriumok sokféle egyedi sémáját, az Omeka Dublin Core rekordjait, valamint az Invenio és OJS rendszerekből exportálható MARC-rekordokat a saját Corvina katalógusunkban használt MARC21 szabványnak.

Tanulmányunk az adatkonszolidáció megtervezését, a mapping során felmerülő problémák áthidalását, illetve a metaadatok konverzióinak megvalósítását mutatja be.

**Kulcsszavak:** discovery, repozitórium, metaadat, Dublin Core, MARC21, OAI-PMH

## Abstract

During 2025, the SZTE Klebelsberg Library and Archives developed a discovery-type integrated search system in collaboration with Monguz Kft., which was put into use at the beginning of 2026. As a result of this development, readers will be able to search and browse the contents of all the library's different databases.

The Discovery integrated search engine required uniform records in MARC21 format. The biggest challenge was that the individual data sources contained records with multiple different metadata schemas. In order to ensure joint searchability, we had to match the many different unique schemas of the EPrints repositories, the Omeka Dublin Core records, and the MARC records exportable from the Invenio and OJS systems to the MARC21 standard used in our own Corvina catalog.

Our paper outlines how we planned the data consolidation, overcame the problems that came up during mapping, and carried out the metadata conversions.

**Keywords:** discovery, repository, metadata, Dublin Core, MARC21, OAI-PMH

## A projekt előzményei

A modern könyvtári és tudományos információs rendszerek számára az egyik legnagyobb kihívást a különböző típusú és szerkezetű adatforrások integrációja jelenti. Egyetemi környezetben különösen gyakori, hogy számos, egymástól jelentősen eltérő tartalmú és szerkezetű adatbázist kell egy keresőplatformon szolgáltatni. Egy ilyen platform célközönsége pedig legalább annyira sokszínű, különböző igényeket támasztó felhasználó, mint ahányféle adatforrás fut össze a közös keresőben.

A felhasználók és az őket információval ellátó könyvtárosok számára egyaránt fontos, hogy egy intézmény teljes portfóliója, így például a nyomtatott dokumentumok, elektronikus források, repozitóriumi tartalmak, kutatási adatok és kiadói rendszerek anyagai egyetlen platformon, egységesen kereshetőek legyenek.

Az SZTE Klebelsberg Könyvtár és Levéltár ennek érdekében olyan discovery típusú keresőrendszert alakított ki, amely a különböző adatforrásaiban található minden rekordot egy közös keresési infrastruktúrában kapcsol össze. A fejlesztés megvalósítására az RRF pályázat adott lehetőséget, melynek keretében a Monguz Kft.-vel együttműködve készült el – többek között – az SZTE Discovery platformja.[1]

Tanulmányunk bemutatja az integráció elméleti hátterét, az alkalmazott metaadat-megfeleltetési eljárásokat, valamint az egyes rendszerek – Corvina könyvtári katalógus; EPrints, Omeka Classic és Invenio repozitóriumok, illetve OJS és OMP kiadói rendszerek – esetében megvalósított gyakorlati fejlesztéseket. Az ismertetett megoldások középpontjában a MARC21 metaadat szabvány áll, amely közös nevezőként biztosítja a heterogén adatforrások egységes kezelését.

Az SZTE Discovery platform előzményeként fontos megemlítenünk a Contentas[2]-t, vagyis az SZTE Klebelsberg Könyvtár és Levéltár Contenta repozitóriumainak VuFind-alapú

1 https://discovery.bibl.u-szeged.hu/

2 https://contentas.ek.szte.hu/

közös keresőjét. A Discovery hátterében álló metaadat mapping módszerét a Contentas kereső fejlesztésekor dolgoztuk ki, amit kollégáink a 2020-as Networkshop konferencián be is mutattak.[3]

## Adatforrásaink

A Discovery rendszer kiinduló adatforrását a könyvtárunk Corvina katalógusa adja, ami gyakorlatilag két, egymástól független adatbázist jelent.

A jateDB a nyomtatott forrásaink leírásait tartalmazza, beleértve az SZTE Klebelsberg Könyvtár és Levéltár törzsállományát, valamint a megmaradt egyetemi hálózati könyvtárak rekordjait is. A dokumentumok formai és tartalmi feltárását könyvtárosaink alapos körültekintéssel, autopszia alapján készítik, így az ebben található rekordokat tekintettük etalonnak a metaadat mapping összeállításakor. A jateDB rekordjai a MARC21 szabvány szerint készülnek, ezért a többi adatforrásunk által használt egyéb metaadatsémát is a MARC21-nek feleltettük meg.

Az ejouDB elektronikus forrásokat integrál, az ebben található rekordok kisebb részben nyílt közgyűjteményi teljes szövegű adatbázisokból, nagyobb részben pedig az általunk előfizetett online adatbázisokból származnak. Az előbbi kategóriába tartozó dokumentumokról könyvtárosaink manuálisan készítenek részletesebb leírásokat, az utóbbi forrásból származó jóval felületesebb rekordokat pedig tömegesen, csomagokban töltjük be a katalógusba. Az adatállomány túlnyomó része az évente megújuló előfizetéseink nyomán folyamatosan változik, ugyanakkor itt is a MARC21 szabvány biztosítja az egységes leírást.

Az SZTE Discovery platformon a könyvtári katalógus mellett a Contenta repozitóriumcsalád kap helyet. Három különböző repozitóriumi szoftverből, összesen tizenegy különálló repozitóriumból integráltunk adatokat: kilenc EPrints, egy Omeka Classic és egy Invenio rendszerből.

3 Farkas, R., & Sándor, Á. (2020). Digitalizált tartalmaink közös keresője VuFind alapokon az SZTE Klebelsberg Könyvtárban. In Networkshop 2020. Országos Online Konferencia. 2020. szeptember 2–4. (pp. 22–32). http://doi.org/10.31915/NWS.2020.3


*1. ábra: Contenta repozitóriumok az SZTE Discovery platformon*

Minden repozitóriumunk más más gyűjtőkörrel rendelkezik.[4] Túlnyomó részük szöveges dokumentumokat tartalmaz PDF formátumban, de vannak kifejezetten multimédiás tartalmakat gyűjtő repozitóriumaink (SZTE Elektronikus Tananyag Archívum – EPrints, SZTE Képtár és Médiatéka – Omeka Classic), valamint egy kutatási adatokat tartalmazó archívumunk (SZTE Adatrepozitórium – Invenio) is.

A repozitóriumokban található metaadat sémákról elmondhatjuk, hogy legalább annyira különbözőek, mint a gyűjtőkörük. Az EPrints nem kötelez egy konkrét szabvány használatára, igény szerint testreszabhatók benne az egyes adatmezők. Az Omeka aktuális verziója

4 A Contenta repozitóriumokról bővebben a https://www.ek.szte.hu/kezdooldal/mit-keres/contenta-repozitoriumok/ oldalunkon írunk.

a kibővített Dublin Core (DCMI) sémát használja. Az Invenio alapbeállításban a kutatási adatok leírására készült DataCite sémát tartalmazza.

A könyvtári katalóguson és a Contenta repozitóriumokon túl két kiadói rendszerből, az SZTE OJS[5] (Open Journal System) és az SZTE EBook[6] (Open Monograph Press) platformokról is integráltunk rekordokat a Discovery közös keresőbe. OJS-ből az ott megjelenő folyóiratok cikkeinek analitikus rekordjai, OMP-ből pedig könyvek leírásai érkeznek a Discoverybe. Alapvetően mindkét kiadói rendszer saját metaadat struktúrával dolgozik, a feldolgozást pedig nem könyvtárosok, hanem a szolgáltatásunkat igénybe vevő szerkesztőségek munkatársai végzik.

## Metaadat megfeleltetések elméleti kérdései

A heterogén adatforrások integrációjának kulcskérdése a különböző metaadatsémák összehangolása. Az SZTE Discovery platform keresőmotorja számára szükséges közös nevező szerepét a MARC21 szabvány tölti be. A különböző rendszerekben jelen lévő saját metaadatsémák, a Dublin Core és a DataCite adatmezőit mind MARC21 mezőknek feleltettük meg.

A metaadat-konverziók előkészítéséhez online megosztott mapping táblázatot készítettünk, amelyben együtt dolgoztak a könyvtárosok, a könyvtár informatikusai, valamint a Discoveryt fejlesztő cég munkatársai. A táblázatban kerültek meghatározásra az egyes forrásrendszerek mezőinek MARC21 megfelelői, valamint a sémák közti eltérések mentén felmerülő problémák feloldására alkalmazott áthidaló megoldások.

Könyvtárunk mintegy húsz éve épít EPrints-alapú repozitóriumokat. A tekintélyes időtáv, valamint az EPrints kötetlenül testreszabható adatmezőinek következtében ezeket a repozitóriumokat volt a legnehezebb közös nevezőre hoznunk. Az egyes repozitóriumokban sokféle, az eltérő gyűjtőkörökhöz igazított adatmezőket használunk, amelyeket sok esetben nehéz volt a megfelelő MARC21 mezőre lefordítani.

Csak hogy néhány kiragadott példát említsünk: az EPrints repozitóriumokban legalább 4 különböző mezőt (pages, pagerange, page_range, page_notes) használunk egy dokumentum terjedelmének meghatározására, ezeket mind a MARC 300 $a almező hivatott megjeleníteni. Ennek épp az ellenkezője áll fenn szerzők tekintetében: az EPrints repozitóriumokban a szerzőket és azok szerzői minőségét (szerző, szerkesztő, fordító stb.) egyetlen összetett mező (creators) tartalmazza, ezt kellett szétbontani MARC 100 $a, 245 $c, 700 $a mezőkre.

A marc rekordok fejlécének meghatározásakor sem volt mindig egyértelmű dolgunk. A MARC leader 06 és 07 pozíciók értékével jóval kevesebb rekordtípus fejezhető ki, mint

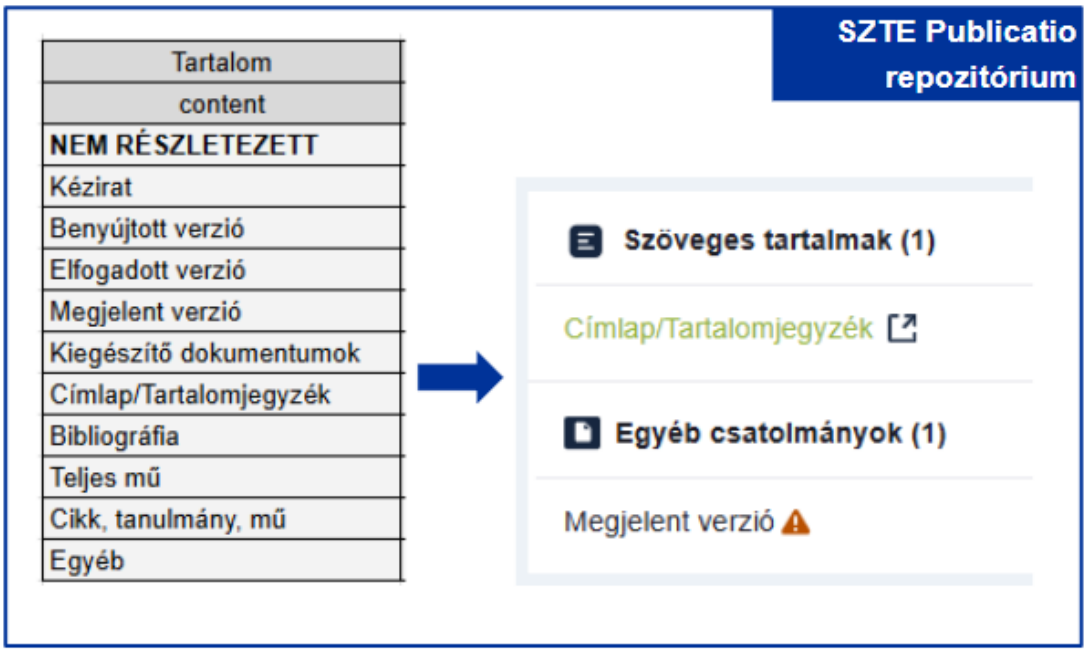
5 https://www.ek.szte.hu/kezdooldal/kutatastamogatas/ojs-folyoirat-platform/

6 https://www.ek.szte.hu/kezdooldal/kutatastamogatas/ebook-platform/, https://ebook.ek.szte.hu/

amennyi az EPrints repozitóriumokban előfordul. A legszemléletesebb példa erre a diploma és doktori dolgozatok típusának kérdése: az SZTE Diploma repozitórium szigorúan korlátozott hozzáférésű szakdolgozatokat tartalmaz (EPrints típus: thesis), az SZTE Doktori repozitórium pedig nyílt hozzáférésű PhD disszertációkat (EPrints típus: thesis_diss). Mindkét típus "ta" fejléccel fordul át MARC-fejlécbe, amely alapján nem különíthetők el megfelelően sem keresőmezőben, sem facettában a Discovery platformon. A probléma feloldására be kellett iktatnunk egy extra lépést, melynek keretében MARC 516 $a almezőben szövegesen megadjuk, hogy az adott rekord diploma vagy doktori dolgozatot ír le.

A repozitóriumi rekordok metaadatainak integrációja mellett külön kihívást jelentett a feltöltött fájlok metaadatainak átvitele a közös keresőbe. A felhasználók szempontjából különösen fontos, hogy már a Discovery felületén megfelelően jelenjen meg egy adott fájl tartalmának leírása, valamint a hozzáférés szintje. A leírás (EPrints content mező) tájékoztatja a felhasználót, hogy milyen típusú dokumentum van a link mögött, amiből következtethet arra, hogy valóban a keresett teljes szöveget fogja elérni, vagy csak valamilyen járulékos dokumentumot. A hozzáférés szintje pedig azt határozza meg, hogy meg tudja-e nyitni az adott fájlt a Discovery saját megjelenítőjében (amennyiben a hozzáférés szintje nyílt), vagy át lesz irányítva az adott repozitóriumba további autentikáció szükségessége miatt (amennyiben a hozzáférés szintje valamilyen módon korlátozott).


*2. ábra: Az SZTE Doktori repozitóriumi fájlok megjelenítése az SZTE Discovery platformon*




*3. ábra: Az SZTE Publicatio repozitóriumi fájlok megjelenítése az SZTE Discovery platformon*

Az SZTE Képtár és Médiatéka rekordjainak konverziójához a Library of Congress által készített Dublin Core–MARC21 megfeleltetést[7] vettük alapul. A Dublin Core séma mezőinek flexibilitása miatt azonban ezt a megfeleltetést is testre kellett szabnunk, hogy megfeleljen a saját feldolgozási hagyományainknak.

Az SZTE Adatrepozitórium esetében az Invenio saját MARC exporteréből indultunk ki, amely szintén némi módosításra szorult. Egyrészt kissé elavult MARC-verziót használ, másrészt esetenként téves mező-megfeleltetéseket alkalmaz. Például eredetileg a DataCite licensz mezőjét MARC 650 $a mezőre fordította le, ezt javítanunk kellett 540 $a mezőre.

Az OJS és OMP platformok szintén rendelkeznek saját MARC-exporterrel, ebbe azonban jelentősen bele kellett nyúlnunk ahhoz, hogy a nekünk megfelelő MARC-mezőkbe érkezzenek az adatok. A problémák az elavult MARC-verzión felül a kiadói rendszerek sajátosságaiból fakadtak. Többek között: egy kiadói rendszerben fontos követni egy cikk életútját, ami többféle dátum (benyújtás dátuma, revíziók dátumai, elfogadás dátuma, online megjelenés dátuma, stb.) rögzítésével lehetséges. A Discovery rekordokban viszont egységesen egyféle dátumot, a kiadás évét jelenítjük meg.

7 Dublin Core to MARC Crosswalk, 2008-04-23: https://www.loc.gov/marc/dccross.html

## Metaadat konverziók a gyakorlatban

### EPrints

A projekt gyakorlati részének egyik legfontosabb eleme az EPrints repozitóriumok MARC-alapú exportjának kialakítása és egységesítése volt. Az export alapját egy EPrintsMARC nevű plugin[8] adta, amely eredetileg parancssoros eszközként szolgált rekordok importálására és exportálására. Ezt a kiegészítőt fejlesztettük tovább annak érdekében, hogy az OAI-kimeneten keresztül is elérhetővé váljanak a rekordok, így a Discovery rendszerünk automatikusan aratni tudja az adatokat. Bár az eredeti fejlesztő már nem frissíti a plugint, az mégis stabil alapot biztosított ahhoz, hogy a saját igényeinkhez igazítsuk.

A munka egyik legnagyobb kihívását az jelentette, hogy összesen 9 különböző EPrints szervert üzemeltetünk, amelyek eltérő feladatokat látnak el. Emiatt olyan adatkonverziós megoldások kialakítására volt szükség, amelyek biztosítják, hogy ugyanaz az exporteszköz minden szerverhez kompatibilisen működjön.

Ennek során több metaadatmezőt is egységesíteni kellett. Például a Diploma repozitóriumban módosítani kellett a karokat tartalmazó mező típusát, hogy az megfeleljen a többi repozitórium struktúrájának. Az adatmennyiség miatt ez a folyamat egy speciálisan erre írt Perl script segítségével több, mint 10 órán keresztül futott. Külön figyelmet kellett fordítani arra, hogy a szerver terhelése ne befolyásolja a napi működést, mivel a rendszereket egyszerre használják dolgozók, hallgatók és külső szolgáltatások is.

Az utóbbi időszakban jelentősen megnőtt az AI-alapú botok által indított aratások száma is, ami további terhelést jelent az EPrints szerverek számára.

Bizonyos rendszereknél kompromisszumokat kellett kötni. A Publicatio repozitóriumban például az MTMT-importok kompatibilitása miatt néhány mezőt – például a dokumentumban előforduló nyelveket tartalmazó language mezőt – szabad szöveges típusként kellett meghagyni, miközben más repozitóriumokban már előre definiált értéklistás mezőket használunk.

Jelenleg is vannak még folyamatban lévő feladatok, például a fájlok MIME-típusainak egységesítése és a hiányzó nyelvi adatok automatikus kitöltése. Az EPrints a MIME-típusokat alapértelmezetten automatikusan állapítja meg a fájl kiterjesztése alapján, majd szükség esetén a karakterkódolási információkat is eltárolja. Ez azonban problémát okozhat a MARC 857 $q mező exportjánál, mert olyan többletinformációk kerülhetnek át a Discovery rendszerbe, amelyek hibás adatfeldolgozást eredményeznek. Ezért olyan, hosszú távon is fenntartható megoldást keresünk, amely a jövőbeli frissítésekkel is kompatibilis marad.

8 https://github.com/eprintsug/EPrintsMARC

```
85740 $cSZTE Publicatio Repozitórium $uhttp://publicatio.bibl.u-szeged.hu/37672/7/36323927.pdf
      $qapplication/pdf $zNyilvános $3Megjelent verzió $70
85740 $cSZTE Publicatio Repozitórium $uhttp://publicatio.bibl.u-szeged.hu/37672/1/Nagy.pdf
      $qapplication/pdf; charset=binary $zNyilvános $3Benyújtott verzió $70
```

*4. ábra: MIME-type értékek a 857 $q almezőben*

A nyelvfelismerés automatizálására a Python-alapú langdetect könyvtárat kezdtük használni. Az eszköz előnye, hogy néhány soros kóddal gyorsan képes meghatározni egy dokumentum nyelvét, és a tesztek során nem tapasztaltunk hamis pozitív találatokat. Ugyanakkor nagy méretű, több száz megabájtos PDF-ek esetében a feldolgozás jelentősen lelassult, illetve bizonyos dokumentumokat nem tudott megfelelően feldolgozni.

Ezért egy hibrid megoldást alakítottunk ki: kisebb fájlok esetében a teljes dokumentumot elemezzük, nagyobb állományoknál pedig csak a dokumentum közepéből olvasunk be egy nagyobb szövegrészletet. Amennyiben a program nem tudott nyelvet meghatározni, a rekordhoz kapcsolódó egyéb metaadatok alapján töltöttük ki a hiányzó információt. A folyamat jelenleg is fejlesztés alatt áll, mivel az egyes szerverek eltérő finomhangolást és időzítést igényelnek.

**Omeka Classic**

A Mediatéka repozitórium esetében a fejlesztési feladat középpontjában az OAI-PMH szolgáltatás bővítése állt. Az Omeka rendszer alapértelmezetten nem biztosított MARC21 formátumú exportot, ezért az OaiPmhRepository plugin[9] működését kellett kiegészíteni PHP nyelven úgy, hogy a rekordok MARC-kompatibilis kimenetben is elérhetővé váljanak.

Az Omeka rendszerben a metaadatok tárolására a DublinCoreExtended plugint[10] használjuk. A fejlesztés során ezeknek a mezőknek az értékeit kellett megfeleltetni a MARC21 szabvány megfelelő mezőinek és almezőinek, hogy az exportált rekordok könyvtári rendszerek és discovery szolgáltatások számára is feldolgozhatóak legyenek.

A gyakorlati megvalósítás során több problémát is kezelni kellett. Az egyik leggyakoribb nehézséget az absztraktok és leírások szöveges mezői jelentették, mivel ezekben gyakran előfordultak sortörések, HTML-elemek és speciális karakterek. Ezek tisztítása és normalizálása szükséges volt ahhoz, hogy a MARC-rekordok szabványos és hibamentes formában kerüljenek exportálásra.

Szintén fontos feladat volt a dátum mezők egységesítése. Az egyes rekordokban eltérő formátumú dátumok szerepeltek, ezért olyan átalakítási logikát kellett kialakítani, amely

9 https://omeka.org/classic/plugins/OaiPmhRepository/

10 https://github.com/omeka/plugin-DublinCoreExtended

egységes, szabványosított formában tudta előállítani a MARC21-fejléc és -mezők számára megfelelő dátumértékeket.

A fejlesztés során kiemelt szempont volt, hogy a létrehozott megoldás hosszú távon is karbantartható maradjon, valamint kompatibilis legyen az Omeka Classic és a hozzá kapcsolódó pluginek későbbi frissítéseivel is.

### Open Journal System

Végezetül az OJS rendszerek esetében már egy jóval egyszerűbb feladatot kellett megoldani. Az OJS alapvetően rendelkezett megfelelő OAI-kimenettel, ezért itt nem teljesen új exportfunkció fejlesztésére volt szükség, hanem a meglévő kimenet testreszabására.

A megoldás során az OJS oaiMetadataFormats pluginjét használtuk fel, amely lehetőséget biztosít különböző metaadatformátumok kezelésére és megjelenítésére. A szükséges módosításokat egy template fájlban kellett elvégezni, amely egy .tpl kiterjesztésű sablonfájl. Ez a formátum elsősorban webalkalmazásokban, dinamikus tartalom előállítására használatos.

## Konklúzió, további fejlesztési tervek

Tanulmányunkban az SZTE Klebelsberg Könyvtár és Levéltár Discovery platformja alatt összefutó különféle adatforrásokat mutattuk be, különös tekintettel azok metaadat struktúráinak integrációjára. A keresőrendszer a Corvina katalógus, az EPrints, Omeka Classic és Invenio alapú repozitóriumok, valamint az OJS és OMP kiadói platformok rekordjait kapcsolja össze egy közös infrastruktúrában. Az integráció alapját a MARC21 metaadatszabvány adja, amelynek minden egyéb előforduló sémát megfeleltettünk.

A projekt egyik legnagyobb kihívását az EPrints repozitóriumok egységesítése jelentette. Az eltérő mezőstruktúrák miatt számos konverziós és adatnormalizálási megoldást kellett eszközölnünk, különösen a szerzői adatok, dokumentumtípusok és fájlmetaadatok kezelése terén. Az EPrintsMARC plugin továbbfejlesztésével lehetővé vált a MARC-alapú OAI-PMH export, így a Discovery rendszer automatikusan képes a rekordok aratására. Az Omeka Classic esetében a Dublin Core-mezők MARC21-kompatibilis exportját kellett megvalósítani, míg az OJS és OMP rendszereknél a meglévő exportfunkciók testreszabása történt meg. A fejlesztések során fontos szempont volt a hosszú távú karbantarthatóság és a későbbi frissítésekkel való kompatibilitás biztosítása.

Kitértünk a jelenleg is fejlesztés alatt álló részprojektekre is, így például a MIME-típusok egységesítésére, a hiányzó nyelvi adatok automatikus felismerésére és a szerverterhelés optimalizálására. A Discovery platform további adatgazdagítása céljából folyamatban van továbbá a mindenkori előfizetett elektronikus tartalmaink integrációja a Summon API segítségével, erről Hoczopán Szabolcs ír részletesen ebben a kötetben. A projekt hosszú távú célja egy stabil, fenntartható és egységes keresési infrastruktúra szolgáltatása az egyetemi tudásvagyon teljes körű feltárhatóságának biztosítása érdekében.

DOI: 10.31915/NWS.2026.8



# VibeARP: ADATGAZDÁSZ ASSZISZTENS AZ ARP-BAN

*Pataki Balázs*
*HUN-REN Számítástechnikai és Automatizálási Kutatóintézet*
*pataki@sztaki.hu*

## Absztrakt

A HUN-REN Adatrepozitórium Platform (ARP) használata egyszerre jelent kihívást a kezdő és a tapasztalt felhasználók számára: előbbieket a rendszer komplexitása, utóbbiakat az ismétlődő, időigényes adatkezelési feladatok terhelik. A VibeARP egy mesterséges intelligenciára épülő adatgazdász asszisztens, amely ezt a kettős problémát célozza meg azzal, hogy automatizálja a rutin feladatokat, miközben jelentősen lerövidíti a tanulási görbét az új felhasználók számára.

Bemutatjuk a VibeARP "klasszikus" megvalósítását, ami saját ágens ciklussal, eszközökkel és felhasználói felülettel készült, valamit a VibeARP "CAOS" (Coding Agent as Operating System) verzióját, ami a felhasználók gépére telepített meglévő kódoló ágensek, mint Codex vagy Claude Code teszi lehetővé ugyanazt a működést, de lokális módon.

**Kulcsszavak:** HUN-REN Adatrepozitórium Platform, ARP, Kutatástámogatás, Adatgazdászat, MI-ágens, Szoftver automatizáció

## Abstract

Using the HUN-REN Data Repository Platform (ARP) poses challenges for both novice and experienced users: newcomers are burdened by the system's complexity, while advanced users are hindered by repetitive, time-consuming data management tasks. VibeARP is an AI-based data steward assistant that addresses this dual problem by automating routine tasks while significantly shortening the learning curve for new users.

We present the "classic" implementation of VibeARP, built with its own agent loop, tools, and user interface, as well as the VibeARP CAOS (Coding Agent as Operating System) version, which enables the same functionality locally by using existing coding agents installed on users' machines, such as Codex or Claude Code.

**Keywords:** HUN-REN Data Repository Platform, ARP, Research Support, Data Stewardship, AI Agent, Software Automation

## Bevezetés

A HUN-REN Adatrepozitórium Platform (ARP) [1] egységes kutatási adatkezelési infrastruktúrát biztosít a Magyar Kutatási Hálózat (HUN-REN) kutatói számára. A hálózat sokszínűsége – 42 kutatóintézet, több mint száz egyetemi kutatócsoport és a tudományterületek teljes spektruma – összetett, több komponensből álló rendszer kialakítását tette szükségessé. Az ARP központi eleme egy Dataverse-alapú adatrepozitórium, amelyet CEDAR-alapú metaadatséma-regiszter, az RO-Crate [2] szabvány teljes körű támogatása, az AROMA metaadatszerkesztő eszköz, valamint egy országos tudásgráfra épülő egységes keresőrendszer egészít ki.

Bár ezek az eszközök - Dataverse, CEDAR, AROMA [3] - önmagukban jól használhatók, az ökoszisztéma egésze elsőre nehezen áttekinthető lehet. Emellett a tapasztalt felhasználóknak is jelentős manuális munkát igényel a megfelelő metaadatok összegyűjtése különféle forrásokból, például fájlokból, dokumentációból, jegyzetekből vagy webes tartalmakból.

## VibeARP

A VibeARP [4] az ARP automatizált használatára kínál megoldást egy chat-alapú felhasználói felületen keresztül működő MI-ágens segítségével. A felhasználók természetes nyelven írhatják le a létrehozni kívánt adatkészletet, miközben az ágens különféle eszközöket használ: webes keresést végez, releváns oldalakról információt gyűjt, fájlokból metaadatokat nyer ki, majd ezeket a CEDAR-alapú sémákhoz illesztve automatikusan RO-Crate metaadatstruktúrává alakítja. A metaadatok folyamatosan megjelennek egy beágyazott, leegyszerűsített AROMA-szerkesztőben is, amely a manuális finomhangolást is lehetővé teszi.

A rendszer egy integrált böngészőágenst is tartalmaz, amely interaktív webes műveletek végrehajtására képes egy MI által vezérelt, tényleges Chrome böngésző használatával. Ennek segítségével például a felhasználó bejelentkezhet a Dataverse-be, majd a bejelentkezés után a böngészőágens hozzáférhet ahhoz az azonosítási tokenhez, amelynek segítségével az elkészült RO-Crate csomag feltölthető a Dataverse-be, létrehozva a kívánt adatcsomagot.

A VibeARP általános feladata, hogy az adatgazdásszal beszélgetve, az ő instrukciói alapján összegyűjtse a szükséges metaadatokat (direkt inputból, webes kereséssel, vagy fájlokból kinyerve), majd azokat az adott RO-Crate-hez megadott metaadat sémákhoz és profilokhoz illessze, végül ezek alapján a végleges RO-Crate JSON-t előállítsa.

A konkrét lépések, amiket az ágens végez, egy előre definiált munkafolyamatot alkotnak:

1. **Profil felfedezés.** A munkamenet első lépése az aktív profilok és a profil megkötések megismerése. Ez a lépés biztosítja, hogy az ágens ne spekuláljon a metaadatstruktúráról, hanem a tényleges profil megkötések alapján dolgozzon.

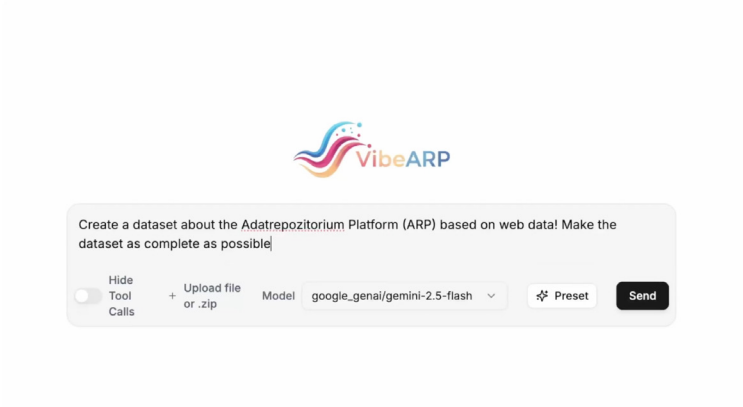
2. **Mezőtervezés.** A profil szabályok alapján az ágens osztályozza a mezőket: kötelező, ajánlott, opcionális, és nem engedélyezett mezőkre. Létrehoz egy explicit tervet, amely meghatározza, milyen értékek hiányoznak még, és milyen sorrendben lesznek kitöltve. A profil illesztési szabályzat előírja, hogy kötelező mezőket először, ajánlott opcionális mezőket másodszor, majd a többi engedélyezett mezőt utoljára kell kezelni.
3. **Adatgyűjtés.** Ha metaadat értékek ismeretlenek és nem vezethetők le lokálisan, az ágens bekérheti az adatokat az adatgazdásztól, vagy webes keresést végezhet, szükség esetén a talált oldalakat is letöltve és elemezve. Az ágens az elsődleges forrásokat részesíti előnyben, mint hivatalos intézményi oldalak, kanonikus dokumentáció, szabványok, stb.
4. **Entitás létrehozása**. Miután minden az aktuális profilnak megfelelő adatot összegyűjtött az ágens, ezekből létrehozza a profilnak megfelelő mezőadatokat és RO-Crate entitásokat.
5. **RO-Crate JSON előállítás.** Az LLM-ek önmagukban is képesek helyes RO-Crate JSON-t létrehozni, ha az adatok már rendelkezésre állnak, azonban ez meglehetősen token-igényes művelet, így ez jelentősen meg tudja drágítani az ágens működését. Ehelyett speciális RO-Crate szerkesztő eszközt adunk az ágens kezébe, amivel részletekben tudja a JSON-t szerkeszteni: hozzáadni, módosítani, törölni. Ezek a módosítási műveleteket egy speciális eszköz (tool) segítségével kerülnek végrehajtásra, és ezzel épül és változik az RO-Crate JSON.
6. **Ellenőrzés.** Mivel az LLM-ek működése nem determinisztikus, akármilyen pontosan adjuk is meg a szabályokat, az elvégzett műveletek után folyamatos ellenőrzésre és visszajelzésre van szükség. A validálás az RO-Crate profilok és az RO-Crate alapját képező JSON-LD szabványnak való megfelelés alapján történik. Ha a validálás során hibát tapasztalunk, arról tudatjuk az ágenst, amely azt egy következő körben képes javítani.

A VibeARP implementációjához a LangGraph [5] ágens keretrendszert használtuk. Ebben valósítottuk meg a saját ágens ciklusunkat, valamint a működéshez szükséges eszközöket (toolokat).

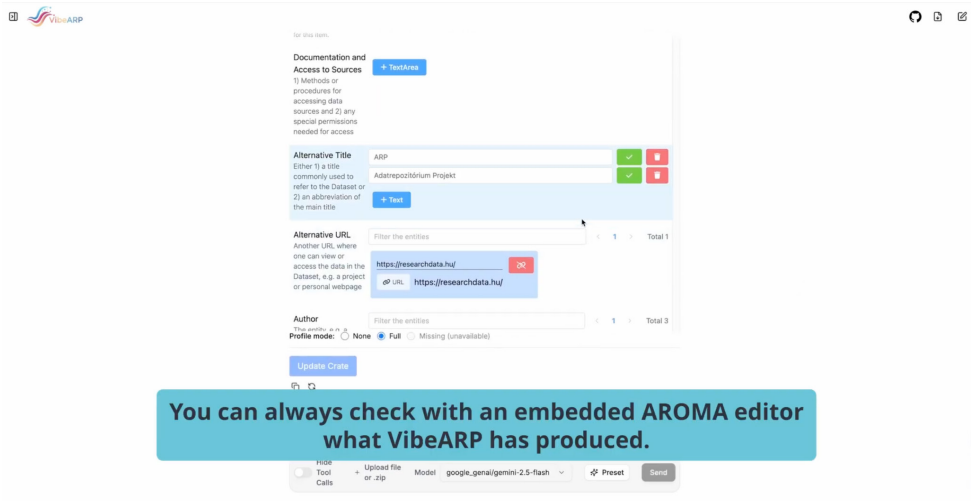
A felhasználói felületet a LangChain Agent chat UI" [6] megoldására alapoztuk. Ez egy React-alapú chatalkalmazás, amely saját komponensekkel bővíthető és a nyílt „Agent Protocol"-on [7] keresztül képes kommunikálni a LangGraphban futó ágenssel.

A VibeARP böngésző-automatizációs megoldása a „VibeARP Browser AI". Ez egy Browser Use [8] -alapú ágens, amihez egy olyan webes megoldást készítettünk, amely a browser-use által irányított Chrome böngészőt integráltan jeleníti meg a chatfelületen. Az eredmény egy böngészőben futó, interaktív böngészőélmény. Ezt jellemzően a „VibeARP Browser AI"

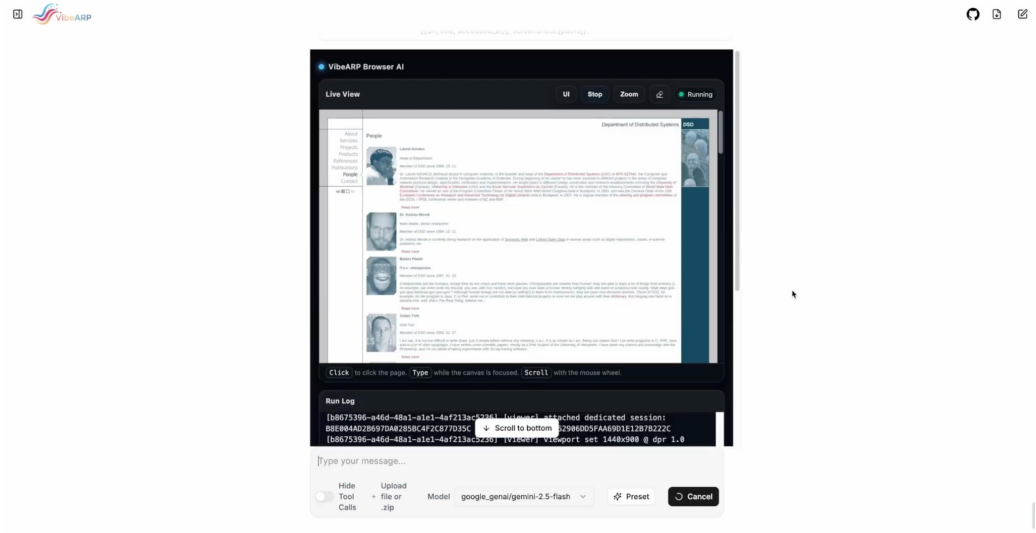
irányítja, szükség esetén azonban az irányítás átadható a felhasználónak is, például jelszó megadásához.




*1. ábra: A VIbeARP chat felülete*




*2. ábra: A VIbeARP-ba integrált AROMA RO-Crate szerkesztő az RO-Crate metaadatok megjelenítésére és szerkesztésére*




*3. ábra: A VIbeARP Browser AI-ágens egy beépülő böngésző, amin az ágens interaktív weboldalakkal tud feladatokat elvégezni, pl. szerzői adatokat kigyűjteni weboldalból*

## VibeARP CAOS

A VibeARP CAOS VibeARP koncepció továbbfejlesztése, amely alapvetően megváltoztatja az implementációs megközelítést: ahelyett, hogy saját MI-ágens ciklust építenénk, a szabványos Model Context Protocolra (MCP) és a kereskedelmi forgalomban kapható kódoló ágensekre támaszkodunk. A névben a CAOS is erre utal: *Coding Agent as Operating System*. Az elnevezést az indokolja, hogy az AI kódoló ágensek, mint Codex, Calude code, Gemini, Kilo stb., valójában már nem csak kódolási feladatokra alkalmasak, hanem általános ágens keretrendszerként is értelmezhetők. Ebből a szempontból egy operációs rendszerre hasonlítanak, ami az alap intelligencia működést biztosítja (LLM-mel való kommunikáció, alap shell-alapú eszközhasználat stb.), és amelyekre különböző módokon saját applikációkat telepíthetünk.

A kódoló ágenseket az alábbi módokon bővíthetjük, vagy szabhatjuk testre:

1. **Prompt-alapú bővítés (CLAUDE.md, AGENTS.md)**

   A legegyszerűbb bővítési forma a projektkönyvtárban elhelyezett Markdown-fájlok, amelyeket a kódoló ágens automatikusan beolvas és a kontextusába injektál. Claude Code a CLAUDE.md, az OpenAI Codex CLI az AGENTS.md stb. A tartalom tetszőleges szöveg lehet: kódolási konvenciók, architekturális döntések, projektspecifikus iránymutatások. Például egy CLAUDE.md tartalmazhatja, hogy „Mindig TypeScriptet használj, sosem JavaScriptet”, vagy „Az API-végpontok a /api/v2/ prefixszel kezdődnek”.

Előnyök: csak egy szövegfájl, a legegyszerűbb módja annak, hogy befolyásoljuk az ágens viselkedését.

Hátrányok: nem ad hozzá új képességeket az ágenshez, csak azt befolyásolja, hogyan viselkedjen, nem azt, mit tudjon. A metaadat-szerkesztés kontextusában például egy CLAUDE.md megmondhatja az ágensnek, hogy „ne szerkeszd közvetlenül az ro-crate-metadata.json fájlt", de nem ad neki eszközt arra, hogy validálja a csomagot vagy feltöltse a Dataverse-be. Emellett fogyasztja a kontextusablakot, és a modell figyelmen kívül is hagyhatja az utasításokat.

2. **Skillek (Skills / Slash Commands)**

   A skill egy újrafelhasználható, feladat-specifikus ágenscsomag, amely instrukciókat, munkafolyamatot, kontextust, referenciákat és opcionálisan futtatható segédprogramokat tartalmaz. Az instrukciókat egyszerű szöveges Markdown-fájlokként lehet megadni, és mellé megadott módon lehet csatolni azokat a segédprogramokat, amiket a skill instrukciói alapján használatba tud venni az ágens. Például, ha van egy képzeletbeli extract_metadata_from_pdf programunk, és ennek működését leírjuk egy skillben, onnantól, ha a felhasználó olyan feladatot akar végezni, amihez egy PDF-ből kell metaadatokat kinyerni, akkor az ágens tudni fogja, hogy ezt hogyan kell csinálnia, és miképpen kell paramétereznie és futtatnia az extract_metadata_from_pdf programot.

   Előnyök: az ágens működése az AGENT.md-nél sokkal jobban testreszabható, különösen a csatolt programok felhasználásával

   Hátrányok: bonyolultabb karbantartani, mert külön skill-struktúrát, leírást, verziózást és esetenként futtatható segédprogramot igényel

3. **Model Context Protocol (MCP)**

   A Model Context Protocol [9] egy 2024 vége felé az Anthropic által publikált nyílt szabvány, amely JSON-RPC 2.0-ra épül. Kliens-szerver architektúrát definiál, ahol egy MCP gazdaalkalmazás (host), például Claude Code, Codex stb. MCP szerverekhez csatlakozik. A szerverek három dolgot nyújthatnak:

   Eszközök (Tools) – hívható függvények, amelyeket a modell meghívhat. Például: „validáld ezt az RO-Crate-et", „keress rá erre a kifejezésre a weben".

   Erőforrások (Resources) – olvasható adatforrások, amelyeket a modell lekérhet. Például: „add vissza a jelenlegi séma-regiszter tartalmát".

   Promptok (Prompts) – újrafelhasználható prompt-sablonok, amelyeket a szerver kínál.

   Előnyök: nyílt szabvány, iparági szintű hordozhatóság. Strukturált, típusos eszköz-interfész (JSON Schema-alapú bemeneti validáció). Új, végrehajtható képességeket

ad az ágensnek. Minimális kontextusköltség (csak igény szerint, eszközhíváskor). A szerver állapotot tarthat fenn (profil-kontextus cache, séma-regiszter).

Hátrányok: szerverfolyamat futtatását igényli (operációs terhelés). A gazdaalkalmazásnak támogatnia kell az MCP-t.

Bár implementáció szempontjából az MCP igényli a legtöbb munkát, a VibeARP CAOS-hoz mégis ezt a megoldást alkalmaztuk, mert ez tűnik végső soron a legkompatibilisebbnek és legflexibilisebbnek. A VibeARP MCP a már előzőleg ismertetett munkafolyamatot megvalósító eszközöket implementálja és teszi az MCP protokollon keresztül elérhetővé a kódoló ágenseknek.

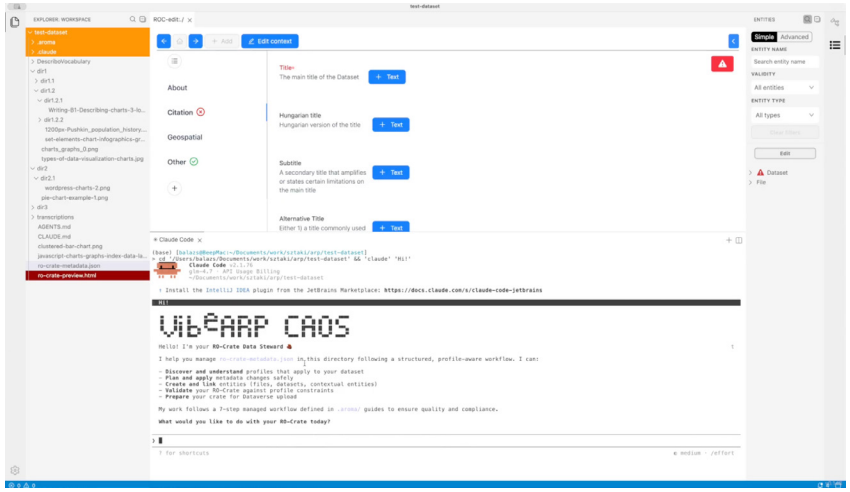
A VibeARP CAOS egy nagyobb projekt, az AROMA-2 keretében került implementálásra. Az AROMA első változata az ARP-ben a Dataverse-ben lévő adatcsomagok RO-Crate szerinti szerkesztését teszi lehetővé, szorosan integrálódva a Dataverse-szel.

Az AROMA-2 célja egy olyan RO-Crate szerkesztői "IDE" létrehozása, amivel a kutatók a saját fájlrendszerükben vagy egyéb forráshelyeken elérhető fájljaikat tudják az RO-Crate szerint metaadatolni, majd a végén a kész RO-Crate csomagot a kiválasztott repozitóriumba feltölteni. Ehhez egy Theia keretrendszeren alapuló új AROMA-megoldást hoztunk létre és a VibeARP CAOS ennek az egyik eleme, egyben ez adja azt az RO-Crate grafikus felületet, amit a klasszikus VibeARP-ban az AROMA beépülő verziója adott. Vagyis a VibeARP CAOS által végzett változtatásokat az AROMA-2-ben követhetik nyomon a felhasználók, és magát a VibeARP CAOS-t is onnan tudják indítani.

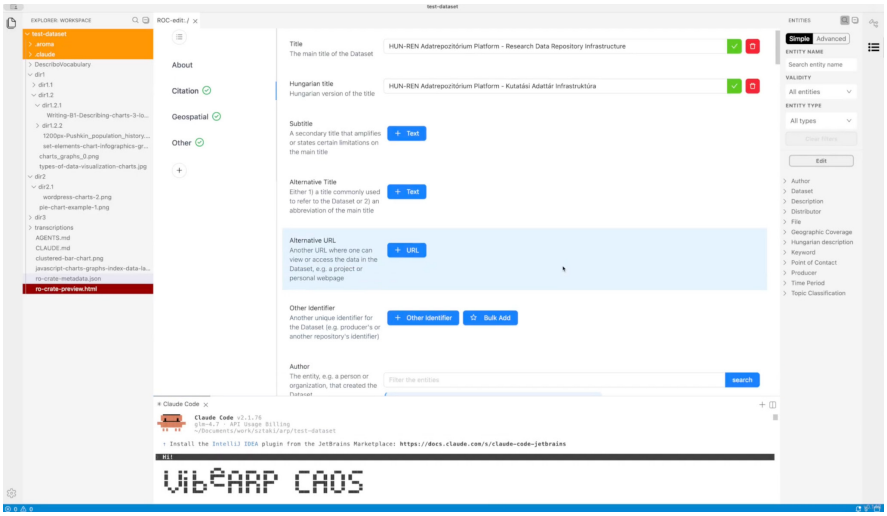
A VibeARP CAOS ágensek kétféleképpen futtathatók: vagy terminálban, a megszokott CLI eszközzel vagy beépülve az AROMA-2-be. Az első megoldás ugyanazt a felhasználói élményt nyújtja, mintha a felhasználó maga indítaná el terminálból a Claude-ot vagy Codexet: ugyanazt a terminál UI-t használhatja, amit már megszokott, és amit esetleg már testre is szabott. A AROMA-2-be beépülő megoldásnál az AROMA-2 adja a chat felületet, de a háttérben ugyanúgy a kiválasztott ágenssel (jelenleg Claude vagy Codex) kommunikál.

A *CAOS*-módszer alkalmazása azért tűnik előremutatónak, mert jó pár problémát megold a VibeARP első megoldásához képest:

- nem kell az ágens ciklussal és az alacsony szintű LLM-kommunikációval és eszköz használattal nekünk foglalkozni,
- a kódoló ágensek önmaguktól is sok hasznos képességgel rendelkeznek, amiket nem nekünk kell megvalósítani,
- rendelkeznek olyan beépített korlátokkal, amiket a VibeARPban még meg sem valósítottunk,
- a felhasználó a saját előfizetésével tudja használni, így a költségeket is ő állja, nem pedig a VibeARP szolgáltatásnak kell központilag finanszíroznia.




*4. ábra: A Claude Code, mint a VibeARP CAOS megvalósítása az AROMA-2-ben egy terminál ablakban nyílik meg az RO-Crate szerkesztőfelület alatt*




*5. ábra: A VibeARP CAOS eredményét az RO-Crate szerkesztőfelületén lehet megnézni és szerkeszteni*

## Összefoglalás

A VibeARP és VibeARP CAOS eszközök jelenleg kísérleti-fejlesztési fázisban vannak és azon dolgozunk adatgazdászokkal együttműködve, hogy olyan eszköz készüljön, ami az ARP-ben hatékonyabbá teheti az adatgazdászi munkát, levéve a repetitív és unalmas munkák terhét róluk. A következő lépésként elkezdjük az eszközt adatgazdászok közreműködésével tesztelni és a visszajelzéseik alapján elkészíteni a VibeARP végleges változatát, amit elérhetővé teszünk az APR felhasználói számára.

## Bibliográfia


**Kovács, L. Az ELKH Adatrepozitórium Platform projekt a kutatási adattárolás és megosztás országos föderált rendszere** 2023 MAGYAR TUDOMÁNY, 184(7)., pp. 839-851. ISSN 0025-0325

**Research Object Crate. Research Object Crate (RO-Crate).** https://www.researchobject.org/ro-crate/

**Pataki, B. Dataverse, CEDAR and RO-Crate: The building blocks of ARP.** YouTube videó. https://www.youtube.com/watch?v=0_ENdITtIQg

**Pataki, B. VibeARP: An experimental AI agent for the Adatrepozitórium Platform.** YouTube videó. https://www.youtube.com/watch?v=ZMvsET-5S5Y

**LangChain. LangGraph**: Agent Orchestration Framework for Reliable AI Agents. https://www.langchain.com/langgraph

**LangChain AI. agent-chat-ui**: Web app for interacting with any LangGraph agent via a chat interface. GitHub-repozitórium.: https://github.com/langchain-ai/agent-chat-ui

**LangChain AI. agent-protocol**. GitHub-repozitórium. https://github.com/langchain-ai/agent-protocol

**Browser Use.** browser-use: Make websites accessible for AI agents. GitHub-repozitórium. https://github.com/browser-use/browser-use

**Model Context Protocol.** https://modelcontextprotocol.io/

DOI: 10.31915/NWS.2026.9



# HORDOZHATÓ KURZUSTARTALMAK GENERATÍV MESTERSÉGES INTELLIGENCIÁVAL TÁMOGATOTT ELŐÁLLÍTÁSA

# PRODUCTION OF PORTABLE COURSE CONTENT SUPPORTED BY GENERATIVE ARTIFICIAL INTELLIGENCE

*Ujhelyi Gábor*
*Eszterházy Károly Katolikus Egyetem*
*ug@mensa.hu*
*ORCID: 0009-0003-3693-100X*

**Absztrakt**

A generatív mesterséges intelligencia gyors fejlődése az online kurzusfejlesztésben is új lehetőségeket teremt. A tanulmány azt vizsgálja, hogyan állíthatók elő alacsony belépési küszöb mellett olyan hordozható, multimodális kurzustartalmak, amelyek nem kötődnek szorosan egyetlen tanulásmenedzsment-rendszerhez, mégis beágyazhatók, továbbíthatók és részben offline módon is felhasználhatók. A bemutatott megközelítés nem az intelligens oktatórendszerekkel vagy komplex LMS-megoldásokkal kíván versenyezni, hanem a tanári tartalomfejlesztést támogatja felhőalapú generatív MI-eszközökkel. A cikk áttekinti a hordozhatóság és a multimodalitás pedagógiai jelentőségét, az eszközválasztás szempontjait, valamint a tartalom-előállítás gyakorlati munkafolyamatát. Kiemelt szerepet kap a forrásanyag pontos kijelölése, a tananyag kisebb kanonikus blokkokra bontása, a modalitáshoz illeszkedő szolgáltatás kiválasztása és az oktatói validálás. A vizsgált modalitások közé tartozik az írott magyarázó szöveg, a vers, a felolvasott szöveg, a podcast-jellegű párbeszéd, a dal, az infografika, a gondolattérkép, a prezentáció és a rövid videó. A tanulmány következtetése szerint a generatív MI jól használható idő- és költséghatékony kurzustartalom-előállításra, de a folyamat csak oktatói kontrollal, dokumentált forráskezeléssel és minőségbiztosítással tekinthető pedagógiailag megbízhatónak.

**Kulcsszavak:** generatív mesterséges intelligencia, nagy nyelvi modellek, multimodalitás, online oktatás, kurzusfejlesztés

**Abstract**

The rapid development of generative artificial intelligence creates new opportunities for online course development. This paper examines how portable, multimodal course materials can be produced with a low entry threshold, without being tightly bound to a single learning management system. The proposed approach does not aim to replace intelli-

gent tutoring systems or complex LMS solutions; instead, it supports teachers in creating embeddable, shareable and partly offline content elements with cloud-based generative AI tools. The paper discusses the pedagogical relevance of portability and multimodality, the main criteria of tool selection, and a practical workflow for content production. Special attention is paid to defining source materials, splitting the curriculum into canonical content blocks, choosing modality-specific services, and validating AI-generated outputs by the teacher. The discussed modalities include explanatory written text, poems, text-to-speech audio, podcast-style dialogues, songs, infographics, mind maps, presentations and short videos. The study concludes that generative AI can support time- and cost-efficient course content production, but pedagogical reliability requires human supervision, documented source management and quality assurance.

**Keywords:** generative artificial intelligence, large language models, multimodality, online education, course development

## 1. Bevezetés

A generatív mesterséges intelligencia az elmúlt években az oktatás egyik meghatározó technológiai témájává vált. A nagy nyelvi modellek és a hozzájuk kapcsolódó multimodális szolgáltatások már nem csupán szöveggenerálásra használhatók, hanem képi, hangalapú, prezentációs és videós tartalmak előállítását is támogatják. Ez a változás különösen fontos az online oktatásban, ahol a hallgatók egyre változatosabb eszközökön, élethelyzetekben és tanulási preferenciákkal találkoznak a kurzustartalmakkal. A terület megítélése ugyanakkor megosztó: a lehetőségek mellett komoly kérdéseket vet fel a pontosság, a szerzői felelősség, a forráskezelés és az oktatói kontroll [4], [5].

A jelen tanulmány azt a kérdést járja körül, hogyan lehet generatív MI segítségével hordozható, multimodális kurzustartalmakat előállítani. A cél nem az oktató kiváltása, hanem egy olyan munkafolyamat bemutatása, amelyben a tanár a forrásanyag, a pedagógiai cél és a minőségbiztosítás felett megtartja a kontrollt, miközben a gépi támogatás csökkentheti az időráfordítást és bővítheti a létrehozható tartalomtípusok körét.

## 2. Elméleti háttér

A mesterséges intelligencia oktatási alkalmazásait a szakirodalom régóta vizsgálja, az intelligens oktatórendszerektől a tanulási analitikán át az automatikus értékelésig [5]. A nagy nyelvi modellek megjelenése azonban új helyzetet teremtett: a modellek kevés példából is képesek feladatmegoldásra, szövegalkotásra és instrukciókövetésre [2]. A foundation modellek általánosítható képességei egyszerre nyitnak meg pedagógiai lehetőségeket és kockázatokat, ezért alkalmazásuknál a cél, a kontextus és a korlátok tudatos meghatározása szükséges [1].

A multimodális modellek fejlődése tovább bővíti a felhasználási lehetőségeket. A szöveg, kép, hang és más reprezentációk közötti átjárás lehetővé teszi, hogy ugyanaz a tananyagrész többféle formában jelenjen meg [3]. Ez nem azt jelenti, hogy minden hallgatóhoz külön, teljesen személyre szabott rendszert kell építeni, hanem azt, hogy a tanár ugyanabból a kanonikus forrásból többféle tanulási útvonalat támogató tartalomelemeket készíthet. A generatív MI oktatási szerepe ezért nem kizárólag technológiai, hanem tanulásszervezési és módszertani kérdés is.

## 3. A hordozható kurzustartalom fogalma

A tanulmányban hordozható kurzustartalomnak azokat a tartalomelemeket tekintem, amelyek nem kizárólag egy adott LMS belső funkciójaként léteznek, hanem egyszerűen exportálhatók, beágyazhatók, linkelhetők, megoszthatók vagy akár offline is használhatók. Ide tartozhat például egy PDF, egy hangfájl, egy képként vagy HTML-ként megjeleníthető infografika, egy prezentáció, egy videó vagy egy egyszerűen küldhető szöveges tananyagegység.

Ez a megközelítés tudatosan nem kíván versenyezni a komplex tanulásmenedzsment-rendszerekkel vagy az intelligens oktatórendszerekkel. Ezek a rendszerek értékesek lehetnek, de fejlesztésük, bevezetésük és fenntartásuk gyakran jelentős erőforrást, szervezeti feltételeket és előzetes tudást igényel. A hordozhatóság ezzel szemben az egyszerűséget, a platformfüggetlenséget és az alacsony belépési küszöböt helyezi előtérbe. Különösen azokban a felsőoktatási környezetekben lehet hasznos, ahol a tanár rendelkezik szakmai tartalommal, de nincs nagy fejlesztői csapata vagy külön multimédiás stúdiókapacitása, esetleg gazdag funkcionalitású keretrendszere.

## 4. Multimodalitás és célcsoport

A multimodalitás indokoltsága több szinten jelenik meg. Egyrészt a tananyagok jellege eltérő: más forma alkalmas fogalmak magyarázatára, folyamatok bemutatására, definíciók rögzítésére vagy összefüggések vizualizálására. Másrészt a hallgatók eltérő helyzetekben tanulnak. Van, aki utazás közben hanganyagot hallgatna, más vizuális összefoglalót vagy gondolattérképet használna ismétlésre, megint más részletesebb írott magyarázatot igényel. A cél tehát nem a tanulótípusok leegyszerűsítő kategorizálása, hanem a hozzáférés és a tanulási helyzetek változatosságának támogatása.

A felhasználói célcsoport elsősorban (kartól és szaktól függetlenül) felsőoktatási kurzusok hallgatóiból áll. A felsőoktatásban résztvevő hallgatókra azért esett a választás, mert esetükben már szükségszerűen megjelenik egyfajta nagyobb önállóság a tanulási módszerekben, amely az önszabályozó tanulással társulva jól használhatóvá teszi számukra az ilyen jellegű online tananyagrészeket. A készítői célcsoport ugyanakkor a tanári oldal: olyan

oktatók, akik jelentős informatikai előképzettség, nagy költségvetés vagy speciális eszközpark nélkül szeretnének többféle formátumú tananyagot készíteni. A megközelítés gyakorlati értékét az adja, hogy a felhőalapú, magyar nyelvet is támogató szolgáltatások már viszonylag alacsony költséggel vagy korlátozott ingyenes keretek között is használhatók.

## 5. Javasolt munkafolyamat

A hordozható multimodális kurzustartalmak előállításának első lépése **a forrásanyag pontos meghatározása.** Ez lehet jegyzet, tankönyvrészlet, előadásvázlat, diasor vagy más oktatói anyag. A generatív MI használata során különösen fontos, hogy a modell ne általános, ellenőrizetlen tudásból dolgozzon, hanem a tanár által kijelölt forrásra támaszkodjon. Ezt a konkrét előállítandó tematika megadását megelőzően pre-promptolással, majd fájlfeltöltéssel, illetve a kapcsolódó adatforrásokból merített adatok felhasználásával történő tartalomelőállítással, azaz retrieval augmented generation (RAG) jellegű megoldással lehet támogatni.

A második lépés a **darabolás.** A hosszú tananyagokat célszerű kisebb, tematikusan koherens blokkokra bontani. Ezek a blokkok kanonikus forrásként működnek: minden modalitás ugyanabból a jóváhagyott egységből készül. Ez csökkenti annak kockázatát, hogy a különböző tartalomformák egymásnak ellentmondjanak, és megkönnyíti az oktatói ellenőrzést is. A darabolás egyben technikai kényszer is lehet, mert a modellek kontextusablaka, fájlkezelése és munkamenet-stabilitása korlátozott.

A harmadik lépés a **modalitásfüggő eszközválasztás**. Más szolgáltatás lehet alkalmas jó minőségű magyar nyelvű szöveg készítésére, más hangelőállításra, képgenerálásra, prezentációra vagy videóra. A szolgáltatások aktuális minősége gyorsan változik, ezért az eszközlisták csak pillanatfelvételként kezelhetők. Tartósabb szempont a nyelvi támogatás, a költség (ingyenes, és/vagy alacsony fix díjjal, pl. freemium modellel elérhető), a fájlformátum, a területi elérhetőség (pl. EU-n belül elérhető), mennyiségi limitációk (pl. napi/havi kérések maximális száma) az exportálhatóság, a formátum szerkeszthetősége (pl. pptx a pdf-prezentáció helyett), az adatkezelési feltételek és a tanári munkafolyamatba illeszthetőség. Például – a teljesség igénye nélkül – az általános modellek közül a Chat GPT, Google Gemini. DeepSeek, Mistral, Grok, vagy funkció specifikusak közül a Nano Banana, Sora, NotebookLM, Gamma, Leonardo, Tengr.ai.

A negyedik lépés a különböző tartalomformákhoz szükséges **promptok** manuálisan vagy metapromptinggal (nagy nyelvi modell segítségével közvetve) történő **előállítása és a sablonosítás** (a különböző tartalomváltozatok konzisztenciájának biztosítása érdekében közös forrástartalom megadása, valamint az azonos formátumok esetén az egységesség azonos instrukciókkal történő biztosítása). A tanár nem egyszerűen tartalmat kér a modelltől, hanem szerepet, célt, célcsoportot, terjedelmet, hangnemet, szerkezetet és ellenőrzési szempontokat ad meg. Például más instrukció szükséges egy tömör vizsgaismétlő infog-

rafikához, egy podcast-jellegű párbeszédhez vagy egy 10 diás prezentációhoz. A jól megírt metapromptok újrahasználhatók, így a folyamat a későbbi tananyagrészeknél gyorsítható.

## 6. Modalitások és felhasználási lehetőségek

Az írott szöveg a legkézenfekvőbb kimenet: alkalmas magyarázat, összefoglaló, fogalomtár, ellenőrző kérdéssor vagy tanulási útmutató készítésére. A vers vagy ritmizált szöveg nem minden tudományterületen központi forma, de memorizálást segítő, motivációs vagy figyelemfelkeltő szerepe lehet. A felolvasott szöveg, azaz a TTS (text-to-speech) -alapú hanganyag különösen azoknak hasznos, akik utazás, munkavégzés vagy ismétlés közben audioformátumban szeretnének tanulni.

A podcast-jellegű párbeszéd formátum lehetőséget ad fogalmak ütköztetésére, kérdés-válasz formájú magyarázatra vagy gyakori félreértések tisztázására. A dal inkább kiegészítő, motivációs vagy emlékeztető elemként értelmezhető. A képi modalitások közül az infografika rövid, vizuálisan rendezett összefoglalásra, a gondolattérkép pedig fogalmi kapcsolatok feltárására alkalmas. A prezentáció a tanári előadás és az önálló tanulás közötti átmeneti forma lehet. A rövid videó a képi, hangalapú és szöveges elemeket egyesíti, de előállítása jelenleg nagyobb utómunka- és ellenőrzési igénnyel jár.

## 7. Gyakorlati és minőségbiztosítási szempontok

A generatív MI alkalmazásának egyik legfontosabb gyakorlati előnye az időmegtakarítás. Ugyanakkor a gyors előállítás nem azonos a kész, ellenőrzött oktatási anyaggal. A modellek tévedhetnek, túláltalánosíthatnak, pontatlanul hivatkozhatnak, vagy a forrástól eltérő állításokat is beemelhetnek. Ezért minden kimenetet tanári validálásnak kell követnie. A minőségbiztosítás része a szakmai pontosság, a nyelvi érthetőség, a tanulási célhoz való illeszkedés, a formai minőség és a hozzáférhetőség vizsgálata.

A költséghatékonyság szintén fontos tényező. A gyakorlatban hasznos lehet a tartalom-előállítás blokkokba szervezése, mert így az előfizetéses szolgáltatások időszakosan, célzottan használhatók. Figyelembe kell venni az ingyenes keretek korlátait, a fájlfeltöltési limiteket, az exportlehetőségeket és az adatvédelmi feltételeket is. Oktatási környezetben különösen fontos, hogy személyes hallgatói adatok ne kerüljenek indokolatlanul külső rendszerekbe.

## 8. Következtetések és további kutatási irányok

A bemutatott megközelítés alapján a generatív MI alkalmas arra, hogy a tanárok alacsony belépési küszöb mellett, több modalitásban állítsanak elő hordozható kurzustartalmakat. A technológia jelenlegi állapota mellett azonban kompromisszumokra van szükség: a szolgáltatások minősége gyorsan változik, a magyar nyelvű kimenetek színvonala modalitásonként eltérő, és a multimodális tartalmak gyakran utómunkát igényelnek. A fő pedagó-

giai érték nem önmagában a gépi generálásban, hanem a strukturált, forrásalapú, oktatói kontrollal végzett tartalomfejlesztésben rejlik.

A következő kutatási lépés egy kontrollcsoportos pilot akciókutatás, amely három dimenzióban vizsgálja a megközelítés hatását: az előállítási időben, a tanulási eredményességben és a hallgatói visszajelzésekben. Így empirikusan is értékelhetővé válik, hogy a multimodális, hordozható tartalomelemek nemcsak gyorsabban előállíthatók-e, hanem támogatják-e a tanulási folyamatot legalább a hagyományos online kurzustartalmak szintjén. A kutatás tágabb kérdése az, hogyan lehet a generatív MI-t úgy beépíteni a felsőoktatási kurzusfejlesztésbe, hogy az egyszerre legyen hatékony, átlátható és pedagógiailag felelős.

## Irodalomjegyzék


[1] Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

[2] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901.

[3] Clark, C., Radford, A., Wu, J., Child, R., Amodei, D., & Sutskever, I. (2021). Multimodal few-shot learning with Frozen Transformers. Advances in Neural Information Processing Systems, 34, 200–212.

[4] Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., ... & Williams, M. D. (2023). Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice, and policy. International Journal of Information Management, 66, 102642.

[5] Popenici, S. A. D., & Kerr, S. (2017). Exploring the impact of artificial intelligence on teaching and learning in higher education. Research and Practice in Technology Enhanced Learning, 12(1), 22.

[6] Ragab, M., Ashary, E. B., Kateb, F., Hakeem, A., Mosli, R., Albogami, N. N., & Nooh, S. (2025). Classification of human-written and AI-generated sentences using a hybrid CNN-GRU model optimized by the spotted hyena algorithm. Alexandria Engineering Journal, 126, 116-130. https://doi.org/10.1016/j.aej.2025.04.071

[7] Toldi, L., & Lengyelné Molnár, T. (2024). Development options for artificial intelligence supported intelligent tutoring systems via the integration of the ChatGPT. Magyar Nyelvőr, 148, 618-627. https://doi.org/10.38143/Nyr.2024.5.618

DOI: 10.31915/NWS.2026.10



# E-LEARNING ALAPÚ KUTATÁSTÁMOGATÁS PHD-HALLGATÓK, OKTATÓK ÉS DOLGOZÓK SZÁMÁRA KÖNYVTÁRI KÖRNYEZETBEN[1]

# E-LEARNING-BASED RESEARCH SUPPORT FOR PHD STUDENTS, LECTURERS AND STAFF IN A LIBRARY ENVIRONMENT

*Fülöp Tiffany*
*SZTE Klebelsberg Könyvtár és Levéltár*
*tiffany.fulop@ek.szte.hu*
*ORCID: 0000-0002-2219-8455*

## Absztrakt

A Szegedi Tudományegyetemen 2022 és 2026 között futó Komplex Digitális Modellváltás – intelligens fokozatváltás projektje keretében a hallgatói és dolgozói készségek fejlesztését célzó tevékenységek részeként az SZTE Klebelsberg Könyvtár és Levéltár öt e-learning kurzust hozott létre.

A kurzusok a könyvtár kutatástámogató szolgáltatásait, valamint a tudományos munka szempontjából kiemelt témákat dolgozzák fel online, könnyen feldolgozható formában, különböző mélységben, az egyetem PhD-hallgatói, oktatói és dolgozói számára.

Az előadás az ismeretanyag e-learning **környezetbe történő átültetésének folyamatát mutatja be a koncepcióalkotástól és szövegírástól kezdve a technikai megvalósításon át az oktatási gyakorlatig.**

**Kulcsszavak:** kutatástámogatás, könyvtári e-learning, élethosszig tartó tanulás, digitális kompetenciák

## Abstract

Within the framework of the Complex Digital Model Change – Intelligent Gear Change project, running at the University of Szeged between 2022 and 2026, the SZTE Klebelsberg Library and Archives developed five e-learning courses as part of the activities aimed at improving students' and staff members' skills.

1 Az oktatóanyag fejlesztését az RRF-2.1.2-21-2022-00012 azonosítószámú, Komplex Digitális Modellváltás – intelligens fokozatváltás projekt támogatta. A projekt Magyarország Helyreállítási és Ellenállóképességi Tervének keretében valósul meg.

The courses present the library's research support services, as well as key topics related to scholarly work, in an online and easily accessible format, at different levels of depth, for the university's PhD students, lecturers and staff.

The presentation introduces the process of transferring this body of knowledge into an e-learning environment, from concept development and text writing through technical implementation to educational practice.

**Keywords**: research support, library e-learning, lifelong learning, digital competencies

## Pályázati lehetőség és előzmények

A Szegedi Tudományegyetem 2022-2026 között részt vett az RRF-2.1.2-21-2022-00012 azonosító számú *Komplex digitális modellváltás – intelligens fokozatváltás* projektben[2], amelynek céljai között szerepelt többek között a gyakorlati oktatást közvetlenül segítő digitális infrastruktúra fejlesztése, digitális eszközök beszerzése, egyetemi tudásbázis létrehozása és az egyetemi polgárok digitális készségeinek fejlesztése is. Ezen belül a projektben tervezett tevékenységek egyrészt kifejezetten az egyetemi kurzusok rövid tananyagokkal történő továbbfejlesztését célozták meg, amelyek meghatározott készségek fejlesztésére irányultak, valamint olyan rövid e-learning tananyagok létrehozását is szorgalmazták, amelyek önállóan meghirdethetők online tréning formájában nem csak egyetemi hallgatók, hanem kifejezetten egyetemi dolgozók és oktatók számára is.

A pályázatleírást áttekintve úgy ítéltük meg, hogy ez kiváló lehetőséget adna számunkra, hogy már meglévő tudásunkat egy strukturáltabb, szélesebb körben elérhető formába szervezzük, ezzel tovább erősítve az SZTE Klebelsberg Könyvtár és Levéltár kutatástámogatást célzó törekvéseit, szolgáltatását, munkáját. Az oktatók és kutatók támogatása egyre jelentősebb könyvtári feladattá válik, miközben nő az azonnali, könnyen feldolgozható információ iránti igény is. Szolgáltatásainknak erre a változó igényre is reagálniuk kell, és a projekt során elkészített anyagok is ezt a célt szolgálják.

Könnyítette munkánkat, hogy e-tananyag készítésben már volt korábbi gyakorlatunk a régebben a könyvtár által tanított, osztálytermi keretek között zajló és a bölcsész képzésben részt vevő BA-hallgatók számára kötelező könyvtárhasználati alapismeretek kurzus online tananyaggá történő alakítása kapcsán.[3] Az első online kurzus létrejöttét a pandémia által is felerősített változó oktatási igények, a kollégák munkaterhének csökkentése, valamint az információ szélesebb körű és folyamatos elérhetővé tétele ösztönözte.

2 A projekt tartalmáról részletesebben itt olvashatnak: https://u-szeged.hu/pmi/rrf-2-1-2-21-2022-00012/rrf-2-1-2-21-2022-00012

3 Az erről szóló tapasztalatainkat a 2021-es Networkshop konferencián mutattuk be, a tanulmány itt tekinthető meg: https://ocs.mtak.hu/index.php/nws/2021/paper/view/74 (Fülöp Tiffany, Nagy Gyula: Az online oktatás könyvtári támogatása a Szegedi Tudományegyetemen)

Ezen túl szakterületspecifikus kollégáink a PhD-hallgatók számára tartottak a tavaszi szemeszterben felvehető tantermi órát, ami a kutatómunkát segítő alapvető könyvtári ismereteket és a kutatók munkáját támogató témákat (pl.: MTMT, tudománymetria, OA) is érintett. Ennek a szolgáltatásnak nehézségei voltak, hogy több kolléga idejét is jelentősen lefoglalta viszonylag kevéssé változó információkat tartalmazó témák bemutatásával, a terem függvényében csak korlátozott számú jelentkező tudott részt venni az órákon, ráadásul a jelenléti oktatás a PhD-hallgatók kutatásközpontú képzésébe nehezebben is illeszkedett be. Úgy láttuk, hogy egy hasznos szolgáltatás viszonylag nehézkesen érte el a célközönséget, így a pályázat felvetésében határozottan fejlődési lehetőséget láttunk.

Mivel szolgáltatásaink jelentős részét képezi a kutatástámogatás, ami az egyetem oktatóit, kutatóit célozza meg, ezért a pályázaton belül kifejezetten felkeltette a figyelmünket az intézmény dolgozóinak továbbképzését szorgalmazó vonal. Sok egyetemi könyvtárhoz hasonlóan, mi is létrehoztunk egy tudásbázist Szerzői eszköztár[4] névvel, ami az összes különböző típusú szolgáltatásunk és eszközünk útmutatóit, leírásait és tájékoztatóit egy jól strukturált honlap keretei között jeleníti meg. Ezeknek a támogatása jelentős kapacitást igényel, ami tényleges működtetésük mellett az állandó tájékoztatását is jelenti ennek a nagy mennyiségű információnak, aminek a megtalálása, feldolgozása és megértése nem feltétlenül csak az oktatók időhiánya, hanem gyakran az információk komplexitása miatt is nehézséget okoz. Ezért láttunk megoldást, segítséget az önálló, komplex, kisméretű információs csomagok feldolgozhatóbb formájában, ami az önálló megértés támogatásának a lehetőségét kínálta skálázható tudásátadási formaként.

## Témaválasztás

Nem törekedtünk a Szerzői eszköztár teljes tartalmának e-tananyaggá alakítására, mert ez ismét nehezen áttekinthető anyagmennyiséget eredményezett volna. Ehelyett olyan támogató tananyagegységeket kívántunk létrehozni, amelyek a gyakran felmerülő kérdésekre, a nehezebben értelmezhető információkra és azokra a témákra reagálnak, ahol a szöveges útmutatóknál többet tudtunk nyújtani.

Mindezek mentén úgy döntöttünk, hogy készítünk egy 6 x 45 perces *Kutatástámogatás PhD hallgatóknak* című kurzust[5], aminek alapjául a már meglévő PhD-kurzusunk szolgált. Annak tartalmi egységeit emeltük át és dolgoztuk fel e-learning formában, ahol elsősorban a stabilabb, kevésbé gyorsan változó tartalmi elemekre fókuszáltunk. Ezt a kurzust kifejezetten alapszintűnek szántuk, amit azért fontos kiemelni, mert a témaválasztás miatt számítanunk kellett átfedésekre a doktori disszertációjuk elkészítésén dolgozó kezdő kutatóknak szánt anyagok és a kutatóként dolgozó oktatók számára szükséges információk

4 http://szerzoknek.ek.szte.hu/

5 https://www.edulib.ek.szte.hu/courses/kutatastamogatas-phd-hallgatoknak/

kapcsán. Egy skálázhatóbb, differenciáltabb tudásanyag létrehozásában reménykedtünk, ahol mindenki azokat az információkat tudja gyorsabban és könnyebben megtalálni és megérteni, amire kifejezetten szüksége van.

Kialakítottunk két 2 x 45 perces haladó szintűnek tekintett kurzust az oktatók részére, a *Tudománymetriai adatbázisok*[6] és az *MTMT, repozitóriumok*[7] oktatói képzést azzal a céllal, hogy a publikálást támogató szolgáltatásaink közül kiemeljünk néhány kisebb, önállóan is értelmezhető modult, ami számukra kiemelten fontos.

Mindezek mellé két darab 2 x 45 perces általánosabb ismereteket tartalmazó anyagot is összeállítottunk *Open Access publikálási lehetőségek*[8] és *Publikálást támogató könyvtári szolgáltatások*[9] címen, amelyek a hallgatók, oktatók és kutatók mellett az ő munkájukat segítő egyetemi dolgozók, kollégák számára is informatív tudással szolgálhatnak.

A kurzusok kialakítása, a skálázhatóság lehetőséget biztosított a számunkra ahhoz, hogy a komplexebb tudásanyagokat feldolgozhatóbb formába emeljük át, valamint azt, hogy megkülönböztessük a kevésbé változó anyagrészeket azoktól, amelyek nagyobb figyelmet igényelnek, ezzel is könnyítve a tananyagok naprakészen tartását.

## Munkafolyamat

A tananyagok elkészítésére körülbelül két hónap állt a rendelkezésünkre, amit az előkészítési és az utómunkálatok egy kicsit meghosszabbítottak. A projektben öt fő vett részt, egymást kiegészítő feladatkörökben. Az adminisztratív felelős Várnai-Vígh Adrienn volt, aki a pályázati anyag elkészítését, az elvárások nyomon követését és a határidők betartását koordinálta. A szakmai koncepció összeállításáért és a tematika kidolgozásáért két szakmai felelős felelt: a kutatástámogató szolgáltatások területén Zeller Rozália, míg a publikálást támogató szolgáltatások esetében Hoczopán Szabolcs. A megvalósítási feladatokat, így a médiaelemek elkészítését és véglegesítését, valamint a szövegek honlapos formába történő átültetését Fekete Franciska és én végeztem, többek között az én feladatom volt az e-tananyagok kialakításának koordinálása is.

A munkafolyamat négy fő szakaszra tagolódott, ezek azonban nem egymást követő, lezárt munkafázisokként valósultak meg. A szűk határidő miatt az ideálisabb kivitelezéshez képest az egyes témák és azok alegységei párhuzamosan kerültek kidolgozásra: amint egy-egy téma esetében lezárult az előző szakasz, az adott anyag továbbléphetett a következő munkafázisba, függetlenül attól, hogy a többi téma ugyanebben az időben melyik szakaszban tartott. Természetesen ennek a típusú munkavégzésnek az egyik legnagyobb nehéz-

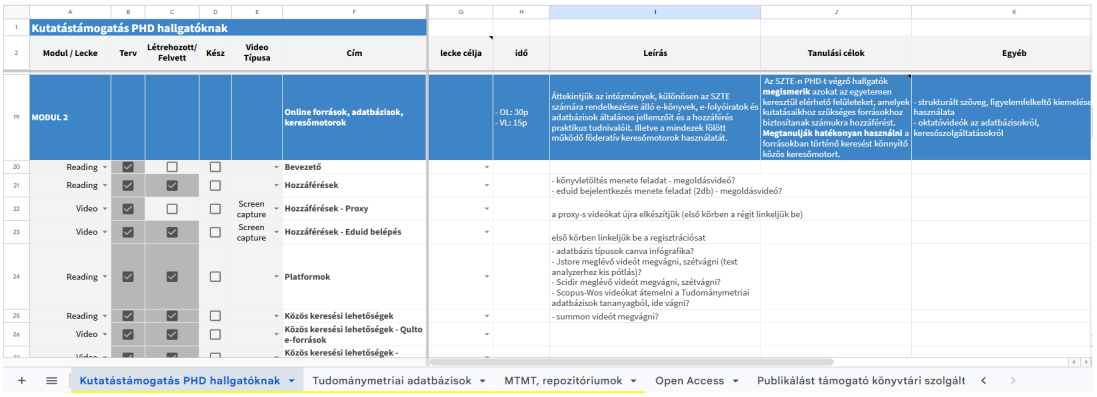
6 https://www.edulib.ek.szte.hu/courses/tudomanymetriai-adatbazisok/

7 https://www.edulib.ek.szte.hu/courses/mtmt-repozitoriumok-oktatoi-kepzes/

8 https://www.edulib.ek.szte.hu/courses/open-access-publikalasi-lehetosegek/

9 https://www.edulib.ek.szte.hu/courses/publikalast-tamogato-konyvtari-szolgaltatasok/

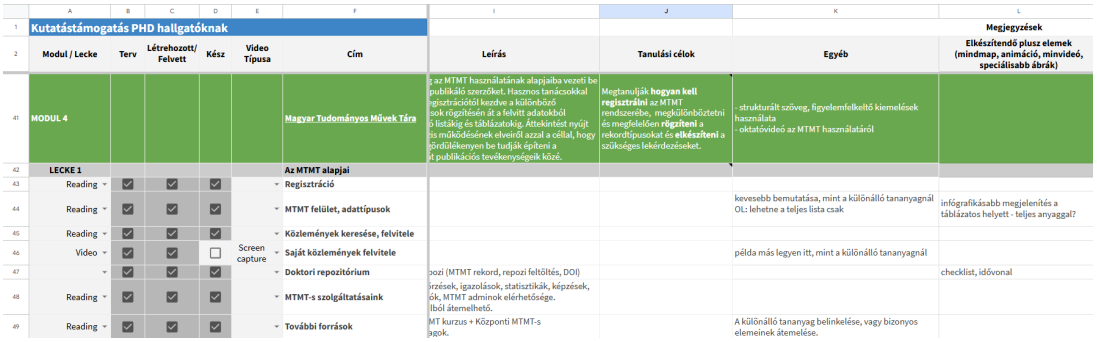
sége a koordinálás volt, amit azzal igyekeztünk kezelni, hogy a Coursera kurzustervezési jógyakorlatát vettük mintának, és egy olyan online táblázatba vezettük a kurzusok haladását, ahol az egyes anyagokat kisebb, jól követhető elemekre bontottuk.



| | Kutatástámogatás PHD hallgatóknak | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | Modul / Lecke | Terv | Létrehozott/ Felvett | Kész | Video Típusa | Cím | lecke célja | idő | Leírás | Tanulási célok | Egyéb |
| 19 | MODUL 2 | | | | | Online források, adatbázisok, keresőmotorok | | - OL: 30p<br>- VL: 15p | Áttekintjük az intézmények, különösen az SZTE számára rendelkezésre álló e-könyvek, e-folyóiratok és adatbázisok általános jellemzőit és a hozzáférés praktikus tudnivalóit. Illetve a mindezek fölött működő föderatív keresőmotorok használatát. | Az SZTE-n PHD-t végző hallgatók **megismerik** azokat az egyetemen keresztül elérhető felületeket, amelyek kutatásaikhoz szükséges forrásokhoz biztosítanak számukra hozzáférést. **Megtanulják hatékonyan használni a** forrásokban történő keresést könnyítő közös keresőmotort. | - strukturált szöveg, figyelemfelkeltő kiemelések használata<br>- oktatóvideók az adatbázisokról, keresőszolgáltatásokról |
| 20 | Reading | ☑ | ☐ | ☐ | | Bevezető | | | | | |
| 21 | Reading | ☑ | ☑ | ☐ | | Hozzáférések | | | - könyvletöltés menete feladat - megoldásvideó?<br>- eduid bejelentkezés menete feladat (2db) - megoldásvideó? | | |
| 22 | Video | ☑ | ☐ | ☐ | Screen capture | Hozzáférések - Proxy | | | a proxy-s videókat újra elkészítjük (első körben a régit linkeljük be) | | |
| 23 | Video | ☑ | ☑ | ☐ | Screen capture | Hozzáférések - Eduid belépés | | | első körben linkeljük be a regisztrációsat | | |
| 24 | Reading | ☑ | ☑ | ☐ | | Platformok | | | - adatbázis típusok canva infógrafika?<br>- Jstore meglévő videót megvágni, szétvágni (text analyzerhez kis pótlás)?<br>- Scidir meglévő videót megvágni, szétvágni?<br>- Scopus-Wos videókat átemelni a Tudománymetriai adatbázisok tananyagból, ide vágni? | | |
| 25 | Reading | ☑ | ☑ | ☐ | | Közös keresési lehetőségek | | | - summon videót megvágni? | | |
| 26 | Video | ☑ | ☑ | ☐ | | Közös keresési lehetőségek - Quilto e-források | | | | | |
| | | | | | | Közös keresési lehetőségek - | | | | | |

Kutatástámogatás PHD hallgatóknak | Tudománymetriai adatbázisok | MTMT, repozitóriumok | Open Access | Publikálást támogató könyvtári szolgált

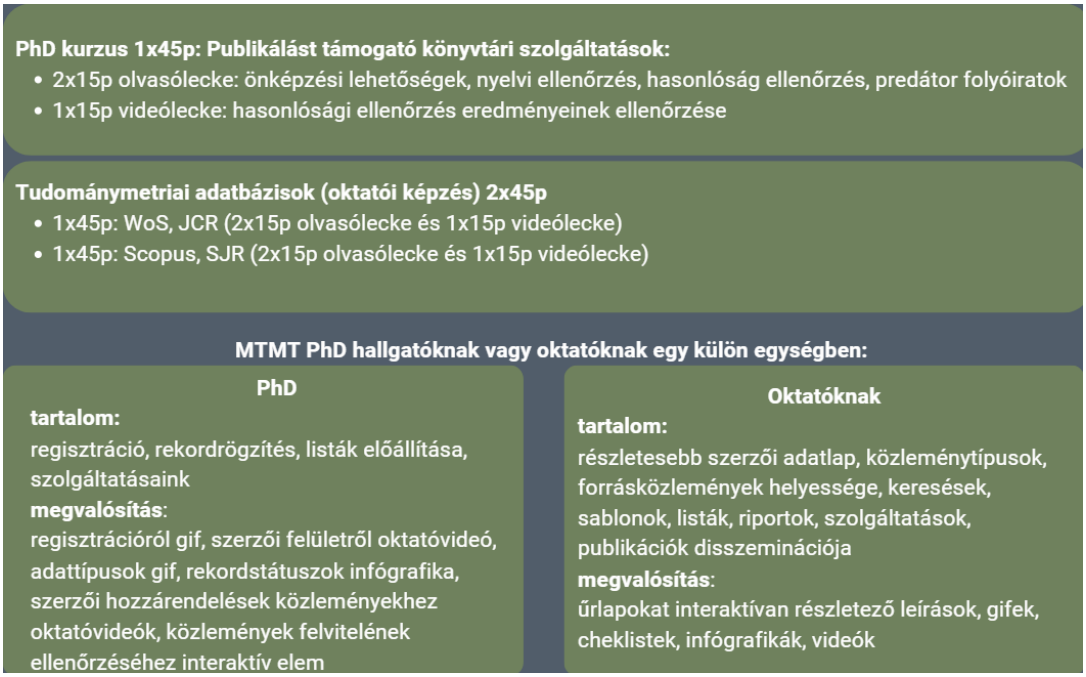
*1. ábra*



| | Kutatástámogatás PHD hallgatóknak | | | | | | | | | Megjegyzések |
|---|---|---|---|---|---|---|---|---|---|---|
| 2 | Modul / Lecke | Terv | Létrehozott/ Felvett | Kész | Video Típusa | Cím | Leírás | Tanulási célok | Egyéb | Elkészítendő plusz elemek (mindmap, animáció, minvideó, speciálisabb ábrák) |
| 41 | MODUL 4 | | | | | Magyar Tudományos Művek Tára | z az MTMT használatának alapjaiba vezeti be publikáló szerzőket. Hasznos tanácsokkal egisztrációtól kezdve a különböző sok rögzítésén át a felvitt adatokból ó listákig és táblázatokig. Áttekintést nyújt is működésének elveiről azzal a céllal, hogy ördülékenyen be tudják építeni a t publikációs tevékenységeik közé. | Megtanulják **hogyan kell regisztrálni** az MTMT rendszerébe, megkülönböztetni és megfelelően **rögzíteni** a rekordtípusokat és **elkészíteni** a szükséges lekérdezéseket. | - strukturált szöveg, figyelemfelkeltő kiemelések használata<br>- oktatóvideó az MTMT használatáról | |
| 42 | LECKE 1 | | | | | Az MTMT alapjai | | | | |
| 43 | Reading | ☑ | ☑ | ☑ | | Regisztráció | | | | |
| 44 | Reading | ☑ | ☑ | ☑ | | MTMT felület, adattípusok | | | kevesebb bemutatása, mint a különálló tananyagnál<br>OL: lehetne a teljes lista csak | infógrafikásabb megjelenítés a táblázatos helyett - teljes anyaggal? |
| 45 | Reading | ☑ | ☑ | ☑ | | Közlemények keresése, felvitele | | | | |
| 46 | Video | ☑ | ☑ | ☐ | Screen capture | Saját közlemények felvitele | | | példa más legyen itt, mint a különálló tananyagnál | |
| 47 | | ☑ | ☑ | ☑ | | Doktori repozitórium | pozi (MTMT rekord, repozi feltöltés, DOI) | | | checklist, idővonal |
| 48 | Reading | ☑ | ☑ | ☑ | | MTMT-s szolgáltatásaink | rzések, igazolások, statisztikák, képzések, ók, MTMT adminok elérhetősége. lból átemelhető. | | | |
| 49 | Reading | ☑ | ☑ | ☑ | | További források | MT kurzus + Központi MTMT-s 920k. | | A különálló tananyag belinkelése, vagy bizonyos elemeinek átemelése. | |

*2. ábra*

A munkafolyamat első szakasza a *témakijelölés* volt. Bár a pályázat benyújtásához szükségünk volt egy tematikai elképzelésre, de a munka valójában azzal kezdődött, hogy a tananyagegységeket altémákra, az altémákat pedig tanulási elemekre bontottuk. Mindemellett a kezdő és haladó ismeretanyagot is már előre külön kellett választani, hogy a későbbiek során ne legyen önismétlő az anyag a kurzus tartalmi részeinek megírásakor. Ennek a részletesebb felbontása látható a fentebbi 1. és 2. ábrán is, de a táblázat előzményének is tekinthető összegző kiindulási alap a 3. ábra felső részén olvasható.




*3. ábra*

A második szakasz a *megvalósítás és készségfejlesztés tervezett módja* volt, ahol a témafelelősök és a megvalósítók részben már közösen, a részletesebb témalisták és a Szerzői eszköztáron található szövegek alapján összegyűjtötték azokat az ötleteket, elemeket, kivitelezési lehetőségeket, amik a megértést, feldolgozást segíthetik (különböző médiaelemek: infógrafikák, oktatóvideók, gifek, ábrák, interaktív feladatok). Ez előzetes tervként szolgált a további munkaszakaszokhoz.

A következő szakasz munkafolyamatához részben az e-tananyagok (és a már kiválasztott médiaelemek, pl.: infógrafika adatai, oktatóvideó script) *szövegeinek megírása* tartozott, amit a technikai felelősök az áttekintés után további médiaelemekkel láttak el. Innentől egy folyamatos revízió zajlott a szövegírók és a megvalósítók között, hiszen az ötletek megvalósításához szövegekre volt szükség, az elkészült elemeket pedig ellenőrizni kellett, hogy megfelelően lettek-e összerakva, létrehozva. Ebben a szakaszban készültek el a *médiaelemek és a narrációval ellátott oktatóvideók* is. A szakmai felelősök által készített minták és magyarázatok alapján a megvalósítók elkészítették a videók scriptjét, majd a felvételeket megvágták és értelmezést segítő elemekkel, például kiemelésekkel, zoomolásokkal és magyarázó szövegekkel látták el. Ugyanebben a szakaszban kezdték el a technikai megvalósítók a folyószövegek strukturált *honlaptartalommá történő alakítását* is, ami a médiaelemek feltöltése mellett egy olyan dizájn alkalmazását is jelentette, ami könnyíti a szövegértelmezést, a megértést és a tanulást. A munka másik jelentős részét a keretrendszerbe történő elhelyezés során az interaktív elemek, feladatok létrehozása jelentette, mert ezek-

nek jelentősebb részét a technikai megvalósítóknak kellett kitalálnia, mivel ehhez ismerni kellett azt a technikai eszközkészletet, amibe beültethető volt az anyag.

Az utolsó szakaszban többféle típusú *ellenőrzés* is történt. 1) A végleges anyagot elsőként a szakmai felelős ellenőrizte, annak érdekében, hogy az elkészült tartalom megfelelően került-e kialakításra, illetve nem keletkezett-e tartalmi pontatlanság a szövegek keretrendszerbe történő átültetése során. 2) A létrehozott oldalakat tartalmilag: érthetőség szempontjából, és technikailag: működés szempontjából is ellenőrizte a munkában nem részt vevő külső kolléga, ami alapján javításra kerültek a szükséges részek. 3) Természetesen a pályázatnak volt egy ellenőrzőlistája, aminek minden tananyagnak meg kellett felelnie.

## Technikai megoldás

A megvalósítás során a már meglévő, általunk üzemeltetett webes rendszerre építettünk, mivel a célunk nem egy zárt, kizárólag tanrendi keretek között használható egyetemi oktatási felület létrehozása volt, hanem egy nyilvánosan elérhető, oktatásra, olvasók kiszolgálására és információszolgáltatásra egyaránt alkalmas eszköz kialakítása. Saját LMS rendszer telepítését és üzemeltetését továbbra sem tartottuk célszerűnek, mivel annak fenntartása jelentős technikai, adminisztratív és humánerőforrást igényelt volna. Ehelyett a könyvtár által már működtetett weboldalak funkcionalitását terjesztettük ki, és ezekben igyekeztünk LMS jellegű, tanulástámogató struktúrát létrehozni.

Az e-tananyagok technikai alapját egy WordPress tartalomkezelő rendszer adta, amelyet LearnPress bővítménnyel[10] egészítettünk ki, aminek felépítése az LMS rendszerek funkcionalitását jelenítette meg. A tananyagok vizuális kialakításában az Elementor builder bővítményt[11] és annak kiegészítőit használtuk, míg az interaktív elemek és feladatok létrehozását a H5P-bővítmény[12] tette lehetővé.

A tananyagokhoz kapcsolódó médiatartalmak elkészítése szintén többféle eszköz bevonásával történt. Az infógrafikák, illusztrációk és egyes vizuális elemek készítéséhez elsősorban a Canvát[13] használtuk, míg az oktatóvideók és gifek elkészítéséhez, illetve szerkesztéséhez a Camtasia videószerkesztő program[14] állt rendelkezésünkre. Ezek az elemek különösen fontos szerepet kaptak abban, hogy a komplexebb, sokszor nehezebben átlátható információk ne kizárólag szöveges formában jelenjenek meg, hanem vizuálisan és interaktív módon is segítsék a felhasználók önálló feldolgozását.

10 https://wordpress.org/plugins/learnpress/

11 https://wordpress.org/plugins/elementor/

12 https://wordpress.org/plugins/h5p/

13 https://www.canva.com/

14 https://www.techsmith.com/camtasia/

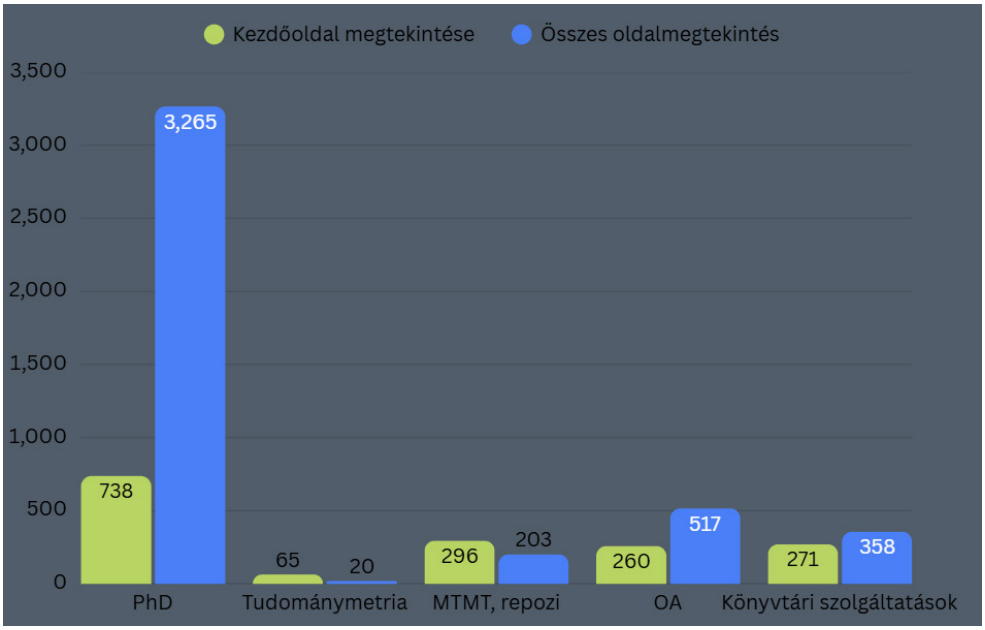
## Eredmények és tapasztalatok

A 2024/25-ös tanév első félévében kezdtük el oktatni a kurzusokat. A PhD-kurzusokat[15] a hallgatók az egyik félévben személyes oktatási formában, a másik félévben pedig az online platformon keresztül hallgathatják, ahol kollégáink továbbra is rendelkezésükre állnak a felmerülő kérdések megválaszolásában, csak ebben az esetben ez elsődlegesen online történik. Bár jelenleg még csak három félév adatai állnak rendelkezésünkre, az eddigi tapasztalatok alapján elmondható, hogy az online kurzusok részvételi aránya magasabb a jelenléti képzésekéhez képest. Míg utóbbiakon átlagosan 5–7 hallgató vesz részt, addig az online kurzusokon eddig 63, 7 majd 17 fővel számolhattunk. Az első oktatott félév végén 20 hallgató anonim módon kitöltött véleménye alapján a kurzus jól fejlesztette a kurzusleírásban szereplő elsajátítható képességeket és ismereteket, a rendelkezésre álló segédanyagok pedig megfelelően támogatták a tananyag feldolgozását.

A szöveges értékelések alapján a hallgatók hasznosnak ítélték a kurzus anyagát, jól felépített, tanulható és használható tananyagnak látták. Kifejezett előnyeként emelték ki az online, önállóan elérhető és saját ütemben feldolgozható formát, valamint a gyakorlati elemek jelenlétét. Negatív kritikaként ugyanakkor felmerült, hogy a gyakorlati alapú megközelítés lehetne még hangsúlyosabb része a tananyagnak, ahogy a kurzus teljesítéséhez szükséges teszt is életszerűbb lenne több gyakorlati feladat beépítésével. Ennek mi is tudatában vagyunk: az első körben elkészített tudásösszegző anyagrészek mellett idő- és kapacitáshiány miatt valóban nem tudtuk teljes mértékben kiaknázni a gyakorlati feladatokban rejlő lehetőségeket, ezért ezek bővítése határozottan a jövőbeli fejlesztések részét képezheti.

Az oktatóknak és dolgozóknak készített kurzusok elvégzésében is segítjük a felhasználókat, hiszen az online kurzusok célja az egyéni ütemezésű, tempójú tanulás támogatása, és a feldolgozás könnyítése, aminek továbbra is része a tanulást segítő oktató jelenléte. A könyvtári tájékoztatási munka során is érzékelhető, hogy a témákban nőtt az általános tájékozottság, ami a weboldal megtekintési adataival is összhangban áll, ahogyan a 4. ábrán látható.

15 https://www.ek.szte.hu/kezdooldal/konyvtarhasznalat/felhasznalokepzes/kutatastamogatas-kurzus/




*4. ábra*

A projekt egyik legfontosabb eredményének azt tekintjük, hogy nem csupán önállóan feldolgozható tananyagok jöttek létre, hanem olyan újrahasznosítható tartalmi egységek is, amelyek később más képzési helyzetekben, tájékoztatási folyamatokban vagy könyvtári szolgáltatások támogatásában is felhasználhatóak. A kurzusok kialakítása lehetőséget adott arra is, hogy a tudásátadás skálázhatóbb formában valósuljon meg, vagyis ugyanazok az információk szélesebb vagy szűkebb célközönség számára, rugalmasabban és a résztvevők egyéni időbeosztásához jobban igazodva váljanak elérhetővé.

A megvalósítás ugyanakkor több kihívást is magával hozott. Ezek közül az egyik legfontosabb a tartalom naprakészen tartása, hiszen a kutatástámogatáshoz és publikáláshoz kapcsolódó szolgáltatások, rendszerek és elvárások folyamatosan változnak. Emiatt az elkészült anyagok nem tekinthetők lezárt produktumoknak, hanem olyan folyamatos gondozást igénylő tudáselemek, amelyek rendszeres felülvizsgálat nélkül gyorsan veszíthetnek aktualitásukból.

A létrejött modell jól mutatja, hogy az egyetemi könyvtárak szerepe a hagyományos információszolgáltatáson túl a digitális tudásátadás és az oktatástámogatás irányába is tovább erősíthető. Az ilyen típusú fejlesztések nem feltétlenül csak egy-egy konkrét kurzus vagy tananyag létrehozását jelentik, hanem egy olyan szemlélet kialakítását is, amelyben a könyvtári tudásanyagok strukturált, újrahasznosítható és skálázható formában kapcsolódhatnak be az intézményi képzési és kutatástámogatási folyamatokba. A projekt így nemcsak tananyagokat hozott létre, hanem egy olyan skálázható tudásátadási modellt is, amely más intézményi környezetben is adaptálható lehet.

DOI: 10.31915/NWS.2026.11



# A PEDAGÓGUSI ADMINISZTRÁCIÓ DIGITÁLIS TRANSZFORMÁCIÓJA: MUNKAFOLYAMAT-AUTOMATIZÁLÁS ÉS RPA ALKALMAZÁSI LEHETŐSÉGEI A KÖZNEVELÉSBEN

# DIGITAL TRANSFORMATION OF TEACHER ADMINISTRATION: WORKFLOW AUTOMATION AND RPA APPLICATION OPPORTUNITIES IN PUBLIC EDUCATION

*Csernai Zoltán*
*Eszterházy Károly Katolikus Egyetem, Digitális Technológia Intézet, Humáninformatika Tanszék*
csernai.zoltan@uni-eszterhazy.hu

**Absztrakt**

Az oktatási intézmények digitális transzformációja során a fókusz gyakran a módszertani megújulásra irányul, miközben az iskolai adminisztratív terhek elérik a kritikus szintet. A legfrissebb nemzetközi kutatások (Delello és mtsai., 2025) rávilágítanak, hogy a repetitív, szabályalapú feladatok jelentősen hozzájárulnak az intézményi túlterheltséghez és a produktivitás csökkenéséhez. Jelen kutatás célja a robotizált folyamatautomatizálás (RPA) és az alacsony kódú (low-code) platformok, így például a Make.com vagy a Power Automate pedagógiai környezetben való alkalmazhatóságának feltérképezése. A vizsgálat során kiemelt figyelmet fordítok az automatizálható kulcsfontosságú munkafolyamatok, így többek között a jelentkezéskezelés és az adatalapú riportkészítés azonosítására, valamint az AI-alapú ökoszisztémákba integrált automatizmusok hatékonyságának mérésére az intézményi időmenedzsment tükrében. A kutatás elemzi a „human-in-the-loop" modell megvalósíthatóságát, vagyis azt, miként szabadíthatók fel kreatív energiák az ismétlődő műveletek gépiesítésével, továbbá vizsgálja az agilis oktatási (Agile Education) keretrendszerek és az automatizáció szoros kapcsolatát. A kutatás szisztematikus szakirodalmi áttekintésen alapul (Rajini és mtsai., 2025; Marchena Sekli és Godo, 2025), bemutatva a nemzetközi tapasztalatokat és a lehetséges hazai adaptációs irányokat. Az eredmények rávilágítanak arra, hogy az automatizáció nem csupán technológiai innováció, hanem a fenntartható, modern iskolai ökoszisztéma működésének alapfeltétele.

**Kulcsszavak:** RPA, munkafolyamat-automatizálás, digitális transzformáció, szervezeti hatékonyság, MI az oktatásban

**Abstract**

During the digital transformation of educational institutions, the focus is often on methodological renewal, while the administrative burden of schools reach critical levels. The latest international research (Delello et al., 2025) highlights that repetitive, rule-based tasks contribute significantly to teacher burnout and reduced productivity. The aim of this research is to explore the applicability of robotic process automation (RPA) and low-code platforms, such as Make.com or Power Automate, in an educational environment. In this study, I focus on identifying key work processes that can be automated, such as enrollment management and data-based reporting, as well as measuring the effectiveness of automations integrated into AI-based ecosystems in terms of institutional time management. The research analyzes the feasibility of the „human-in-the-loop" model, i.e., how creative energies can be unleashed by automating repetitive tasks, and examines the close relationship between Agile Education frameworks and automation. The research is based on a systematic review of the literature (Rajini et al., 2025; Marchena Sekli & Godo, 2025), presenting international experiences and possible directions for adaptation in Hungary. The results highlight that automation is not just a technological innovation, but a fundamental requirement for the functioning of a sustainable, modern school ecosystem.

**Keywords:** RPA, workflow automation, digital transformation, organizational efficiency, AI in education

## 1. Bevezetés

Napjaink köznevelési rendszerében a digitális transzformáció elsősorban a tanítási-tanulási folyamat megújítására koncentrál, miközben a pedagógusokat érintő adminisztratív kötelezettségek mértéke kritikussá vált. Az oktatókra egyre kiterjedtebb, szoros értelemben vett pedagógiai munkát nem igénylő feladatok hárulnak, mint például a jelenléti ívek vezetése, az osztályzatok manuális feldolgozása, a szülői kommunikáció kezelése vagy a komplex órarend-készítési folyamatok.

A probléma súlyát nemzetközi kutatások is alátámasztják: az automatizáció nem csupán időt takarít meg, hanem növeli a pedagógusok autonómiaérzetét, ami kulcsfontosságú védőfaktor a kiégés ellen (Delello és mtsai., 2025). Bizonyos adminisztratív munkafolyamatok célzott automatizálásával pedig akár 90% feletti időmegtakarítás is elérhető (Bhardwaj és Kumar, 2025).

Jelen tanulmány célja a robotizált folyamatautomatizálás és a generatív mesterséges intelligencia integrált alkalmazási lehetőségeinek bemutatása a köznevelési adminisztrációban. A kutatás során három pilot projekt technikai architektúrájának és hatékonyságának bemutatására kerül sor. Bár a fejlesztések hosszú távú célja a pedagógusok kognitív tehermentesítése, a bemutatott munkafolyamatok közvetlenül az iskolai adminisztráció

(a titkárság és az iskolavezetés) operatív feladatköreibe ágyazódnak be, bizonyítva, hogy az alacsony kódszintű („low-code") platformok alkalmazása képes az intézményi működés hatékonyságát radikálisan növelni.

## 2. Elméleti és módszertani háttér

A robotizált folyamatautomatizálás (RPA) lényege olyan szoftveres ágensek alkalmazása, amelyek képesek a digitális környezetben végzett emberi interakciók és szabályalapú folyamatok utánzására (Sekli és Godo, 2025). Az oktatási szférában az RPA jelentősége abban rejlik, hogy képes áthidalni a különböző szoftveres rendszerek (pl. KRÉTA, egyedi adatbázisok, e-mail kliensek) közötti integrációs réseket programozói beavatkozás nélkül is.

A kutatás módszertani keretét két meghatározó paradigma adja:

- „Human-in-the-loop" modell: A technológiai automatizmusok kognitív tehermentesítést végeznek, de a szakmai döntéshozatal és a végső validáció minden esetben a pedagógus kezében marad. Ez a megközelítés garantálja, hogy a mesterséges intelligencia támogató asszisztensként, nem pedig az emberi kontroll helyettesítőjeként jelenjen meg a pedagógiai munkafolyamatokban.
- Agilis oktatási keretrendszer: Az automatizációt nem egy merev, megváltoztathatatlan programként, hanem rugalmas építőkockákból álló, iteratív rendszerként kezeljük, amelyet a tanári visszajelzések alapján folyamatosan finomíthatunk. Ez teszi lehetővé a dinamikus alkalmazkodást a tanév hullámzó igényeihez (pl. beiratkozási időszak vagy vizsgák adminisztrációja), ahol az RPA környezete képes a változó terheléshez igazodni (Almeida és mtsai., 2025).

A megvalósításhoz használt „low-code" platformok (pl. Microsoft Power Automate, Make.com) grafikus felületükkel teszik lehetővé ezen komplex munkafolyamatok gyors felépítését és módosítását.

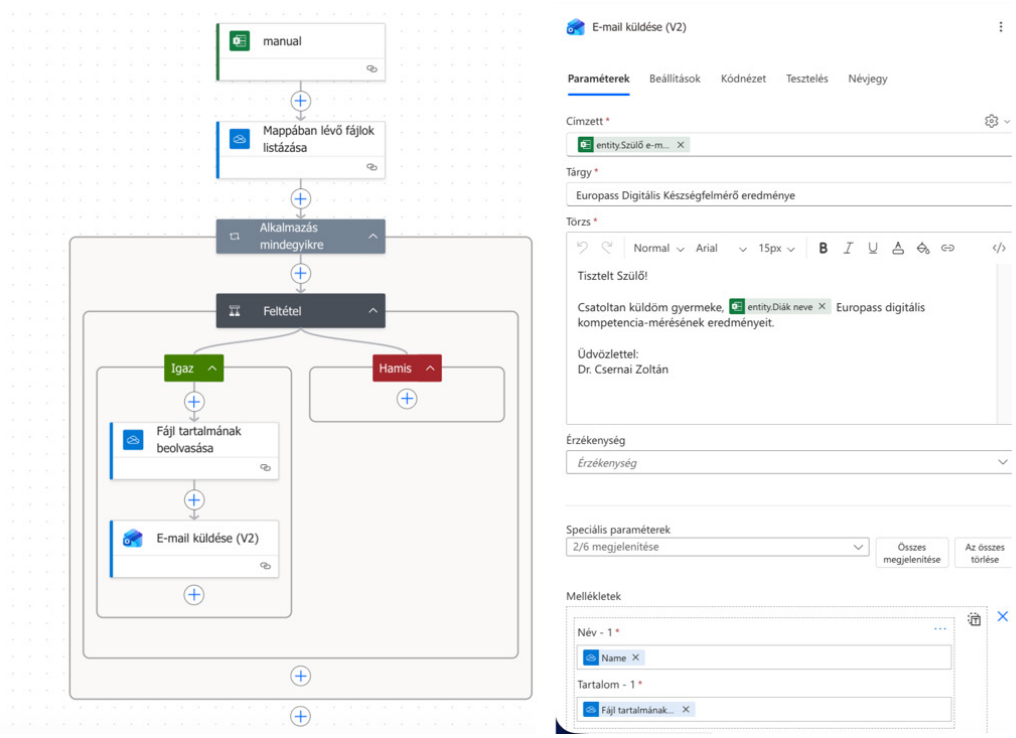
## 3. Az automatizáció gyakorlati megvalósítása: Pilot projektek

A kutatás során három különböző komplexitású munkafolyamat került megvalósításra, amelyek lefedik a pedagógiai adminisztráció legfontosabb területeit: az értékelést, a kommunikáció-szervezést és a dokumentum-feldolgozást.

### 3.1. Digitális Kompetencia-monitor: Automatizált és személyre szabott visszacsatolás

A bemutatott első projekt célja a tanulók digitális készségeinek Europass keretrendszer alapján történő nyomon követése és az eredmények szisztematikus eljuttatása az érintettekhez. A munkafolyamat bemeneti oldalán a diákok egy önértékelő tesztet töltenek ki, amelynek eredményét PDF formátumban töltik fel egy tanár által megosztott felhőalapú tárhelyre (Microsoft OneDrive). Az automatizáció bevezetése előtt ez a fázis jelentette a folyamat legidőigényesebb részét: minden egyes dokumentumot manuálisan kellett megnyitni, azonosítani a tanulót és a szülőt, majd egyéni e-mailek formájában továbbítani az adatokat.

A technikai megvalósítás során a Microsoft Power Automate platformja szolgált alapul. A rendszer adatforrásként egy központi Excel-munkafüzetet használ, amely tartalmazza a tanulók és szülők e-mail címeit, valamint a fájlok elérhetőségi útvonalait. A folyamat architektúrája biztosítja a korábban részletezett „human-in-the-loop" kontrollt: az automatizmus nem fut le automatikusan minden fájlra, hanem a pedagógus az Excel-táblázatban manuálisan jelöli ki a feldolgozandó sorokat, így ő határozza meg a kiküldés pontos idejét.




*1. ábra. A Digitális Kompetencia-monitor Power Automate munkafolyamata*
*A kép forrása: Saját forrás*

Az algoritmus lelke egy logikai párosítás: a rendszer listázza a felhőtárhelyen található dokumentumokat, majd egy „Alkalmazás mindegyikre" ciklus segítségével végigfut az állományokon. Egy feltételalapú vizsgálat során összeveti a mappában lévő fájlok nevét az adatbázisban tárolt egyedi azonosítókkal. Sikeres egyezés esetén a rendszer beolvassa a PDF-fájl tartalmát, és összeállítja a személyre szabott üzenetet. Az e-mail minden eleme (címzett, tárgy, megszólítás, diák neve) dinamikus paraméterként kerül beemelésre az adatforrásból. Ez a felépítés garantálja az abszolút pontosságot és kizárja a manuális hibázás lehetőségét.

### 3.2. Intelligens Pedagógiai Munkaszervező: NLP-alapú e-mail osztályozás és útválasztás

A második pilot projekt egy magasabb technológiai szintet képvisel, ahol a mesterséges intelligencia kognitív képességeit hívjuk segítségül a strukturálatlan információk feldolgozásához. Sok köznevelési intézményben a központi e-mail cím információs tölcsérként működik, ahol a beérkező üzenetek manuális szűrése és szortírozása jelentős időt és kognitív energiát emészt fel.

A megvalósítás során a Make.com platform és az OpenAI GPT-4 alapú természetes nyelvfeldolgozó (NLP) modelljének integrálására valósult meg. Ahhoz, hogy a mesterséges intelligencia „valódi munkatársként" tudjon részt venni a folyamatban, precíz rendszerszintű utasításokra volt szükség. A modell paraméterezése az OpenAI Playground felületén történt, ahol megtörtént az MI digitális személyiségének definiálása: a rendszer az „osztályfőnök digitális asszisztenseként" hozza meg döntéseit. A rendszerszintű utasítások kidolgozása során szigorú döntési logika került kialakításra: az MI a beérkező szövegeket négy kategóriába (Logisztika, Pályaorientáció, Szervezés vagy Egyéb) sorolja be.

A tévedések minimalizálása érdekében minden kategóriához konkrét példák kerültek hozzárendelésre a tanítási fázisban. Technikai szempontból meghatározó a kimeneti adatstruktúra rögzítése: a rendszer kizárólag érvényes JSON-formátumban válaszolhat, magyarázó szöveg nélkül. Ez a JSON-formátum biztosítja azt a technológiai hidat, amelyen keresztül a mesterséges intelligencia kognitív döntése konkrét gépi parancsokká alakítható.


*2. ábra. Az intelligens e-mail kategorizálás és útválasztás technikai architektúrája*
*A kép forrása: Saját forrás*

A munkafolyamat architektúrája a Make.com felületén egy „Gmail" modullal indul, amely meghatározott időközönként monitorozza az olvasatlan üzeneteket. Mivel egyszerre több levél is érkezhet, egy „Array Aggregator"[1] és egy „Iterator"[2] modul segítségével történik az adatok előkészítése. Ezek a modulok felelősek azért, hogy a levelekből csak a releváns tartalmat emeljék ki, és azokat egyenként, tisztított formában adják át az intelligens elemzésnek. Az MI által generált nyers JSON-objektumot egy „JSON Parser"[3] modul alakítja strukturált adatokká, amelyeket végül egy „Router" (útválasztó) modul kezel. Ez az útválasztó a beállított szűrők alapján a megfelelő ágra irányítja az üzenetet, automatikusan generálva a továbbítást a területért felelős kolléga e-mail címére.

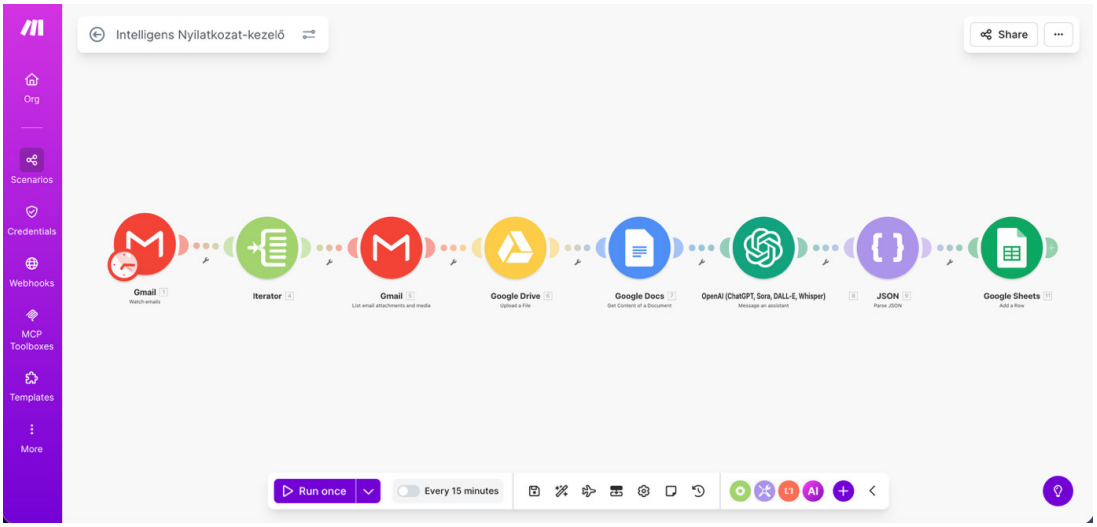
### 3.3. Intelligens Nyilatkozat-kezelő: Multimodális adatkinyerés és bináris döntéshozatal

A harmadik pilot projekt a köznevelési adminisztráció egyik legidőigényesebb és leginkább hibaérzékeny területét, a papíralapú szülői nyilatkozatok digitális feldolgozását célozza meg. A kiindulópontot olyan strukturálatlan adatok jelentik, mint a szkennelt PDF-dokumentumok vagy képformátumú fotók. Hagyományos úton ezek feldolgozása rendkívül nehézkes, mivel a tartalom manuális rögzítést igényel az intézményi adatbázisokba.

1 Az „Array Aggregator" elnevezésű tömb-aggregáló modul a különálló adategységeket egyetlen strukturált listába gyűjti össze a hatékonyabb, csoportos feldolgozás érdekében.

2 Az „Iterator" néven futó ismétlő modul a kapott adatlistát egyedi elemeire bontja, majd azokat egymás után, sorban továbbítja a következő munkafolyamat-fázisnak.

3 A „JSON Parser" funkcióját tekintve egy olyan adatértelmező összetevő, amely a mesterséges intelligencia által generált nyers, szöveges kódformátumot fordítja le géppel könnyen olvasható, rendszerezett és szűrhető egyedi adatmezőkké.




*3. ábra. A nyilatkozat-kezelő end-to-end munkafolyamata a Make.com felületén*
*A kép forrása: Saját forrás*

A megvalósítás során a Make.com felületén egy úgynevezett end-to-end automatizáció került kialakításra, amely a beérkező e-mail csatolmányoktól a végleges adatrögzítésig minden lépést emberi beavatkozás nélkül kezel. Bár a folyamatvázlat lineárisnak tűnik, a fejlesztés során a legnagyobb kihívást a szkennelt dokumentumok szöveges tartalmának precíz kinyerése jelentette. A megoldás kulcsa a „Google Drive" és a „Google Docs" modulok közötti szoros integráció: a rendszer a feltöltött állományokat automatikus konvertálással szerkeszthető dokumentummá alakítja. Ez a technológiai híd biztosítja, hogy a „Get content" funkcióval kinyerhető legyen a teljes szöveges állomány, amelyet a mesterséges intelligencia már értelmezni tud.

A folyamat során egy OpenAI GPT-alapú digitális asszisztens alkalmazására került sor, azonban a feladatkör itt jóval összetettebb. A multimodális modell feladata a karakterfelismert nyilatkozatok adatainak precíz kinyerése és rendszerezése. A technológia nemcsak a betűket ismeri fel, hanem értelmezi a dokumentum vizuális szerkezetét is, azonosítva a releváns mezőket (diák neve, szülő neve, dátum). A rendszerszintű utasításokban definiált döntési logika fókuszában a bináris döntéshozatal áll: a mesterséges intelligenciának meg kell állapítania, hogy a szülő hozzájárul vagy nem járul hozzá az adott tevékenységhez.

Külön kiemelendő a fejlesztés hibakezelési logikája: amennyiben a forrásdokumentum hiányos vagy olvashatatlan, a modell nem találgat, hanem „Ismeretlen" értékkel jelöli meg az adott mezőt. Ez a momentum alapozza meg a „human-in-the-loop" kontrollt, hiszen az oktató a kimeneti Google Táblázatban azonnal azonosíthatja a további beavatkozást igénylő eseteket. A folyamat zárásaként a „JSON Parser" által elemeire bontott adatok automatikusan, strukturált sorként kerülnek rögzítésre.

## 4. Eredmények és konklúziók

A bemutatott három pilot projekt tapasztalatai rávilágítanak arra, hogy a robotizált folyamatautomatizálás és a generatív mesterséges intelligencia integrált alkalmazása meghatározó változást hozhat a köznevelési adminisztrációban. A vizsgált folyamatok során elért 90% feletti időmegtakarítás messze túlmutat a puszta statisztikai eredményeken; ez a javulás közvetlenül járul hozzá az iskolai adminisztráció napi terhelésének radikális csökkenéséhez, ami áttételesen a pedagógusok tehermentesítését és közvetett jóllétét is támogatja. A kognitív kapacitások felszabadításával az oktatók és az intézményvezetés figyelme visszaterelhető az intézményi működés valódi prioritásaira: a diákok egyéni támogatására és a folyamatos szakmai-módszertani megújulásra.

A fejlesztési folyamat egyúttal igazolta a „human-in-the-loop" modell életképességét is, amely garantálja, hogy a technológia kizárólag támogató asszisztens maradjon. A mesterséges intelligencia előkészíti és rendszerezi az adatokat, de a végső szakmai döntéshozatal és a validáció minden ponton az emberi kontroll alatt marad. A kialakított architektúra emellett bizonyította, hogy a modern technológiai megoldások maradéktalanul összeegyeztethetők a szigorú GDPR adatvédelmi előírásokkal is. A jövőbeli kutatások számára a bemutatott keretrendszer intézményi szintű skálázása és a prediktív analitikai megoldások integrálása szolgálhat alapul, tovább segítve a fenntartható, modern digitális oktatási környezet kialakítását.

## 5. Felhasznált irodalom


Almeida, D., de Souza, M. C., da Silva, B. P., Paes, R. L., Soliman, M., Costa, S. B., & Bertochi, H. (2025). Agile Methodologies in the Educational Context: A Systematic Literature Review. *Sustainable Engineering and Innovation, 15*, 351–358.

Arunfred, N., Bini Marin, V., Shenbagaraman, V. M., & Saravanan, P. (2025). Teacher Readiness and Perceptions Towards AI-Based Smart Teaching Tools. *Frontiers in AI and Applications, 414*, 138–144.

Bazyl, O., Abilova, O., Karpenko, O., Mierienkov, H., & Poliakova, A. (2025). Assessing the impact of artificial intelligence integration on educational processes in higher education institutions of Ukraine and Kazakhstan. *Sustainable Engineering and Innovation, 7*(1), 97–116.

Bhardwaj, V., & Kumar, M. (2025). Transforming higher education with robotic process automation: Enhancing efficiency, innovation, and student-centered learning. *Discover Sustainability, 6*(1).

Chen, C.-H., Fei, H.-Y., & Tsai, C.-C. (2026). Hierarchical analysis of in-service teachers' barriers to technology-integrated instruction: A review of 2000–2024 publications. *Computers and Education*, 242.

Dave, B., Martin, P., David, S. S., Kumar, S., & Chakraborty, T. (2026). Enhancing healthcare worker mental health via artificial intelligence-driven work process improvements: A scoping review. *International Journal of Medical Informatics, 205*.

Delello, J. A., Sung, W., Mokhtari, K., Hebert, J., Bronson, A., & De Giuseppe, T. (2025). AI in the Classroom: Insights from Educators on Usage, Challenges, and Mental Health. *Education Sciences, 15*(2).

Halder, D., Al Bastaki, E. M., Suleymanova, S., Muhammad, N., & Purushothaman, A. (2024). Agile Blended Learning: A Promising Approach for Higher Education in the UAE. *SN Computer Science, 5*(5).

Jacobs, J., Scornavacco, K., Clevenger, C., Suresh, A., & Sumner, T. (2024). Automated feedback on discourse moves: Teachers' perceived utility of a professional learning tool. *Educational Technology Research and Development, 72*(3), 1307–1329.

Marchena Sekli, G. F., & Godo, A. (2025). Transforming Teaching and Learning with Robotic Process Automation: A Systematic Review of Pedagogical Applications. *Journal of Information Technology Education: Research, 24*.

Orsdemir, A., Tilki, G., & Altinay, F. (2019). Evaluation by Teachers of "Use of Influence in Agile Management" by School Administration. *International Journal of Disability, Development and Education, 66*(6), 577–589.

Pathania, M., Singh, C. P., Kaur, D. P., & Mantri, A. (2025). Effects of Self-Adaptive Approach of Iterative Game Based Learning on Performance and Satisfaction of Elementary School Students in Mathematics: An Action Research Field Experiment. *SN Computer Science, 6*(5).

Reid, P. (2017). Supporting instructors in overcoming self-efficacy and background barriers to adoption. *Education and Information Technologies, 22*(1), 369–382.

Sutipitakwong, S., & Jamsri, P. (2020). The effectiveness of RPA in fine-tuning tedious tasks. Scopus. Proceedings - 2020 6th International Conference on Engineering, *Applied Sciences and Technology, ICEAST 2020*.

DOI: 10.31915/NWS.2026.12



# MEGÚJULÁS UTÁN TOVÁBBFEJLESZTVE: KERESÉSI FUNKCIÓ ÉS PUBLIKÁCIÓFELTÖLTŐ ŰRLAP FEJLESZTÉSE A TUDÓSTÉRBEN

# FURTHER DEVELOPMENT AFTER THE RENEWAL: ENHANCING SEARCH AND THE PUBLICATION SUBMISSION FORM IN TUDÓSTÉR

*Bor Balázs*
*Debreceni Egyetem Egyetemi és Nemzeti Könyvtár*
*bbor@lib.unideb.hu*
*ORCID: 0000-0001-8141-1896*

## Absztrakt

A tanulmány a Debreceni Egyetem Egyetemi és Nemzeti Könyvtára által működtetett Tudóstér 2025 eleji technológiai megújulását követő fejlesztési szakaszt mutatja be. Két, a mindennapi kutatástámogatási gyakorlatban meghatározó területre fókuszál: a keresési funkció célzott továbbfejlesztésére és a publikációfeltöltő űrlap megújítására. Az új keresési működés a korábbi, általános bibliográfiai visszakereséshez képest jobban támogatja a publikációs teljesítmény áttekintését, a listák és kimutatások összeállítását, valamint az adatok gyors ellenőrzését. A publikációfeltöltő űrlap fejlesztése a korábbi, nehezebben továbbfejleszthető technológiai megoldás kiváltásával és egy szolgáltatásorientáltabb adminisztrációs workflow kialakításával jár. A tanulmány azt mutatja be, hogyan épülhetnek be az éles használat tapasztalatai a további fejlesztésekbe, és hogyan tartható fenn a szolgáltatás stabil működése a változtatások mellett.

**Kulcsszavak:** kutatástámogatás; Tudóstér; repozitórium; Open Science; Elasticsearch; intézményi CRIS

## Abstract

This paper presents the development phase following the technological renewal of Tudóstér, the research support platform operated by the University and National Library of the University of Debrecen, in early 2025. It focuses on two areas central to everyday research support practice: the targeted improvement of the search function and the renewal of the publication submission form. Compared to the former solution, which primarily supported general bibliographic retrieval, the new search functionality better supports the overview of publication output, the creation of lists and reports, and quick data verification. The renewed submission form replaces a less easily maintainable tech-

nological solution and introduces a more service-oriented administrative workflow. The paper shows how experience from live operation can inform further development, while maintaining the stable operation of the service during ongoing changes.

**Keywords**: research support; Tudóstér; repository; Open Science; Elasticsearch; institutional CRIS

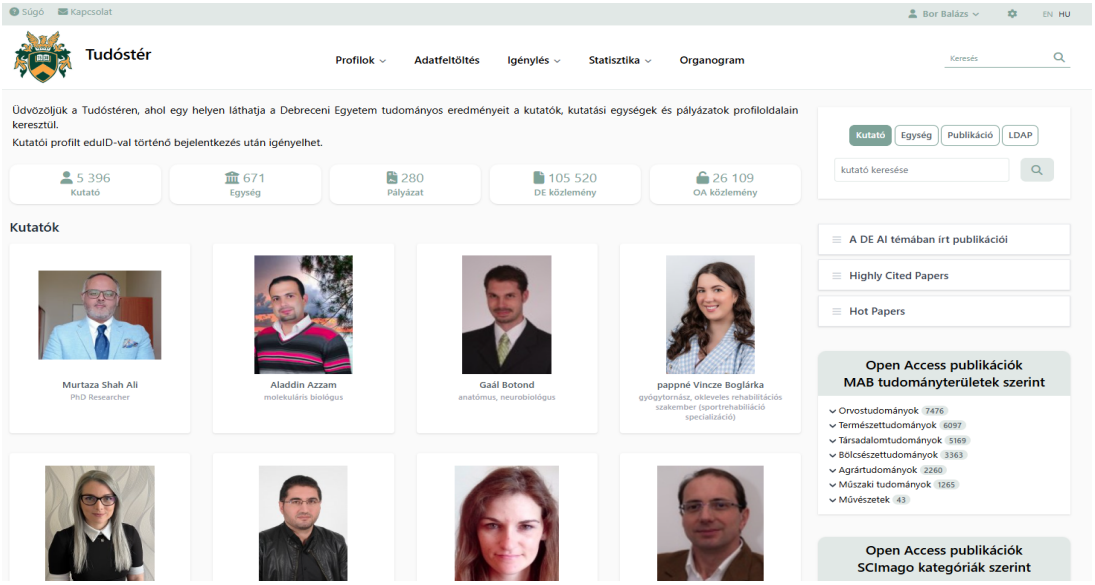
## Bevezetés

A Debreceni Egyetem publikációs és kutatástámogatási szolgáltatásai hosszú intézményi fejlődési folyamat eredményeként alakultak ki. A Tudóstér mai formája nem önállóan létrejött informatikai alkalmazás, hanem több évtizedes bibliográfiai, repozitóriumi és szolgáltatásszervezési tapasztalat folytatása. A rendszer célja, hogy a kutatói profilok, a publikációs adatok, a teljes szövegű dokumentumok és a kapcsolódó könyvtári szolgáltatások egykapus, átlátható környezetben jelenjenek meg.[1]

A Networkshop 2025 konferencián elhangzott előadás és az abból készült tanulmány a Tudóstér stratégiai szerepét, valamint az egykapus kutatástámogatási ökoszisztéma felépítését mutatta be.[2] Jelen tanulmány erre az előzményre építve a megújult Tudóstér két gyakorlati fejlesztési területét vizsgálja: az integrált keresési infrastruktúrát és a publikációfeltöltő űrlap továbbfejlesztését. A hangsúly nem pusztán a technológiai megoldáson van, hanem azon is, hogy az új rendszer miként teszi egyszerűbbé a későbbi fejlesztéseket, a szolgáltatások összekapcsolását és az adminisztratív feldolgozást.

1 Tudóstér. https://tudoster.unideb.hu/ [Megtekintve: 2026.06.18.]

2 Bor B., Nyitrai N., Fazekas-Paragh J. (2025): Egykapus kutatástámogatás a gyakorlatban: A Tudóstér új dimenziói. Networkshop 2025, 39-46. DOI: 10.31915/NWS.2025.5




*1. ábra: A Tudóstér nyitófelülete*

## Történeti előzmények és fejlesztési igények

A Debreceni Egyetem publikációs adatvagyonának kezelése a papíralapú egyetemi bibliográfiák hagyományára épült. Ezek a kiadványok intézményi szinten dokumentálták a kutatói teljesítményt, de használatuk és aktualizálásuk a nyomtatott forma korlátaihoz kötődött. A 90-es évek közepétől a publikációs adatok egyre inkább a Corvina integrált könyvtári rendszerben, MARC-rekordokként jelentek meg. Ez a lépés a bibliográfiai adatok egységesebb, kereshetőbb és könyvtári szempontból ellenőrizhetőbb kezelését tette lehetővé.

A következő jelentős állomás a Debreceni Egyetem elektronikus Archívumának, a DEA-nak az indulása volt 2007-ben. A repozitóriumi környezet a teljes szövegű dokumentumok intézményi archiválását és szolgáltatását biztosította. 2008-tól már feltöltőűrlap is segítette a publikációk beküldését, amire azért volt szükség, mert a publikációból egyrészt katalógusrekord készült, másrészt a dokumentum teljes szövege minősített Dublin Core metaadatokkal a repozitóriumba került. Az űrlap ezt a két irányt kapcsolta össze: egyszerre támogatta a bibliográfiai rekord létrehozását és a repozitóriumi feltöltést. A több helyen kezelt metaadatokat és magát a dokumentum teljes szövegét a 2012-től Drupal-alapon működő Tudóstér fogta össze, amely kutatói profilokkal, intézményi megjelenítéssel és publikációs listákkal bővítette ezt az ökoszisztémát, és a kutatók számára könnyebben használható portálfelületet kínált.

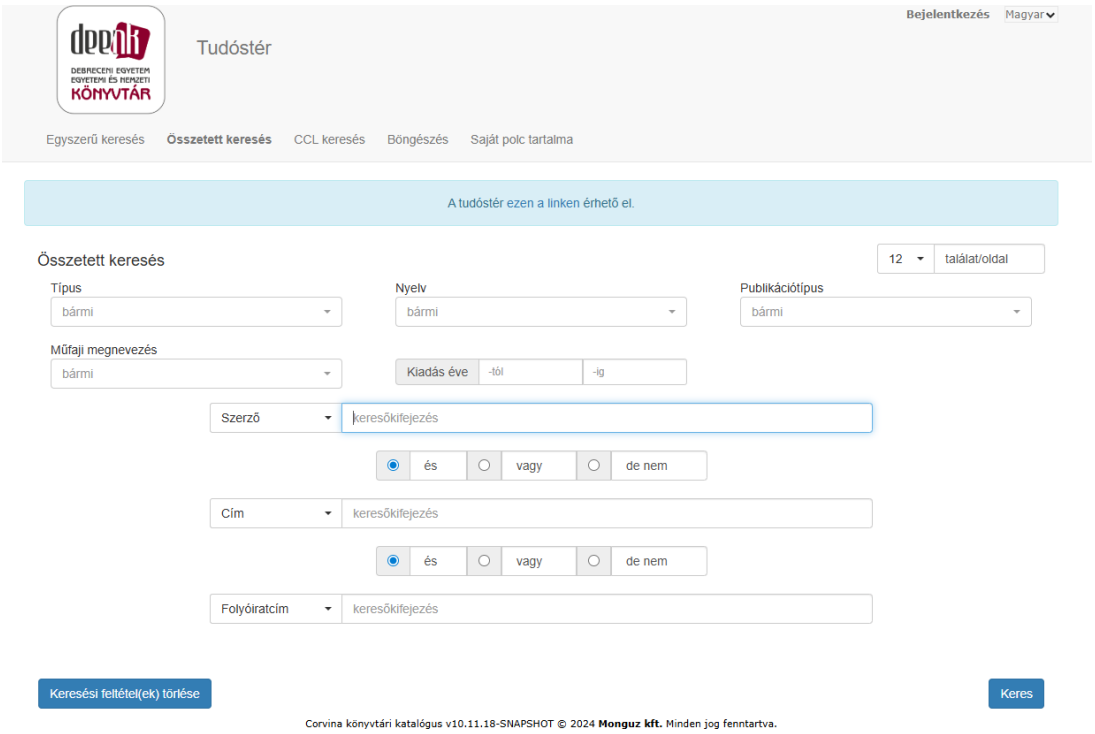
A Drupal-alapú működés hosszú ideig megfelelő keretet adott, de a technológiai környezet változása, a kutatói igények növekedése és a szolgáltatási folyamatok összetettebbé válása miatt egyre erősebben jelentkezett az újratervezés igénye. A Tudóstér 2025 februárjától, egy hosszabb fejlesztőmunka eredményeként Laravel-alapokra került. Az új architektúra

nemcsak modernebb felületet és stabilabb működést biztosít, hanem olyan fejlesztési alapot is, amelyen a keresés, a feltöltés és az adminisztrátori munkafolyamatok könnyebben továbbépíthetők.

A fejlesztési igények között kiemelt szerepet kapott a felhasználói belépési pont egyszerűsítése, a kézi adatmozgatás csökkentése, az adminisztrátori feldolgozás átláthatóbbá tétele és az adatok későbbi kereshetőségének javítása. A cél nem a könyvtári szakmai ellenőrzés kiváltása volt, hanem annak jobb előkészítése és támogatása.

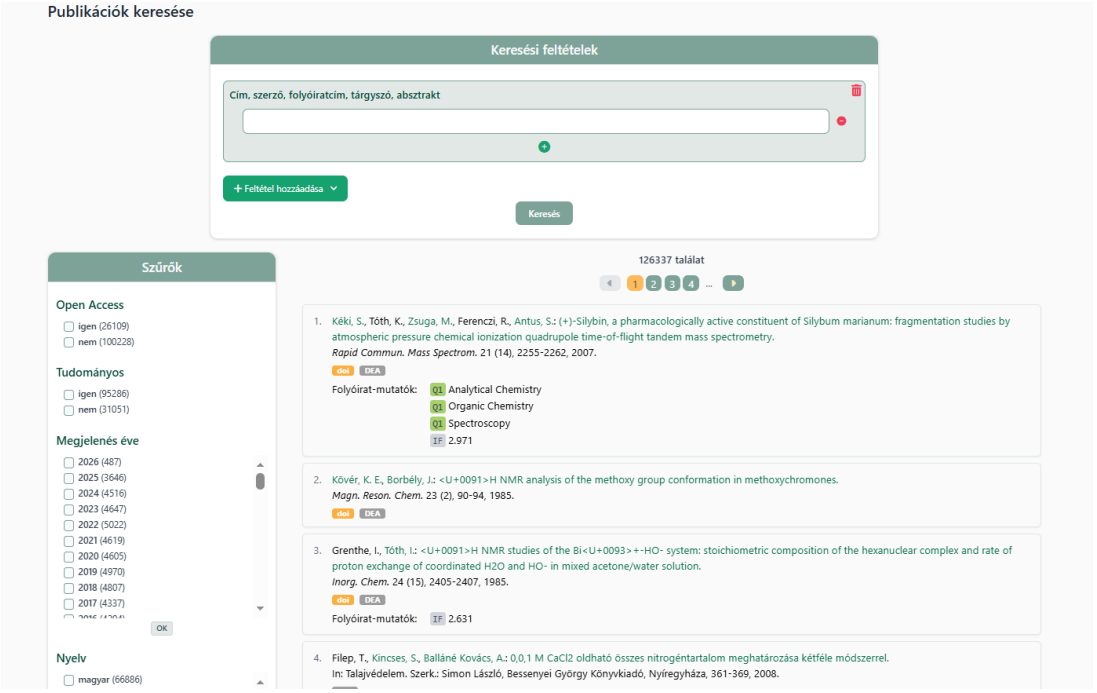
## Az integrált keresési infrastruktúra

A megújult Tudóstér egyik, már a felhasználói oldalon is közvetlenül érzékelhető fejlesztési területe a keresés volt. Az új technológiai alapok lehetővé tették, hogy a korábbi háttérbeli modernizáció után olyan funkciók is hatékonyan továbbfejleszthetők legyenek, amelyek a mindennapi használatban is érzékelhető változást hoznak. A korábbi kereső, amely a Corvinában kezelt publikációs adatbázis webes katalógusfelületeként működött, jól használható alapot adott az általános bibliográfiai visszakereséshez, de a kutatástámogatási szempontok nehezebben jelentek meg benne: a releváns részhalmaz eléréséhez több lépésre volt szükség, a találatok kevésbé közvetlenül kapcsolódtak a Tudóstér szolgáltatásaihoz, és a napi ellenőrzési feladatokhoz is kevésbé célzott visszakeresést tett lehetővé.




*2. ábra: A korábbi Corvina-alapú összetett kereső felülete*

Az új keresési működés központi eleme az Elasticsearch-alapú réteg, amely nem kizárólag a keresőfelületet szolgálja ki, hanem a Tudóstér több megjelenítési és lekérdezési funkciójának közös adatlekérdezési alapja. A Laravel-alapú felület a Corvina, a DEA és a Tudóstér saját adatbázisának adatait aggregált formában használja fel, majd ezekből Elasticsearch-indexek épülnek. Az egyik index JSON-szerkezetű publikációs rekordokat, egy másik authority jellegű, például szerzői adatokat tartalmaz. A kutatói profilok publikációs listái, a statisztikai lekérdezések és a találati megjelenítések is erre az adatstruktúrára támaszkodnak. Az új keresőfelület erre a már meglévő lekérdezési rétegre épül rá, és azt teszi felhasználói oldalon irányíthatóvá: a keresési feltételek és szűkítések módosításával közvetlenül alakítható, milyen részhalmaz jelenjen meg az Elasticsearch-indexekből.

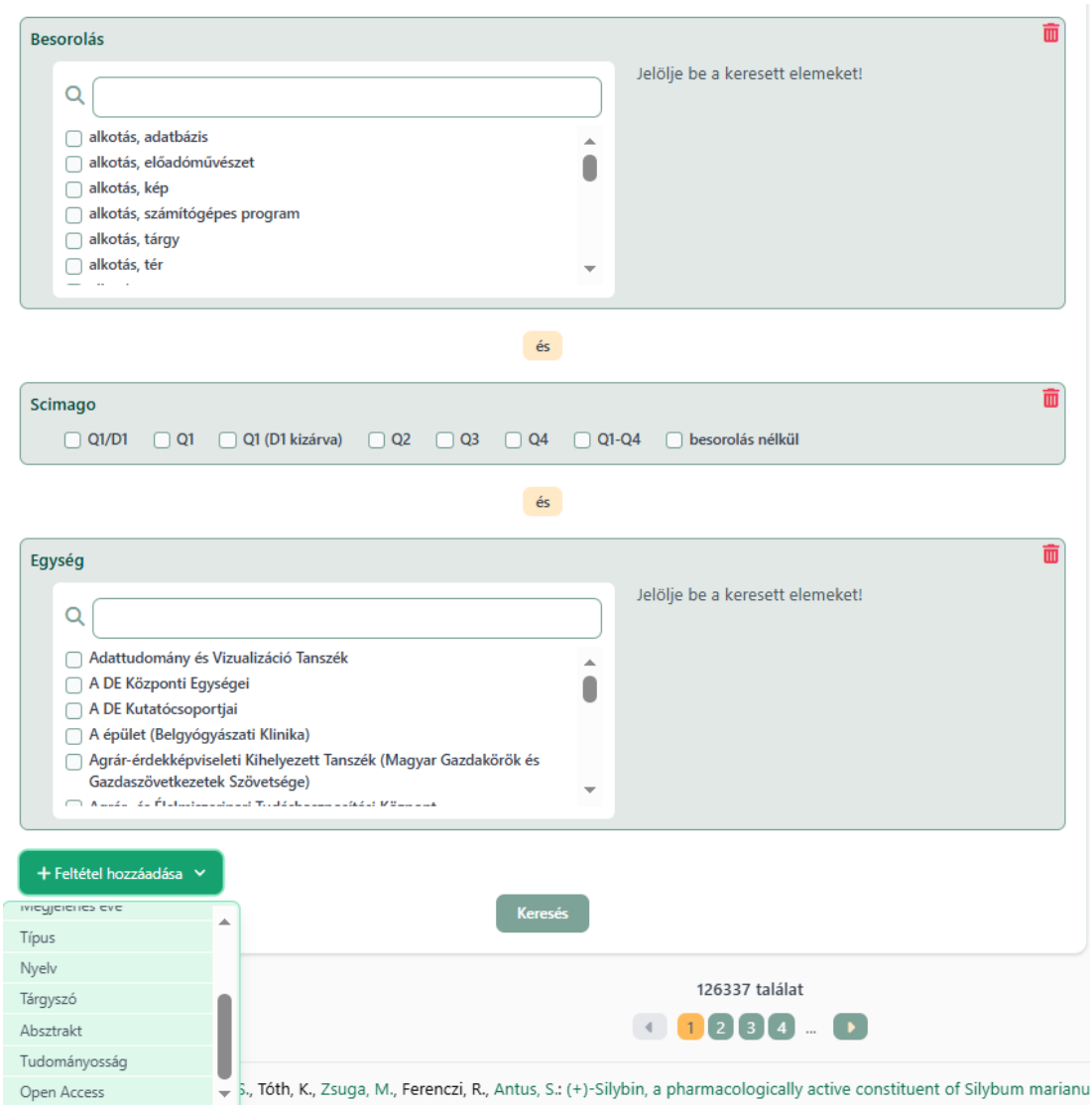



*3. ábra: Az új Tudóstér publikációkereső találati és szűrési felülete*

A keresési infrastruktúra egyik fontos eredménye, hogy a publikáció nem pusztán bibliográfiai tételként jelenik meg, hanem összetett rekordként, amelyben a könyvtári, repozitóriumi és szolgáltatási információk együtt használhatók. A DEA-ból átvett adatok között szerepelhet például az open access státusz vagy a tudománymetriai mutatókkal kapcsolatos információk. Ezek az adatok nemcsak a megjelenítést gazdagítják, hanem a keresési és szűrési lehetőségeket is bővítik.

Az új felület a szűkítést is szolgáltatásközelibbé teszi. Az összetett keresési feltételek és a facettás szűrés együttesen használhatók, így a találati lista tudományterület, szervezeti

egység, év vagy besorolás szerint gyorsan releváns részhalmazra szűkíthető. A keresés logikája ezáltal nemcsak bibliográfiai kérdésekre, hanem konkrét kutatástámogatási és adminisztrátori ellenőrzési helyzetekre is reagál.




*4. ábra: Összetett keresési feltételek és szűkítések az új keresőben*

## A publikációfeltöltés korábbi logikája

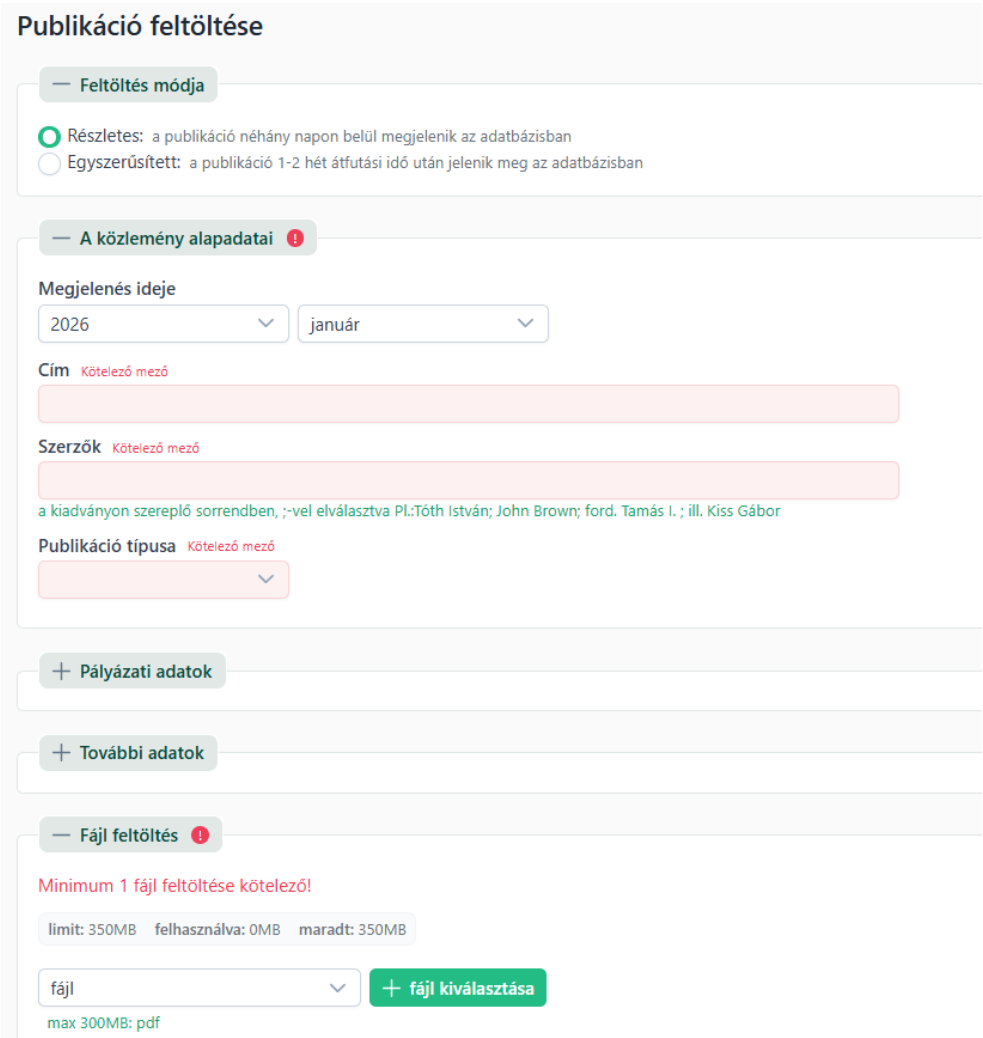
A teljes szövegű publikációk beküldésének intézményi támogatása már a korábbi feltöltőűrlappal is működött. A korábbi, Google Web Toolkit (GWT) technológiával készült feltöltőűrlap hosszú ideig kiszolgálta a folyamatot, de a megújítás idejére egyre nehezebben volt továbbfejleszthető és fenntartható. A kötöttebb felépítés, a technológiai karbantarthatóság korlátai és a munkafolyamat támogatásához rendelkezésre álló szűkebb mozgástér miatt a változtatások bevezetése körülményesebbé vált.

A korábbi folyamatban ugyanakkor már jelen volt az a szakmai logika, amelyet az új rendszernek meg kellett őriznie. A feltöltés során a rendszer rekordazonosítókat kért a Corvinától, a dokumentum teljes szövege a DEA-ba került, onnan visszaérkezett a handle azonosító, majd ez beépült a megfelelő MARC-rekordba, és megtörtént a Corvina-betöltés, valamint a két rendszer azonosítókon keresztüli összekapcsolása. A fejlesztés ezért nem a folyamat megszüntetéséről, hanem annak korszerűbb, követhetőbb és rugalmasabb újraépítéséről szólt.

## Az új publikációfeltöltő űrlap és az adminisztráció

Az új publikációfeltöltő űrlap fejlesztésének kiindulópontja az volt, hogy a kutató számára a teljes szöveg beküldése egyszerű, szolgáltatásszerű folyamatként jelenjen meg, miközben a háttérben a könyvtári feldolgozás számára minden szükséges adat strukturáltan keletkezzen. A feltöltés tehát nem egyszerű fájlküldés, hanem egy repozitóriumi és bibliográfiai feldolgozási folyamat első lépése.

Az új űrlap fejlesztése nem elsősorban formai átalakítást jelentett, hanem a korábbi működés részletes feltérképezésére épülő, szakaszos fejlesztési és tesztelési folyamatot. Az átláthatóbb szerkezet és blokkos felépítés jobban követhető adatbevitelt tesz lehetővé, miközben a rugalmasabb fejlesztési környezet a szolgáltatási munkafolyamathoz is jobban illeszkedik. A cél egyszerre volt a kutatói és a könyvtárosi használhatóság javítása. Az űrlap a kézirat készítésekor tesztfázisban működik, éles bevezetése a tesztelés lezárását követően, várhatóan 2026 nyár végén történhet meg.




*5. ábra: Az új publikációfeltöltő űrlap blokkos szerkezete*

Az űrlaphoz kapcsolódó adminisztrációs felület egyik fontos újdonsága, hogy a beküldött adatok nem kerülnek azonnal betöltésre sem a Corvinába, sem az intézményi repozitóriumba. A feltöltés először a Tudóstérben tárolódik: itt kezelhető a teljes szövegű dokumentum és a Corvina felé előkészített, MARCXML-formátumban[3] tárolt bibliográfiai rekord. A rendszer továbbra is két irányt kapcsol össze: a publikációból MARC-alapú katalógusrekord készül, a teljes szöveg pedig Dublin Core-alapú repozitóriumi metaadatokkal a DEA-ba kerül. A kapcsolatot a katalógusazonosító és a repozitóriumi handle biztosítja.

A Tudóstérben tárolt rekorddal többféle művelet végezhető: ellenőrizhető, javítható, szükség esetén pedig az adatok visszaadhatók az űrlapos felületnek további feldolgozásra. Ez

3 Library of Congress (2026): MARC 21 XML Schema. https://www.loc.gov/standards/marcxml/ [Megtekintve: 2026.06.18.]

azért jelent előrelépést, mert bizonyos javítások nem igényelnek MARC-formátumot ismerő feldolgozókönyvtárosi gyakorlatot, hanem a kutatástámogatási munkában részt vevő kollégák is elvégezhetik őket. A javítások után indítható a rekord betöltése a Corvinába és a teljes szöveg elhelyezése a DEA-ban. Így a rendszer megőrzi az integrált könyvtári rendszer és az intézményi repozitórium előnyeit, miközben a munkafolyamat a helyi kutatástámogatási szempontokhoz is jobban igazíthatóvá válik.

## Szolgáltatási hatások és további fejlesztési irányok

A Tudóstér fejlesztése ott válik igazán szolgáltatásfejlesztéssé, ahol a technológiai változás a napi könyvtári és kutatói munkafolyamatokban is érzékelhetővé válik. A kutató számára az új rendszer egyszerűbb felületet és követhetőbb ügymenetet kínál a publikációkhoz kapcsolódó feladatokhoz, a rögzítéstől a visszakeresésen és a teljes szöveg beküldésén át az intézményi kutatástámogatási ökoszisztémában elérhető szolgáltatások igénybevételéig. A folyamatot támogató könyvtáros az automatizálható lépések magasabb szintű támogatásával segítheti a kutatói munkát, miközben a strukturáltabb munkafolyamat a könyvtári oldalon is rugalmasabb feldolgozást tesz lehetővé.

A további fejlesztési irányok között szerepel az ORCID-integráció megvalósítása, valamint a Tudóstérben elérhető teljes szövegű tartalmak jobb hasznosítása. A rendszer egyik erőssége, hogy egyre több publikációhoz kapcsolódik nyíltan elérhető dokumentum, ezért fontos cél a tartalmi feltárás részletesebbé tétele. Mivel ez jelentős munkaráfordítást igényel, vizsgáljuk annak lehetőségét, hogy a metaadatolási és tartalmi feltárási folyamatok egy része automatizálható-e, akár mesterséges intelligencia bevonásával.

A fejlesztések másik iránya az adatkapcsolatok erősítése. Mivel a Tudóstér jelenleg nem tartalmaz önálló hivatkozási adatokat, vizsgálható a hivatkozások beemelése nagy tudományos adatbázisok (például a Scopus, a Web of Science vagy az MTMT) API-kapcsolatain keresztül. Emellett fontos együttműködési irány az MTMT-vel való kapcsolat fejlesztése is, mivel a könyvtár vállalja, hogy a Tudóstérbe feltöltött publikációkat a kutatástámogatással foglalkozó könyvtárosok az MTMT-be is áttöltik. A jelenleg részben automatizált folyamat hosszabb távon API-alapú adatátadással válhatna hatékonyabbá.

## Összegzés

A Tudóstér új keresési infrastruktúrája és publikációfeltöltési munkafolyamata egy hosszabb intézményi fejlődési ív legújabb állomása. A papíralapú egyetemi bibliográfiáktól a Corvina MARC-rekordjain, a DEA repozitóriumi szolgáltatásain, a 2008-tól működő feltöltőűrlapon és a 2012-től Drupal-alapú Tudóstéren át vezetett az út a 2025 februárjától Laravel-alapokon működő rendszerig.

A fejlesztés tapasztalata alapján egy intézményi CRIS-rendszer értéke nem önmagában

az új technológiában mutatkozik meg, hanem abban, hogy a korábbi könyvtári adatkezelési gyakorlatokat, a repozitóriumi szolgáltatásokat és a kutatói igényeket alakítható szolgáltatási térbe rendezi. A Tudóstér megújítása ezért nem lezárt projektként, hanem továbbépíthető szolgáltatási alapként értelmezhető, amelyre újabb intézményi igények és kutatástámogatási feladatok építhetők. A cél olyan rendszer fenntartása, amely az egyetemi kutatástámogatási ökoszisztéma részeként hosszú távon is képes követni a változó igényeket, és alkalmazkodni az intézményi működés, a technológia és a tudományos kommunikáció átalakulásához.

## Irodalom


Bor B., Nyitrai N., Fazekas-Paragh J. (2025): Egykapus kutatástámogatás a gyakorlatban: A Tudóstér új dimenziói. In: Oktatási, kutatási és közgyűjteményi infrastruktúrák és tartalmak: digitális transzformáció felsőfokon. NETWORKSHOP 2025. Hungarnet Egyesület, Budapest, 39-46. DOI: 10.31915/NWS.2025.5

Elastic (2026): Elasticsearch documentation. Elastic. https://www.elastic.co/docs

Library of Congress (2026): MARC 21 XML Schema. https://www.loc.gov/standards/marcxml/

DOI: 10.31915/NWS.2026.13



# A KÖNYVTÁR MINT IRÁNYTŰ A MESTERSÉGES INTELLIGENCIA KORÁBAN

## *ÚJ SZEREPEK, KOMPETENCIÁK ÉS OKTATÁSTÁMOGATÓ FUNKCIÓK A DIGITÁLIS TUDÁSTÁRSADALOMBAN*

# THE LIBRARY AS A COMPASS IN THE AGE OF ARTIFICIAL INTELLIGENCE: NEW ROLES, COMPETENCIES, AND EDUCATIONAL SUPPORT FUNCTIONS IN THE DIGITAL KNOWLEDGE SOCIETY

*Gerencsér Judit*
*doktorandusz, Eszterházy Károly Katolikus Egyetem*

*elnök, Magyar Könyvtárosok Egyesülete*
*stratégiai és nemzetközi kapcsolatok szakértő,*

*MNMKK Magyar Nemzeti Múzeum*

*ORCID: 0009-0001-9293-6599*

### Bevezetés

A mesterséges intelligencia gyors ütemű fejlődése alapvető változásokat idéz elő a könyvtárak működésében, stratégiai szerepében és társadalmi küldetésében. A könyvtárak – amelyek hagyományosan az emberi tudás megőrzésének és közvetítésének meghatározó intézményei voltak – napjainkban olyan technológiai és társadalmi környezetben működnek, amely szükségessé teszi funkcióik és kompetenciarendszerük újraértelmezését. A mesterséges intelligencia nemcsak az információhoz való hozzáférés és a tudáselőállítás módját alakítja át, hanem a szolgáltatások, az oktatás, a kutatás és a tanulási folyamatok szerkezetére is jelentős hatást gyakorol. Ennek következtében a könyvtáraknak nem csupán működésüket és technológiai infrastruktúrájukat kell fejleszteniük, hanem új szemléleti keretben kell meghatározniuk saját szerepüket, feladataikat a digitális átállás során.

Az tanulmány négy egymással szorosan összefüggő tematikus terület köré szerveződik. Vizsgálja a 21. századi munkaerőpiaci trendeket és kompetenciaelvárásokat, elemzi a könyvtárak és könyvtárosok átalakuló szakmai szerepeit, bemutatja a könyvtárak oktatástámogató funkcióinak változását, valamint foglalkozik a könyvtárosok technológiai felké-

szültségének és digitális kompetenciáinak kérdéskörével. A tanulmány célja annak bemutatása, hogy a könyvtárak miként válhatnak a digitális korszak emberközpontú, adaptív és etikus tudásközpontjaivá.

**Kulcsszavak**: mesterséges intelligencia, Library 5.0, digitális transzformáció, digitális kompetenciák, MI-műveltség, információs műveltség, könyvtárak, könyvtáros szakma, oktatástámogatás, élethosszig tartó tanulás, technológiaelfogadás, ember–MI együttműködés.

## Abstract

The rapid development of artificial intelligence is bringing about fundamental changes in the operation, the strategic role and the social mission of libraries. Traditionally regarded as key institutions for the preservation and dissemination of human knowledge, libraries now operate within a technological and social environment that requires the redefinition of their functions and competency frameworks. Artificial intelligence is transforming not only the ways in which information is accessed and knowledge is produced, but also the structure of services, education, research and learning processes. As a result, libraries must not only develop their services and technological infrastructures, but also redefine their role within a new conceptual framework shaped by the ongoing digital transformation.

The study is organized around four closely interconnected thematic areas. It examines 21st-century labour market trends and competency expectations, analyses the changing professional roles of libraries and librarians, explores the evolving educational support functions of libraries and addresses the issues of technological preparedness and digital competencies among librarians. The aim of the study is to demonstrate how libraries can become human-centred, adaptive and ethical knowledge hubs in the digital age.

**Keywords**: artificial intelligence, Library 5.0, digital transformation, digital competencies, AI literacy, information literacy, libraries, librarianship, educational support, lifelong learning, technology acceptance, human–AI collaboration.

## Munkaerőpiaci trendek és a 21. századi kompetenciák

A 21. századi munkaerőpiacot egyre erőteljesebben formálja a technológiai fejlődés gyors üteme, a folyamatos digitális transzformáció, valamint a kompetenciaelvárások jelentős átalakulása (World Economic Forum, 2023; OECD, 2023; Mhaske et al., 2025). A 21. század munkaerőpiacán való sikeres helytállást már nem kizárólag a szakmai és technikai tudás határozza meg, hanem olyan komplex kompetenciák együttese, amelyek magukban foglalják a digitális, kognitív, szociális és érzelmi készségeket is (CEDEFOP, 2024). A gyorsan változó társadalmi és technológiai környezetben kiemelt jelentőségűvé vált az alkalmazkodóképesség, a folyamatos tanulás, valamint az interdiszciplináris együttműködésre való nyitottság, amelyek a modern munkaerőpiac alapvető elvárásaiként jelennek meg (OECD, 2023; OECD, 2025).

A mesterséges intelligencia térnyerése által a munkaerőpiac egyre inkább az ember és a technológia együttműködésére épül. Az OECD elemzése szerint a mesterséges intelligencia elsősorban nem megszünteti a munkaköröket, hanem átalakítja az egyes feladatokat és a szükséges kompetenciákat (Green, 2024). Ennek következtében felértékelődnek azok az emberi készségek, amelyeket az MI nehezen képes helyettesíteni, így különösen a kreativitást, a kritikus gondolkodást, az empátiát, az etikai érzékenységet és az összetett problémamegoldást.

A World Economic Forum előrejelzései szerint a jövőálló karrier alapja nem csak szakmai tudás, hanem a tanulási képesség, az alkalmazkodás és a MI-műveltség, vagyis a mesterséges intelligencia működésének és társadalmi hatásainak megértése (World Economic Forum, 2025). A technológiai tudás és a humán kompetenciák integrációja egyre inkább meghatározza a szakmai sikerességet a digitális korban.

A nemzetközi trendek szerint a technológiai fejlődés üteme sok esetben gyorsabb, mint az emberi alkalmazkodásé, ami bizonytalanságot, mentális terhelést és szervezeti feszültségeket eredményezhet. A kutatások ugyanakkor arra is rámutatnak, hogy a mesterséges intelligencia önmagában nem képes értéket teremteni: a sikeres digitális transzformáció alapja az ember és az MI együttműködése, amelyben a technológiai rendszerek hatékony működését az emberi értelmezés, az etikus döntéshozatal és a folyamatos tanulás támogatja (Harvard Business Review, 2026).

## A könyvtárak és könyvtárosok új szerepei a 21. századi trendek tükrében

A mesterséges intelligencia térnyerése alapvetően újra értelmezi a könyvtárak és a könyvtárosok szerepét a digitális tudástársadalomban. Az a kérdés egy egyetemi környezetben, hogy „ha mindent meg lehet kérdezni az MI-től, miért van még szükség könyvtárra?", valójában nem pusztán technológiai, hanem társadalmi, oktatási és kulturális kérdés is. A könyvtárak hagyományos információközvetítő funkciója mellett egyre hangsúlyosabbá válik az emberközpontú tudásközvetítés, a hiteles információk biztosítása, valamint a kritikus gondolkodás és az információs műveltség fejlesztése. A Library 5.0 szemlélet a könyvtárakat olyan etikus, inkluzív és közösségi tudásökoszisztémákként értelmezi, amelyekben a technológia nem helyettesíti az embert, hanem támogatja az emberi fejlődést, a közösségi részvételt és a tanulási folyamatokat. A koncepció középpontjában az ember és a technológia közötti egyensúly, valamint a közösségi és kulturális értékek megőrzése áll (Passalacqua et al., 2025).

A nemzetközi trendek azt mutatják, hogy a könyvtárak szerepe egyre inkább túlmutat a hagyományos gyűjteményszervezési és dokumentumszolgáltatási funkciókon. A digitális transzformáció, a mesterséges intelligencia és az adatalapú működés következtében a könyvtárak a digitális írástudás, az Open Science, az élethosszig tartó tanulás és a közösségi részvétel meghatározó intézményeivé válnak. Az IFLA Trend Report 2024 kiemeli, hogy a

tudásgyakorlatok átalakulnak, az információs rendszerek komplexebbé válnak, miközben a társadalomban növekszik az igény a hiteles információkra és a közösségi kapcsolódásra (IFLA, 2024). Ebben a környezetben a könyvtárosok szerepe is kibővül: a jövő könyvtárosa egyszerre digitális mentor, oktató, technológiai közvetítő és etikai facilitátor, aki támogatja a felhasználókat az egyre összetettebb digitális és mesterséges intelligenciával támogatott információs környezetekben.

## A könyvtárak új feladatai az oktatás támogatásában

A 21. századi oktatás alapvető átalakuláson megy keresztül, amelyet a digitalizáció, a mesterséges intelligencia térnyerése és a személyre szabott tanulási modellek terjedése egyaránt meghatároz. A tanulás folyamata egyre kevésbé kötődik kizárólag a hagyományos tantermi keretekhez, miközben előtérbe kerülnek a rugalmas, interaktív és hibrid tanulási környezetek. Ebben a megváltozott oktatási ökoszisztémában a könyvtárak szerepe jelentősen felértékelődik: a könyvtárak már nem pusztán információforrások, hanem olyan komplex tanulási és közösségi terek, amelyek támogatják a digitális kompetenciák, az információs műveltség, valamint a kritikus gondolkodás fejlődését. A könyvtárak a hagyományos pedagógiai megközelítések és az innovatív, technológia-alapú tanulási formák közötti híd szerepét is betöltik, miközben biztonságos és inkluzív környezetet biztosítanak az együttműködéshez, az önálló tanuláshoz és a kreativitás kibontakoztatásához.

A könyvtárak oktatástámogató funkciója a digitális korszakban több dimenzióban is kibővül. A könyvtár mint „harmadik hely" olyan befogadó közegként jelenik meg, ahol a formális és informális tanulás találkozik, és amely elősegíti az élethosszig tartó tanulást, valamint az egyenlő hozzáférést a digitális tudáshoz. Ezzel párhuzamosan a könyvtáros szerepe is átalakul: a hagyományos információközvetítő funkció mellett egyre hangsúlyosabbá válik oktatói, digitális facilitátori és pedagógiai partnerszerepe (IFLA, 2024; EBLIDA, 2024). A könyvtárosok aktívan részt vesznek az információs műveltség, a digitális készségek és az etikus mesterségesintelligencia-alkalmazás fejlesztésében, továbbá támogatják a projektalapú és kollaboratív tanulási folyamatokat digitális tartalmak, workshopok és innovatív oktatási megoldások révén (IFLA, 2024; UNESCO, 2021).

A könyvtárak oktatási szerepének erősödése ugyanakkor szorosan összefügg a könyvtárosok digitális és technológiai felkészültségével. A mesterséges intelligencia által támogatott tanulási környezetek sikeres kialakítása és működtetése nagymértékben függ attól, hogy a könyvtárosok milyen attitűdökkel, kompetenciákkal és technológiaelfogadási jellemzőkkel rendelkeznek. Jelen tanulmány szerzőjének doktori kutatása is e kérdéskör vizsgálatára irányul, különös figyelmet fordítva a könyvtárosok mesterségesintelligencia-technológia elfogadására, valamint arra, hogy ezek a tényezők miként befolyásolják a könyvtárak tanulási környezetté válását és innovációs képességét. A kutatás várható eredményei hozzájárulhatnak a könyvtári kompetenciafejlesztési programok, a továbbképzési rendszerek, valamint a mesterséges intelligencia könyvtári alkalmazását támogató stratégiai fejlesztések megalapozásához.

## A könyvtárosok technológiai felkészültsége és a digitális kompetenciák fejlesztése

A mesterséges intelligencia, az adatalapú rendszerek és a digitális platformok térnyerése alapvetően alakítja át a könyvtári szakmai gyakorlatot, amely új kompetenciák és szemléletmódok kialakulását teszi szükségessé a könyvtáros szakmában. A digitális transzformáció sikeressége már nem kizárólag a technológiai infrastruktúrán múlik, hanem jelentős mértékben a könyvtárosok technológiai felkészültségén, attitűdjén, alkalmazkodóképességén és innovációra való nyitottságán. A nemzetközi trendek rámutatnak arra, hogy a könyvtári innováció egyik legfontosabb hajtóereje a folyamatos szakmai fejlődés, valamint a támogató szervezeti kultúra megléte, amely elősegíti az új technológiák – különösen a mesterséges intelligencia, az adatelemzés és az immerzív digitális technológiák – integrációját a könyvtári szolgáltatásokba. Ebben a környezetben a könyvtárosok szerepe fokozatosan kibővül: digitális közvetítőként, oktatóként, technológiai facilitátorként és etikai iránytűként támogatják a felhasználókat az egyre összetettebb információs ökoszisztémákban.

A könyvtáros szakma kompetencia- és szerepátalakulása szorosan összefügg az MI-műveltséggel, az adatműveltség és a digitális kritikai gondolkodás fejlődésével. A generatív és agentikus mesterséges intelligencia megjelenése új kihívásokat és egyben új szakmai lehetőségeket teremt a könyvtárak számára, hiszen ezek a rendszerek már nem csupán információkat szolgáltatnak, hanem komplex folyamatokat terveznek, rendszerek között kommunikálnak és strukturált eredményeket hoznak létre. A könyvtárosoknak ezért egyre nagyobb szükségük van olyan kompetenciákra, mint az etikus mesterségesintelligencia-használat, az MI-tartalmak kritikus értékelése, a prompttervezés, valamint a technológiai és kiberbiztonsági tudatosság. A jövő könyvtárosa egyszerre válik digitális mentorrá, tanulástámogató szakemberré, adatkurátorrá, közösségfejlesztővé és innovációs katalizátorrá, miközben továbbra is meghatározó szerepet tölt be az élethosszig tartó tanulás és a digitális befogadás támogatásában. A Library 5.0 szemléletben a könyvtár így nem csupán alkalmazkodik a technológiai változásokhoz, hanem aktív alakítója is a digitális tudástársadalom fejlődésének.

A magyar könyvtári szakterületen is egyre hangsúlyosabban jelennek meg azok a fejlesztési és stratégiai törekvések, amelyek a digitális kompetenciák fejlesztését, a mesterséges intelligencia felelős alkalmazását és a tanulástámogató szolgáltatások megújítását célozzák. E folyamatokat szakmai szervezetek, intézmények és képzési programok egyaránt támogatják. A Magyar Könyvtárosok Egyesületének megújított stratégiája és 2026. évi munkaterve is kiemelt prioritásként kezeli a digitális transzformációhoz kapcsolódó kompetenciafejlesztést, valamint a mesterséges intelligencia könyvtári alkalmazásának szakmai támogatását, amely képzések, műhelymunkák és tudásmegosztó kezdeményezések formájában is megjelenik.

## Összegzés

A 21. századi könyvtárak szerepe alapvetően átalakul a mesterséges intelligencia és a digitális transzformáció hatására. A könyvtárak egyre inkább emberközpontú, adaptív és közösségi tudásökoszisztémákként működnek, amelyek egyszerre támogatják az oktatást, a digitális kompetenciafejlesztést, az élethosszig tartó tanulást és a társadalmi részvételt. A technológiai innováció nem csökkenti a könyvtárak jelentőségét, hanem új funkciókkal és felelősségekkel ruházza fel őket.

A Library 5.0 szemléletben a könyvtár nem pusztán alkalmazkodik a változó világhoz, hanem aktív alakítója a digitális tudástársadalomnak. Ebben a folyamatban a könyvtárosok szerepe is új dimenzióval bővül: a jövő könyvtárosa nem csupán információközvetítő, hanem digitális mentor, etikai közvetítő és a tudatos technológiahasználat támogatója. A könyvtár így továbbra is meghatározó intézmény marad a tudás, a közösség és az emberi kapcsolatok megőrzésében és fejlesztésében a mesterséges intelligencia korszakában.

Az elemzett nemzetközi és hazai trendek alapján megállapítható, hogy a könyvtárak a digitális tudástársadalom meghatározó szereplőiként továbbra is kulcsszerepet tölthetnek be a digitális kompetenciafejlesztés, az információs műveltség támogatása és az etikus technológiahasználat előmozdítása területén. E szerepkör sikeres betöltése ugyanakkor egyre szorosabban összefügg a könyvtárosok digitális kompetenciáival, technológiaelfogadási attitűdjeivel és innovációs nyitottságával, amelyek meghatározó tényezői lehetnek a mesterséges intelligenciára épülő szolgáltatások és tanulási környezetek sikeres bevezetésének.

A szerző álláspontja szerint a mesterséges intelligencia nem a könyvtárak jelentőségének csökkenését, hanem szerepük újraértelmezését hozza magával. A jövő könyvtára olyan tudás- és tanulási ökoszisztémaként működhet, amelyben az emberi szakértelem, a technológiai innováció és a közösségi értékteremtés egymást erősítve járul hozzá a digitális tudástársadalom fejlődéséhez. Ennek megvalósításában meghatározó szerep jut a könyvtárosok kompetenciáinak, attitűdjeinek és folyamatos szakmai fejlődésének, amelyek a könyvtári innováció egyik legfontosabb erőforrását jelentik.

Meggyőződésem, hogy a 21. századi könyvtár nem csupán reagál a technológiai változásokra, hanem aktív szerepet vállal azok társadalmi és oktatási hasznosításában is. A könyvtárak jövője ezért nem elsősorban technológiai kérdés, hanem annak felismerése, hogy az emberi tudás, a közösségi értékek és az innováció együttesen képesek fenntartható és befogadó tanulási környezeteket létrehozni a mesterséges intelligencia korszakában.

## Hivatkozások


CEDEFOP (2024) *Skills Forecast: Trends and Challenges to 2030*. Luxembourg: Publications Office of the European Union. Available at: https://www.cedefop.europa.eu (2026. május 15.)

EBLIDA (2024) *EBLIDA Strategy 2025–2028*. The Hague: EBLIDA. Available at: https://eblida.org/about-eblida/strategic-plan/ (2026. május 15.)

GREEN, A. (2024) *Artificial intelligence and the changing demand for skills in the labour market*. OECD Artificial Intelligence Papers, No. 14. Paris: OECD Publishing. https://www.oecd.org/content/dam/oecd/en/publications/reports/2024/04/artificial-intelligence-and-the-changing-demand-for-skills-in-the-labour-market_861a23ea/88684e36-en.pdf (2026. május 15.).

Harvard Business Review: 9 Trends Shaping Work in 2026 and Beyond – https://hbr.org/2026/02/9-trends-shaping-work-in-2026-and-beyond (2026. május 15.)

IFLA (2024) *IFLA Trend Report 2024: Facing the Future of Information with Confidence*. The Hague: International Federation of Library Associations and Institutions. https://www.ifla.org/wp-content/uploads/ifla-trend-report-2024.pdf (2026. május 15.)

MHASKE, P., Bhattacharjee, B., Haldar, N., Upadhyay, P. and Mandal, A. (2025) *Bridging digital skill gaps in the global workforce: A synthesis and conceptual framework building*. Research in Globalization, 11, 100311. https://doi.org/10.1016/j.resglo.2025.100311 (2026. május 15.)

OECD (2023) *OECD Employment Outlook 2023: Artificial Intelligence and the Labour Market*. Paris: OECD Publishing. Available at: OECD Employment Outlook 2023 (2026. május 15.)

OECD (2025) *OECD Skills Outlook 2025: From Skills to Labour Market Opportunities*. Paris: OECD Publishing. https://doi.org/10.1787/26163cd3-en

Passalacqua, M., et al. (2025) *Human-centred artificial intelligence in Industry 5.0: A systematic review*. International Journal of Production Research. https://doi.org/10.1080/00207543.2024.2406021 (2026. május 15.)

UNESCO (2021) *Reimagining our futures together: A new social contract for education*. Paris: UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000379707.locale=en (2026. május 15.).

World Economic Forum (2023) *The Future of Jobs Report 2023*. Geneva: World Economic Forum. Available at: https://www.weforum.org/reports/the-future-of-jobs-report-2023 (2026. május 15.)

World Economic Forum (2025) *The Future of Jobs Report 2025*. Geneva: World Economic Forum. Available at: https://www.weforum.org/publications/the-future-of-jobs-report-2025/ (2026. május 15.)

DOI: 10.31915/NWS.2026.14



# ELEKTRONIKUS LEVELEZÉSKIADÁS MINT SZEMANTIKUS HÁLÓZAT

# ELECTRONIC CORRESPONDENCE PUBLISHING AS A SEMANTIC NETWORK

*Szemigán Dorottya Henrietta*
*Digitális Örökség Nemzeti Laboratórium*
*szemigan.dorottya@btk.elte.hu*
*ORCID: 0009-0007-5116-838X*

*Dobás Kata*
*ELTE HTK Irodalomtudományi Kutatóintézet*
*dobas.kata@abtk.hu*
*ORCID: 0009-0009-7632-8276*

*Palkó Gábor*
*ELTE HTK Irodalomtudományi Kutatóintézet*
*palko.gabor@abtk.hu*
*ORCID: 0000-0002-4394-8577*

*Fellegi Zsófia*
*ELTE HTK Irodalomtudományi Kutatóintézet*
*fellegi.zsofia@abtk.hu*
*ORCID: 0000-0001-9199-1759*

*Nemes-Jakab Éva*
*ELTE HTK Irodalomtudományi Kutatóintézet*
*jakabev@gmail.com*
*ORCID: 0009-0001-4516-0935*

*Sárközi-Lindner Zsófia Mónika*
*ELTE HTK Irodalomtudományi Kutatóintézet*
*zsofia.monika@gmail.com*
*ORCID: 0000-0002-2558-0633*

## Absztrakt

A tanulmány Kántor Lajos kolozsvári irodalomtörténész e-mail-hagyatékának feldolgozásán keresztül mutatja be, miként alakítható ki egy born-digital irodalmi hagyaték teljes feldolgozási workflow-ja az archiválástól és a hosszú távú megőrzés biztosításától az MBOX-alapú feldolgozáson át a TEI XML-ben megvalósuló, szemantikus adatbázishoz kapcsolt digitális e-mail-kiadásig. Kutatási projektünk fókuszában Kántor Lajos és Ilia Mihály elektronikus levelezésének digitális tudományos kiadásra való előkészítése áll. Esettanulmányunkon keresztül bemutatjuk a korpusz kijelölésének szempontjait, a metaadatok strukturálását, az ITIdata szemantikus adatbázissal való összekapcsolást, a levelek e-mail-specifikus TEI XML-struktúráját, adatvédelmi és anonimizálási elveinket, valamint a lektorálást támogató ellenőrző- és munkafelületünket.

**Kulcsszavak:** e-mail, szemantikus adatbázis, colab notebooks, digitális kiadás, e-mail-hagyaték, GDPR, ePrivacy

## Abstract

The paper presents how a complete workflow for processing a born-digital literary archive can be developed through the processing of the e-mail legacy of Lajos Kántor, a literary historian from Cluj-Napoca: from archiving and ensuring long-term preservation, through MBOX-based processing, to a digital e-mail edition implemented in TEI XML and connected to a semantic database. The focus of our research project is the preparation of a digital scholarly edition of the electronic correspondence between Lajos Kántor and Mihály Ilia. Through this case study, we present the criteria for defining the corpus, the structuring of metadata, integration with the ITIdata semantic database, the structure of e-mail-specific TEI XML files for the letters, our principles for data protection and anonymization, and the review interface developed for proofreading.

**Keywords:** e-mail, semantic database, colab notebooks, digital edition, e-mail heritage, GDPR, ePrivacy

## 1. Bevezetés:

A 20–21. századra az e-mail a kulturális, tudományos és intézményi kommunikáció, így a kortárs hagyatékok egyik meghatározó born-digital történeti forrástípusává vált. Az e-mail, amely egyszerre szöveganyag és metaadatokkal rendelkező rekord, nem tekinthető pusztán a papíralapú levél digitális megfelelőjének, hiszen a levéltest mellett fejlécadatokat, címzési mezőket, időbélyegeket, válaszkapcsolatokat, karakterkódolási sajátosságokat, továbbítási információkat és csatolmányokat is tartalmazhat. Feldolgozása ezért csak a technikai, szerkezeti és relációs rétegek együttes figyelembevételével végezhető el.

Az e-mailek megőrzése, feltárása és hozzáférhetővé tétele összetett technikai, intézményi és jogi feladat (Task Force on Technical Approaches for Email Archives, 2018). A szakirodalom az e-mailarchiválást életciklus-alapú kurátori folyamatként írja le, amelyben a létrehozás, használat, exportálás, megőrzés, leírás és hozzáférhetővé tétel egymással összefüggő munkafázisokként jelennek meg (Pennock, 2006; Prom, 2019). Jaillant érvelésével összhangban a technikai megmentés azonban önmagában még nem jelent kutatói hozzáférhetőséget. A megfelelő leírás, kereshetőség, jogosultságkezelés és jogilag szabályozott hozzáférési modell[1] nélkül az e-mailgyűjtemény továbbra is zárt, a kutatás és a nyilvánosság számára ténylegesen nem hozzáférhető archívum maradhat (Jaillant, 2019).

Bár a TEI Correspondence SIG és a Correspondence Metadata Interchange Format fontos mintát adnak a levelezési események strukturált leírására (Dumont, 2016; TEI Correspondence SIG, n.d.), az e-mail esetében a pontos időbélyegek, a többes címzés, a válaszláncok, a csatolmányok, a Message-ID és a nem publikálható technikai metaadatok is a kiadás módszertani döntéseinek részévé váltak. Jelen tanulmányunk bemutatja, hogyan dolgoztuk ki a fentiek figyelembevételével a digitális bölcsészeti területen még egyedülálló digitális emialkiadás módszertani alapjait.

## 2. A Kántor-e-mail-hagyaték feldolgozás teljes workflow-ja

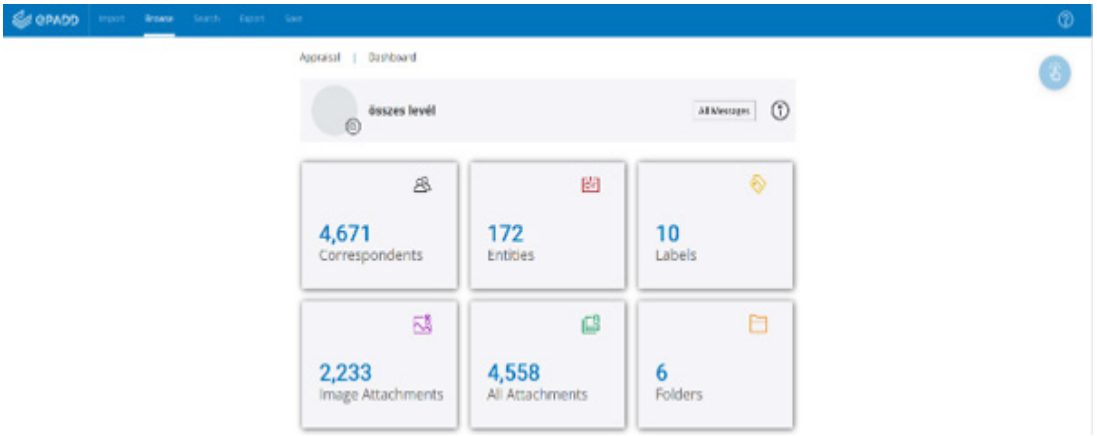
E-mail-hagyaték feldolgozás lehetséges módszertanának létrehozásával foglalkozó kutatási projektünk Kántor Lajos kolozsvári irodalomtörténész, filológus, kritikus és szerkesztő elektronikus levelezésének archiválását, feldolgozását és részleges forráskiadását célozta meg. A munkafolyamat első szakasza az e-mailállomány exportjára, integritásbiztosítására, szabványos archiválására és hosszú távú megőrzésére irányult; ennek módszertani tanulságait korábbi tanulmányunk foglalta össze (Alföldi et al., 2023a). A második szakasz az MBOX-formátum feldolgozására, kereshetővé tételére, tisztítására és vizualizációs előkészítésére koncentrált, amelynek eredményeit az NLP4DH 2023 konferencián mutattuk be (Indig et al., 2023). A jelen írás erre a két előzményre építve a harmadik munkafázist, a Kántor Lajos és Ilia Mihály közötti elektronikus levelezés TEI XML-alapú, szemantikus adatbázishoz kapcsolt digitális forráskiadásának létrehozását tárgyalja.[2]

---

1 A hozzáférés e-mail-hagyatékok esetében jogi és etikai kérdés is, mivel ezek harmadik személyek adatait, nem nyilvános elérhetőségeket, érzékeny információkat és visszaazonosításra alkalmas technikai metaadatokat tartalmazhatnak.

2 Az e-mail TEI-alapú reprezentációjára lásd Syd Bauman és Elisa Beshero-Bondar kísérletét a TEI-L nyilvános LISTSERV-archívumával. A vállalkozás fontos párhuzam, de nem közvetlen előzmény: a két projekt párhuzamosan zajlott, a TEI-L-modell kísérleti jellegű, és egy nyilvános levelezőlista archiválására, nem pedig személyes irodalmi e-mail-hagyaték digitális kritikai forráskiadására irányul. A jelen projektben a TEI XML-specifikációt a szemantikus adatbázis-kapcsolat, az anonimizálás, a hozzáférés-szabályozás és a technikai metaadatok adatvédelmi szűrése is alakította. Magyar kontextusban Kalcsó Gyula és Szűcs Kata Ágnes tanulmánya, valamint az MNMKK OSZK DBK Katalist-projektje említendő: ezek fő hangsúlya azonban az e-mailek hosszú távú közgyűjteményi megőrzésén, nem a digitális kiadási modell kialakításán van.

## 3. A forráskiadás korpusza




*1. ábra: a teljes e-mailállomány az ePADD szoftver megjelenítésében*

A digitális forráskiadás korpuszát a teljes hagyaték egy főként irodalomtörténetileg releváns része, Kántor Lajos és Ilia Mihály levelezése adja. A hagyaték hibrid jellegű, a benne foglalt teljes e-mailállomány 11 246 digitális objektumot[3] és 1424 papír alapon fennmaradt, kinyomtatott e-mailt tartalmaz[4]. A forráskiadásra kijelölt Ilia Mihály és Kántor Lajos levelezésére vonatkozó korpusz 69 nyomtatott e-mailből, 5 kéziratból és 360 digitális levélből, e-mailből áll, a kiadás így összesen 434 levelet közöl 2003–2017 közötti időszakból. A korpusz kiválasztását a technikai vagy mennyiségi szempontok mellett az határozta meg, hogy a Kántor és Ilia levelezése egy adott történeti pillanatban a magyar[5] irodalmi közeg, szerkesztői és tudományos kapcsolatrendszer olyan szeletét teszi láthatóvá, amely egyszerre kínál irodalom- és intézménytörténeti, valamint művelődés- és politikatörténeti kutatási lehetőségeket.

## 4. Korpusz feldolgozása, tömeges betöltés, TEI XML e-mail-specifikációk

### 4.1 Adatfeldolgozás

A hagyaték hibrid jellege az adatfeldolgozás szempontjából külön módszertani kihívást jelentett. Az e-maileket ezért három különböző bibliográfiai leírással rendelkező e-mail-típusként kezeltük, melyek metaadatait ennek mentén tabuláris formában rendszerez-

3 A 2002–2023 közötti időszakból, ti. az e-mailcím a Kántor 2017-ben bekövetkezett halála után is, az archiválásig aktív volt.

4 Az 1997–2016 közötti évekből származó e-mailek még kiegészülnek adott esetben a nyomtatott e-mailekre rögzített válaszlevelek kéziratával is.

5 Magyar nyelvű, azaz belföldi és határon túli egyaránt.

tünk.[6] Ennek több előnye is kirajzolódott, egyfelől a katalogizálás/leltározás funkciója, az adatok strukturálása, valamint a tömeges betöltéshez való előkészítése.

A kinyomtatott e-mailek szkennelt másodpéldányainak optikai karakterfelismerése a Transkribus szoftverével történt, az MBOX-állomány feldolgozásához pedig a Stanford University által fejlesztett, open source mbox-megjelenítő ePADD és a full MIME support-tal rendelkező Windows MBox Viewer szoftvereket használtuk. Ez utóbbival több karakterkódolási problémát is sikerült áthidalni.

### 4.2 Tömeges betöltés, ITIdata

A kinyert metaadatok strukturálását két célkitűzés mentén végeztük el a projekt során: 1. a létrehozott specifikáció mentén megtörténhessen a félautomatikus betöltés a QuickStatements eszköz[7] segítségével az ITIdata[8] elnevezésű, az ELTE HTK Irodalomtudományi Kutatóintézet szemantikus adatbázisába, 2. a TEI fájlok headerjeinek félautomatikus létrehozását is támogassa.

A specifikáció kidolgozásánál a schema.org e-mailekre kidolgozott leírását vettük alapul, valamint figyelembe vettük az ITIdata már meglévő adatainak struktúráját, elsősorban a klasszikus, papíralapú levelezéskiadásokhoz (Arany János és Jókai Mór levelezése) készült tulajdonságok (property) listáját. A metaadatokat excel táblázatokban rögzítettük, formázásukat a QuickStatements-el való betöltéshez igazítottuk.

A betöltés előkészítésekor megtörtént a nevek azonosítása, névtérre fektetése, a részletes annotációk elkészítése, amelynek során az e-mailekben említett nevek, művek, földrajzi nevek és intézmények entitásazonosítása is végbement.

Mivel a Kántor-Ilia-e-mail-hagyaték az e-mailezésnek egy köztes állapotát rögzíti, amelyben az e-mailek még kinyomtatásra is kerültek, és feladójuk sokszor ezekre rögzítette válaszát (kézírással), ezért a szemantikus kapcsolatok kidolgozásánál ezt is figyelembe kellett vennünk.[9]

---

6 A digitális formában fennmaradt leveleket e-mailként, a kinyomtatott e-maileket e-mailmásolatként, míg a kinyomtatott e-mailekre kézírással írt válaszleveleket kéziratként kezeltük.

7 https://www.wikidata.org/wiki/Help:QuickStatements/hu

8 Az adatbázisról: Dobás, Kata és Fellegi, Zsófia és Palkó, Gábor (2023) *A kis gömböc meséje – az ITIdata irodalomtudományos adatbázis fejlesztése 2022–2023-ban.* In: Új technológiákkal, új tartalmakkal a jövő digitális transzformációja felé. HUNGARNET Egyesület, Budapest, pp. 192–198. Online hozzáférés: https://real.mtak.hu/182937/ Az adatbázis elérhetősége: https://itidata.abtk.hu/wiki/Main_Page

9 https://itidata.abtk.hu/wiki/Item:Q492429 – az e-mail nyomtatott és elektronikus változata közötti szemantikus kapcsolatokat létrehoztuk, a kapcsolat fajtáját minősítőben rögzítettük. A teljes korpusz elérhetősége az ITIdata-ban: https://tinyurl.com/27dyceu6

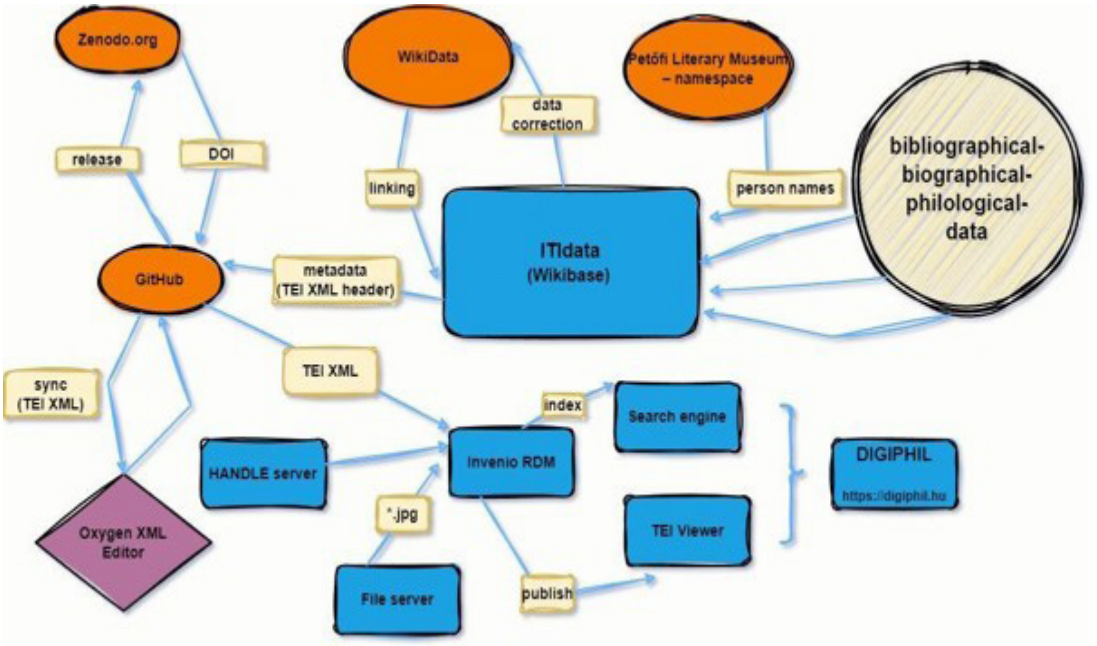

*2. ábra részlet az ITIdata egyik rekordjának szemantikus kapcsolatairól*

### 4.3 TEI XML e-mail-specifikációk

Az e-mailek publikálásának PDF-formátumú általános gyakorlatával szemben digitális kiadásunk – a tudományos igényesség jegyében – a TEI XML szabványos, interoperábilis, ember és gép számára egyaránt olvasható, a hosszútávú megőrzést, összetett keresést is biztosító formátum mellett döntött.

A teiHeaderben szerepet kap több az e-mailre vonatkozó metaadat, az egyes levelek részletes bibliográfiai adatait tartalmazó ITIdata rekodok azonosítói, a küldés pontos idejére, valamint a levelek közötti kapcsolatokra vonatkozó információk és az egyes e-mailek PID azonosítója. Az elektronikus levelek címét egységes szintaxis alapján alakítottuk ki: a cím a küldő nevéből, a címzett nevéből, valamint a küldés dátumából és pontos idejéből áll – a küldő és a címzett nevét gondolatjel választja el egymástól. A TEI XML-struktúrában a levél címe, az e-mail eredeti tárgya, valamint a levél szövege a <body> elemben kapott helyet. A levelek adatgazdagítása során a szövegben szereplő személyneveket, településeket, intézményeket, műveket, műcímeket és dátumokat, továbbá a bibliográfiai adatokhoz, a kritikai apparátushoz és a különböző szövegváltozatokhoz kapcsolódó információkat a megfelelő TEI-elemekkel jelöltük.

A korpusz részletes annotálása számos lehetőséget kínál a levelezés további vizsgálatára, és új dimenziókat nyit meg annak digitális bölcsészeti eszközökkel és módszerekkel történő tartalmi feltérképezésében. A két szerző levelezésében említett, irodalomtörténetileg releváns szereplők azonosítása révén létrehozhatóvá válhat egy meghatározott korszak irodalmi közegének hálózati reprezentációja, hasonlóan a levelezésben említett művek kapcsolatrendszeréhez. A törzsszövegben előforduló numerikus dátumok és szöveges időutalások annotálása szintén fontos eleme volt a feldolgozásnak, mivel ezek az e-mail időbélyegével és a küldés dátumával összevetve lehetővé teszik annak vizsgálatát, hogy a felek milyen gyakorisággal tematizálnak aktuális eseményeket. Ezáltal elkülöníthetővé válhat a jelenre és közelmúltra, illetve a távolabbi múltra vonatkozó párbeszéd.




*3. ábra: DigiPhil infrastruktúra*

## 5. Adatkezelési stratégiák

Az Ilia Mihály és Kántor Lajos közötti e-mailváltás kritikai forráskiadása – a levéltest és a metaadatok adatszerkezetének feltárása, szemantikus adatbázisban való rögzítése, valamint TEI XML-kódolása mellett – az adatkezelés jogi és etikai kereteinek újragondolását is szükségessé tette.

Az e-mail metaadatokat tartalmazó réteg (pl. e-mail-címek, szerverinformációk, IP-címek, továbbítási láncok, időbélyegek, CC/BCC mezők, csatolmányok és rejtett címzettek) adott esetben érzékenyebb információkat hordozhatnak, mint a tartalmi szöveg. Az Európai Unió jogértelmezése szerint ezek az adatok az elektronikus hírközlési adatok körébe tartoznak, így a GDPR mellett az ePrivacy-szabályozás hatálya alá is esnek.

Ebből következően az e-mailkiadás során nem volt elegendő a levélszöveg klasszikus anonimizálása vagy pszeudonimizálása, hanem különös figyelmet kapott a Message-ID, a routing lánc, az IP-címek, a levelezőszerver-információk, valamint a kézbesítési és fejléc-metaadatok kezelése, és ezek publikálhatóságának korlátozása.

A kutatás adatkezelési protokollja a tudományos feldolgozás céljainak és a személyes adatok védelmének egyensúlyára épül. Ennek során figyelembe vettük a 2016/679/EU rendeletet (GDPR), különösen a tudományos kutatási célú adatkezelésre vonatkozó 89. cikket, a 2002/58/EK irányelvet (ePrivacy), valamint az információs önrendelkezési jogról szóló 2011. évi CXII. törvényt (Infotv.). A jogi keretet a hagyaték jogtulajdonosaival – Ilia Mihály

esetében, valamint a Kántor Lajos-hagyatékot gondozó Minerva Művelődési Egyesülettel – kötött szerződéses megállapodások egészítik ki.

A hozzáférés szabályozása két szinten történt. A nyilvános repozitóriumba kizárólag válogatott és anonimizált e-mailek kerültek, míg a teljes, feldolgozatlan korpusz zárt környezetben marad, amelyhez kizárólag a jogtulajdonosok és a kutatásban részt vevő szakemberek férhetnek hozzá, kutatási engedély alapján.

Az anonimizálási és adatminimalizálási elvek egyrészt a levéltestre, másrészt a TEI-rekord metaadatokat tartalmazó <teiHeader> elemére terjednek ki. A TEI fájlnak a levél szövegét közlő <body> elemében strukturált módon, a <gap> elem két megkülönböztetett típusát használtuk e célra: személynevek esetében <gap reason="anonymised" unit="persName" extent="1"/>, elérhetőségi és kommunikációs adatok (e-mail, telefonszám, lakcím) esetében <gap reason="anonymised" unit="email/phoneNumber/address" extent="1"/>, míg érzékeny vagy nem rekonstruálható tartalmi egységek esetében <gap reason="anonymised" unit="word" extent="2"/> került alkalmazásra.

A <teiHeader> sem az eredeti e-mail fejléceket reprezentálja, hanem azok kutatási célú, adatvédelmi szempontok szerint szűrt és normalizált megfelelőjét. A nyilvánosan elérhető rekordokban a kommunikációs infrastruktúrára vonatkozó érzékeny vagy potenciálisan visszaazonosításra alkalmas adatok (például szerver-útvonalak, IP-alapú információk) nem jelennek meg.

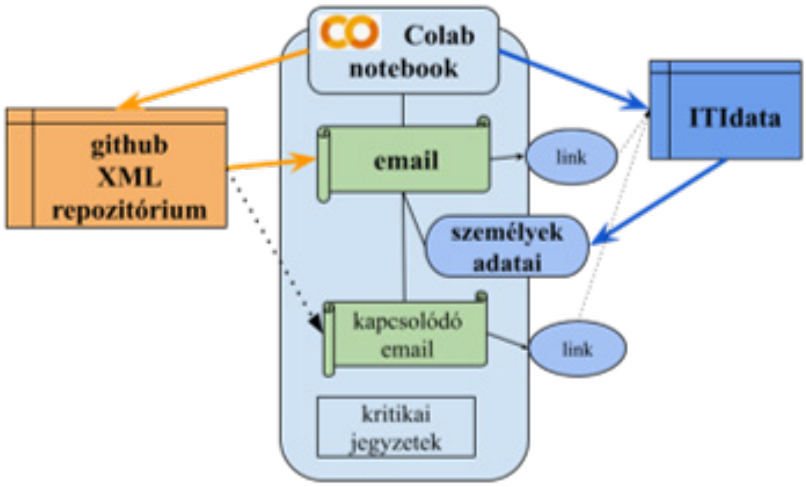
A Kántor Lajos és Ilia Mihály közötti e-mailváltás kritikai forráskiadása ezen gyakorlati szempontok bevezetésével Magyarországon elsőként valósít meg olyan publikációs modellt, amelyben a filológiai, információtechnológiai és adatvédelmi szempontok egységes rendszerben kerülnek integrálásra.

## 6. Lektorálási munkafolyamat

A lektorálási folyamathoz egy összetett, kollaboratív környezetet kellett létrehozni, ugyanis rugalmasan és egyidőben rendelkezésre kellett állnia a levelekről készült friss TEI XML-fájloknak, a levelek Wikibase reprezentációjának (ITIdata), valamint az azonosított entitások adatlapján szereplő tulajdonságoknak. Ennek összefogására és a lektor felé való világos és letisztult előkészítésére a Google Colab-ot mint ingyenes és testreszabható eszközt alkalmaztuk. A Google Colab notebookok böngészőből futtatható, felhőalapú Python-környezetet[10] biztosítanak, amelyben a kódcellák, szöveges magyarázatok és vizualizációk egyetlen dokumentumban kezelhetők. A felhasználók telepítés nélkül futtathatnak

10 A Python egy könnyen tanulható, magas szintű programozási nyelv, amely egyszerű szintaxisával és jó olvashatóságával támogatja a gyors fejlesztést. https://docs.python.org/3/

Python-kódot, külső könyvtárakat használhatnak, fájlokat tölthetnek be.[11] Az e-mailek XML-jeit tartalmazó GitHub repozitórium integrációja lehetővé tette, hogy a lektor közvetlenül az eredeti fájlok legfrissebb verziójával dolgozhasson. Minden munkamenet a repozitórium installálásával és a Python programkód törzsének aktiválásával kezdődik, amely egy-egy kódcella futtatását jelenti, egyszeri műveletként elvégezhető. A notebookkal való további interakciók sem igényelnek programozói tudást, az egyes levelek lektorálásához szükséges információk előkészítését ugyanis az aktuális XML-fájlnév megadása után elvégzi a program.

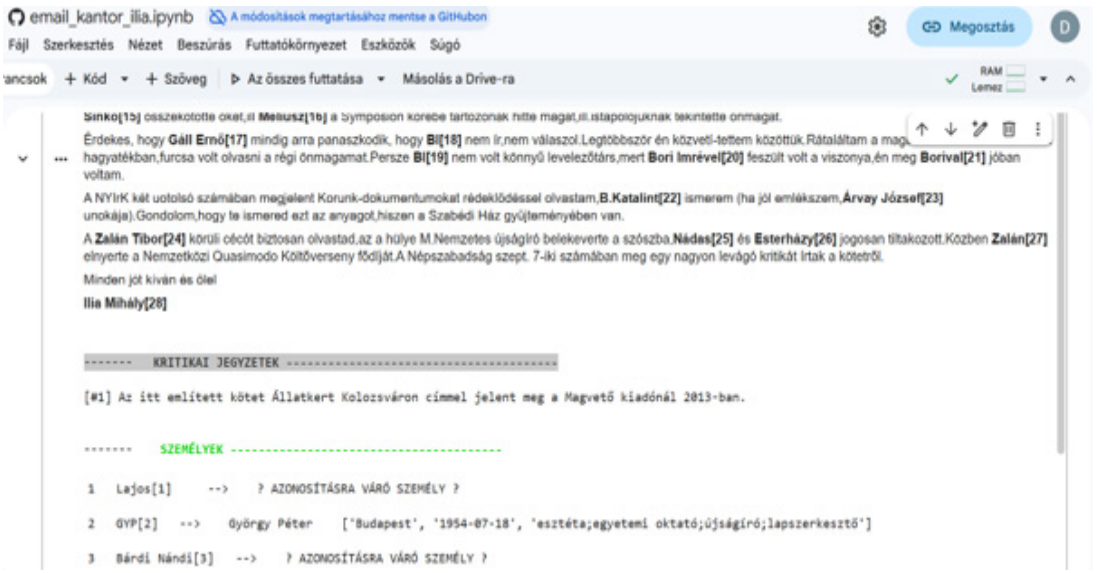



*4. ábra: Colab notebook infrastruktúra*

Futtatás után a lektor számára a notebookban megjelenik a levél szövege, illetve az ITIdata adatlapjára mutató kattintható link. A szöveg az XML-címkéket elhagyva látható, ugyanakkor az azonosított entitások félkövérrel és számozással, a kritikai jegyzetek számozással, az anonimizált részek külön kiemeléssel jelennek meg benne. A kritikai jegyzetek szövegét – amennyiben tartozik a levélhez – közvetlenül a levél alá írja ki a notebook, így ellenőrizhető a helyességük, illetve a lektor további jegyzeteket, vagy ezek kiegészítését javasolhatja. Az azonosítandó személyekről táblázatos formában olvashatóak az ITIdata adatlapjukról valós időben lekérdezett, kulcsfontosságú állítások (születési hely, születési idő, foglalkozás), melyek alapján a lektor értékelheti, hogy helyes volt-e az előzetes azonosítás, avagy javasolhat módosítást, azonosíthatja az entitást, amennyiben előzetesen üresen lett

11 Továbbá igény szerint GPU- vagy TPU-erőforrásokat is igénybe vehetnek. Ez különösen alkalmassá teszi a Colabot például reprodukálható kísérletek, adatfeldolgozási feladatok bemutatására. Kifejezetten támogatja a kollaboratív munkát, hiszen hasonlóan a Google Docs-hoz a notebookok megoszthatók a jogosultságok differenciálásával. Az alkalmazás további lehetőségeiről lásd: https://research.google.com/colaboratory/faq.html

hagyva. Mindezeket kontextusként segíti a kapcsolódó levelek szövegének egyszerű megjelenítése, illetve ITIdata rekordjukra mutató link megjelenítése.




*5. ábra: Lektorálási felület*

## Összegzés

A Kántor–Ilia-levelezés digitális forráskiadása több szinten is modellkísérletnek tekinthető. A részletes annotáció, az ITIdata adatbázissal való összekapcsolás, a TEI XML-alapú reprezentáció és a Colab-alapú lektorálási munkafolyamat együttesen egy módszertani javaslatot kínál arra, hogyan válhat az elektronikus levelezés a kortárs hagyatékok tudományosan kiadható, szemantikusan kereshető és adatvédelmi szempontból felelősen kezelhető forrástípusává.

## Bibliográfia

Alföldi István – Szemigán Dóra Hédi – Palkó Gábor – Fellegi Zsófia. „Kutatói e-mail hagyaték archiválása és feldolgozása." In *Új technológiákkal, új tartalmakkal a jövő digitális transzformációja felé*, szerk. Tick József – Kokas Károly, 199–207. Budapest: HUNGARNET Egyesület, 2023.

Alföldi István – Szemigán Dóra Hédi – Palkó Gábor – Fellegi Zsófia. *Kántor Lajos e-mail levelezése*. Zenodo, 2023. DOI: 10.5281/zenodo.7825184.

Bauman Syd – Beshero-Bondar Elisa. „Preserving a LISTSERV Archive in TEI: The Case of TEI-L." *Journal of the Text Encoding Initiative* 19 (2025). DOI: 10.4000/15pss.

Dumont Stefan. „Correspondence Metadata Interchange Format: The Correspondence Metadata of Digital Editions." *Journal of the Text Encoding Initiative* 10 (2016).

Indig Balázs – Horváth László – Szemigán Dóra Hédi – Nagy Mariann Eszter. „Emil.RuleZ! – An Exploratory Pilot Study of Handling a Real-Life Longitudinal Email Archive." In *Proceedings of the 3rd International Conference on Natural Language Processing for Digital Humanities*. Association for Computational Linguistics, 2023.

Jaillant Lise. „After the Digital Revolution: Working with Emails and Born-Digital Records in Literary and Publishers' Archives." *Archives & Manuscripts* 47, 3. sz. (2019): 285–304. DOI: 10.1080/01576895.2019.1640555.

Kalcsó Gyula – Szűcs Kata Ágnes. „Az e-mailek hosszú távú megőrzése." *Könyvtári Figyelő* 68, 3. sz. (2022): 403–410.

MNMKK OSZK Digitális Bölcsészeti Központ. „E-mailek archiválása az MNMKK OSZK Digitális Bölcsészeti Központjában." *dHUpla – Digitális Bölcsészeti Platform*. https://dhupla.hu/page/kreativ/email-archivalas. Hozzáférés: 2026. május 10.

Pennock Maureen. „Curating E-mails: A Life-Cycle Approach to the Management and Preservation of E-mail Messages." *DCC Digital Curation Manual*. Digital Curation Centre, 2006.

Prom Christopher J. *Preserving Email*. 2. kiadás. Digital Preservation Coalition Technology Watch Report. Digital Preservation Coalition, 2019. DOI: 10.7207/twr19-01.

Schneider J. et al. „Appraising, Processing, and Providing Access to Email in Contemporary Literary Archives." *Archives & Manuscripts* 47, 3. sz. (2019): 305–326. DOI: 10.1080/01576895.2019.1622138.

Task Force on Technical Approaches for Email Archives. *The Future of Email Archives: A Report from the Task Force on Technical Approaches for Email Archives*. Washington, DC: Council on Library and Information Resources, 2018.

TEI Correspondence SIG. *Correspondence Metadata Interchange Format / TEI Correspondence SIG Documentation*. é. n.

DOI: 10.31915/NWS.2026.15



# ÚJ TELJESÍTMÉNYÉRTÉKELÉSI KERETRENDSZER FEJLESZTÉSE A HAMVAS BÉLA PEST MEGYEI KÖNYVTÁRBAN

# DEVELOPING A NEW PERFORMANCE APPRAISAL FRAMEWORK AT THE HAMVAS BÉLA COUNTY LIBRARY OF PEST

*Nagy Andor*
*Hamvas Béla Pest Megyei Könyvtár, módszertani és hálózati osztály vezetője*
*nagy.andor@hpmk.hu*

*Tóth Máté*
*Hamvas Béla Pest Megyei Könyvtár, igazgató*
*toth.mate@hpmk.hu*

## Absztrakt

A Hamvas Béla Pest Megyei Könyvtár 2021 óta tudatosan építi azt a munkahelyi környezetet, amely a könyvtáros pálya külső nehézségei ellenében is képes megtartani a kollégákat. A munkatársi elégedettségmérések és a Charles Handy modellje szerinti szervezetikultúra-felmérés jelezték, hogy a 2017 óta működő, kétoldalú teljesítményértékelési mechanizmus nem tölti be a feladatát: a munkatársak formálisabb, többirányú visszajelzési rendszerre vágytak. A piacon elérhető 360 fokos megoldások nem illeszkedtek a szervezeti struktúránkhoz és az adatvédelmi igényeinkhez, ezért saját webalkalmazást fejlesztettünk. A tanulmány bemutatja a fejlesztés szakmai hátterét, a rendszer felépítését, a tervezési döntéseket, az anonimitás kezelését, valamint a tesztidőszak első tanulságait.

**Kulcsszavak:** 360 fokos teljesítményértékelés, könyvtári humánerőforrás-menedzsment, szervezeti kultúra, munkahelyi jóllét, webalkalmazás-fejlesztés

## Abstract

Since 2021 the Hamvas Béla County Library of Pest has been building a workplace environment that retains colleagues despite the structural difficulties of the librarian profession. Staff satisfaction surveys and an assessment of organisational culture based on Handy's model signalled that the two-way appraisal mechanism in place since 2017 was not functioning well: staff desired a more formal, multi-source feedback system. Available commercial 360-degree solutions did not fit our organisational structure or data-protection

requirements, so we developed our own web application. This paper presents the professional background of the development, the architecture of the system, the design decisions, the handling of anonymity, and the first lessons from the pilot phase.

**Keywords**: 360-degree performance appraisal, library human resource management, organisational culture, workplace well-being, web application development

## Bevezetés

A Hamvas Béla Pest Megyei Könyvtárban – hasonlóan más intézményekhez – hosszú időn keresztül küzdöttünk a magas fluktuációval, illetve a szakképzett munkaerő felvételének és megtartásának nehézségeivel. Az alacsony fizetések, a szakma alacsony presztízse, a kiszámítható jövőkép hiánya mind egy olyan környezetet határoznak meg, amely negatívan hat a könyvtárosok motiváltságára. 2021 óta szisztematikusan mérjük a munkatársak elégedettségét, a közérzetüket befolyásoló jelenségeket, a munkahelyi légkör általános alakulását, a dolgozók stressz-szintjét. Az eredmények figyelembe vételével igyekszünk egy olyan munkahelyi légkört teremteni, amely mindezen külső tényezők ellenében képes támogatni a kollégáink pályán – és a könyvtárunkban – maradását.

A munkatársi elégedettségmérések eredményei alapján felmerülő nehézségekre, problémákra egyrészt igyekszünk azonnal reagálni, de tekintettel a kérdés súlyára valamint az első elégedettségmérés eredményei alapján kibontakozó általános rossz hangulatra, 2021-ben elindítottunk egy munkatársak jóllétét segítő programot, amelyben a magánélet és munka egyensúlyának megteremtése, a közösséghez tartozás erősítése, a szakmai fejlődés lehetőségének biztosítása valamint az általános munkavállalói jóllét voltak a fejlesztendő területek.

A program öt év alatt látványos és érezhető eredményeket hozott mind a dolgozók elégedettségét, mind pedig a szervezet teljesítményét tekintve. 2021 és 2025 között minimálisra szorítottuk a fluktuációt, a dolgozók elégedettsége valamennyi területen radikálisan nőtt, a munkatársak által átélt stressz-szintje jelentős mértékben csökkent. Az intézkedések pozitívan hatottak arra, ahogyan a használók megítélik munkatársainkat. Ezzel az innovációval nyertük el 2026-ban a Könyvtári Minőség Díjat. (KMD, 2026) Jelen tanulmányunkban bemutatott fejlesztésnek ez a program jelenti a kontextusát.

## A teljesítmény értékelése

A programunk lényege, hogy a munkatársainknak igyekszünk pozitív megerősítéseket adni, érzékeltetni velük, hogy fontosak a szervezetnek, szükségünk van a könyvtár működésével kapcsolatos visszajelzéseikre és hosszú távon is számítunk a munkájukra. Éppen ezért valamennyi munkatársi visszajelzést megfontolunk, mindegyikre igyekszünk reagálni.

A könyvtárban 2017 óta működött egy teljesítményértékelési mechanizmus, amelynek az

volt a lényege, hogy egy papír alapú pontozólapon az igazgató értékelt minden munkatársat, ezzel párhuzamosan pedig az illető saját magát. Ezt követően egy személyes megbeszélésen keresték, hogy mi lehet a különbségek oka. A munkatársak világossá tették, hogy ezt az értékelési rendszert nem tartják megfelelőnek. Nem motiválta őket, kínos beszélgetések részeseinek érezték magukat, amely nem jelentett érdemi visszajelzést, így nem is segítette a teljesítményük javítását. A visszajelzések alapján igyekeztünk új, alternatív megoldásokat találni. A szervezeti struktúra átalakításával új vezetőket neveztünk ki, akik kevesebb munkavállalóval vannak közvetlen munkakapcsolatban, így a visszajelzéseket is azonnal, kevésbé formális módon tudják átadni számukra. A formálisabb visszajelzésekhez bevezettük a 10, 20 és 30 éves munkaviszony után járó bronz, ezüst és arany hűségjutalmakat, valamint az év dolgozója díjat, amelyet a munkatársak szavazatai alapján kerülnek kiosztásra a karácsonyi ünnepség során.

A későbbi munkatársi elégedettségmérések alapján azonban kiderült, hogy ez nem elegendő a munkavállalóink számára. A vezetők habitusától és a munkavállalók igényeitől függően többen is úgy érezték, hogy a vezetőkkel való közvetlen munkakapcsolat során ugyan nagyon sok visszajelzést kapnak, de vágynának egy formálisabb, objektív szempontok alapján végzett értékelési rendszerre, amelyben nem csak a vezető értékelheti a beosztottját, hanem fordítva is, sőt, a dolgozók egymást is. A munkatársainkban felmerült az igény arra, hogy egymásnak, illetve egymásról is adhassanak pozitív visszajelzéseket, felhívhassák a társaik és a vezetők figyelmét egy-egy olyan teljesítményre, amely a környezetükben előfordult, de megítélésük szerint nem kapott kellő elismerést.

Miután 2023-ban a Charles Handy-féle modell alapján felmértük a könyvtárban uralkodó szervezeti kultúrát, világossá vált a számunkra, hogy mind a teljesítmény, mind pedig a személyes kapcsolatok nagyon hangsúlyos értékeket jelentenek a munkatársi kollektívában. A legerősebbeknek a feladatkultúra és a személykultúra jegyei bizonyultak a felmérés alapján. Előbbiben jellemző, hogy a befolyás elsősorban a szaktudásból és a feladatok magas szintű elvégzéséből fakad és nem a pozícióból vagy hatalomból. A személykultúrában jellemző, hogy közeli vagy akár baráti kapcsolatok is vannak az alkalmazottak között a szervezeten belül, amelynek nyomán a kölcsönös bizalom és az egymás munkája iránti elkötelezettség alakul ki. A motiváció a személy- és a feladatkultúrában egyaránt belülről fakad, amelyhez pedig kulcsfontosságú, hogy a dolgozók rendszeres visszajelzést kapjanak és adjanak egymásnak az elvégzett feladatokról. (Bakacsi 2015)

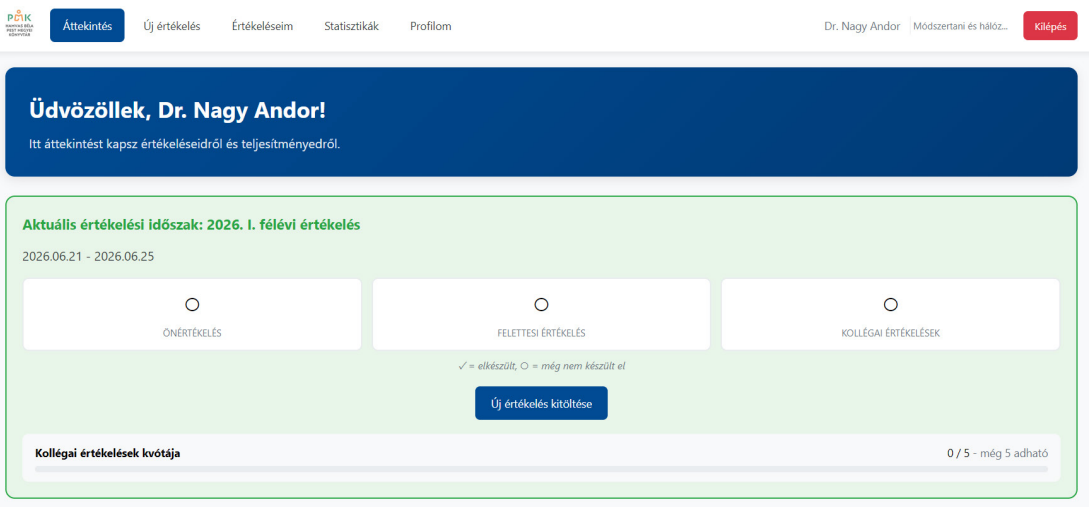
Egy valóban 360 fokos értékelési rendszer tervei kezdtek körvonalazódni, amelyben előre meghatározott szempontok mentén – beállításoktól függően – lényegében bárki, bárkit értékelhet a dolgozók közül.

## Szakirodalmi háttér

A többforrású értékelés kiterjedt empirikus megalapozottsággal bír. Smither et al. (2005) metaelemzése szerint a 360 fokos visszajelzés akkor hatásos, ha a kitöltő szükségesnek érzi a változást, pozitívan reagál rá, és konkrét fejlesztési célokat tűz ki. Kluger és DeNisi (1996) elmélete szerint a visszajelzés önmagában nem mindig javít: az esetek mintegy egyharmadában rontotta a teljesítményt, jellemzően akkor, ha a kapott információ a személyt és nem a feladatot helyezte fókuszba. Bracken és Rose (2011) négy feltételt nevez meg a működő 360 fokos folyamatra: releváns kompetenciák, hiteles adatok, számonkérhetőség, és a teljes szervezetre kiterjedő részvétel. Atwater és Yammarino (1992) az önismereti rést – az önértékelés és a mások értékelése közötti eltérést – emelte be a vezetésfejlesztés szempontjai közé. Mindennek előfeltétele a kollégák közötti pszichológiai biztonság (Edmondson, 1999): annak közös meggyőződése, hogy egy kritikus vélemény megfogalmazása nem jár utólagos hátránnyal. A teljesítményértékelés fejlesztő hatása a szervezeti kultúrától is függ (Bakacsi, 2015; Karoliny & Poór, 2017); esetünkben a feladat- és személykultúra erős jelenléte (Handy, 1993) a többirányú megoldás mellett szólt.

## Saját fejlesztésű webalkalmazás

A piacon elérhető 360 fokos megoldások jellemzően angol nyelvűek, költséges előfizetést igényelnek, vagy nem igazodnak egy magyar közintézmény belső struktúrájához és adatvédelmi kereteihez. A könyvtár saját szerverén futó, PHP-nyelvű webalkalmazás mellett döntöttünk: az adatok a könyvtár infrastruktúráján maradnak, a változtatások bármikor beépíthetők. Az 56 kérdésből álló kérdésegyüttest a könyvtár Minőség Irányítási Tanácsa dolgozta ki, és vezetői körben közösen véglegesítettük. Az alkalmazás két felületre tagolódik.

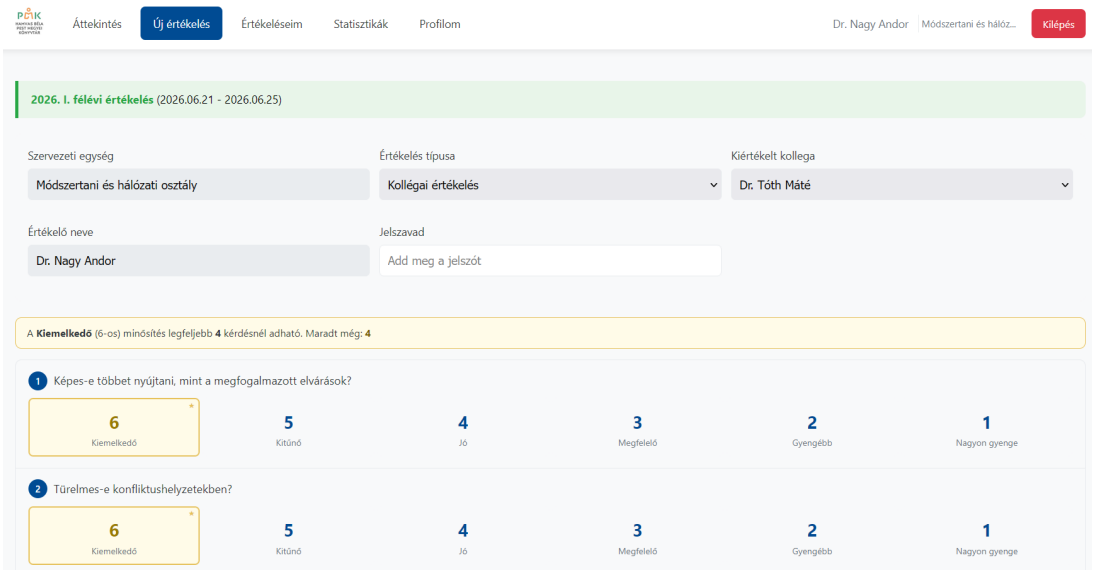



*1. ábra: A munkatársi felület főoldala bejelentkezés után*

A munkatársi felületen minden kolléga elvégezheti az aktív értékeléseket és megtekintheti saját korábbi eredményeit. A vezetői felületet az osztályvezetők, az igazgatóhelyettes és az igazgató éri el; itt kezelhetők az időszakok, az összesített statisztikák, a trendek és az automatizált figyelmeztetések. A rendszer háromféle értékelési típust kezel – önértékelést, felettesit és kollégait –, és időszakonként szabályozható, hogy ki kit értékelhet.

## Az értékelőlap és a tervezési döntések

Az értékelőlap 56 kérdésből áll, három kategóriába sorolva: személyes tulajdonságok, munkahelyi viselkedés és teljesítmény. Minden kérdés egy hatfokú skálán értékelhető.




*2. ábra: Érékelés beküldése a munkatársi felületen*

Két olyan tervezési döntést hoztunk, amelyek a szakirodalomból (Bracken & Rose, 2011; Smither et al., 2005) is levezethetők. Az első: a kérdések véletlenszerű sorrendben jelennek meg, és a kitöltő nem látja, melyik kategóriába tartoznak – így nem tudja utólag „kompenzálni" az egyik blokkban szándékoltnál alacsonyabbra sikerült értékelését. A második: az összesített pontszám és a kategóriánkénti átlag csak az értékelés elküldése után válik láthatóvá. Ha a kitöltő végigkövetheti az aktuális átlagot, hajlamos lehet a szám felé igazítani az utolsó pár pontot.

### Anonimitás és bizalom

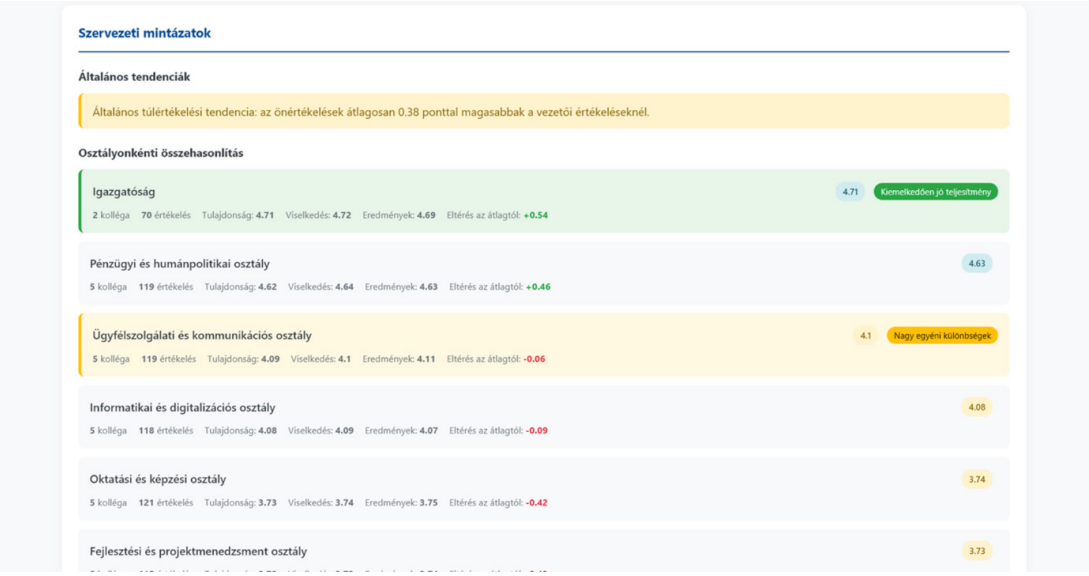
Az anonim értékelés alapértelmezésben be van kapcsolva, de a kitöltő dönthet úgy is, hogy névvel adja le. A megkapott értékelések a munkatársi felületen véletlenszerű sorrendben, időbélyeg nélkül jelennek meg – a beérkezések sorrendje és időpontja gyakran elég ahhoz, hogy egy-egy értékelő utólag azonosítható legyen. A pszichológiai biztonság (Edmondson, 1999) tehát konkrét tervezési döntéseken is múlik.

### A munkatársi és vezetői nézőpont

Az eredményoldalt a munkatárs nézőpontjából terveztük. Mindenki látja saját értékeléseit – az önértékelését, a felettesétől kapottat és a kollégai értékeléseket – kategóriánkénti bontásban, az előző időszakokhoz viszonyított trenddel együtt.


*3. ábra: Automatikusan létrehozott diagramok és statisztikák (részlet) az igazgatói felületen*

A vezetői nézet összesített statisztikákat, osztályonkénti összehasonlítást, trendeket és radardiagramokat ad. Az egyik legfontosabb funkció az automatizált figyelmeztetések rendszere: a szoftver jelzi, ha az önértékelés és a felettesi értékelés között a küszöbértéket meghaladó az eltérés, ha valakinek az átlaga kiemelkedően alacsony, vagy ha az értékelései között a szórás nagy.




*4. ábra: Szervezeti mintázatok az igazgatói felületen (részlet)*

Ezek nem ítéletek, hanem munkaeszközök – arra hívják fel a figyelmet, hogy melyik kollégával érdemes leülni egy hosszabb beszélgetésre. A logika az önismereti rés (Atwater & Yammarino, 1992) koncepciójához illeszkedik: a beszélgetés indulópontját nem a pontszám, hanem az eltérés adja.

## Technikai háttér

A rendszer PHP-ban íródott, SQL adatbázist használ, a könyvtár saját szerverén fut, mobilon is teljes értékűen kezelhető. Minden értékelés elküldéséhez jelszavas megerősítés szükséges, a felület brute-force és több másik technikai védelemmel, valamint mindenre kiterjedő naplózással rendelkezik. Az adatok CSV-formátumban is kimenthetők. Az adminisztrációs felület minden lényeges paramétert testre szabhatóvá tesz – a szervezeti egységeket, a vezetői szerepköröket, az értékelési típusokat és időszakokat, a kérdéseket –, így más, hasonló méretű intézmények saját nevükkel és struktúrájukkal is használhatják.

## Tesztelési tapasztalatok és tanulságok

A rendszer jelenleg tesztelés alatt áll, éles bevezetésére még nem került sor. A szakirodalom szerint az egyik legnagyobb buktató, ha a szervezet túl korán, hibás kérdésekkel vagy hibás kommunikációval indítja el a folyamatot (Bracken & Rose, 2011; Smither et al., 2005). A próbaüzem három tanulsága mégis kirajzolódik. Az első: néhány kérdés többféleképpen volt értelmezhető, ezeket az éles bevezetés előtt pontosítjuk – a tételek egyértelműsége a 360 fokos rendszer hatékonyságának egyik feltétele (DeNisi & Kluger, 2000).

A második: az eredmények kommunikálása legalább annyit számít, mint maga az értékelés. Ha a kollégák kitöltenek egy kérdőívet, de utána semmi nem történik, a következő körben kisebb erőfeszítést tesznek a részletes válaszadásra. A harmadik: a 360 fokos értékelés nálunk azért működhet, mert öt éven át tudatosan építettük a bizalmat. A szoftver a folyamatot támogatja, de a bizalmat nem hozza létre; csak hasznosíthatóvá teszi azt, amit a szervezet már felépített (Edmondson, 1999).

## Záró megjegyzések

A bemutatott rendszer nem újdonság a 360 fokos értékelés szakirodalmában, de az a kontextus, amelyben létrejött, igen: egy magyar közgyűjtemény, ahol az értékelés nem önálló projektként, hanem egy ötéves jólléti program szerves részeként került napirendre. Az alkalmazás más, hasonló adottságú intézmények számára is hasznosítható.

## Hivatkozások


[1] Atwater, L. E., & Yammarino, F. J. (1992). Does self–other agreement on leadership perceptions moderate the validity of leadership and performance predictions? Personnel Psychology, 45(1), 141–164. https://doi.org/10.1111/j.1744-6570.1992.tb00848.x

[2] Bakacsi Gy. (2015). A szervezeti magatartás alapjai: Alaptankönyv bachelor hallgatók számára. Semmelweis Kiadó.

[3] Bracken, D. W., & Rose, D. S. (2011). When does 360-degree feedback create behavior change? And how would we know it when it does? Journal of Business and Psychology, 26(2), 183–192. https://doi.org/10.1007/s10869-011-9218-5

[4] DeNisi, A. S., & Kluger, A. N. (2000). Feedback effectiveness: Can 360-degree appraisals be improved? Academy of Management Executive, 14(1), 129–139. https://doi.org/10.5465/AME.2000.2909845

[5] Edmondson, A. C. (1999). Psychological safety and learning behavior in work teams. Administrative Science Quarterly, 44(2), 350–383. https://doi.org/10.2307/2666999

[6] Handy, C. B. (1993). Understanding organizations (4th ed.). Penguin Books.

[7] Karoliny M.-né, & Poór J. (Szerk.). (2017). Emberi erőforrás menedzsment kézikönyv: Rendszerek és alkalmazások (6. átdolg. kiad.). Wolters Kluwer.

[8] KMD [Könyvtári Minőségi Díj]. (2026, január 20). Átadták a könyvtári és a közművelődési minőségügyi díjakat a Magyar Kultúra Napján. Országos Széchényi Könyvtár. https://oszk.hu/hirek/atadtak-konyvtari-es-kozmuvelodesi-minosegugyi-dijakat-magyar-kultura-napjan_260120

[9] Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. Psychological Bulletin, 119(2), 254–284. https://doi.org/10.1037/0033-2909.119.2.254

[10] Smither, J. W., London, M., & Reilly, R. R. (2005). Does performance improve following multisource feedback? A theoretical model, meta-analysis, and review of empirical findings. Personnel Psychology, 58(1), 33–66. https://doi.org/10.1111/j.1744-6570.2005.514_1.x

DOI: 10.31915/NWS.2026.16



# MESTERSÉGES INTELLIGENCIA ASSZISZTENS FELHŐALAPÚ RAG ARCHITEKTÚRÁBAN

# ARTIFICIAL INTELLIGENCE ASSISTANT IN A CLOUD-BASED RAG ARCHITECTURE

*Kelemen László Ádám*
*Programozó, Azure Cloud & AI Engineer*
*Szegedi Tudományegyetem, Informatikai Fejlesztési Igazgatóság*
*kelemen.laszlo.adam@szte.hu*

## Absztrakt

A nagy nyelvi modellek fejlődése lehetővé tette olyan intelligens asszisztens rendszerek létrehozását, amelyek természetes nyelven támogatják a felhasználókat az információkeresési és ügyintézési feladatok során. A kizárólag generatív modellekre épülő megoldások ugyanakkor intézményi környezetben korlátozott megbízhatóságúak lehetnek, mivel pontatlan vagy nem naprakész válaszokat is előállíthatnak. A tanulmány egy felhőalapú *Retrieval-Augmented Generation* (RAG) architektúrára épülő mesterséges intelligencia asszisztens keretrendszert mutat be, amelyet a Szegedi Tudományegyetem számára fejlesztettünk. A rendszer Azure felhőszolgáltatásokra épül, és hibrid keresést, vektoralapú dokumentum-visszakeresést, szemantikus újrarangsorolást, valamint többfázisú dokumentumfeldolgozást alkalmaz. A dokumentumfeldolgozás során jelenleg két chunking stratégia támogatott: a standard *paragraph-first chunking* és a fejezetszintű struktúrát figyelembe vevő *hierarchical chunking*. A rendszer enterprise retrieval optimalizációkat, jogosultságkezelést és biztonsági mechanizmusokat is tartalmaz. Az eredmények alapján az architektúra alkalmas nagyméretű intézményi dokumentumállományok feldolgozására és kontextusérzékeny válaszok generálására, miközben csökkenti a hallucinációs kockázatot.

**Kulcsszavak:** mesterséges intelligencia, Retrieval-Augmented Generation, RAG, hibrid keresés, vektorkeresés, dokumentum-visszakeresés, felhőalapú architektúra, Microsoft Azure

## Abstract

The advancement of large language models has enabled the development of intelligent assistant systems that support users in natural language information retrieval and administrative tasks. However, solutions relying exclusively on generative models may have lim-

ited reliability in institutional environments, as they can produce inaccurate or outdated responses. This paper presents a cloud-based *Retrieval-Augmented Generation* (RAG) artificial intelligence assistant framework developed for the University of Szeged. The system is built on Microsoft Azure cloud services and employs hybrid search, vector-based document retrieval, semantic reranking, and a multi-stage document processing pipeline. During document processing, the system currently supports two chunking strategies: standard *paragraph-first chunking* and *hierarchical chunking*, which preserves chapter-level semantic structure. The architecture also includes enterprise retrieval optimizations, access control, and security mechanisms. The results indicate that the system is suitable for processing large institutional document collections and generating context-aware responses while reducing hallucination risk.

**Keywords:** artificial intelligence, Retrieval-Augmented Generation, RAG, hybrid search, vector retrieval, document retrieval, cloud architecture, Microsoft Azure

## 1. Bevezetés

A mesterségesintelligencia-alapú asszisztensek fejlődése szorosan kapcsolódik a nagy nyelvi modellek és a Transformer architektúra elterjedéséhez (Vaswani, és mtsai., 2017). Ezek a rendszerek természetes nyelven képesek kommunikálni, kérdésekre válaszolni, valamint információkeresési és ügyintézési folyamatokat támogatni. Intézményi környezetben azonban a kizárólag generatív modellekre épülő megoldások nem minden esetben elegendőek, mivel válaszaik nem feltétlenül ellenőrizhetőek, naprakészek vagy forrásokkal alátámasztottak.

A problémára megoldást jelenthetnek a *Retrieval-Augmented Generation* (RAG) architektúrák, amelyek a generatív modelleket dokumentumalapú információ-visszakereséssel egészítik ki (Lewis, és mtsai., 2020). A RAG-rendszerek a felhasználói kérdéshez releváns dokumentumrészleteket keresnek vissza, majd ezeket kontextusként adják át a nyelvi modellnek. Ez csökkentheti a hallucinációk előfordulását, és lehetővé teszi naprakész intézményi tudásbázisok használatát.

Jelen tanulmány a Szegedi Tudományegyetem számára fejlesztett, felhőalapú mesterséges intelligencia asszisztens keretrendszert mutatja be. A rendszer Microsoft Azure szolgáltatásokra épül, és komplex RAG pipeline-t valósít meg dokumentumfeldolgozással, vektoralapú indexeléssel, hibrid kereséssel, szemantikus újrarangsorolással, valamint enterprise szintű jogosultságkezeléssel. A fejlesztés célja olyan architektúra kialakítása volt, amely nagyméretű intézményi dokumentumállományok esetén is skálázható, biztonságos és alacsony hallucinációs kockázattal működik.

A tanulmány külön hangsúlyt helyez a dokumentumfeldolgozási stratégiákra. A rendszer két chunking (szeletelés) módszert támogat: a *paragraph-first chunking* megközelítést,

amely a paragrafusokat tekinti elsődleges szemantikai egységnek, valamint a *hierarchical chunking* stratégiát, amely a dokumentum fejezetszintű szerkezetét is figyelembe veszi. A további fejezetek bemutatják a rendszer architektúráját, a dokumentumfeldolgozási láncot, a *retrieval pipeline* (visszakeresési folyamat) működését, valamint a biztonsági és enterprise komponenseket.

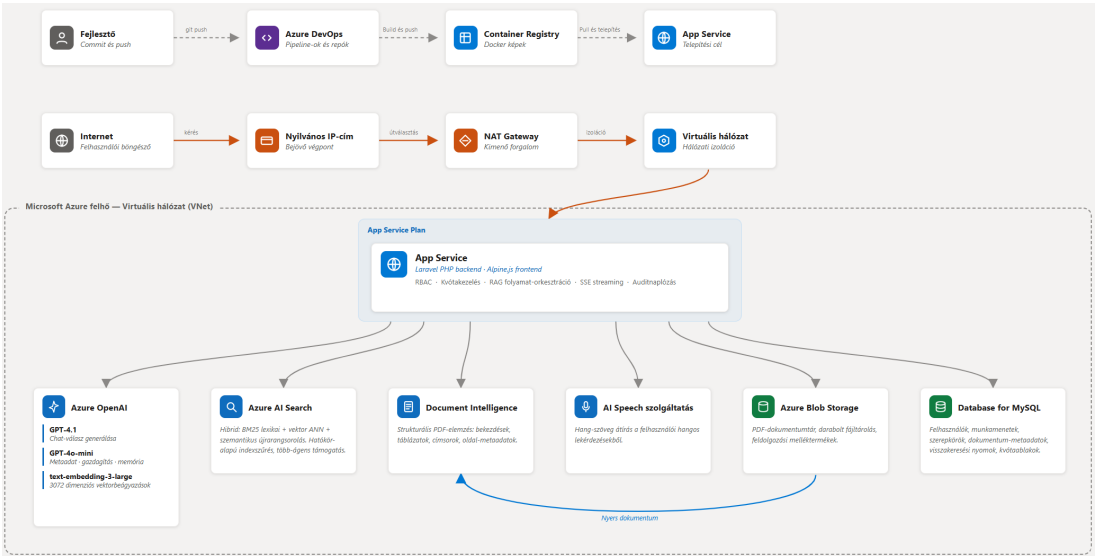
## 2. A rendszer architektúrája

A fejlesztett AI-asszisztens keretrendszer egy felhőalapú, többkomponensű RAG-architektúrára épül. Elsődleges célja intézményi dokumentumállományok természetes nyelvű lekérdezése és forrásalapú válaszok generálása. A rendszer kialakításánál kiemelt szempont volt a skálázhatóság, a modularitás, a dokumentum-hozzáférések kontrollja és az enterprise környezetben elvárt biztonság.

Az infrastruktúra Microsoft Azure szolgáltatásokat használ. A válaszgenerálást Azure OpenAI biztosítja GPT-alapú modellekkel, a dokumentum-visszakeresést Azure Search végzi, míg a PDF dokumentumok strukturális feldolgozását Azure Document Intelligence támogatja. A dokumentumok és kapcsolódó feldolgozási állományok Azure Blob Storage-ban tárolódnak. Az alkalmazás backend rétege Laravel alapú PHP-környezetben, a felhasználói felület pedig Alpine.js és Tailwind CSS technológiákkal készült. A rendszer deployment folyamata konténerizált üzemeltetési modellt alkalmaz, amely Azure Web App környezetben biztosítja a teszt- és éles infrastruktúra skálázható működését. A rendszer strukturált adatainak perzisztens tárolását Azure Database for MySQL szolgáltatás biztosítja.

A rendszer működésének központi eleme a dokumentumfeldolgozási és visszakeresési folyamat. A feltöltött dokumentumok metaadat-generáláson és strukturális elemzésen mennek keresztül, majd szeletelési stratégiák alapján dokumentumszeletekre bomlanak. A szeletek *embedding* (vektoros) reprezentációi a text-embedding-3-large modell segítségével készülnek (OpenAI, 2024), majd Azure Search indexbe kerülnek. A felhasználói kérdések feldolgozásakor a rendszer hibrid keresést alkalmaz, amely a BM25-alapú lexikális keresést, a vektoralapú szemantikus keresést és a szemantikus újrarangsorolást kombinálja (Robertson & Zaragoza, 2009).

Az architektúra több intelligens ágens párhuzamos működését támogatja. Az ágensek eltérő dokumentumkészletekhez, keresési scope-okhoz (hatókör) és jogosultsági szintekhez rendelhetők, így különböző intézményi területek elkülönítetten kezelhetők. A szerepalapú hozzáférés-vezérlés biztosítja, hogy a visszakeresési folyamat csak a felhasználó számára engedélyezett dokumentumokon fusson.




*1. ábra: Az MI-alapú RAG-asszisztens rendszerarchitektúrája Microsoft Azure környezetben.*

## 3. Dokumentumfeldolgozás és chunking (szeletelési) stratégiák

A RAG-rendszerek hatékonyságát jelentősen befolyásolja az indexelt dokumentumszeletek minősége. A túl nagy szeletek ronthatják a visszakeresés pontosságát, a túl kicsik pedig kontextusvesztést okozhatnak. Ezért a rendszerben a dokumentumfeldolgozás nem pusztán technikai előkészítő lépés, hanem a visszakeresés minőségét meghatározó komponens.

A feldolgozási folyamat a PDF-állományok feltöltésével indul. A dokumentumok Azure Blob Storage-ba kerülnek, majd automatikus metaadat-generálás történik, amely a dokumentum címét és rövid leírását állítja elő. A strukturális elemzést Azure Document Intelligence végzi, amely paragrafusokat, táblázatokat és egyéb logikai egységeket azonosít. Ezekhez oldalszám, szerkezeti szerep és egyéb metaadatok társulhatnak.

A rendszer jelenleg két szöveges szeletelési stratégiát támogat. A standard *paragraph-first chunking* a paragrafusokat önálló szemantikai egységként kezeli. A módszer nem alkalmaz oldalhatár-alapú szeletelést, így az összetartozó tartalom akkor is egy chunkban maradhat, ha több oldalra tördelődik. Ez különösen hasznos szabályzatok, tanulmányi dokumentumok és ügyintézési leírások esetén, ahol a fejezetcím és a hozzá tartozó tartalom gyakran eltérő oldalon jelenik meg.

A paragraph-first stratégia puha és kemény karakterkorlátokat használ. A puha korlát a preferált szeletméretet jelöli, míg a kemény korlát a rendkívül hosszú paragrafusok kezelését biztosítja. A rendszer átfedéses szeletelést is alkalmaz, amelyben a szomszédos chunkok bizonyos közös paragrafusokat tartalmazhatnak, ezzel javítva a kontextus folytonosságát.

A *hierarchical chunking* stratégia a dokumentum fejezetszintű szerkezetét is figyelembe veszi. A módszer az azonos logikai egységhez tartozó paragrafusokat közösen kezeli, és a

chunkokhoz fejezethierarchiát kapcsolhat. Ez különösen előnyös nagyméretű, strukturált dokumentumok esetén, ahol a fejezetszintű kontextus megőrzése javíthatja a visszakeresés relevanciáját.

A táblázatos tartalmak feldolgozására külön mechanizmus szolgál. A táblázatok soralapú szeletekre bonthatók, miközben a fejlécinformációk ismétlése biztosítja, hogy az önállóan visszakeresett szeletek értelmezhetők maradjanak. Ez tantervi követelményeket, kreditstruktúrákat vagy adminisztratív adatokat tartalmazó dokumentumok esetén kiemelten fontos.

A feldolgozott dokumentumszeletek vektoros formában kerülnek indexelésre (OpenAI, 2024). Az vektoros generálás kötegelt módon történik, és gyorsítótárazás támogatja, amely csökkenti az ismételt API-hívások számát. A többféle szeletelési stratégia célja, hogy eltérő dokumentumszerkezetek esetén is szemantikailag koherens és hatékonyan visszakeresheto egységek jöjjenek létre.

### 3.1 A két szeletelési stratégia összehasonlítása

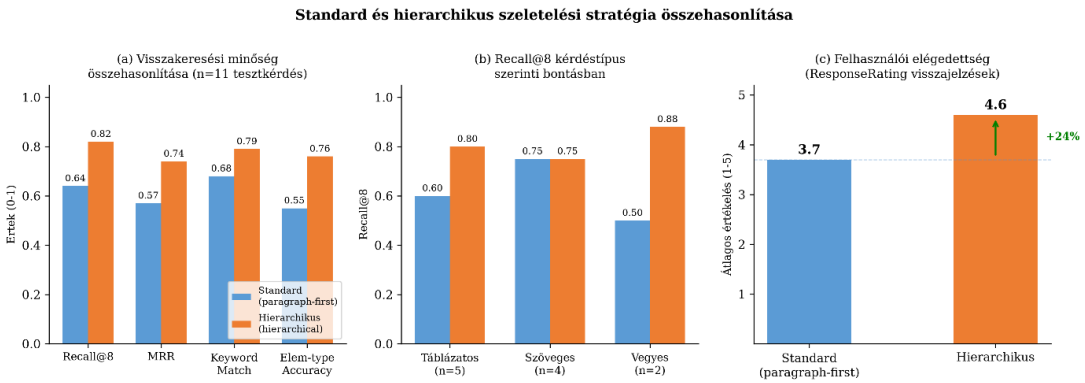
A 2. ábra a standard bekezdésalapú (paragraph-first) és a hierarchikus (hierarchical) szeletelési stratégia összehasonlítását mutatja be három dimenzióban, 11 reprezentatív tesztkérdés alapján.

Az a) diagram négy visszakeresési mutatót szemléltet 0–1 skálán: a *Recall@8* a helyes dokumentum top 8 találat közötti előfordulási arányát; az *MRR* (Mean Reciprocal Rank) a helyes találat átlagos ranghelyének reciprokát (1,0 = mindig első); a *Keyword Match* az elvárt kulcsszavak visszakeresési lefedettségét; az *Element Type Accuracy* a visszaadott tartalom típusának (tábla/szöveg) helyességét méri. A hierarchikus stratégia mindegyik mutatón felülmúlta a standard megközelítést.

A b) diagram a *Recall@8* értékét bontja le kérdéstípusonként: táblázatos lekérdezéseknél a hierarchikus chunking 0,60-ról 0,80-ra javította a visszakeresési arányt, mivel a fejezetszintű kontextus megőrzése révén a táblázatfejlécek és a kapcsolódó sorok szemantikai összefüggése nem vész el az indexelés során, míg szöveges kérdéseknél mindkét stratégia azonosan teljesített (0,75).

A c) diagram a felhasználói visszajelzéseken alapuló elégedettségi átlagot mutatja 1–5 skálán: a hierarchikus stratégia bevezetésével az átlagos értékelés 3,7-ról 4,6-re nőtt, ami 24%-os relatív javulást jelent.

Az eredmények azt mutatják, hogy a hierarchikus szeletelési stratégia különösen hatékony az intézményi környezetre jellemző, erősen strukturált dokumentumok – tantervek, szabályzatok, mobilitási táblázatok – visszakeresésénél, míg szöveges tartalmak esetén a két módszer teljesítménye egyenértékű.




*2. ábra: Szeletelési stratégiák összehasonlítása*

## 4, Retrieval pipeline (visszakeresési folyamat) és válaszgenerálás

A rendszer központi komponense a *Retrieval-Augmented Generation* (Lewis, és mtsai., 2020) pipeline, amely a dokumentum-visszakeresést és a válaszgenerálást egységes folyamatban valósítja meg. Célja, hogy a felhasználói kérdéshez releváns dokumentumszeleteket rendeljen, majd ezek alapján forrásalapú választ generáljon.

A visszakeresési folyamat első lépése a felhasználói kérdés vektoros reprezentációjának előállítása. Ez a vektor az Azure Search hibrid lekérdezésében kerül felhasználásra. A hibrid keresés több módszert kombinál: a BM25-alapú lexikális keresés a kulcsszavas egyezéseket rangsorolja, míg a vektoralapú keresés a kérdés és a dokumentumszeletek szemantikai hasonlóságát vizsgálja (Robertson & Zaragoza, 2009) (OpenAI, 2024). A két megközelítés kombinálása lehetővé teszi a pontos kifejezések és a jelentésalapú egyezések együttes kezelését.

A visszakeresett találatokra szemantikus újrarangsorolás épül. Az Azure Search beépített újrarangsorolója mellett a rendszer saját optimalizációs réteget is alkalmaz, amely figyelembe veszi a relevanciapontszámokat, a dokumentumszeletek metaadatait és a találatok diverzitását. A diverzitási mechanizmus csökkenti annak esélyét, hogy a végső kontextus túl sok, azonos oldalról vagy dokumentumrészből származó szeletet tartalmazzon.

A visszakeresési folyamat dokumentumtípus-specifikus optimalizációkat is támogat. A rendszer képes különbséget tenni szöveges és táblázatos jellegű kérdés-válaszok között, így a strukturált adatokra irányuló lekérdezések más súlyozással kezelhetők. Enterprise környezetben további funkciók is elérhetők, például a lekérdezés-gazdagítás (*query enrichment*), a dinamikus paraméterhangolás és az iteratív visszakeresés. Ezek rövid, többértelmű vagy összetettebb kérdések esetén javíthatják a visszakeresett dokumentumszeletek relevanciáját.

A válaszgenerálás során a kiválasztott dokumentumszeletek kontextusblokként kerülnek a nyelvi modell elé. A rendszer minden szelethez hivatkozási metaadatokat rendel, így a

válaszok nemcsak generált szövegként, hanem konkrét dokumentumforrásokhoz kapcsolva jelennek meg. Ez intézményi környezetben különösen fontos, mert a válaszok pontossága mellett azok visszakövethetősége is alapkövetelmény.

A generált válaszok Azure OpenAI-alapú nyelvi modellekkel készülnek. A rendszer streaming-alapú válaszküldést is támogat, amely a szöveg fokozatos megjelenítésével javítja a felhasználói élményt. A pipeline kialakításának fő célja a hallucinációs kockázat csökkentése: a modell minden esetben explicit dokumentumkontextus alapján állítja elő válaszát.

### 5. Biztonsági és enterprise mechanizmusok

Intézményi környezetben az AI-asszisztensek esetén alapvető követelmény a dokumentumhozzáférések szabályozása, az erőforrás-felhasználás kontrollja és a felhasználói interakciók monitorozása. A fejlesztett rendszer ezért több enterprise szintű biztonsági mechanizmust tartalmaz.

A hozzáférés-kezelés szerepalapú modellre épül. A felhasználók hallgatói, oktatói, adminisztrátori vagy egyéb jogosultsági szintekhez rendelhetők. A dokumentumhozzáférést hatókör-alapú szűrés egészíti ki, amely biztosítja, hogy a visszakeresési folyamat kizárólag a felhasználó számára elérhető dokumentumállományban keressen. A keresési szűrők whitelist (engedélyezési lista)-alapú validációt használnak, így csak előre engedélyezett mezőkön keresztül hajthatók végre.

A hitelesítés elsődlegesen Microsoft Entra ID-alapú egyszeri bejelentkezéssel történik. A rendszer emellett API-alapú hozzáférést is támogat, amely JWT tokenekkel vagy partnerintegráció esetén API-kulccsal működhet. A prompt injection jellegű mintázatok felismerésére bemeneti szűrés szolgál, amely naplózza a gyanús lekérdezéseket.

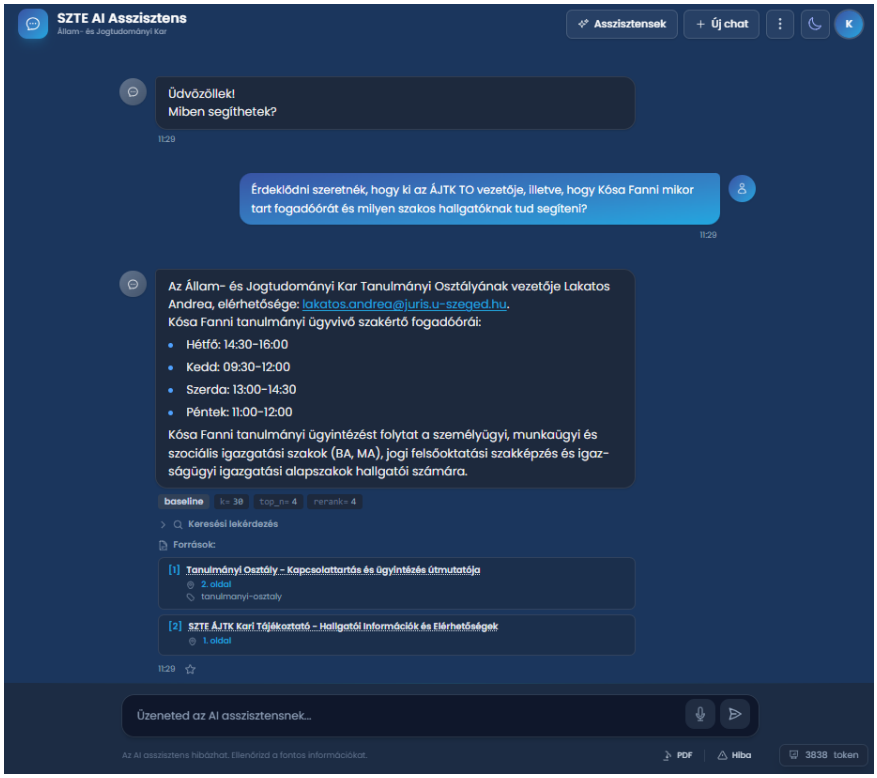
A működési stabilitást tokenkvóta-alapú erőforrás-szabályozás támogatja. A kvóták globális, szerepköri vagy felhasználói szinten határozhatók meg, így megelőzhető a túlzott modellhasználat. Az enterprise pipeline része a visszakeresési nyomkövetés is, amely a keresési paramétereket, iterációkat, időigényt és tokenfelhasználást rögzíti. Ez támogatja a teljesítményelemzést és a későbbi optimalizációt.

A rendszer naplózás-szanitizálási megoldásokat is alkalmaz, amelyek eltávolítják a naplókból az érzékeny adatok egy részét, például e-mail-címeket vagy telefonszámokat. Az adminisztratív műveletek auditnaplózása biztosítja a dokumentumkezelési, jogosultságkezelési és konfigurációs változások visszakövethetőségét. Ezek a mechanizmusok együttesen támogatják a biztonságos és kontrollált intézményi működést.

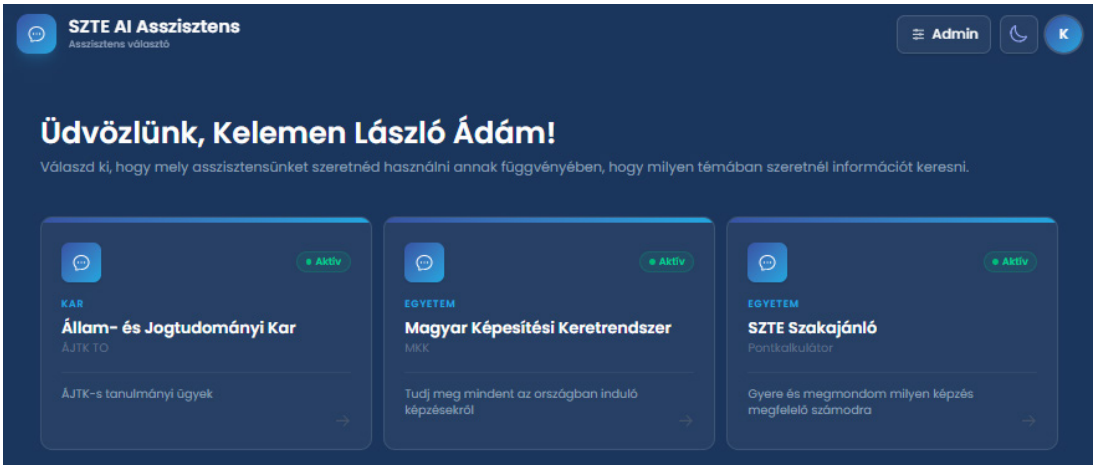
## 6. A rendszer felhasználói felülete

Az alábbi ábrák a fejlesztett mesterséges intelligencia asszisztens keretrendszer egyes felhasználói felületi komponenseit és működési példáit mutatják be. A képernyőképek szemléltetik a természetes nyelvű lekérdezések kezelését, a dokumentumalapú válaszgenerálást és az intézményi környezetben alkalmazott funkcionalitásokat.

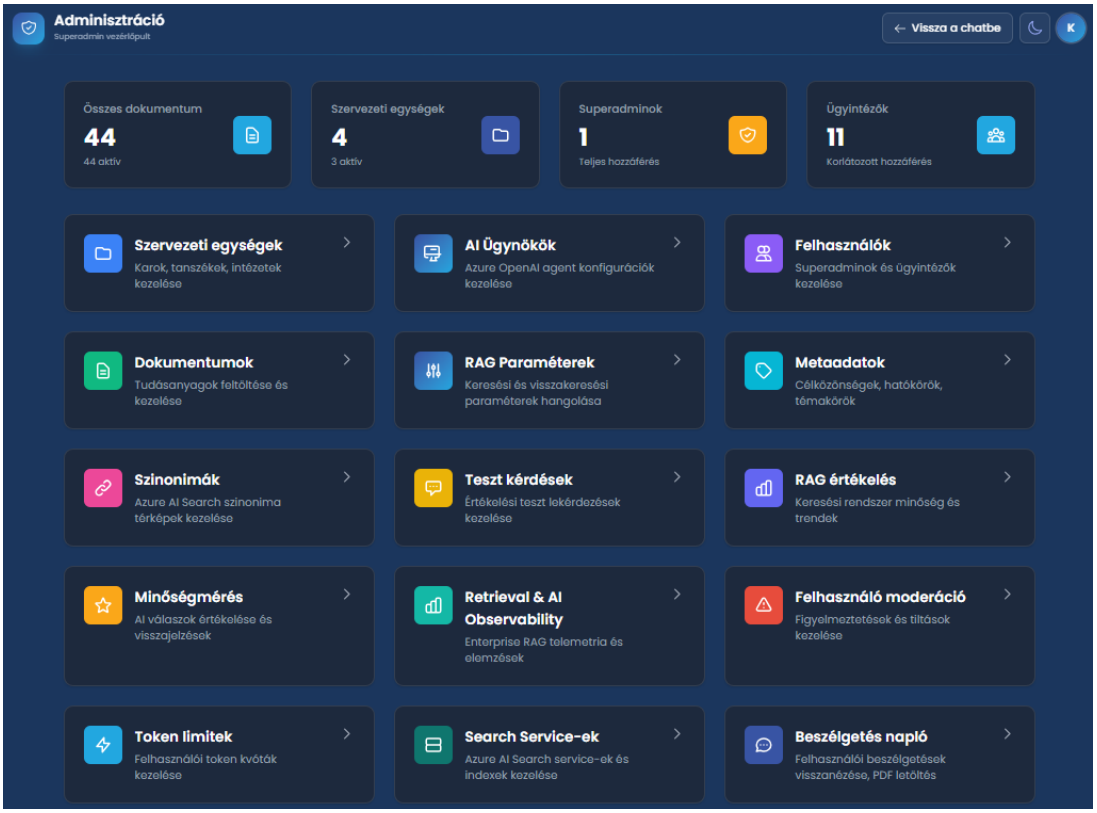

*3. ábra: Belépési felület*




*4. ábra: Dokumentumalapú válaszgenerálás és forráshivatkozások a felhasználói felületen*




*5. ábra: Asszisztensválasztó felület*




*6. ábra: Az AI-asszisztens adminisztrációs és visszakeresési menedzsmentfelülete*

## 7. Következtetések

A tanulmány egy felhőalapú *Retrieval-Augmented Generation* architektúrára épülő mesterséges intelligencia asszisztens keretrendszert mutatott be, amely intézményi dokumentumkezelési és információ-visszakeresési feladatokat támogat. A rendszer hibrid keresést, vektoralapú szemantikus visszakeresést, szemantikus újrarangsorolást és többféle szeletelési stratégiát alkalmaz.

A fejlesztés egyik fontos eredménye, hogy a rendszer dokumentumfeldolgozási lánca képes eltérő szerkezetű intézményi dokumentumok kezelésére. A *paragraph-first chunking* a paragrafusszintű szemantikai koherenciát, míg a *hierarchical chunking* a fejezetszintű kontextus megőrzését támogatja. A visszakeresési folyamat enterprise optimalizációi – például a dinamikus paraméterhangolás és az iteratív visszakeresés – tovább javíthatják a releváns dokumentumszeletek kiválasztását.

A biztonsági komponensek, köztük a szerepalapú hozzáférés-kezelés (RBAC), a hatókör-alapú dokumentumszűrés, a kvótakezelés és a visszakeresési nyomkövetés alkalmassá teszik az architektúrát intézményi környezetben történő használatra. Az eredmények alapján a RAG megközelítés hatékonyan alkalmazható egyetemi dokumentumállományok

természetes nyelvű lekérdezésére, miközben csökkenti a generatív modellek hallucinációs kockázatát és javítja a válaszok visszakövethetőségét.

A jövőbeni fejlesztési irányok közé tartozik a visszakeresési folyamatok további optimalizálása és a találatok relevanciájának növelése, különös tekintettel a különböző szeletelési stratégiák kvantitatív összehasonlítására és a visszakeresési paraméterek automatikus adaptációjára. A fejlesztés további célja az egyetemi tanulmányi rendszerekkel történő mélyebb integráció kialakítása, amely lehetővé teheti adminisztratív és tanulmányi folyamatok intelligens támogatását. Emellett a keretrendszer új funkcionális iránya a hagyományos kérdés-válasz alapú működés kiterjesztése párbeszédorientált, ajánlórendszer jellegű interakciókra. Ennek egyik példája a fejlesztés alatt álló szakajánló asszisztens, amely természetes nyelvű párbeszéd segítségével képes a felhasználói preferenciák és érdeklődési körök alapján releváns képzési ajánlásokat megfogalmazni.

## Irodalomjegyzék


Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., . . . Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Advances in Neural Information Processing Systems 33. 33.* NeurIPS. Forrás: https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html?utm_source=chatgpt.com

OpenAI. (2024). *Vector embeddings*. Letöltés dátuma: 2026. 05 07, forrás: OpenAI API Documentation: https://platform.openai.com/docs/guides/embeddings

Robertson, S., & Zaragoza, H. (2009). The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends in Information Retrieval, 3*. Forrás: https://dl.acm.org/doi/10.1561/1500000019

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., . . . Polosukhin, I. (2017). Attention Is All You Need. *Advances in Neural Information Processing Systems 30. 30.* Long Beach, California: NeurIPS. Forrás: https://papers.nips.cc/paper/7181-attention-is-all-you-need

DOI: 10.31915/NWS.2026.17



# AZ AARC-TREE PROJEKT EREDMÉNYEI – AZ AZONOSÍTÁS ÉS JOGOSULTSÁGKEZELÉS TOVÁBBFEJLESZTÉSE A KUTATÁSI INFRASTRUKTÚRÁK KÖZÖTTI EGYÜTTMŰKÖDÉSHEZ

# RESULTS OF THE AARC-TREE PROJECT – ADVANCING AUTHENTICATION AND AUTHORIZATION FOR COLLABORATION ACROSS RESEARCH INFRASTRUCTURES

*Mohácsi János*
*azonosítási és jogosultsági szolgáltatás felelős, nemzetközi K+F szakértő*
*Pro-M Zrt, GÉANT*
*mohacsi.janos@pro-m.hu*

**Absztrakt**

A modern kutatási folyamatok alapvető feltétele a határokon átívelő, multidiszciplináris együttműködés. Ebben a környezetben a kutatóknak számos különböző erőforráshoz (számítási kapacitás, specifikus szoftverek, adattárolók, adatrepozitóriumok) kell hozzáférniük. Az AARC (Authentication and Authorization for Research and Collaboration) projekt sorozat, és annak legújabb tagja, az AARC-TREE (Authentication and Authorisation for Research and Collaboration – Technical Revision to Enhance Effectiveness) projekt, a Blueprint Architektúra (BPA) továbbfejlesztésével és a szakpolitikai keretrendszer harmonizációjával válaszol ezekre a kihívásokra. A cikk bemutatja az AARC BPA evolúcióját, az interoperabilitást segítő technikai és szabályozási irányelveket, valamint az AARC-kézikönyv szerepét a közösségi tudásmegosztásban.

**Kulcsszavak:** Hitelesítés, jogosultságkezelés, bizalom, kutatási infrastruktúrák, digitális infrastruktúra, identitáskezelés, nyílt tudomány, kutatási együttműködés

**Abstract**

Modern research processes fundamentally depend on cross-border, multidisciplinary collaboration. In this environment, researchers must access a wide range of resources, including computational capacity, data storage, and specialized software. The AARC (Authenti-

cation and Authorization for Research and Collaboration) project series—and its latest step, the AARC-TREE (Authentication and Authorisation for Research and Collaboration – Technical Revision to Enhance Effectiveness) project—addresses these challenges through the further development of the Blueprint Architecture (BPA) and the harmonization of policy frameworks. This paper presents the evolution of the AARC BPA, outlines the technical and regulatory guidelines that support interoperability, and highlights the role of the AARC handbook in fostering community knowledge sharing.

**English keywords:** Authentication, Authorisation, Trust, Research Infrastructures, Digital Infrastructure, Identity management, Open Science, Research Collaboration

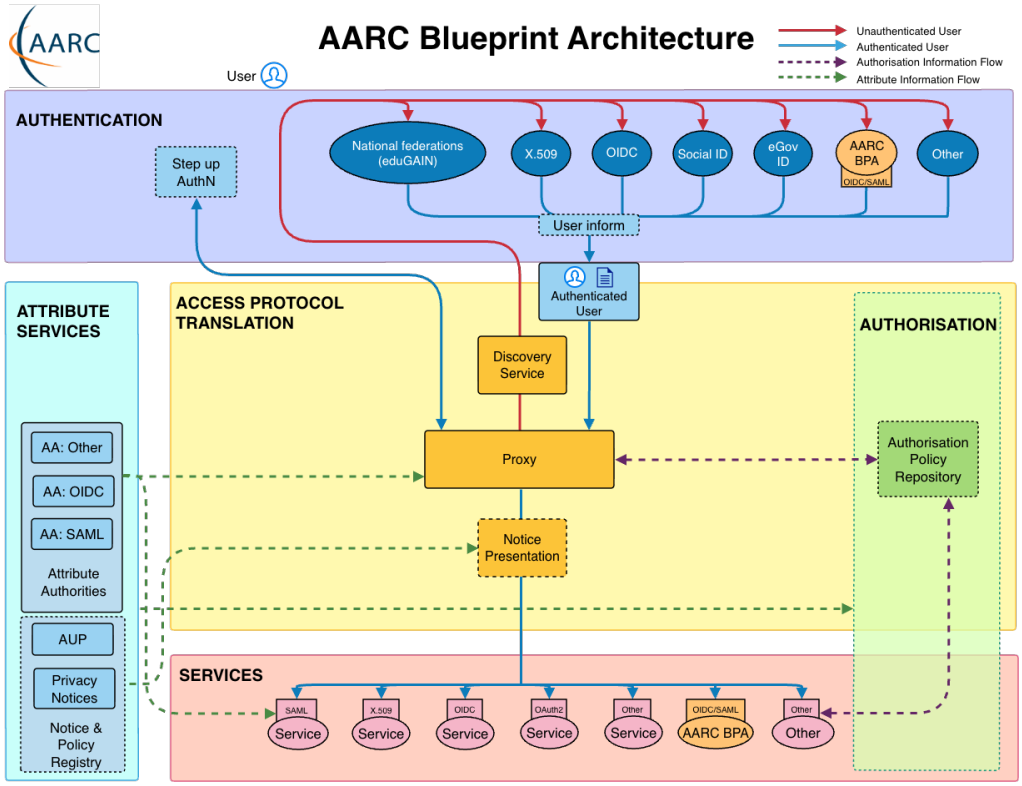
## 1. Bevezetés: A föderatív/szövetségi alapú azonosítás jelentősége

A kutatási ökoszisztémában az adatok és az erőforrások elszórtan, különböző országokban és intézményekben helyezkednek el. A kutatók számára kritikus, hogy saját intézményi azonosítójukkal (Single Sign-On (SSO) – Egyszeri bejelentkezési rendszer) érhessék el a nemzetközi kollaborációk eszközeit.

Az eduGAIN és a kapcsolódó nemzeti azonosítási szövetségek (mint a magyarországi eduID.hu) megteremtették az alapot ehhez a bizalmi hálózathoz (trust network). Azonban a kutatási infrastruktúrák specifikus igényei – például a csoporttagság-kezelés, a különböző szintű biztonsági követelmények és a szolgáltatások közötti adatátvitel – túlmutatnak az alapvető föderatív képességeken. Erre a problémára kínál rendszerszintű megoldást az AARC kezdeményezés.

## 2. Az AARC Blueprint Architektúra (BPA) evolúciója

Az AARC projekt legfontosabb eredménye a **Blueprint Architecture (BPA)**, amely egy referenciamodellként szolgál az azonosítási és jogosultságkezelési infrastruktúrák (AAI – Authentication and Authorization Infrastructure) tervezéséhez.




*1. ábra: AARC BPA architektúra (forrás AARC TREE projekt)*

### 2.1. A BPA rétegei

A BPA architektúra öt fő rétegre tagolódik (1. ábra):

1. **Authentication Layer (Azonosítási réteg) :** A kutatók saját intézményi vagy közösségi azonosítóit (IdP SAML - Security Assertion Markup Language, vagy OIDC -OpenID Connect) kezeli. Tartalmazhat további, AARC-BPA-kompatibilis proxykat és többlépcsős (step-up) hitelesítési mechanizmusokat.
2. **Attribute Service Layer (Attribútumkezelési réteg):** Itt történik a közösség-specifikus attribútumok (pl. kutatói csoporttagságok) és a policy kezelése.
3. **Access Protocol Translation Layer (Hozzáférési protokollfordítási réteg) :** Magában foglalja az SP–IdP–Proxy komponenseket és a Discovery Service-t. Kezeli az értesítések megjelenítését és a protokollok közötti fordítást.
4. **Authorisation Layer (Autorizációs réteg)**: Szabályozza a szolgáltatásokhoz való hozzáférést, alkalmazhat hozzáférési szabályzat-nyilvántartást.

5. **Service Provider (SP) Layer (Szolgáltatási réteg):** A tényleges szolgáltatások (számítási klaszterek, repozitóriumok), amelyek a Proxy-tól kapott információk alapján engedélyezik a hozzáférést. Tartalmazhat AARC-BPA-kompatibilis proxykat láncolt működéshez.

A kutatási infrastruktúrák AAI rendszerei ennek a moduláris architektúrának egyes komponenseit implementálják, hogy a kutatók egyetlen digitális azonosítóval hozzáférjenek különböző országok és kutatási infrastruktúrák szolgáltatásaihoz, valamint a kutatási infrastruktúrák azonosítási és jogosultságkezelési szinten egységes és szabványos rendszerben tudjanak működni.

### 2.2. Miért volt szükség a továbbfejlesztésre (AARC-TREE)?

A technológia fejlődésével és az európai adatterek (pl. EOSC – European Open Science Cloud) megjelenésével új kihívások merültek fel. Az AARC-TREE projekt célja a BPA modernizálása, hogy támogassa a több-proxys modelleket, a magasabb szintű biztonsági elvárásokat és a skálázhatóságot. A legfontosabb új kihívások a következők voltak:

- Felhasználóközpontúság - felhasználói identitás a hitelesítés időpontjában teljes mértékben független bármely konkrét együttműködési kontextustól.
- OIDC-támogatás megvalósítása – modern föderációs megoldás kialakítása, mely interaktív és nem interaktív (gépi) működést is támogat, továbbá a rendszer biztonsága magas szintű lesz.
- Skálázhatóság – kutatási infrastruktúrák és közösségek támogatása az azonosítási kontextusok segítségével.
- Moduláris kialakítás – kutatási infrastruktúrák és közösségek igényeinek figyelembevételével csak a szükséges komponenseket szükséges implementálni
- Kompatibilitás – együttműködési képesség a közösségi központú AARC-architektúrával és a jelenleg fejlesztési alatt álló adattárca megoldásokkal.

A fejlesztés 2 fő irányvonal mentén haladt: új technikai megoldások kialakítása és a szakpolitikai megfelelés lehetővé tétele.

## 3. Szakpolitikai és technikai irányelvek

Az interoperabilitás nem csupán technikai kérdés; egységes szabályozási környezetre (Policy) is szükség van. Az AARC-közösség az AEGIS (AARC Engagement Group for Infrastructures) csoporton keresztül koordinálja ezeket a tevékenységeket.

### 3.1. Technical Working Group

A technikai munkacsoport felelős az implementációs útmutatókért. Olyan kérdésekkel foglalkoznak, mint:

- Az azonosítók perzisztenciája (persistence) és egyedisége (uniqueness);
- Token-alapú hitelesítés (OpenID Connect / OAuth2) integrálása a hagyományos SAML-környezetbe;
- Attribútum-kiadási politikák finomhangolása;
- Egységes technikai szabályok kialakítása.

A technikai munkacsoport munkájába bármely érdeklődő bekapcsolódhat a munkacsoport levelezési listáján keresztül: https://lists.geant.org/sympa/info/aarc-architecture .

### 3.2. Policy Working Group

A szabályozási munkacsoport az etikai és jogi megfelelőséget biztosítja. Kiemelt terület a **Sirtfi (Security Incident Response Trust Framework for Federated Identity)**, amely incidenskezelési keretrendszert ad a föderáció tagjainak. Az irányelvek célja, hogy a szolgáltatók bízni tudjanak a kutatói közösségek által szolgáltatott adatokban.

A szabályozási munkacsoport munkájába bármely érdeklődő bekapcsolódhat a munkacsoport levelezési listáján keresztül: https://lists.geant.org/sympa/info/aarc-na3

## 4. Az AARC Kézikönyv: A tudásbázis

Az AARC-TREE egyik újdonsága az **AARC Handbook**, amely egy strukturált gyűjtemény a bevált gyakorlatokról (Best Practices) és az aktuális AAI-implementációkról. Ez az adatbázis segít az új kutatási projekteknek abban, hogy ne kelljen „újra feltalálniuk a kereket", hanem már bizonyított architektúrákat adaptálhassanak.

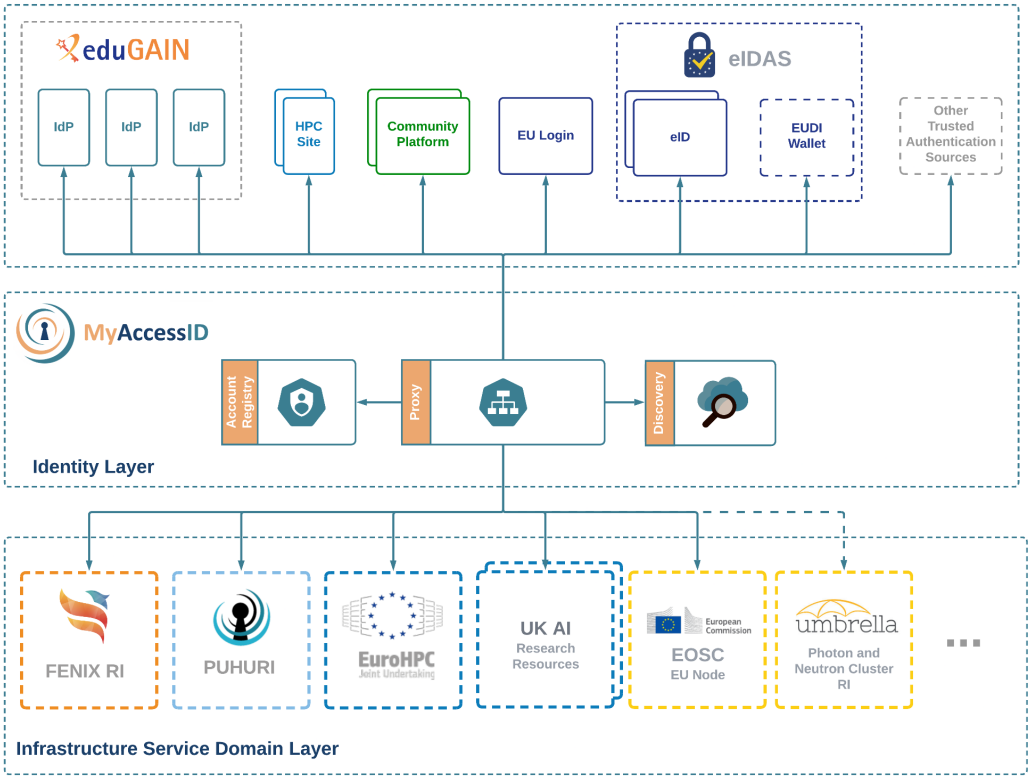
A Kézikönyv tartalma:

- AARC BPA áttekintése és információ arról, hogy miért érdemes alkalmazni;
- AARC-ajánlások és azok kapcsolódásai;
- A meglévő kutatási infrastruktúrák (pl. EGI, ELIXIR, Fenix) AAI-implementációinak áttekintése;
- Gyakori hibaforrások és azok megoldásai;
- Szakpolitikai sablonok;
- Gyakorlati útmutató AARC-kompatibilis azonosítási és jogosultság kezelési infrastruktúra implementációjára.

## 5. Gyakorlati alkalmazási példák

Az AARC BPA-t ma már számos területen alkalmazzák sikerrel (lásd 2. ábra):

- **EOSC (European Open Science Cloud):** Az európai tudományos felhő az AARC modelljére építi saját azonosítási és jogosultságkezelési keretrendszerét – alapvető komponensként építve a GÉANT (nemzetközi oktatási-kutatási hálózat) és oktatási és kutatási hálózatok (NREN – National Research and Educational Network) közössége által fejlesztett MyAccessID megoldásra.
- **Erasmus+ és Európai Egyetemi Szövetségek:** A hallgatói mobilitás és a közös egyetemi kurzusok során is az AARC elveit követik az azonosításnál, szintén építve a GÉANT és NREN-ek megoldásaira.
- **EuroHPC szövetségi szolgáltatási platform:** Az EuroHPC FP (EuroHPC szövetségi szolgáltatási platform) az azonosítás és jogosultságkezelés tekintetében szintén a GÉANT és NREN-nel AARC BPA-kompatibilis MyAccessID-re épít.
- **Hosztolt/Menedzselt megoldások:** Mivel az AAI üzemeltetése komplex feladat, egyre több kutatócsoport választ előre konfigurált, hosztolt proxyszolgáltatásokat, ami növeli a biztonságot és csökkenti a költségeket.
- Magyarország 2025-ben Pro-M Zrt. (2024-ben a KIFÜ) égisze alatt vett részt a AARC TREE projekt megvalósításban és jelenleg vesz részt MyAccessID (GÉANT Core AAI) fejlesztésében, amely az AARC TREE-ben kidolgozott architektúra implementációját jelenti. A MyAccessID alkalmazhatóságát, GÉANT eduGAIN OIDFED-t (AARC-G100 skálázási teszt) a Pro-M a kutatóintézetekkel és egyetemekkel együttműködésben **teszteli a hazai oktatás-kutatás számára.**




*2. ábra: MyAccessID és különböző kutatási infrastruktúrák kapcsolata (forrás: GN5-2 projekt)*

## 6. Összegzés és jövőkép

Az AARC-TREE projekt eredményei kritikus fontosságúak a jövő kutatási infrastruktúrái számára. Az egységesített BPA modell és a hozzá kapcsolódó szabályozási keretrendszer lehetővé teszi, hogy a kutatók ne az adminisztratív akadályokkal, hanem magával a tudományos munkával foglalkozhassanak.

Az AARC-TREE projekt ajánlásokat 15 ország 22 partnere dolgozta ki, melyeket Európán belül mintegy 40 ország, de azon kívül még további 40+ ország (USA, Kanada, Ausztrália, Japán, Szingapúr, Malajzia, India, Brazília, afrikai országok stb.) fog alkalmazni a saját kutatástámogatási rendszereikben. Európai szinten ezt az architektúrát alkalmazza már az EOSC (2028-ra), EuroHPC (már működik) és a hallgatói és oktató mobilitást támogató Erasmus programban is teljeskörűvé válik 2030-ra.

Magyarország számára az AARC-TREE legfontosabb eredménye, hogy a hazai kutatási és felsőoktatási infrastruktúrák felkészülhetnek az európai kutatási térben egyre inkább egységessé váló digitális azonosítási, hozzáférés-kezelési és együttműködési modellekre, amely épít a 2012 óta folyamatosan bővülve 79 taggal működő eduID.hu magyar azonosítási szövetség eredményeire.

Ennek gyakorlati hatása az, hogy a magyar intézmények technológiailag és szolgáltatás szempontjából alkalmassá válnak arra, hogy a jövőben ne egyedi integrációs megoldásokon keresztül, hanem egységes európai mechanizmusok mentén kapcsolódjanak a nemzetközi kutatási szolgáltatásokhoz és együttműködésekhez.

A jövő iránya a még szorosabb integráció a különböző európai adatterek irányába (pl. Egészségügyi adattér – Health Data Space), a felhasználó-központú ellenőrizhető hitelesítési adatok (verifiable credentials) és digitális adattárcákat (digital wallet) alkalmazó technológiák folyamatos beépítése a kutatási folyamatokba. Az AARC-közösség továbbra is nyitott, és várja mindazon infrastruktúrák csatlakozását, amelyek BPA-kompatibilis azonosítást kívánnak megvalósítani.

## 7. Referenciák


[1] AARC Project Website: https://aarc-project.eu

[2] GÉANT eduGAIN Service: https://edugain.org

[3] AARC Blueprint Architecture Guidelines (2025/2026 update): - https://aarc-community.org/Architecture/

[4] AARC Kézikönyv: https://wiki.geant.org/spaces/AARC/pages/127860738o/AARC+-Handbook

[5] AARC Engagement Group for Infrastructures (AEGIS) dokumentáció: https://wiki.geant.org/spaces/AARC/pages/123765937/AEGIS

DOI: 10.31915/NWS.2026.18



# 5G SA FUNKCIÓ MÉRÉSI MÓDSZERE

# MEASUREMENT METHODOLOGY FOR 5G STANDALONE FUNCTIONS

*Soós Gábor*
**BME-TMIT –Távközlési és Mesterséges Intelligencia Tanszék**
soos@tmit.bme.hu

*Ormos Pál*
*HUN-REN SZTAKI – Department of Network Security and Internet Technologies*
ormos.pal@sztaki.hun-ren.hu

*Kis-Bogdán Kolos*
*HUN-REN SZTAKI – Department of Network Security and Internet Technologies*
kis-bogfan.kolos@sztaki.hun-ren.hu

*Kovács Bertalan*
*HUN-REN SZTAKI – Department of Network Security and Internet Technologies*
kovacs.bertalan@sztaki.hun-ren.hu

*Kozma Márk*
*MAGYAR TELEKOM NYRT. –Technology Unit*
kozma.mark1@telekom.hu

*Seres Tamás*
*MAGYAR TELEKOM NYRT. –Technology Unit*
seres.tamask@ext.telekom.hu

*Leleszi András*
*MISKOLCI EGYETEM*
andras.leleszi@uni-miskolc.hu

**Absztrakt**

A publikáció egy reprodukálható mérési eljárást mutat be az 5G standalone (5G SA) hálózatszeletelés (slicing) működésének validálására nagy forgalmú környezetben, továbbá röviden összegzi a helyszíni tesztek eredményeit. A cikkben összefoglaljuk az 5G SA slicing mobilhálózati architektúráját, a slice/service type (SST/SD) szerepét, valamint a konfigurált Quality of Service (QoS) paraméterek és a felhasználói élményt kifejező Quality of

Experience (QoE)-mutatók közötti kapcsolatot. Ezt követően célzott irodalomkutatásra alapozva meghatározunk egy mérési architektúrát, amelyet felhasználva aktív méréseket végzünk végberendezés-oldali és a valós forgalmazáshoz közeli mérőjellel.

**Kulcsszavak:** 5G SA, mobilhálózat, mérési módszertan

## Abstract

This paper presents a reproducible measurement methodology and corresponding results for verifying the operation of 5G SA network slicing in high-traffic environments. It also provides a brief overview of the findings from on-site measurements. The paper outlines the architecture of 5G SA network slicing, the role of SST/SD, and the relationship between configured Quality of Service (QoS) parameters and Quality of Experience (QoE) indicators.

Based on targeted literature review, a measurement architecture is proposed that relies on active measurements performed at the end-device side, using traffic patterns closely resembling real user behavior.

**Keywords:** 5G SA, mobile network, measurement methodology

## I. Bevezetés

Az 5G network slicing (hálózatszeletelés) egyre inkább megkerülhetetlen mind az ipar, mind a lakosság számára nyújtott szolgáltatás területén [1].

Jól megfigyelhető trend, hogy a mobilhálózatok fejlődésével egyre inkább a hangalapú kommunikáció háttérbe szorulásával az állandó jelenlét és az egyre nagyobb hálózati forgalom lesz az, ami új kihívások elé állítja a szolgáltatókat. Ahhoz, hogy ezeket a hálózati igényeket a különböző szereplők ad-hoc vagy állandó igényeihez tudjuk igazítani, elengedhetetlenné vált a network slicing funkció kialakítása. Ez viszont csak 5G SA esetében lehetséges.

## II. Alapfogalmak

### II.1. 5G – Non standalone (NSA)

Az 5G NSA-architektúra olyan átmeneti 5G-hálózati megoldás, amelyben az 5G New Radio (NR) hozzáférési technológia a meglévő 4G Long Term Evolution (LTE) maghálózatára, azaz az Evolved Packet Core (EPC) hálózatra támaszkodik. Ebben a modellben a vezérlési sík, a Control Plane (CP), továbbra is a 4G-hálózaton keresztül működik, míg a felhasználói adatforgalom, a User Plane (UP), részben vagy egészben 5G-kapcsolaton valósul meg.

Az NSA célja a gyors 5G bevezetés, mivel nem igényel teljesen új maghálózatot, hanem a 4G és az 5G együttműködésére (dual connectivity) épít. Ennek a megoldásnak azonban

megvan az a hátránya, hogy így nem lehet teljes mértékben kihasználni az 5G nyújtotta képességeket, mint például az különösen alacsony késleltetést (ultra low latency).

### II.2.5 G – Standalone (SA)

Az 5G SA-architektúra olyan teljes értékű 5G-hálózati megoldás, amelyben az 5G NR hozzáférési technológia egy dedikált 5G-maghálózatra, azaz az 5G Core (5GC) hálózatra épül függetlenül a korábbi 4G-infrastruktúrától. Ebben a modellben mind a CP, mind a UP teljes mértékben 5G-hálózaton keresztül valósul meg.

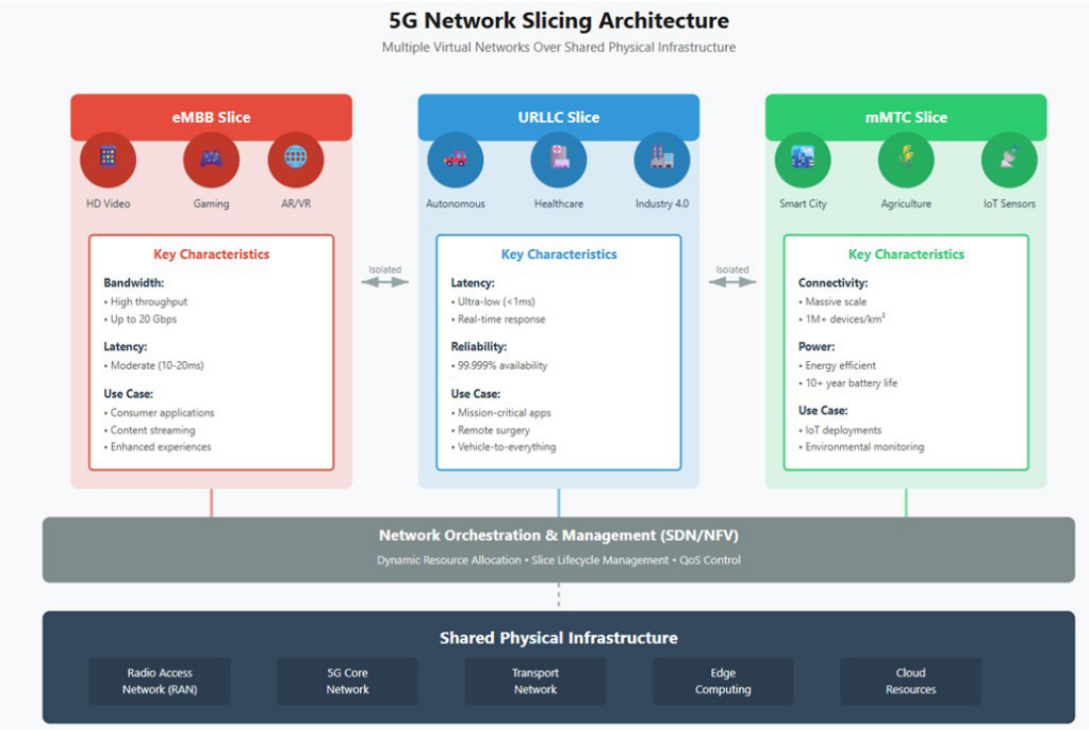
Az SA lehetővé teszi a 5G teljes funkcionalitásának kihasználását. Támogatja az olyan fejlett szolgáltatásokat, mint a hálózati szeletelést (network slicing) és a különösen alacsony késleltetést, az Ultra-Reliable Low-Latency Communications-t (URLLC).

## II.3. Network Slicing

Az 5G network slicing olyan hálózati funkcionalitás, amely lehetővé teszi egy fizikai 5G-infrastruktúrán belül több, egymástól logikailag elkülönített, végponttól végpontig (end-to-end) működő virtuális hálózat (esetünkben szelet) létrehozását. Ezek a szeletek különböző szolgáltatási követelményekhez (pl. késleltetés, sávszélesség, megbízhatóság) optimalizálhatók, miközben közös fizikai erőforrásokat használnak.

Lehetővé teszi különböző szolgáltatástípusok párhuzamos kiszolgálását – pl. enhanced Mobile Broadband (eMBB), URLLC, massive Machine-Type Communications (mMTC) – ld. II.4. és II.5. fejezetek –, amely felépítését az 1. ábra írja le.

A szoftveralapú hálózati megoldásokkal – mint a Software-Defined Networking (SDN) és a Network Functions Virtualization (NFV) – dinamikusan létrehozható és menedzselhető hálózati szeletek mindegyike testre szabott QoS-paraméterekkel rendelkezik. Ennek a rugalmasságnak köszönhetően ez a technológia kulcsszerepet játszik az olyan kritikus ipari alkalmazásokban, mint például az okosgyárak és az autonóm járművek működtetése.




*1. ábra 5G network slicing archtitektúra* [2]

### II.4. enhanced Mobile Broadband (eMBB)

Az 5G SA-hálózatokban az eMBB szelettípus a nagy sávszélességű, adatintenzív szolgáltatások kiszolgálását célozza, különös tekintettel az olyan helyszínekre, ahol az egységnyi területre jutó kapcsolódó készülékek száma extrém magas, például tömegrendezvények. A mobilhálózatok működése alapvetően kapacitásorientált, ahol a hálózat a magas adatsebesség és a stabil kapcsolat biztosítására optimalizál, miközben a késleltetési követelmények jellemzően mérsékeltebbek. Ennek megfelelően az eMBB tipikusan az általános mobil szélessávú forgalom kiszolgálására szolgál, és sok esetben alapértelmezett szeletként jelenik meg a nem kritikus szolgáltatások számára.

### II.5. Ultra-Reliable Low-Latency Communications (URLLC) és massive Machine-Type Communications (mMTC)

Az eMBB mellett a másik két alapvető 5G-szolgáltatási kategória az mMTC és az URLLC, amelyek eltérő szolgáltatási követelményeket fednek le. Az mMTC célja nagyszámú, alacsony adatforgalmú eszköz skálázható kiszolgálása, elsősorban Internet of Things (IoT) alkalmazások esetén, ahol az energiahatékonyság és a hálózati kapacitás eszközszám szerinti optimalizálása kerül előtérbe. Ezzel szemben az URLLC a kritikus, alacsony késleltetést és nagy megbízhatóságot igénylő szolgáltatásokat támogatja, szigorú QoS-köve-

telmények mellett. A három szelettípus egymást kiegészítve teszi lehetővé a különböző szolgáltatási igények kiszolgálását, amely network slicing alkalmazásával logikailag elkülönítve, közös fizikai infrastruktúrán valósítható meg.

### II.6. Slice/Service Type (SST)

5G SA esetében az SST egy 8 bites azonosító, amely a hálózati szelet szolgáltatási típusát jelöli, és megjelöli a szelet alapvető jellemzőit.

Az SST alapján különböztethetők meg például az eMBB, URLLC és mMTC típusú szeletek.

### II.7. Slice Differentiator (SD)

Az SD több fő funkciót valósít meg az 5G SA esetében. Ezek a következők:

- Ugyanazon SST-hez tartozó szeletek (slice-ok) között tesz különbséget, így több, egymástól független szelet is létrehozható azonos szolgáltatási kategórián belül.
- Szolgáltatások finom szétválasztása, amely lehetővé teszi, hogy azonos típusú szeleteken belül eltérő QoS-igényű szolgáltatások külön kezelhetők legyenek.
- A szelet kiválasztás segítése során a hálózat és a végberendezés az SD segítségével tudja kiválasztani a megfelelő szeletet.
- A szolgáltatói testreszabás támogatásával az operátor külön szeleteket alakíthat ki például különböző ügyfeleknek vagy felhasználási esetekre, még azonos SST mellett is. Erőforrás-kezelés támogatásával hozzájárul ahhoz, hogy a hálózat a különböző szeleteket eltérő módon kezelje (pl. prioritás, késleltetés, sávszélesség).

### II.8. Quality of Experience (QoE)

5G SA esetében a QoE a felhasználó által ténylegesen érzékelt szolgáltatásminőséget jelenti, vagyis azt, hogy mennyire jó a felhasználói élmény. A QoE azt méri, hogy az adott szolgáltatás (pl. videó, játék, hanghívás, ipari alkalmazás) mennyire felel meg a felhasználói elvárásoknak valós körülmények között.

Az 5G SA-hálózatokban a QoE értékelése során a felhasználói élményt közvetlenül befolyásoló, mérhető paraméterek kerülnek előtérbe, amelyek a hálózati QoS-jellemzők hatását tükrözik az adott alkalmazás szintjén. A QoE-paraméterek jellemzően alkalmazásfüggők, azonban általánosan az alábbi fő mérőszámokkal írhatók le:

- **Észlelt késleltetés (end-to-end latency):** A felhasználó által érzékelt válaszidő, amely különösen kritikus valós idejű alkalmazások (pl. hang- és videóhívás, URLLC szolgáltatások) esetén.

- **Késleltetés-ingadozás (jitter):** A késleltetés változékonysága, amely a folyamatos médiaátvitel (pl. videóstreaming, Voice over Internet Protocol (VoIP)) minőségét befolyásolja.
- **Csomagvesztés hatása:** Az elveszett adatcsomagok aránya és annak felhasználói élményre gyakorolt hatása, például kép- és hangkimaradások formájában.
- **Hasznos adatsebesség (goodput):** A ténylegesen elérhető átbocsátóképesség a felhasználó számára, amely meghatározza például a letöltési időket és a videó felbontását.
- **Szolgáltatás-folytonosság (service continuity):** A kapcsolat megszakítás-mentessége, beleértve az újracsatlakozási eseményeket és az esetleges szolgáltatáskimaradásokat.
- **Indítási késleltetés (startup delay):** Az alkalmazás vagy szolgáltatás elindításához szükséges idő (pl. videólejátszás kezdete).
- **Pufferelési események (stalling):** A lejátszás közbeni megakadások száma és időtartama, különösen streaming szolgáltatások esetén.
- **Felhasználói minőségi mutató (pl. Mean Opinion Score (MOS)):** Összetett, gyakran modellalapú mutató, amely a felhasználó szubjektív élményét számszerűsíti.

Ezen paraméterek együttesen írják le a QoE-t, és lehetővé teszik a különböző 5G-szeletek (eMBB, mMTC, URLLC) teljesítményének összehasonlítását felhasználói szempontból. A QoE-mutatók alkalmazása különösen fontos a hálózati szeletelés környezetben, mivel közvetlen visszacsatolást adnak a konfigurált QoS-paraméterek gyakorlati hatásairól.

## III. Kihívások

### III.1. Tesztelésre használt eszközök

A teszteléshez többféle eszközt és szoftvert használtunk. A tesztelésre használt hálózatok közül azonban némely eszköz nem tudott felcsatlakozni az 5G SA-ra, csak 5G NSA-ra, így a további teszt lépések nem voltak futtathatók a tervek alapján.

#### III.1.1. Mobilkészülékek

A tesztelésre az elterjedtebb gyártók mobiltelefonjait terveztük használni. A könnyen beszerezhető típusokból több Android és iOS-rendszerű készüléket megvizsgálva, megállapítottuk, hogy csak erősen limitált számú készülék képes felcsatlakozni az 5G SA-hálózatra. A várakozásokkal ellentétben az egyik kevésbé elterjedt és már nem forgalmazott régebbi Android típus minden probléma nélkül csatlakozott mind az 5G NSA-, mind az 5G SA-hálózatra.

### III.1.2. 5G-képes routerek

Összehasonlítás céljából méréseket végeztünk, ill. terveztünk méréseket végezni különböző 5G-képes mobilhálózati routerekkel is.

Látókörünkbe került több ipari kivitelű mobilhálózati eszköz, azonban ezekkel sajnos nem sikerült csatlakozni az 5G SA-hálózatra. Ezzel szemben tesztelésre gyártott nem ipari kivitelű 5G modem HAT for Raspberry Pi [6] és egy tesztelési célú 5G routerrel sikerült méréseket végezni mind az 5G SA-, mind az 5G NSA-hálózaton.

### III.1.3. Szoftverek

Egy fizetős, mind Android, mind iOS által támogatott komplett mérő szoftvert használtunk, de mivel a fentebb ismertetett problémákba ütköztünk a mobil telefonok kapcsán, sajnos ez a szoftver nem volt használható, mert a tesztelésre használt Androidban található chipsethez nincs gyártói támogatás és a routereken sem használható.

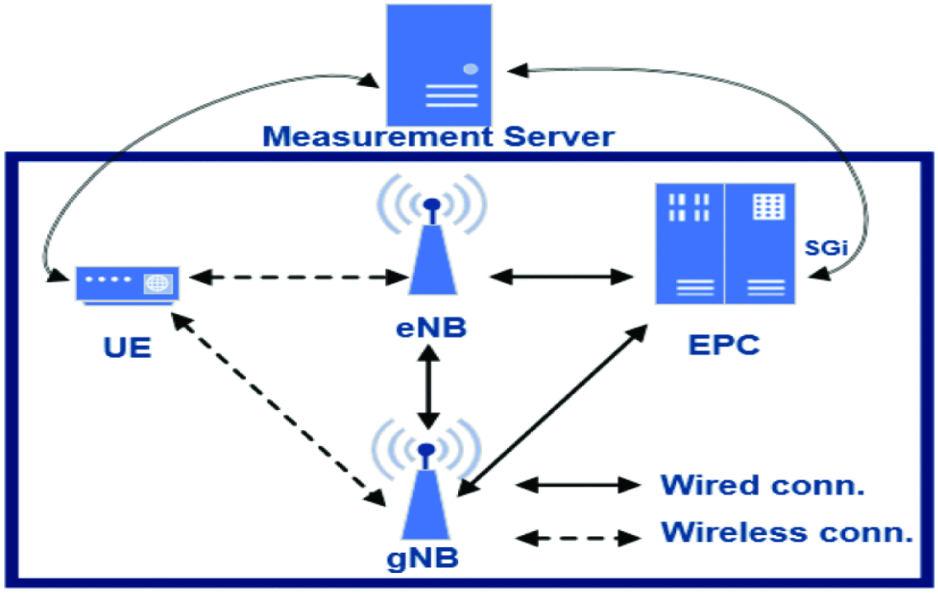
Mindezk miatt az ingyenes szoftverek működését és együttműködését vizsgáltuk. A mérési tervekhez a már általunk jól ismert iPerf3 és Speedtest mellé a Flent (The FLExible Network Tester (https://flent.org/)) szoftverre esett a választás. A következőkben ismertetésre kerülő mérési eljárásban már ezeket alkalmaztuk. Ezeket sikerült mind a routerekkel mind a telefonnal integrálva használni.

## IV. Mérési módszerek

### IV.1. Korábban alkalmazott módszertan

A [3] forrás szerzői a méréseiket laboratóriumi körülmények között végezték, ahol csak a tesztrendszerben részt vevő eszközök képezték a hálózat szerves részét, a natív háttérforgalmat egy dedikált berendezés segítségével generálták. A teljesítménymérések elvégzésekor maga az 5G NSA-hálózat (User Equipment (UE), eNodeB (eNB), gNodeB (gNB) és Evolved Packet Core (EPC)) töltötte be a Device Under Test (DUT) szerepét és egy szeparált mérőszerver hajtotta végre az egyes elemi méréseket (lásd a 2. ábrán). A mérési környezetben a mérőszerver egyik interfészét kábelen keresztül csatlakoztatták az UE-hez, míg a másik interfészét az EPC SGi interfészéhez kapcsolták.

A mérőszerver a küldött és fogadott csomagok alapján kiszámította a legfontosabb teljesítménymutatókat, köztük a késleltetést (latency), az átbocsátóképesség (throughput) és a csomaghibaarányt (packet error ratio). Ezen mutatók megméréséhez elsősorban az iPerf, az iPerf3 és a Cisco TRex szoftvereszközöket használták, az eredményeket pedig a Wireshark csomagelemző segítségével ellenőrizték. Az értékelés keretében mind az UDP-, mind a TCP-adatfolyamok átvitelét megvizsgálták, emellett a TRex szoftver segítségével magas frekvenciával IP-csomagokat is küldtek.

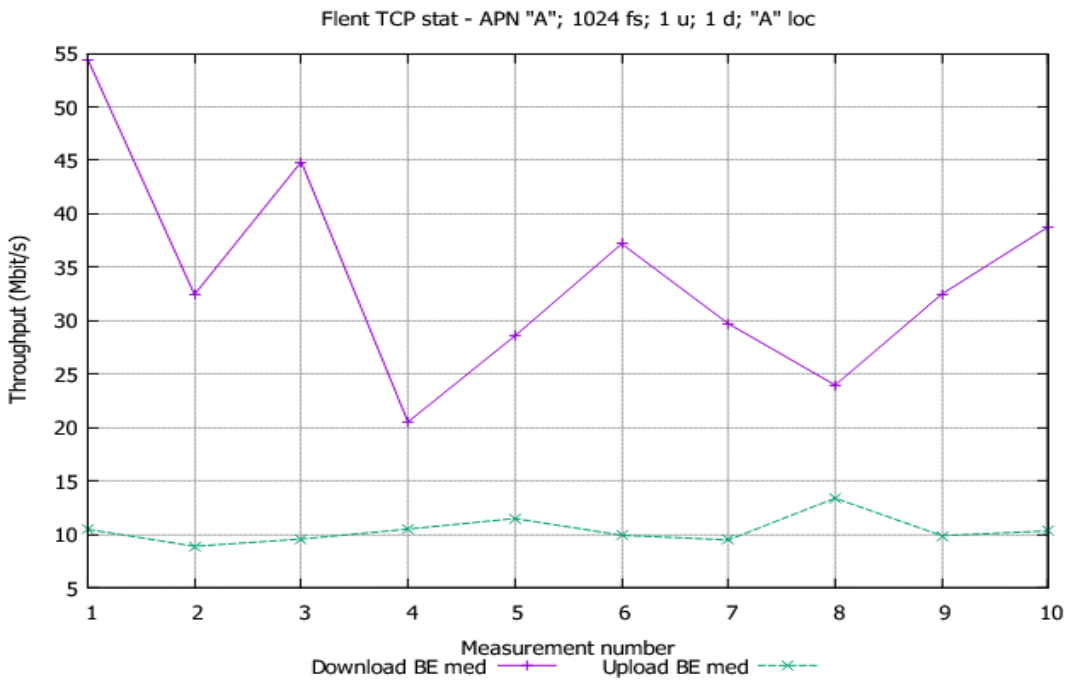



*2. ábra. Az 5G-tesztkörnyezet architektúrája* [3]

**IV.2. Javasolt mérési módszertan**

A korábban bemutatott, szigorúan laboratóriumi körülmények között végrehajtott vizsgálatokkal szemben méréseiket zárt tesztrendszerben, dedikált háttérforgalommal végezték. Ezzel szemben az általunk kidolgozott új mérési módszertan egy kiterjesztett, valós körülményeket is figyelembe vevő paramétertérben vizsgálja az 5G hálózatok teljesítményét.

A mérési eljárások és a kiértékelt teljesítménymutatók tekintetében a módszertanunk szorosan épít a [3] által is alkalmazott protokollokra, azonban jelentősen kibővíti azokat az RFC 2544 hálózati tesztelési szabvány irányelvei alapján. Az RFC 2544 az IETF által definiált, globálisan elfogadott benchmark módszertan, amely egységes és objektív keretrendszert biztosít a hálózati eszközök teljesítményének mérésére [4]. Legfőbb előnye a gyártófüggetlenség és a magas fokú reprodukálhatóság, mivel olyan egzakt paraméterek vizsgálatára összpontosít, mint az átbocsátóképesség, a késleltetés és a csomagvesztés. Alkalmazása lehetővé teszi az adatátviteli képességek precíz, összehasonlítható dokumentálását.




*3. ábra. Az első helyszínen mért mérési eredmény kimutatása*

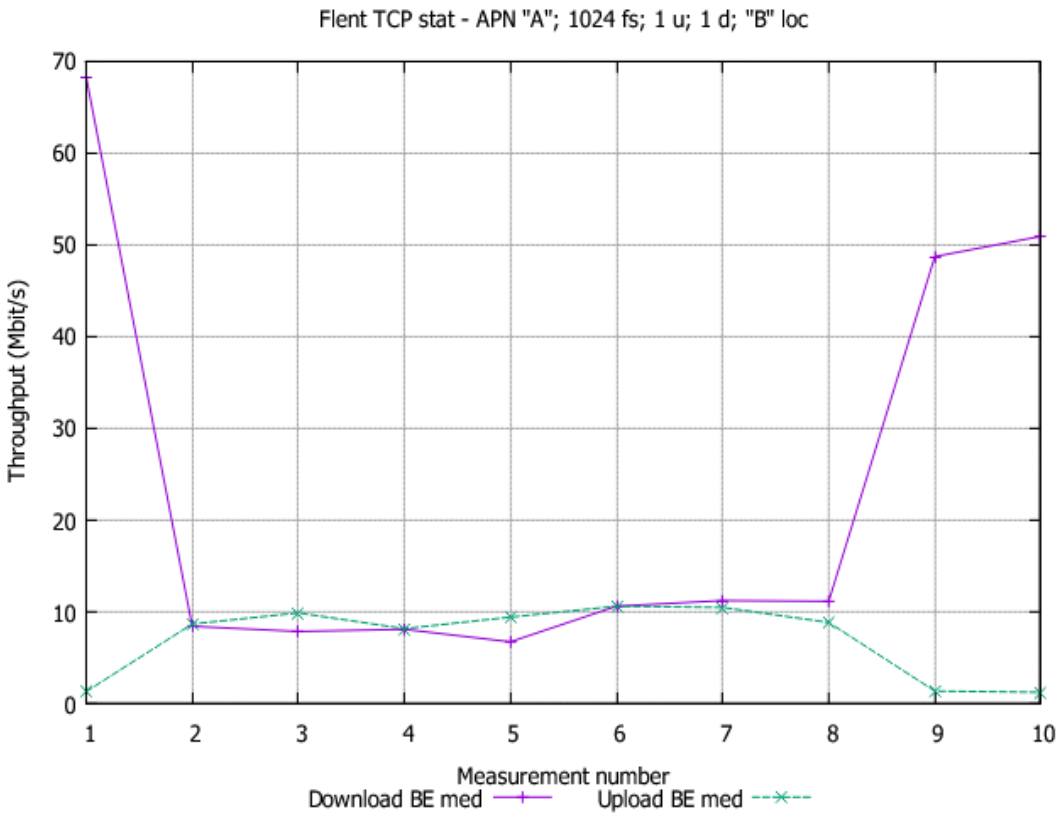
A komplex értékelés érdekében egy háromtengelyes mérési mátrixot definiáltunk. A mátrix első tengelye az időkorrelációs vizsgálatokat foglalja magában: ennek során ugyanazon a bázisállomáson (Base Station Subsystem (BSS)) gyűjtünk adatokat eltérő időpontokban, például különböző napokon, vagy éppen hálózati csúcsterhelést jelentő események alkalmával. A második tengelyt a helykorrelációs mérések adják, amelyek jól definiált geográfiai mérési pontokon, valós térben vizsgálják a szolgáltatás minőségét. Végül a harmadik tengely a hálózati szeletelés gyakorlati képességeit elemzi, ahol a hálózatra csatlakozó végberendezések különböző hálózati szeletek kihasználásával kerülnek tesztelésre.

Az RFC 2544 ajánlásait szigorúan követve a teszteket különböző, standardizált keretméretekkel hajtjuk végre, emellett a forgalmat dedikáltan vizsgáljuk felfelé irányuló (uplink), lefelé irányuló (downlink), valamint fel-le irányú (bidirectional) terhelés esetén is [4]. A kinyert metrikák az áteresztőképesség és a késleltetés mellett kiegészülnek a csomagkésleltetés-ingadozás (packet delay variation (PDV)) vizsgálatával is, hiszen ez a mutató elengedhetetlen a valós idejű szolgáltatások (pl.: videokonferencia) vizsgálatakor [4].

### IV.3. A mérési módszertan validálása

A mérési módszertan gyakorlati alkalmazhatóságának és helytállóságának igazolására előzetes próbavizsgálatokat végeztünk két különböző földrajzi helyszínen. Az eljárás tesztelése során a hálózati forgalom generálására és a teljesítménymutatók rögzítésére a nyílt forráskódú Flent szoftvereszközt alkalmaztuk. [5]

A teljesítménymérések során a hálózati terhelést egyidejűleg indított fel- és leirányú (uplink/downlink) TCP-adatfolyamokkal vizsgáltuk. A konzisztens és statisztikailag igazolható eredmények érdekében fix 1024 bájtos csomagméretet alkalmaztunk, és mindkét helyszínen 10 egymást követő iterációt hajtottunk végre. A mérések során azonos APN-re kapcsolódott a mérőeszköz, vagyis a két mérés között a fizikai helyszín tért csak el. A két mérési ponton rögzített TCP fel- és letöltési eredményeket a [3.] és a [4.] ábrák szemléltetik.




*4. ábra. A második helyszínen mért mérési eredmény kimutatása*

### IV.4. A mérési eredmények értékelése

A mérési eredmények értékelése alapján éles kontraszt rajzolódik ki a két hálózati végpont TCP-átviteli teljesítménye között: míg a 3. ábrán bemutatott „A" lokáció viszonylag megbízható és stabil kapcsolatot biztosít 20,53 és 54,45 Mbit/s között ingadozó letöltési, valamint egyenletesen 8,91 és 13,41 Mbit/s közötti feltöltési sebességgel, addig a 4. ábrán látható „B" lokáció kifejezetten volatilis és aszimmetrikus hálózati viselkedést mutat. Ez utóbbi helyszínen a kiugróan magas, 48–68 Mbit/s-ot is elérő letöltési tüskék a feltöltési sávszélesség kritikus (1,2–1,4 Mbit/s) visszaesését okozzák.

## V. Összefoglalás

Cikkünkben áttekintő jelleggel ismertettük az 5G NSA- és SA-hálózatok közötti alapvető architekturális és működésbeli különbségeket, valamint bemutattuk az új generációs 5G SA-mobilhálózatokon végzett adatátviteli mérések sajátosságait és legfontosabb kihívásait. Részletesen elemeztük a kutatás során azonosított mérési problémákat, majd ismertettük az ezek kezelésére kidolgozott és a gyakorlatban is megvalósított mérési módszertant. Bemutattuk továbbá azokat a mobilhálózati mérőeszközöket és szoftvereket, amelyek a mérések végrehajtását és az adatok megbízható kiértékelését tették lehetővé. Végezetül ismertettük a kialakított mérési eljárás alkalmazásával kapott eredményeket, valamint azok értékelését és a levont következtetéseket.

## Hivatkozások


[1] Google LLC, "5G network slicing — Android Open Source Project — source.android.com." https://source.android.com/docs/core/connect/5g-slicing, 2025. (Letöltés ideje: 2026. január 16.).

[2] Google LLC, "5g network slicing — tehltechworld." https://techlteworld.com/5g-network-slicing/, 2025. (Letöltés ideje:2026. január 19.)

[3] G. Soóos, D. Ficzere, P. Varga, and Z. Szalay, "Practical 5G KPI Measurement Results on a Non-Standalone Architecture," in NOMS 2020 - 2020 IEEE/IFIP Network Operations and Management Symposium, pp. 1–5, 2020

[4] Benchmarking Methodology for Network Interconnect Devices." RFC 2544,Mar. 1999

[5] T. Høiland-Jørgensen, C. A. Grazia, P. Hurtig, and A. Brunstrom, "Flent:The FLExible Network Tester," in Proceedings of the 11th EAI International Conference on Performance Evaluation Methodologies and Tools, VALUE-TOOLS 2017, (New York, NY, USA), p. 120–125, ssociation for Computing Machinery, 2017.

[6] Google, „SIM8202 Raspberry" https://www.rpibolt.hu/SIM8202G-M2-5G-HAT-for-Raspberry-Pi-5G/4G/3G?srsltid=AfmBOopX2Pb0FuvVCR7EU70JDJjFOVbGXt9TdpknY-VQ59kl0HqfqBah1

DOI: 10.31915/NWS.2026.19



# JOGI ÉS TECHNOLÓGIAI KIHÍVÁSOK AZ ELEKTRONIKUS MEGFIGYELÉSBEN A KÖZNEVELÉSI INTÉZMÉNYEKBEN

# LEGAL AND TECHNOLOGICAL CHALLENGES OF ELECTRONIC SURVEILLANCE IN PUBLIC EDUCATION INSTITUTIONS

**Dr. Albert Ágota LL. M**
*Karinthy Frigyes Gimnázium*
*dralbertagota@gdprszakszeruen.hu*

## Absztrakt

Az elektronikus megfigyelőrendszerek alkalmazása a köznevelési intézményekben az elmúlt években jelentősen megnövekedett, ideértve a kamerás megfigyelés, a biometrikus azonosítás és az automatizált beléptető rendszerek használatát is. De vajon a biztonsági, vagyonvédelmi vagy éppen adminisztratív célok mennyiben igazolják ezt az „innovációt”? Az új technológiák új problémákat teremtenek különösen azért, mivel az érintettek jelentős része gyermek, illetve kiszolgáltatott helyzetben lévő munkavállaló.

Jelen tanulmány áttekinti az elektronikus megfigyelés legfontosabb jogi és technológiai kihívásait a köznevelési intézményekben, különös tekintettel a GDPR alapelveire, a szükségesség és arányosság követelményére, valamint a biometrikus adatkezelések kockázataira. A tanulmány célja annak bemutatása, hogy a technológiai lehetőségek önmagukban nem elegendőek a jogszerűség igazolására, és az oktatási intézmények számára komplex adatvédelmi, adatbiztonsági és szervezeti megfelelési kötelezettségek keletkeznek.

**Kulcsszavak:** GDPR, elektronikus megfigyelés, CCTV, biometrikus adat, köznevelés, adatvédelem

## Abstract

The use of electronic surveillance systems in public education institutions has significantly increased in recent years, including the use of CCTV systems, biometric identification technologies, and automated access control solutions. However, to what extent do security, asset protection, or administrative objectives justify this form of “innovation”? New technologies create new challenges, particularly because a significant proportion of the affected individuals are children or employees in vulnerable positions.

This study reviews the most important legal and technological challenges of electronic surveillance in public education institutions, with special focus on the principles of the GDPR, the requirements of necessity and proportionality, and the risks associated with biometric data processing. The aim of the study is to demonstrate that technological possibilities alone are not sufficient to justify lawfulness, and that educational institutions face complex obligations regarding data protection, information security, and organizational compliance.

**Keywords:** GDPR, electronic surveillance, CCTV, biometric data, education, data protection

## 1. Bevezetés

A digitális technológiák fejlődése az oktatási intézmények életére is kihat, és a köznevelési intézményekben is egyre gyakoribbá váltak a kamerás megfigyelőrendszerek, az elektronikus beléptető rendszerek és bizonyos esetekben a biometrikus azonosítási megoldások alkalmazása. Ezen rendszerek bevezetésének célja jellemzően a vagyonvédelem, a személyi biztonság növelése, az illetéktelen belépések megakadályozása, illetve az intézményi incidensek kezelése.

A köznevelési intézmények azonban sajátos környezetet jelentenek nemcsak alapjogi, hanem adatvédelmi szempontból is. Az érintettek jelentős része gyermek, akik a hatályos adatvédelmi szabályozása alapján fokozott védelemre jogosultak. A gyermekek sok esetben nem képesek teljes mértékben felmérni az adatkezelések hosszú távú következményeit, illetve a jogaik érvényesítése terén is hátrányos helyzetben vannak, ezért az intézményekre fokozott felelősség hárul. A köznevelési intézmények ráadásul olyan intézmények, ahol a jelenlét és a részvétel sok esetben nem önkéntes jellegű, ezért az érintettek mozgástere korlátozott – és ez nemcsak a gyermekekre, illetve azok szüleire jellemző, hanem az intézményekben dolgozó munkavállalókra (pedagógusokra, valamint a pedagógusok munkáját segítő személyekre) is igaz, azaz az ő kiszolgáltatott helyzetükre is figyelemmel kell lenni az új technológiák bevezetésekor.

Az elektronikus megfigyelés technológiai fejlődése az utóbbi években új dimenziókat nyitott. A hagyományos CCTV-rendszereket egyre gyakrabban egészítik ki intelligens elemző funkciók, mozgásérzékelő algoritmusok, rendszámfelismerés, arcfelismerés vagy automatizált viselkedéselemzés. Ezek a rendszerek már nem pusztán rögzítenek, hanem képesek személyek azonosítására, események kategorizálására és automatikus döntések támogatására is úgy, hogy általánosan elérhetővé váltak a széles közönség (azaz jelen esetben az oktatási intézmények) számára is.

A technológiai fejlődés ugyanakkor gyakran gyorsabb, mint az intézmények adatvédelmi és szervezeti felkészültsége. Éppen ezért egy kamerarendszer telepítése első látásra „csak”

technikai projektnek tűnhet, valójában azonban komplex jogi, adatbiztonsági és szervezeti kérdéseket vet fel. Már a rendszer tervezési szakaszában felmerülhetnek olyan problémák, mint például a kamerák látóterének jogszerű kialakítása, a szomszédos ingatlanok érintettsége, a közös adatkezelés kérdése, az adatfeldolgozók megfelelő kiválasztása, illetve az érintetti jogok érvényesítésének módjai.

## 2. Az elektronikus megfigyelés adatvédelmi keretei

Az elektronikus megfigyelőrendszerek működtetése során elsődlegesen az Európai Parlament és a Tanács (EU) 2016/679 rendelete (GDPR) alkalmazandó. Emellett relevánsak a tagállami adatvédelmi szabályok, a köznevelési szektorális jogszabályok, az adatvédelmi hatóság és az Európai Adatvédelmi Testület iránymutatásai, valamint a hazai és uniós hatósági és bírósági jogalkalmazás is.

A GDPR alapján a kamerafelvételek és biometrikus azonosítók személyes adatnak minősülnek, amennyiben azokon/azokkal egy természetes személy azonosítható vagy azonosítottá válik. Az arcfelismerésre vagy ujjlenyomatra épülő rendszerek esetében ráadásul a személyes adatok különleges kategóriájába tartozó adatok kezeléséről beszélünk, amelyre szigorúbb szabályok alkalmazandók. Éppen ezért, amennyiben kamerarendszert kívánunk telepíteni az intézményünkben, számos körülményt nem hagyhatunk figyelmen kívül. Jelen tanulmány célja ezen körülmények elemzése, a terjedelmi korlátok miatt nem a teljesség igényével.

### 2.1. A jogszerűség és átláthatóság követelménye

A GDPR egyik alapelve a jogszerűség, tisztességes eljárás és átláthatóság elve. Ez a gyakorlatban azt jelenti, hogy adatkezelőként jogszerűen kell eljárnunk, valamint pontosan meg kell határoznunk a tervezett adatkezelésünk célját, jogalapját és körülményeit, azaz a megfigyelésünk nem történhet általános, korlátlan („parttalan", „figyeljünk meg mindent és mindenkit") vagy meghatározatlan (homályos, „majd utólag kitaláljuk") célból. Amennyiben csak „biztonság" vagy „saját biztonság" céljából van szükségünk videokamerás megfigyelésre, az nem lesz kellően konkrét megfogalmazás és nem fogja alátámasztani a megfigyelésünk jogszerűségét.

A gyakorlatban ez mit jelent? Azt, hogy előre meg kell határoznunk – többek között, de nem kizárólag –, hogy milyen területeket akarunk megfigyelni, milyen célból történik majd a megfigyelés, kik férhetnek hozzá a felvételekhez, mennyi ideig őrizzük meg a felvételeket (adatokat), és milyen esetekben és hogyan használjuk fel a felvételeket.

Az átláthatóság követelménye alapján az érintetteket megfelelő módon tájékoztatnunk kell. Ez a tájékoztatás azonban nem merülhet ki egy, a bejárat közelében lehelyezett figyelmeztető piktogramban. Adatkezelőként részletes adatkezelési tájékoztatót[1] kell biztosítanunk, amelyből az érintettek (pedagógusok, gyermekek, egyéb, az intézmény területén dolgozó-tartózkodó személyek) és a szülők is pontosan megismerhetik az adatkezelésünk körülményeit, ideértve az adatkezelés céljait, az adatkezelő kilétét és az érintett jogainak meglétére és gyakorlására vonatkozó részletes tudnivalókat, valamint az adatkezelés legnagyobb hatásaival kapcsolatos információkat.

### 2.2. Célhoz kötöttség és funkcióbővülés

A GDPR célhoz kötöttség elvének megfelelve személyes adat csak meghatározott, egyértelmű és jogszerű célból kezelhető. Ez különösen fontos az oktatási intézményekben, ahol a biztonsági célból gyűjtött adatokat könnyen más célokra is fel lehetne használni. Gyakorlati problémát jelenthet például, ha a vagyonvédelmi célból telepített kamerákat munkavállalók vagy a diákok ellenőrzésre használjuk, a tanórákat megfigyeljük, illetve ez a megfigyelés a pedagógusok értékelésének eszközzé válik, valamint a beléptetési adatokat jelenléti vagy teljesítményelemzési célokra használjuk.

A modern rendszerek esetében különösen jelentős kockázat az úgynevezett „function creep", vagyis a funkcióbővülés jelensége. Egy eredetileg egyszerű kamerarendszert – különösebb nehézségek nélkül – később arcfelismerő vagy viselkedéselemző funkciókkal egészíthetjük ki, amely már teljesen más adatvédelmi kockázati szintet jelent. Az, hogy egy CCTV rendszer képes bizonyos funkciókra, még nem jelenti azt, hogy az a funkció jogszerű.

### 2.3. Szükségesség és arányosság

A GDPR egyik legfontosabb adatvédelmi követelménye a szükségesség és arányosság elve. Az érdekek mérlegelése során nem elegendő arra hivatkoznunk, hogy a megfigyelési rendszerünk növeli a biztonságot, azt is bizonyítanunk kell, hogy a választott megoldás valóban szükséges, és nem létezik kevésbé korlátozó, a megfigyeltek privát szférájába kevésbé behatoló alternatíva.

Az intézményünk főbejáratának kamerás megfigyelését indokolhatjuk például vagyonvédelmi vagy biztonsági célokkal, ugyanakkor az osztálytermek folyamatos megfigyelése, a diákok biometrikus azonosítása vagy az automatizált viselkedéselemzés már súlyos alapjogi kérdéseket vet fel, különös tekintettel például a tanítás szabadságára, valamint a biometrikus adatok kezelésének korlátaira.

---

1 lásd EDPB 3/2019. számú iránymutatása a személyes adatok videoeszközökkel történő kezeléséről, az első szintű tájékoztatásnak (figyelmeztető tábla) információ-tartalma (114. pont), https://www.edpb.europa.eu/sites/default/files/files/file1/edpb_guidelines_201903_video_devices_hu.pdf

A szükségességi-arányossági vizsgálat során figyelembe kell vennünk az adatkezelésünk célját, az érintettek körét, a kezelt adatok érzékenységét, az adatkezelésünk időtartamát, a megfigyelésünk intenzitását, valamint az esetleges alternatív megoldásokat. A szükségesség esetében az adattakarékosság elve is mindig irányadó, azaz a gyűjtött személyes adatoknak az adatkezelésünk céljai szempontjából megfelelőnek és relevánsak kell lenniük, és a szükségesre kell korlátozódniuk.

Adatkezelőként kötelesek vagyunk felmérni, hogy hol és mikor szükséges feltétlenül a videokamerás megfigyelés – például más megfigyelést tehet szükségessé a tanítási időn belüli és kívüli időszak, valamint az oktatási intézményben is tilos megfigyelni olyan területeket, amely a munkavállalók pihenésére, kötetlen beszélgetésére szolgál (például a tanári szoba). Hasonlóan azt is át kell gondolnunk, melyek a megfigyelésünk szükséges határai, és a megfigyelésünk szempontjából nem releváns területeket ki kell takarnunk (például mosdókba, öltözőkbe nem láthatunk be). Előfordulhat, hogy a kamerák látószögébe beleesik a szomszédos ingatlan területe, ekkor vagy kitakaráshoz (maszkoláshoz, kikockázáshoz) folyamodhatunk, vagy a megfigyelt terület használójával megállapodásban rögzíthetjük a megfigyelés körülményeit (például hozzájárulás, közös adatkezelésre vonatkozó megállapodás).

## 3. Biometrikus rendszerek és technológiai kockázatok

A biometrikus azonosítási technológiák alkalmazása az egyik legérzékenyebb terület az elektronikus megfigyelésen belül, mivel az arcfelismerés, az ujjlenyomat-azonosítás vagy más biometrikus megoldások olyan különleges személyes adatokat kezelnek, amelyek az érintettek egyedi azonosítására alkalmasak. A GDPR 9. cikkének (1) bekezdése tiltja a biometrikus adatok kezelését, illetve a (2) bekezdése fogalmazza meg ezen tiltás alóli kivételeket. Ezen kívül tagállami jogalkotás is szűkítheti az adatkezelők mozgásterét.

Az európai adatvédelmi gyakorlat több olyan esetet is ismer, ahol oktatási intézmények biometrikus rendszereinek alkalmazását kifogásolták. Egy katalán iskola például arcfelismeréses rendszert vezetett be a diákok jelenlétének ellenőrzésére. Az adatvédelmi hatóság álláspontja szerint ez a rendszer aránytalan volt, mivel a cél hagyományos módszerekkel is elérhető lett volna. Emellett az intézmény nem biztosított megfelelő tájékoztatást a szülők számára, és nem rendelkezett megfelelő adatfeldolgozási megállapodással sem. Hasonló vitákat váltott ki egy francia iskola arcfelismeréses beléptetőrendszere is, ahol az azonosítás egyes esetekben (például ikrek esetében) ujjlenyomat-kezeléssel egészült ki.

Az ilyen, biometrikus azonosítást alkalmazó rendszerek kapcsán felmerül a kérdés, hogy az oktatási intézményekben valóban szükséges-e ennyire intenzív megfigyelési technológiák alkalmazása, és a nemzetközi esetek közös tanulsága az, hogy az ilyen intézményekben a biometrikus adatok kezelése általában nem szükséges és nem arányos, azaz nem elfogadható.

Miért? A biometrikus adatok egyik legnagyobb sajátossága azon kívül, hogy egy adott személyt konkrétan beazonosítanak az, hogy azok nem módosíthatók. Egy jelszó vagy belépőkártya cserélhető, azonban az arckép vagy az ujjlenyomat nem, ezért például egy esetleges adatvédelmi incidens következményei hosszú távon fennmaradhatnak. Éppen ezért, amennyiben biometrikus adatokat kezelő rendszert szeretnénk alkalmazni, tisztában kell lennünk annak jelentős kockázataival, így például a jogellenes adatkezelés lehetőségével, a nem megfelelően alkalmazott jogalappal, a hozzájárulás érvénytelenségével, az elégtelen tájékoztatás kockázataival, az adatbiztonsági hiányosságok, a túlzott adatgyűjtés és az algoritmikus hibák és torzítások magas kockázatával.

Különösen problémás a hozzájárulás kérdése. Az adatkezelők szeretnek hozzájárulásra hivatkozni, azonban az oktatási környezetben nehezen biztosítható a hozzájárulás egyik alapkövetelménye, az önkéntesség, mivel a diákok és szülők sok esetben kiszolgáltatott helyzetben vannak. Hasonló probléma merül fel a pedagógusok és egyéb munkavállalók esetében is, ahol a munkaviszonyból (köznevelési foglalkoztatotti jogviszonyból) eredő alá-fölérendeltség megkérdőjelezheti a hozzájárulás valódi önkéntességét.

Az új generációs megfigyelőrendszerek gyakran mesterséges intelligenciára épülnek, és képesek lehetnek arcok automatikus felismerésére, mozgások elemzésére, viselkedési minták azonosítására, illetve rendellenes események automatikus jelzésére. Ezek a technológiák ugyanakkor számos és igen jelentős hibalehetőségeket hordoznak, például az algoritmusok téves azonosításokat eredményezhetnek, különösen rossz fényviszonyok, tömeges környezet vagy nem megfelelő tanítóadatok esetén. A hibás felismerések diszkriminatív következményekkel is járhatnak, bizalmatlanságot szülve és mérgezve az intézményünk légkörét.

A mesterséges intelligencia alkalmazása további kérdéseket vet fel az automatizált döntéshozatal és az emberi felülvizsgálat, mint követelmény tekintetében is. Különösen aggályos lehet, ha az oktatási intézményekben automatizált rendszerek alapján történik személyek gyanúsnak minősítése vagy bizonyos magatartások előszűrése, éppen ezért nemcsak a GDPR rendelkezéseit kell szem előtt tartanunk, hanem a mesterséges intelligencia alkalmazásáról szóló rendeletet[2] is.

A kockázatok csökkentése érdekében a GDPR 35. cikke szerinti adatvédelmi hatásvizsgálatot kell lefolytatnunk, mely során adott esetben ki kell kérnünk az érintettek képviselőinek véleményét, valamint az adatvédelmi tisztviselő szakmai tanácsát.

2 AZ EURÓPAI PARLAMENT ÉS A TANÁCS 2024. június 13-i (EU) 2024/1689 RENDELETE a mesterséges intelligenciára vonatkozó harmonizált szabályok megállapításáról, valamint a 300/2008/EK, a 167/2013/EU, a 168/2013/EU, az (EU) 2018/858, az (EU) 2018/1139 és az (EU) 2019/2144 rendelet, továbbá a 2014/90/EU, az (EU) 2016/797 és az (EU) 2020/1828 irányelv módosításáról (a mesterséges intelligenciáról szóló rendelet)

## 4. Adatbiztonsági és szervezeti kihívások

Az elektronikus megfigyelőrendszerek működtetése nemcsak adatvédelmi, hanem komoly adatbiztonsági kihívásokat is jelent. A kamerafelvételek és biometrikus adatok érzékeny információknak minősülnek, ezért megfelelő technikai és szervezeti intézkedéseket kell alkalmaznunk.

Az egyik legfontosabb kérdés, hogy ki férhet hozzá a felvételekhez és milyen feltételek mellett. A hozzáféréseket szigorúan korlátoznunk és minden hozzáférést dokumentálnunk kell. Szabályozniuk szükséges például a jogosultsági szinteket, a hozzáférés naplózását, a felvételek exportálását, a törlési folyamatokat, valamint az incidenskezelési eljárásokat.

A megőrzési idő meghatározása szintén kulcsfontosságú kérdés. A felvételek nem őrizhetők meg korlátlan ideig, és előre meg kell határoznunk a szükséges és arányos megőrzési időt.

A gyakorlatban a kamerarendszereket és beléptető rendszereket sok esetben külső szolgáltatók telepítik vagy üzemeltetik. Ez adatfeldolgozói jogviszonyt hozhat létre, amely a GDPR 28. cikkének megfelelő írásbeli szerződéses szabályozást igényel részünkről. További problémákat vethet fel a felhőalapú adattárolás alkalmazása, mely esetben is tisztában kell lennünk az adattárolás helyével, harmadik országba történő adattovábbítással, a szolgáltató biztonsági intézkedéseivel, valamint az incidenskezelési folyamatokkal.

Amennyiben oktatási intézményként nem rendelkezünk megfelelő informatikai és adatvédelmi erőforrásokkal ezen rendszerek teljes körű ellenőrzésére, az további kockázatokat eredményezhet.

## 5. Következtetések

Az elektronikus megfigyelőrendszerek alkalmazása a köznevelési intézményekben összetett jogi, technológiai és etikai kérdéseket vet fel. A gyermekek fokozott adatvédelmi védelme miatt az oktatási intézményeknek különösen körültekintően kell eljárniuk a megfigyelési rendszerek tervezése és működtetése során.

A technológiai fejlődés önmagában nem elegendő az adatkezelések jogszerűségének igazolására, a szükségesség és arányosság követelményének vizsgálata minden adatkezelés esetében elengedhetetlen. Különösen igaz ez a biometrikus rendszerekre és a mesterséges intelligencián alapuló megfigyelési technológiákra.

A jogszerű működéshez az intézményeknek megfelelő adatvédelmi dokumentációval, adatbiztonsági intézkedésekkel, adatkezelési tájékoztatókkal, hatásvizsgálatokkal és szerződéses garanciákkal kell rendelkezniük. Emellett fontos szerepe van az adatvédelmi tisztviselőknek és az intézményi tudatosság növelésének is.

A jövő egyik legfontosabb kihívása annak biztosítása lesz, hogy az oktatási intézmények a biztonsági és szervezeti célok megvalósítása mellett képesek legyenek megőrizni az érintettek alapvető jogait, különösen a gyermekek, a szülők és a munkavállalók (köznevelési foglalkoztatottak és egyéb munkavállalók) magánszférájának védelmét.

DOI: 10.31915/NWS.2026.20



# “EGY JELSZÓ MIND FÖLÖTT”, AVAGY KI VAGYOK ÉN? VAN-E, VÉD-E IGAZÁN A NEM JELSZAVAS AUTENTIKÁCIÓ?

# ONE PASSWORD TO RULE THEM ALL: OR, WHO AM I?

# DOES PASSWORDLESS AUTHENTICATION REALLY EXIST, AND DOES IT TRULY PROTECT?

*Pfeiffer Szilárd*
mailbox@pfeifferszilard.hu

*Balsai Péter*
balsai.peter@mithrandir.hu

**Absztrakt**

A vezető techcégek kommunikációjában ma megszokott kijelentés, hogy a jelszó elavult módszer. Egyetértenek abban, hogy birtokalapú és/vagy biometrikus megoldásokkal kell alkalmazni. A „jelszómentesség” sokszor csak illúzió, a termékeladást segítő marketing-üzenet. Felvetjük, hogy a jelszavak teljes mellőzése hamis dogma, hiszen a biokulcsok és hardvertokenek gyakran megjegyezhető kódokra (PIN, recovery code) támaszkodnak. A CIA-háromszög, a Parkeri hatszög és a Kerckhoffs-elvek használatával mutatjuk be a részletek ismertetése nélkül, hogy a biometria és a birtokalapú megoldások számos biztonsági elvet sértenek/sérthetnek. Azt is bemutatjuk, hogy a jelszó nyújtotta entrópia modern aszimmetrikus protokollokkal (pl.: OPAQUE) ötvözve magasabb szintű kriptográfiai biztonságot nyújtanak az alapvető elvek betartása mellett. Érintjük a szabad, nyílt forráskódú szoftverek szerepét a technológiai szuverenitás és az auditálhatóság megőrzésében.

A valódi kérdés tehát nem az, hogy alkalmazzuk-e a jelszavakat, hanem az, hogy mennyire közvetlen módon alkalmazzuk őket. Kitérünk arra a tapasztalatunkra, hogy az azonosítás és jogosultságvizsgálat a való világból kerül át a digitális térbe, és egyre inkább a digitális térbeli azonosítás határozza meg kilétünket. Így a nemzetállamok jogi és magánszemély-azonosítási monopóliuma a nagy techcégek és partnereik kezébe csúszik át.

**Kulcsszavak**: autentikáció, jelszó, jelszómentesség, biometria, technológiai szuverenitás, opaque protokoll, digitális identitás, CIA-háromszög, Parkeri hatszög, Kerckhoffs-elv

## Abstract

In the messaging of leading tech companies, it is now a commonplace claim that the password is an obsolete method. They agree that it should be used with possession-based and/or biometric solutions. „Passwordlessness" is often only an illusion – a marketing message that helps sell products. We argue that abandoning passwords entirely is a false dogma, since biometric keys and hardware tokens frequently rely on memorable codes (PIN, recovery code). Using the CIA triad, the Parkerian hexad, and Kerckhoffs's principles, we show – without detailing the specifics – that biometric and possession-based solutions violate, or may violate, a number of security principles. We also present that the entropy provided by a password, combined with modern asymmetric protocols (e.g., OPAQUE), delivers a higher level of cryptographic security while respecting the fundamental principles. We touch on the role of free and open-source software in preserving technological sovereignty and auditability.

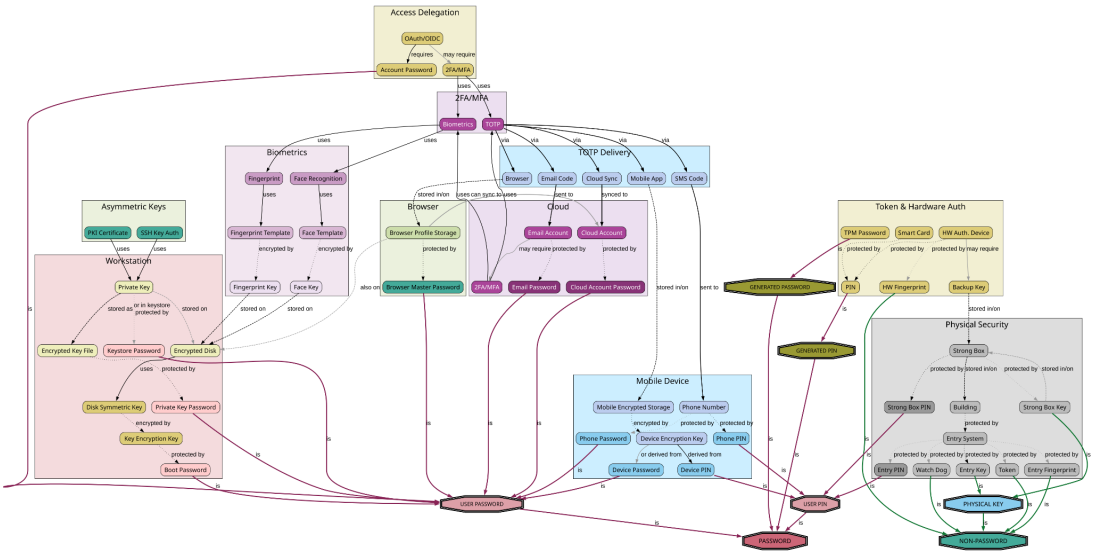
The real question, therefore, is not whether we use passwords, but how directly we use them.

We also reflect on our experience that identification and authorization are moving from the physical world into the digital space, and that digital-space identification increasingly determines who we are. As a result, the nation-states' monopoly on legal and individual identification is slipping into the hands of the large tech companies and their partners.

**Keywords**: authentication, password, passwordless, biometrics, technological sovereignty, opaque protocol, digital identity, CIA triad, Parkerian hexad, Kerckhoffs's principle

## 1. Bevezetés

Az 1. ábra megmutatja, hogy milyen gyakran jutunk – egy vagy több lépésben – jelszavakhoz a különböző azonosítási technológiák során.




*1. ábra: Autentikációs eljárások visszavezethetősége jelszavakra*

### 1.1. A jelszó széleskörű használata

A birtokalapú biztonsági tokenek és eszközök gyakran PIN-kóddal (azaz jelszóval) védettek. A mobilos biometria hasznos, ám nem ritkán csak kényelmi funkció. Meghibásodás esetén viszont  a PIN-kód vagy a valamilyen minta (bit-sor) lesz a tartalékmechanizmus. Egy központi azonosítószolgáltató (identity provider) megoldás is legtöbbször jelszóra, illetve valamilyen visszaállító eljárásra támaszkodik. Léteznek ugyan jelszó nélkül használható, biometriával védett hardveres eszközök (pl.: YubiKey Bio [1]), ám kompromittálódás vagy eszközvesztés esetén a visszaállító kódok (recovery keys) tárolása újfent jelszavakhoz vezet (pl.: széf kód).

A jelszómentes hitelesítés (passwordless authentication) terjedésével ráadásul új kritikus biztonsági kérdés merül fel: a külső hardver, szoftver, internetkapcsolat (azonosítószolgáltató) és/vagy dedikált online szolgáltatók iránti fundamentális igény. Ezeket az infrastrukturális elemeket gyakran az Európai Unió és annak joghatóságán kívül fejlesztik (pl.: Microsoft), gyártják (pl.: RSA Security), illetve üzemeltetik (pl.: Amazon, Google). Ez az ellátási lánc a kritikus infrastruktúráinkat (pl.: közigazgatási rendszerek) olyan adatvédelmi és szuverenitási problémákkal terheli meg, amelyek a jelszó (tudás) esetében nem lennének jelen.

## 2. A személyazonosítás nagyvonalú története

Az azonosítás nem új keletű, csak az informatika túl fiatal.

**Kőkorban és az őskorban** az arc- és hangfelismerés, a ruházat, az ékszerek, a csoporthoz tartozást jelző testfestés és a (heg)tetoválások segítették a felismerést.

**Ókor és középkor** ehhez hozzátette a gyűrűket, címereket, okleveleket [2], pecséteket és egyházi nyilvántartásokat, az előkelőknek a festményeket.

**Újkorban** kötelező lett az egyházi/állami anyakönyv (hazánkban: 1895) és elterjedtek a hivatalos azonosító okmányok, amik tartalmaztak alapvető adatokat (név, születési idő, "apja neve").

**A XIX.–XX.** században már a fénykép (igazolványkép), ujjlenyomat, vércsoport, de akár még a koponyaalak is szolgált azonosítási célokra. **A közelmúltban** kamera, íriszszkenner, vénaszkenner, ujjlenyomat-olvasó, hanganalizátor, DNS-profilozó stb. is használatosak, melyek közös jellemzője, hogy a fizikai valóságot valamilyen digitális kivonattá kódolva és egyúttal veszteségesen tömörítve, tárolják és vizsgálják.

Fentiek mellett mindig volt jelentőségük a hangoknak, jelszavaknak és jelmondatoknak is, melyek a "valakiságot", illetve egy adott csoporthoz való tartozást igazolták.

## 3. Az azonosítás mint állami monopólium megszűnésének jelei

Az Apple – és más cégek – a DUNS-számot (Data Universal Numbering System, amit egy vállalkozás [3] kezel ) használják [4, 5, 6], ami nagy, többségében amerikai cégek által elfogadott módja a cégek azonosításának. Számukra ez a szám digitális tanúsítványként igazolja a vállalkozás jogi létezését[7]. A nemzetállamok most is adószámot, cégjegyzékszámot stb. használnak. A Google a diákok jogviszonyának igazolására SheerID-t kér [8, 9], amit szintén egy vállalkozás [10] kezel. Mind a DUNS, mind a SheerID csak információforrásként használják az állami adatokat, saját döntéseik alapján. Ez a trend számunkra arra utal, hogy az azonosítás monopóliuma csúszik ki az állami szervezetek kezéből. Az állami monopólium eltűnésének további lendületet ad az eIDAS-2 [11] mobilalapú digitális tárcája, melynek folyományaként az európai és más (nem az USA vagy Kína) államok mind a hardver, mind a szoftver feletti uralmat elvesztik. Az első lépcsőben [12] használt chipkártyás azonosítás, még sok állam, vagy szövetség számára nyújtott valamekkora önállóságot, tekintve, hogy a szükséges biztonsági chipet európai gyártók is képesek gyártani. Ráadásul a nyomtatott, különböző biztonsági elemekkel ellátott azonosítókártyák, a külső eszköz (hardver, szoftver) nélküli ellenőrzésre is lehetőséget adnak. A tisztán elektronikus "mobiltárca" mindkét lehetőséget megszünteti, hiszen nincs valódi európai gyártó, alkalmazásbolt, és ha megsérült, vagy lemerült az eszköz, pl. telefon akkor azonosításra nem használható.

## 4. Jelszavak

Míg a hagyományos meghatározások a jelszóra elsősorban az emberek közötti használatot (azonosításra) írják le, míg a szerzők a modern számítógépes világban betöltött szerepet hangsúlyozva az alábbi definíciókat adják.

> *Egy bitsorozat a számítógép számára, melyet az ember – a könnyebb megjegyezhetőség érdekében – nyomtatható karakterekkel és az informatikai rendszerek gyengeségei miatt az angol ábécé betűivel, illetve egyéb nyomtatható karakterekkel – például írásjelekkel – ír be.* – Balsai Péter

> *A jelszó nem más, mint egy – kizárólag az emberi megjegyezhetőség érdekében –, szövegesen ábrázolt véletlen érték, amely az autentikációs algoritmusok számára a szükséges entrópiát szolgáltatja.* – Pfeiffer Szilárd

### 4.1. A jelszó mint entrópiaforrás

Mivel a számítógépek csak álvéletleneket generálnak, a hitelesítés valódi értéke a jelszó magas bemeneti entrópiájában, statisztikai egyediségében rejlik. A jelszó olyan „biológiai entrópia", amit a felhasználó visz be a rendszerbe egy feltörhetetlen kriptográfiai kulcs származtatásához.

Amikor jelszómentes hitelesítésre váltunk, elveszítjük az afeletti kontrollt, melynek szerepe kritikus. A jelszó valódi értéke ugyanis a statisztikai egyediségében rejlik, nem csupán a bitek mennyiségében[1]. A felhasználók többsége prediktálható mintázatokat követ, ami drasztikusan csökkenti a tényleges entrópiát. Ezért fontos említeni (terjedelmi korlátok miatt nem kifejtve) a tipikus jelszóhibákat[13]: évszámok, családtagok, háziállatok nevei, gyakori szavak, kényszeralgoritmusok stb. A jó jelszó jellemzői: egyedi (nem hasonlít másokéra), semleges (a személyhez könnyen nem köthető), több szó / mondat, aminek nem kell komplexnek (pl.: speciális karakter), csak hosszúnak lennie, megjegyezhető. Egy komplett iparági marketing magyarázza mindenkinek, hogy a megfelelő jelszavakat nem lehet megjegyezni, ám ez így ebben a formában nem igaz.

1 A bitmélység (bit depth) a jelszó elméleti keresési terének mérete: egy n hosszú, k elemű ábécéből álló jelszó bitmélysége log2 (kn) bit. Ez az elméleti felső korlát, amely csak akkor valósul meg, ha a jelszó valóban egyenletes eloszlásból származik.

### 4.2. Lehet jó jelszót használni, megjegyezni

Pszichológia- és a memóriakutatások azt mondják [14], hogy

- A rövid távú memóriában teljesen véletlenszerű karakterekből átlagosan 7 ± 2 elemet lehet megjegyezni[14], és ennek a hosszú távú memóriába kerüléséhez már sok ismétlés kell.[2]
- Gyakori (a személy által használt) szavakat mondattá összefűzve, a rövid távú memória „tömbösítve" szavankénti egységként kezeli azokat, és a hosszú távú memóriába kerüléshez pár olvasás kell [15]. Napi legalább egyszeri használat mellett néhány százalék a felejtés esélye.

Vélelmezzük, hogy a hiedelemmel ellentétben **hosszú és jó jelszavak is lehetnek megjegyezhetők az ember számára.** Egy véletlenszerű, személyes kontextushoz nem kötődő jelmondat a gép számára szükséges magas entrópiát biztosítja. Bár lenne megoldás (OPAQUE [16] – lásd később), ami az offline szótáralapú támadást a gyakorlatban kiküszöbölné, de a kiinduló jelszó minősége még így is döntő jelentőségű. Fenntartással kell kezelni az informatikai marketinget, a társtudományokban van/lehet megoldás, ám **a cégek nem tanítani, hanem eladni akarnak.**

## 5. Biztonsági elveknek való megfelelés

A CIA triad, Parkeri hexád [17], Kerckhoffs-elv [18], Zero Trust direktíva [19] nem újdonságok, sok éve léteznek. Figyelembevételük, betartásuk – a szerzők folyamatban lévő kutatásának [20] tapasztalatai alapján – nem igazán jellemző. Nem tapasztaljuk, hogy rendszerek vizsgálatára együtt használják ezeket. Ezért gondolatébresztőként megosztjuk a következő táblázatainkat.

### 5.1. A CIA-triád és Parkeri hexád

| | **Jelszó** | **Birtoklás** | **Biometria** | **OPAQUE** |
|---|---|---|---|---|
| **Bizalmasság** | sérti | *sértheti* | sérti | **nem** |
| Integritás | nem | sérti | sérti | **nem** |
| Rendelkezésre állás | *sértheti* | sérti | sérti | *sértheti* |
| **Hitelesség** | sértheti | sértheti | nem | nem |
| **Birtoklás és kontroll** | sérti | sérti | sérti | **nem** |
| Hasznosság | nem | nem | nem | nem |

2 Egy ma egyre kívánatosabb 20 karakteres jelszó pl f5!Tz9#kL2qP0rWv&n1X) ~85–95% eséllyel megtanulhatatlan, és a megtanulni képes 5–15 % is gyorsan és szinte biztosan gyorsan elfelejti.

**5.2. Kerckhoffs-elv / módszerek**

| | **Jelszó** | **Birtoklás** | **Biometria** | **OPAQUE** |
|---|---|---|---|---|
| **Feltörhetetlenség** | *sértheti* | *sértheti* | *sértheti* | **nem** |
| **Nyilvános specifikáció** | **nem** | sérti | sérti | **nem** |
| **Megjegyezhetőség / csere** | **nem** | sérti | sérti | **nem** |
| **Távíró (hálózat)** | **nem** | **nem** | **nem** | **nem** |
| **Hordozhatóság** | **nem** | *sértheti* | sérti | **nem** |
| **Egyszerűség** | sérti | sérti | sérti | **nem** |

**5.3. Zero trust direktíva**

| | **Jelszó** | **Birtoklás** | **Biometria** | **OPAQUE** |
|---|---|---|---|---|
| **Adatátvitel közben** | *sértheti* | *sértheti* | *sértheti* | **nem** |
| **Feldolgozás közben** | **nem** | sérti | sérti | **nem** |
| **Tárolás közben** | **nem** | sérti | sérti | *sértheti* |

## 6. Az OPAQUE protokoll: Egy jelszó, ezer egyedi kulcs

Részleges választ és javuló megfelelést adna a biztonsági alapelvek előbb mutatott megsértésére az aPAKE (asymmetric Password-Authenticated Key Exchange) protokollok alkalmazása. Ilyen protokoll például az OPAQUE (VOPRF[21] implementációval), ami az alábbi előnyeire való tekintettel sem terjedt el még a gyakorlatban.

1. Zero-Knowledge Proof: A jelszó soha nem utazik a hálózaton. A klienstől a szerverig csak egy vakított (blinded) bizonyíték utazik, a szerver egyedül a hozzáférés érvényességét hagyja jóvá.
2. Cross-site tracking prevenció: A kulcsszármaztatási folyamatba (Key Derivation Function – KDF) az OPAQUE képes bevonni a szerver azonosítóját (Context) is. Ennek eredményeképp ugyanabból a jelszóból a KDF funkciószerverenként eltérő aszimmetrikus kulcspárt fog generálni.
3. Rugalmas paraméterek: a biztonsági szint skálázható, a kulcspár generálásához szükséges számítási-, illetve memóriakapacitás-igény növelésével, miközben a jelszó változatlan.
4. A hardveres tokenekkel ellentétben nem kell új eszközt vásárolni a paraméterek frissítéséhez.

## 7. Technológiai szuverenitás és a nyílt vagy saját forráskód együtt jár

A Cloud-szinkronizált Passkeyek használata — a fentiekkel szemben — azzal jár, hogy az óriáscégek digitális gyámjaink lesznek. A felhőbe például a következő kritikus adatok kerülhetnek: privát kulcs (end-to-end titkosítva, a mesterjelszó védi), metaadatok (melyik szolgáltatáshoz tartozik, létrehozás időpontja), valamint a credential azonosító. A jelszó-visszaállítási mechanizmusok (pl.: elfelejtett Apple ID jelszó esetén) a „Zero Knowledge" szempontjából vetnek fel kérdéseket. A zárt forráskód, illetve az ellenőrizhető audit hiánya miatt a ténylegesen használt titkosítási, hash és egyéb mechanizmusok nem ellenőrizhetők. A kulcsok fizikai helye pedig joghatósági kérdéseket vet fel: az államok kikényszeríthetik az adatok átadását, lehetnek politikai okok miatt beépített „kém megoldások" is. Az akuttá váló – nem javított – sérülékenységek is a gyártóval szembeni súlyos függőséget hoznak létre

Ezek alapján belátható, hogy a szabad és saját fejlesztésű szoftvereknek kiemelkedő jelentősége lehetne.

## 8. Konklúzió

A digitális és a valós tér egybeolvadásával a szuverén, megbízható állami azonosítás ma már kritikus, az alapvető infrastruktúra-szolgáltatás része kell(ene) legyen. Az azonosítás a fizikai és digitális tér építőeleme, ennek megfelelő szemléletet és erőforrásokat igényel problémáinak kezelése.

Nem az a kérdés, hogy alkalmazzuk-e a jelszavakat, hanem az, hogy mennyire közvetlen módon történik az alkalmazásuk — és kitől függünk mindeközben. Valóban jobbak-e, jelszómentesek-e a „jelszómentes" megoldások, vagy csupán más, nehezebben kontrollálható kockázatokat vállalunk be egy illúzióért? **A biometria és a hardver tokenek sem mentesek a biztonsági elvekkel kapcsolatos kompromisszumoktól és geopolitikai függőséget is teremthetnek.**

A jelszó legfőbb hátránya a feloldható a modern azonosítási keretrendszerek (OPAQUE) alkalmazásával. Az egyik legbiztonságosabb kulcsgeneráló erő továbbra is a saját emberi tudatunkból fakadó entrópia marad — és ez az egyetlen autentikációs faktor, amelyhez sem külföldi hardver, sem szoftver, sem szolgáltató nem szükséges, **a jelszó a tudatunkból (**most még**) nem olvasható ki.**

[1] "YubiKey Bio Series," Yubico. Accessed: Jun. 28, 2026. [Online]. Available: https://www.yubico.com/products/yubikey-bio-series/

[2] L. Benedictus, "A brief history of the passport," *The Guardian*, Nov. 17, 2006. Accessed: Jun. 28, 2026. [Online]. Available: https://www.theguardian.com/travel/2006/nov/17/travelnews

[3] "A D-U-N-S számról." Accessed: Jun. 28, 2026. [Online]. Available: https://www.dnb.com/hu-hu/smb/duns.html

[4] "DUNS-szám," Google Ads Súgó. Accessed: Jun. 29, 2026. [Online]. Available: https://support.google.com/google-ads/answer/85961?hl=hu

[5] "KB0756540 - Why is a DUNS number required for me to register?," SAP Ariba Help Center. Accessed: Jun. 29, 2026. [Online]. Available: https://hc.ariba.com/index.html?sap-ui-language=en#/item/KB0756540

[6] "Company account verification requirements – Windows apps," Microsoft Learn. Accessed: Jun. 29, 2026. [Online]. Available: https://learn.microsoft.com/en-us/windows/apps/publish/store-business-verification-reqs

[7] G. Teubner, "Global Bukowina: Legal pluralism in the world society," in *Critical theory and legal autopoiesis*, D. Göbel, Ed., Manchester University Press, 2019. doi: 10.7765/9781526139948.00017.

[8] "Google AI Pro student trial," Google One Help. Accessed: Jun. 29, 2026. [Online]. Available: https://support.google.com/googleone/community-guide/379602230/google-ai-pro-student-trial?hl=en

[9] "Amazon Music FAQ." Accessed: Jun. 29, 2026. [Online]. Available: https://verify.sheerid.com/amazon-music-faq/

[10] "Identity Marketing Platform," SheerID. Accessed: Jun. 28, 2026. [Online]. Available: https://www.sheerid.com/

[11] "AZ EURÓPAI PARLAMENT ÉS A TANÁCS 2024. április 11-i (EU) 2024/1183 RENDELETE a 910/2014/EU rendeletnek az európai digitális személyazonossági keret létrehozása tekintetében történő módosításáról - Hatályos Jogszabályok Gyűjteménye," Hatályos Jogszabályok Gyűjteménye. Accessed: Jun. 29, 2026. [Online]. Available: https://net.jogtar.hu/jogszabaly?docid=a2401183.eup

[12] "AZ EURÓPAI PARLAMENT ÉS A TANÁCS 2014. július 23-i 910/2014/EU RENDELETE a belső piacon történő elektronikus tranzakciókhoz kapcsolódó elektronikus azonosításról és bizalmi szolgáltatásokról, valamint az 1999/93/EK irányelv hatályon kívül helyezéséről - Hatályos Jogszabályok Gyűjteménye," Hatályos Jogszabályok Gyűjteménye. Accessed: Jun. 28, 2026. [Online]. Available: https://net.jogtar.hu/jogszabaly?docid=a1400910.eup

[13] "Miért ne becsüljük le a kisbetűs jelszavakat? 3. rész," Bitport. Accessed: Jun. 28, 2026. [Online]. Available: https://bitport.hu/miert-ne-becsuljuk-le-a-kisbetus-jelszavakat-password-security-3-resz/

[14] G. A. Miller, “Classics in the History of Psychology An internet resource developed by Christopher D. Green York University, Toronto, Ontario (Return to Classics index).” Accessed: Jun. 29, 2026. [Online]. Available: https://psychclassics.yorku.ca/Miller/

[15] C. E. Bonhage, L. Meyer, T. Gruber, A. D. Friederici, and J. L. Mueller, “Oscillatory EEG dynamics underlying automatic chunking during sentence processing,” *NeuroImage*, vol. 152, pp. 647–657, 2017, doi: https://doi.org/10.1016/j.neuroimage.2017.03.018.

[16] D. Bourdrez, H. Krawczyk, K. Lewi, and C. A. Wood, “The OPAQUE Augmented Password-Authenticated Key Exchange (aPAKE) Protocol,” Internet Engineering Task Force, Request for Comments RFC 9807, 2025. doi: 10.17487/RFC9807.

[17] D. B. Parker, *Fighting computer crime: a new framework for protecting information*. New York: Wiley, 1998.

[18] A. Kerckhoffs, “La Cryptographie Militaire (Seconde partie)”.

[19] S. Rose, O. Borchert, S. Mitchell, and S. Connelly, “Zero Trust Architecture,” National Institute of Standards and Technology, Aug. 2020. doi: 10.6028/NIST.SP.800-207.

[20] S. Pfeiffer, “Assessing IT Solutions Against Security Principles,” Szilárd Pfeiffer. Accessed: Jun. 30, 2026. [Online]. Available: https://pfeifferszilard.hu/2026/06/30/assessing-it-solutions-against-security-principles.html

[21] A. Davidson, A. Faz-Hernandez, N. Sullivan, and C. A. Wood, “Oblivious Pseudorandom Functions (OPRFs) Using Prime-Order Groups,” Internet Engineering Task Force, Request for Comments RFC 9497, 2023. doi: 10.17487/RFC9497.

DOI: 10.31915/NWS.2026.21



# A MESTERSÉGES INTELLIGENCIA PEDAGÓGIAI KOMPETENCIÁI

# THE PEDAGOGICAL COMPETENCIES OF ARTIFICIAL INTELLIGENCE

*T. Nagy László*
t.nagy.laszlo@drhe.hu

**Absztrakt**

A mesterséges intelligencia (MI), különösen a nagy nyelvi modellek (Large Language Models, LLM-ek) fejlődése jelentős változásokat eredményez az oktatás és a tanulástámogatás területén. Az olyan rendszerek, mint a ChatGPT, a Gemini vagy a DeepSeek új lehetőségeket teremtenek a személyre szabott tanulás, az automatikus tartalomgenerálás, valamint az adaptív oktatástámogatás területén. Jelen tanulmány célja e három, széles körben hozzáférhető nagy nyelvi modell pedagógiai kompetenciáinak összehasonlító elemzése. A kutatás során a modellek teljesítményét formális és informális tanulási helyzetekben vizsgáltuk, különös tekintettel az életkori adaptivitásra, a kreativitásra, a didaktikai értékre és az oktatói támogatás lehetőségeire. Az eredmények rámutatnak arra, hogy a vizsgált LLM-ek már jelenlegi fejlettségi szintjükön is alkalmasak lehetnek pedagógiai segédeszközként való alkalmazásra, ugyanakkor működésük számos etikai, módszertani és nyelvi problémát is felvet.

**Kulcsszavak:** mesterséges intelligencia, pedagógiai kompetencia, nagy nyelvi modellek, tanulástámogatás, digitális oktatás

The development of artificial intelligence (AI), particularly large language models (LLMs), is bringing about significant changes in the fields of education and learning support. Systems such as ChatGPT, Gemini, and DeepSeek are creating new opportunities in the fields of personalized learning, automatic content generation, and adaptive learning support. The aim of this study is to conduct a comparative analysis of the pedagogical competencies of these three widely accessible large language models. In the research, we examined the models' performance in formal and informal learning contexts, with particular attention to age-appropriateness, creativity, didactic value, and opportunities for teacher support. The results indicate that the examined LLMs may already be capable of being used as pedagogical tools at their current level of development, yet their operation also raises numerous ethical, methodological, and linguistic issues.

**Keywords:** artificial intelligence, pedagogical competence, large language models, learning support, digital education

## 1. Bevezetés

A digitális technológia fejlődése az elmúlt évtizedekben alapvetően alakította át az oktatás szerkezetét, módszertanát és eszközrendszerét. Az oktatási innováció egyik legmeghatározóbb területévé napjainkra a mesterséges intelligencia alkalmazása vált, amely különösen a generatív MI-rendszerek megjelenésével került a pedagógiai kutatások középpontjába. (Holmes et al., 2019)

A nagy nyelvi modellek megjelenése új szintre emelte az ember-gép interakció lehetőségeit. Ezek a rendszerek nem csupán információ-visszakeresésre képesek, hanem komplex nyelvi műveletek végrehajtására, kreatív szövegalkotásra, problémamegoldásra és adaptív kommunikációra is. (McDonough, 2025) Az oktatás szempontjából ez különösen jelentős fejlemény, mivel a modellek potenciálisan képesek támogatni a differenciált oktatást, az egyéni tanulási utak kialakítását és a pedagógusok adminisztratív terheinek csökkentését. (Luckin, 2018)

A mesterséges intelligencia oktatási alkalmazása azonban nem tekinthető problémamentes folyamatnak. Számos kutatás rámutat arra, hogy az MI-rendszerek működése komoly pedagógiai, etikai és társadalmi kérdéseket vet fel. (Zawacki-Richter et al., 2019) Ide tartozik többek között a tanulói autonómia kérdése, az információk megbízhatósága, az algoritmikus torzítás, valamint az a probléma, hogy a tanulók kritikátlanul fogadhatják el a generált válaszokat. (Kasneci et al., 2023)

A generatív MI oktatásban való alkalmazásának egyik legfontosabb kérdése az, hogy ezek a rendszerek milyen mértékben képesek valódi pedagógiai kompetenciák megjelenítésére. A pedagógiai kompetencia fogalma alatt jelen esetben nem pusztán a tantárgyi tudás reprodukcióját értjük, hanem az életkori sajátosságok figyelembevételét, a motiváció fenntartását, a didaktikai érzékenységet, valamint a tanulási folyamat támogatásának képességét is.

A szakirodalomban egyre több kutatás foglalkozik az LLM-ek oktatási potenciáljával. Holmes és munkatársai szerint a mesterséges intelligencia hosszú távon képes lehet személyre szabott tanulási környezetek kialakítására, amelyek figyelembe veszik a tanulók előzetes tudását, haladási ütemét és egyéni igényeit. (Holmes et al., 2019) Luckin hangsúlyozza, hogy az MI nem a pedagógus helyettesítésére, hanem az emberi tanítás támogatására alkalmas eszköz lehet. (Luckin, 2018) Ezzel szemben Zawacki-Richter és munkatársai arra figyelmeztetnek, hogy az oktatási MI-rendszerek kutatása jelenleg erősen technológiaorientált, miközben kevés figyelem irányul a pedagógiai és társadalmi következményekre. (Zawacki-Richter et al., 2019)

Kasneci és szerzőtársai különösen a ChatGPT oktatási alkalmazhatóságát vizsgálták, és arra a következtetésre jutottak, hogy a generatív nyelvi modellek jelentős támogatást nyújthatnak a tanulási folyamatban, ugyanakkor fokozott pedagógiai kontrollt és kritikai médiaműveltséget igényelnek. (Kasneci et al., 2023)

A hazai szakirodalomban Szabóné Balogh Ágota hangsúlyozza, hogy a mesterséges intelligencia az oktatásban elsősorban támogató technológiaként értelmezhető. Véleménye szerint az MI-rendszerek sikeres integrációjához nem csupán technológiai infrastruktúrára, hanem új pedagógiai szemléletre és digitális kompetenciákra is szükség van. (Szabóné 2023)

## 2. Célkitűzések és a kutatás kérdései

Jelen kutatás elsődleges célja annak feltárása volt, hogy a három, széles körben hozzáférhető nagy nyelvi modell, a ChatGPT, a Gemini és a DeepSeek milyen mértékben képes pedagógiai kompetenciák megjelenítésére az oktatás és a tanulástámogatás különböző szintjein. A vizsgálat különös hangsúlyt fektetett az alsó tagozatos korosztályra, mivel ebben az életkori szakaszban kiemelt jelentőséggel bír a magyarázatok érthetősége, az élményszerű tanulás támogatása, valamint az életkori sajátosságokhoz igazított kommunikáció.

A kutatás abból az alapfeltevésből indult ki, hogy az LLM-ek pedagógiai értéke nem kizárólag az információszolgáltatás pontosságában mérhető, hanem abban is, hogy mennyire képesek didaktikailag releváns, motiváló és adaptív válaszok létrehozására. Ennek megfelelően a vizsgálat nem csupán lexikális tudást vagy tárgyi helyességet elemzett, hanem a modellek kommunikációs és pedagógiai működését is.

A kutatás során az alábbi főbb kompetenciaterületek vizsgálatára került sor:

- **Szövegértés és tudásreprezentáció:** A modellek képessége a bemeneti információk megértésére és releváns tudásanyag generálására.
- **Információszolgáltatás és magyarázat:** A kérdésmegválaszolási képesség és a komplex fogalmakat értelmező magyarázatok előállításának a minősége.
- **Kreativitás és támogatás:** A modellek teljesítménye kreatív feladatok (pl. történetírás, játéktervezés) támogatásában.
- **Adaptivitás és személyre szabás:** A figyelem itt arra irányult, hogy egyes modellek miként képesek alkalmazkodni a különböző tanulási szintekhez és egyéni igényekhez.
- **Oktatói támogatás:** Annak felmérése, hogy az LLM-ek hozzájárulhatnak-e az oktatók munkájának segítéséhez és/vagy hatékonyabbá tételéhez.

## 3. Alkalmazott módszertan

A kutatás célkitűzéseinek megvalósítása érdekében olyan komplex módszertani keretrendszert alkalmaztam, amely egyaránt ötvözte a kvalitatív és kvantitatív elemzés eszközeit. A vizsgálat nem csupán a modellek által generált válaszok pontosságát elemezte, hanem azok pedagógiai relevanciáját, didaktikai értékét, kreativitását és életkori adaptivi-

tását is. A módszertan kialakításakor alapvető szempont volt, hogy a kutatás minél inkább tükrözze a nagy nyelvi modellek valós oktatási környezetben történő használhatóságát.

### 3.1. Vizsgálati környezet és kutatási háttér

A kutatás három széles körben elérhető nagy nyelvi modell vizsgálatára épült: ChatGPT v5, Gemini v2.5 és DeepSeek v3.2. Az elemzés során kizárólag az egyes rendszerek ingyenesen hozzáférhető, nyilvános szöveges felületeit használtuk, mivel a cél az volt, hogy a vizsgálat eredményei a felhasználók többsége számára elérhető szolgáltatásokra vonatkozzanak.

A tesztelési folyamat 2025 szeptemberében és októberében zajlott. Eredetileg saját tesztelésben gondolkodtam, azonban rövid időn belül nyilvánvalóvá vált, hogy a modellek válaszait jelentős mértékben befolyásolhatják a korábbi interakciók, a felhasználói előzmények és a kontextuális memóriaelemek. Ennek csökkentése érdekében a vizsgálatba több mint húsz különböző, regisztrált felhasználó (tesztelő) lett bevonva.

A diverz felhasználói kör bevonása lehetővé tette a válaszok konzisztenciájának és stabilitásának vizsgálatát, valamint csökkentette annak kockázatát, hogy az eredményeket egyedi felhasználói profilok vagy korábbi beszélgetések torzítsák. A vizsgálat során törekedtem arra is, hogy ugyanazon kérdések többszöri ismétlésével elemezzem a modellek válaszainak stabilitását. Ez különösen fontos szempontnak bizonyult, mivel a generatív MI-rendszerek működésének egyik alapvető sajátossága, hogy azonos promptok is eredményezhetnek eltérő válaszokat. A kutatás ezért nem egyszeri válaszokat, hanem ún. válaszmintázatokat elemzett.

Az adatgyűjtés volumene jelentősnek tekinthető. Egyetlen tesztelő esetében a három modell összes válasza hozzávetőlegesen 70–100 oldalnyi szöveganyagot eredményezett. Ez összességében nagy mennyiségű adatot biztosított a részletes összehasonlító elemzéshez.

### 3.2. A kérdéssor felépítése és az adatgyűjtés folyamata

A kutatás empirikus alapját egy tizennyolc kérdésből álló pedagógiai kérdéssor képezte. A kérdések összeállításánál elsődleges szempont volt, hogy azok ne kizárólag lexikális tudást vagy egyszerű információ-visszaadást vizsgáljanak, hanem komplex pedagógiai helyzeteket modellezzenek. A kérdéssor több különböző kompetenciaterület mérésére szolgált, és négy nagyobb kategóriába sorolható.

#### 3.2.1. Életkor-specifikus magyarázatok

Ebben a kategóriában a kérdések azt vizsgálták, hogy a modellek mennyire képesek életkori sajátosságokhoz igazított kommunikációra. A feladatok során a modelleknek különböző

életkorú gyermekek számára kellett összetett fogalmakat egyszerűsíteniük és pedagógiailag megfelelő módon értelmezniük

### 3.2.2. Kreatív és élményszerű tartalomgenerálás

A második kategória a modellek kreatív képességeit vizsgálta. Az oktatásban egyre nagyobb szerepet kapnak az élményszerű és motivációközpontú tanulási formák, ezért fontos kérdésnek tekinthetjük, hogy az LLM-ek milyen mértékben képesek ilyen jellegű tartalmak előállítására.

### 3.2.3. Pedagógiai stratégia és módszertani gondolkodás

A harmadik kategória célja annak feltárása volt, hogy a modellek válaszaiban megjelenik-e tudatos pedagógiai szervezőelv vagy módszertani logika. Ez a dimenzió különösen fontosnak bizonyult, mivel a generatív modellek egyik leggyakoribb oktatási felhasználási módja a pedagógiai ötletgenerálás.

### 3.2.4. Reflektív és etikai kérdések

A negyedik kategória az etikai érzékenység és reflektív gondolkodás vizsgálatára irányult. A mesterséges intelligencia oktatási alkalmazása számos társadalmi és pedagógiai kérdést vet fel, ezért fontosnak tartottam annak elemzését, hogy a modellek miként reagálnak érzékeny vagy vitatott témákra.

## 3.3. Az értékelési folyamat és elemzési szempontok

A generált válaszok értékelését két független értékelő végezte. Az elemzés során az értékelési szempontokat előzetesen egyeztettük annak érdekében, hogy minimalizálható legyen a szubjektív értelmezésekből fakadó eltérés.

Az értékelési rendszer kvalitatív és kvantitatív elemekből épült fel.

### 3.3.1. Kvalitatív elemzés

A kvalitatív vizsgálat során a válaszokat az alábbi főbb dimenziók mentén elemeztük:

- érthetőség;
- kreativitás;
- életkori megfelelés;
- pedagógiai és didaktikai érték;

Külön figyelmet fordítottunk arra, hogy a válaszok mennyire támogatják a tanulási folyamatot, illetve mennyiben tekinthetők valóban pedagógiai jellegű kommunikációnak.

### 3.3.2. Kvantitatív értékelés

A kvalitatív elemzést numerikus értékeléssel egészítettük ki. Az egyes válaszokat tízfokozatú skálán pontoztuk. A hagyományos ötfokozatú értékelési rendszer helyett azért alkalmaztunk tízpontos skálát, mert ez lehetővé tette az árnyaltabb különbségek pontosabb megjelenítését.

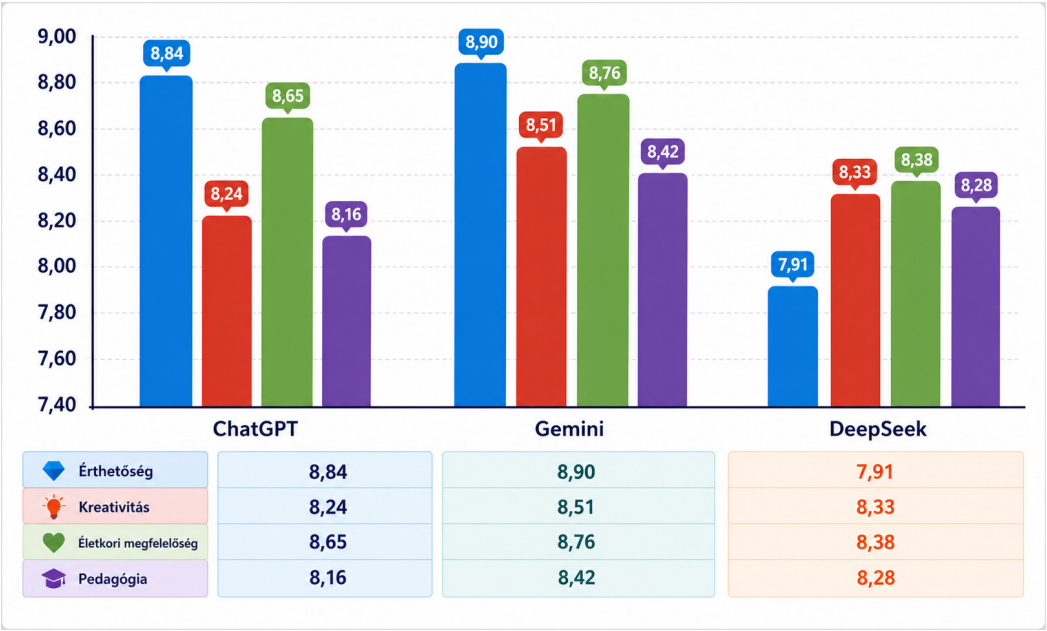
A kettős értékelési rendszer lehetővé tette nem csupán az egyes modellek összehasonlítását, hanem azok specifikus erősségeinek és gyengeségeinek részletesebb feltárását is.

## 4. A vizsgálat eredményei

A kutatás során gyűjtött kvalitatív és kvantitatív adatok elemzése részletes képet adott a vizsgált nagy nyelvi modellek pedagógiai kompetenciáiról, valamint azok eltérő működési sajátosságairól. Az eredmények alapján megállapítható, hogy bár mindhárom modell képes pedagógiailag releváns tartalmak generálására, teljesítményük jelentős különbségeket mutatott az egyes vizsgált dimenziókban.

### 4.1. Kvantitatív összehasonlító eredmények

A 10-es skálán végzett összesített pontozás eredményeit az 1. ábra mutatja be.




*1. ábra. ChatGPT, Gemini, DeepSeek pontszámai*

### 4.1.1 Érthetőség

Az érthetőség vizsgálata során a Gemini és a ChatGPT teljesítménye közel azonosnak bizonyult. A válaszok többsége logikusan strukturált, jól követhető és nyelvileg koherens volt. A modellek jellemzően képesek voltak a komplex fogalmak fokozatos felépítésére és egyszerűsített magyarázatára.

A DeepSeek esetében az érthetőséget több nyelvi probléma gyengítette. Megjelentek magyartalan mondatszerkezetek, pontatlan szóhasználatok, neologizmusok és helyesírási hibák. Ezek a problémák különösen érzékenyen érintik az oktatási felhasználást, mivel a nyelvi pontosság a pedagógiai kommunikáció egyik alapvető követelménye.

### 4.1.2. Kreativitás és pedagógiai tartalomgenerálás

A kreatív feladatok elemzése során mindhárom modell viszonylag magas szintű teljesítményt mutatott, ugyanakkor markáns különbségek váltak láthatóvá a válaszok stílusában és pedagógiai megközelítésében.

A Gemini különösen erősnek bizonyult a játékos és élményszerű tanulási helyzetek kialakításában. Több esetben komplex pedagógiai szituációkat épített fel, amelyek egyszerre tartalmaztak: motivációs elemeket, differenciált feladatokat, interaktív megoldásokat és tanulói aktivitást ösztönző komponenseket. A modell válaszai gyakran közelítettek egy valódi tanórai forgatókönyv logikájához.

A ChatGPT kreatív válaszai strukturáltabbak és visszafogottabbak voltak. A rendszer jellemzően biztonságosabb, kevésbé kockázatos pedagógiai megoldásokat választott. A DeepSeek a kreativitás területén több alkalommal meglepően eredeti válaszokat adott. A modell különösen történetalkotásban és játékos szituációk generálásában mutatott erősségeket, válaszai időnként azonban túlzottan bőbeszédűvé vagy inkoherenssé váltak.

A kreativitás vizsgálata során megfigyelhető volt, hogy az LLM-ek egyik legnagyobb erőssége az oktatási ötletgenerálás támogatása. Ez pedagógiai szempontból különösen értékes eredmény, mivel a kreatív tervezés az egyik legnagyobb idő- és energiaigényű pedagógiai tevékenység.

### 4.1.3. Életkori adaptivitás és differenciálás

A kutatás egyik központi kérdése az volt, hogy a modellek milyen mértékben képesek alkalmazkodni eltérő életkori sajátosságokhoz.

A Gemini ebben a dimenzióban mutatta az egyik legerősebb teljesítményt. A modell több esetben tudatosan egyszerűsítette nyelvezetét, rövid mondatokat alkalmazott, és olyan hétköznapi analógiákat használt, amelyek jól illeszkedtek az adott korosztály tapasztalati világához.

A ChatGPT szintén jól teljesített, de bizonyos esetekben a rendszer inkább leegyszerűsítette a fogalmakat, mintsem valóban életkorhoz igazította volna azokat. A DeepSeek válaszai több esetben kreatívak és motiválóak voltak, ugyanakkor a nyelvi bizonytalanság miatt az életkori megfelelés következetessége kevésbé volt stabil.

A differenciálás kérdésében mindhárom modell képes volt bizonyos szintű adaptív működésre. Azonban az elemzés alapján megállapítható, hogy a vizsgált rendszerek jelenleg inkább nyelvi mintázatok alapján „szimulálják" a pedagógiai differenciálást.

#### 4.1.4. Pedagógiai és didaktikai érték

E szempont elemzése során azt vizsgáltuk, hogy a modellek válaszai mennyire tekinthetők valódi oktatási támogatásnak.

A Gemini ezen a területen mutatta a legkiegyensúlyozottabb teljesítményt. A válaszok gyakran tartalmaztak fokozatosságot, motivációs elemeket, reflektív kérdéseket, tanulói aktivitást ösztönző megoldásokat. Ez arra utal, hogy a modell válaszstruktúrája jelentős mennyiségű pedagógiai jellegű mintát integrált.

A ChatGPT pedagógiai szempontból megbízható, de konzervatívabb működést mutatott. A válaszok sok esetben informatívak és strukturáltak voltak, ugyanakkor kevésbé jelent meg bennük a pedagógiai kreativitás vagy az interaktivitás.

A DeepSeek didaktikai szempontból több alkalommal kifejezetten erős teljesítményt nyújtott. A modell gyakran próbált reflektív vagy empatikus kommunikációt alkalmazni, azonban ezt a nyelvi bizonytalanság időnként gyengítette.

### 4.2. Kvalitatív és egyéb megfigyelések

A kutatás során különösen érdekes eredményeket hozott az etikai kérdések elemzése. A modellek jellemzően óvatos, társadalmilag elfogadható és „felelős" válaszokat generáltak.

A kisgyermekkori képernyőhasználattal, az MI veszélyeivel vagy a pedagógusok helyettesíthetőségével kapcsolatos kérdésekre adott válaszokban mindhárom rendszer hangsúlyozta az emberi pedagógus szerepének fontosságát, a mértékletességet, a kritikai gondolkodás szükségességét, a technológia felelős használatát. Ez egyértelműen jelzi a modellekbe épített biztonsági és etikai szűrők működését. Ugyanakkor megfigyelhető volt az is, hogy a válaszok sok esetben erősen standardizáltak és konfliktuskerülők voltak. A modellek hajlamosak voltak konszenzusos, óvatos álláspontokat megfogalmazni.

A kutatás egyik fontos következtetése, hogy a jelenlegi nagy nyelvi modellek legnagyobb pedagógiai potenciálja nem az autonóm oktatásban, hanem a pedagógusok támogatásában rejlik. A vizsgálat ugyanakkor azt is világossá tette, hogy a modellek alkalmazása

továbbra is pedagógiai kontrollt és kritikai értékelést igényel, különösen a tartalmi pontosság, a nyelvi minőség és az életkori megfelelés területén.

A használat komfortját a válaszadás sebessége adja. Ebben a tekintetben a Gemini volt a leggyorsabb, míg a DeepSeek a leglassabb.

## 5. Összegzés és következtetések

A kutatás eredményei alapján megállapítható, hogy a nagy nyelvi modellek pedagógiai alkalmazhatósága már jelenlegi fejlettségi szintjükön is jelentős potenciált hordoz magában. A vizsgált rendszerek nem csupán információ-visszaadásra bizonyultak alkalmasnak, hanem különböző mértékben képesek voltak pedagógiai szempontból releváns, kreatív és életkorhoz igazított válaszok generálására is. Az összehasonlító elemzés alapján három eltérő modellprofil rajzolódott ki.

A Gemini a vizsgálat során a legkiegyensúlyozottabb teljesítményt nyújtotta. Válaszait magas fokú érthetőség, didaktikai szervezettség és pedagógiai adaptivitás jellemezte. A rendszer különösen erősnek bizonyult a komplex fogalmak életkorhoz igazított magyarázatában, valamint a motiváló és interaktív tanulási helyzetek kialakításában. Eredményei arra utalnak, hogy jelenlegi formájában ez a modell rendelkezik a legnagyobb potenciállal az általános oktatási környezetben történő támogatói alkalmazásra.

A ChatGPT stabil és nyelvileg koherens működést mutatott. A modell erőssége elsősorban a kiszámítható és jól szerkesztett kommunikációban rejlik, miközben kreatív vagy komplexebb didaktikai helyzetekben visszafogottabb teljesítményt nyújtott.

A DeepSeek eredményei sajátos kettősséget mutattak. A modell több alkalommal kimagasló kreativitást és részletgazdag pedagógiai gondolkodást demonstrált, ugyanakkor a magyar nyelvi minőség területén jelentős problémák jelentkeztek. Mindez különösen érzékeny kérdés az alsó tagozatos oktatásban, ahol a nyelvi minták minősége pedagógiai szempontból kiemelt jelentőséggel bír.

A kutatás eredményei megerősítik a nemzetközi szakirodalomban megjelenő azon álláspontokat, amelyek szerint a generatív mesterséges intelligencia legnagyobb oktatási potenciálja jelenleg nem a pedagógus helyettesítésében, hanem a pedagógiai folyamatok támogatásában rejlik. (Luckin, 2018; Holmes et al., 2019) A vizsgált modellek különösen alkalmasnak bizonyultak pedagógiai ötletgenerálásra, kreatív tanulási szituációk kialakítására, differenciált magyarázatok készítésére, szemléltető példák generálására és a tanórai előkészítő munka támogatására.

A kutatás ugyanakkor több fontos korlátozó tényezőre is rámutatott. Az egyik legjelentősebb probléma továbbra is az információk megbízhatóságának kérdése. A modellek időnként pontatlan, leegyszerűsített vagy túlzottan általános válaszokat generáltak, amelyek pedagógiai kontroll nélkül félrevezetővé válhatnak. Ez különösen problematikus lehet

olyan életkori szakaszokban, ahol a tanulók még nem rendelkeznek megfelelő kritikai médiaműveltséggel vagy forráskritikai kompetenciával.

A kutatás eredményeinek értelmezésekor fontos figyelembe venni a vizsgálat korlátait is. Az elemzés kizárólag az egyes rendszerek ingyenesen elérhető verzióira terjedt ki, így a fejlettebb fizetős modellek eltérő teljesítményt mutathatnak.

Mindezek ellenére a kutatás világosan jelzi, hogy a MI pedagógiai szerepe a következő években várhatóan tovább erősödik. Az oktatás jövője valószínűleg nem az emberi pedagógus és a mesterséges intelligencia versenyére, hanem együttműködésére épül majd. A generatív modellek legnagyobb értéke jelenleg nem az autonóm oktatásban rejlik, hanem abban, hogy képesek támogatni a pedagógus kreatív, szervező és differenciáló munkáját. A vizsgálat egyik legfontosabb tanulsága talán éppen az, hogy az oktatás emberi dimenziói, az empátia, a nevelési érzékenység, a motiváció fenntartása és a közösségi kapcsolatok kezelése olyan komplex tényezők, amelyek jelenleg még messze túlmutatnak a generatív rendszerek működésén. A MI egyelőre inkább imitálja a pedagógiai folyamatokat, mint valóban érti.

Összegzésként – a kutatásunk eredményei alapján – megállapíthatjuk, hogy a vizsgált LLM-ek ígéretes technológiát jelentenek az oktatás és tanulástámogatás számára a célzott korcsoportokban, de használatuk körültekintést és az egyes modellek specifikus képességeinek részletesebb ismeretét igényli. A pedagógusok feladata, hogy kritikusan értékeljék ezen eszközöket, és megtalálják azokat a módszereket, amelyekkel az LLM-ek valóban az oktatás és tanulástámogatás hatékony eszközeivé, vagy kreatív asszisztenseivé válhatnak. (Holmes et al., 2019), (Szabóné 2023), (Kasneci et al. 2023)

## Irodalomjegyzék


Holmes, W., Bialik, M., & Fadel, C. (2019). Artificial Intelligence in Education: Promises and Implications for Teaching and Learning. Center for Curriculum Redesign. Boston, 2019.

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., ... & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences, Volume 103, April 2023.

Luckin, Rosemary (2018). Machine Learning and Human Intelligence: The Future of Education for the 21st Century. UCL Institute of Education Press. London, 2018.

Michael McDonough (2025) Large language model (LLM) szócikk, Britannica.com, https://www.britannica.com/topic/large-language-model (Hozzáférés: 2025. 10. 24.)

Szabóné Balogh Ágota (2023): Mesterséges intelligencia az oktatásban. Mesterséges Intelligencia – interdiszciplináris folyóirat, V. évf. 2023/2. szám. 51–61.

T. Nagy László, Németh Áron (2024), A mesterséges intelligencia (MI) teológiai kompetenciái In: Tick, József; Kokas, Károly; Holl, András (szerk.) Az oktatás, a kutatás és

a közgyűjtemények digitális transzformációja felsőfokon: NETWORKSHOP 2024: 33. Országos Informatikai Konferencia: Budapest, HUNGARNET Egyesület (2024) 213 p. pp. 45–54. , 10 p.

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education. International Journal of Educational Technology in Higher Education, 16(1), 1–27.

## Források

ChatGPT, https://chatgpt.com/ (Hozzáférés: 2026. 05. 10.)

DeepSeek, https://www.deepseek.com/en (Hozzáférés: 2026. 05. 10.)

Google Gemini, https://gemini.google.com/ (Hozzáférés: 2026. 05. 10.)

DOI: 10.31915/NWS.2026.22



# TÖMEGES IDÉZÉSFELTÖLTÉS AZ MTMT-BE NAGY NYELVI MODELL SEGÍTSÉGÉVEL

# CITATION BATCH UPLOAD INTO MTMT WITH THE HELP OF AN LLM

*Micsik András, Tanácsi Roland*
*HUN-REN Számítástechnikai és Automatizálási Kutatóintézet*
*Elosztott Rendszerek Osztály (DSD)*

## Absztrakt

A Magyar Tudományos Művek Tára (MTMT) a hazai tudományos eredmények és eredményesség hiteles adatbázisa. A kutatások eredményességének jelentős mutatója az idézettség, és ehhez az MTMT folyamatosan adatokat gyűjt. Az idézési adatokat nem mindig lehet központilag megszerezni vagy megvásárolni, ezért ezek feltöltése tudományterületenként változó arányban a szerzőkre és az MTMT adminisztrátoraira hárul. Az esetenként több ezer új idézési adat feltöltése óriási plusz terhet ró a kutatókra és a kutatást segítőkre. E feladat gépi asszisztálására a HUN-REN SZTAKI Elosztott Rendszerek Osztálya (DSD) már régóta próbálkozik különböző megoldásokkal. Az elmúlt 1–2 év mesterséges intelligencia (MI) fejlesztései már lehetővé teszik, hogy egy nagy nyelvi modell (LLM) intelligens módon gyűjtsön metaadatokat weblapokról. E lehetőség kihasználására fejlesztettünk ki egy parancssorból vezérelhető tömeges idézésfeldolgozó szkriptet, a Metaxa-t (Metadata Extractor), amelynek alapvető működési elveit mutatjuk itt be.

**Kulcsszavak**: MTMT, Tudománymetria, Metaadat-kinyerés, Nagy nyelvi modell, LLM

## Abstract

The Hungarian Science Bibliography (MTMT) is the authoritative database of Hungarian scientific achievements and performance. Citation count is a significant indicator of research performance, and MTMT collects citation data for this purpose. Citation data cannot always be obtained or purchased centrally, so uploading these data often falls to the authors and MTMT administrators. The occasional need to upload several thousand new citation records places a substantial additional burden on researchers and research support staff. To provide machine-assisted support for this task, the Department of Distributed Systems (DSD) of HUN-REN SZTAKI has long been experimenting with various solutions. Advances in artificial intelligence (AI) over the past 1–2 years now make it possible for a large language model (LLM) to intelligently collect metadata from web pages. To

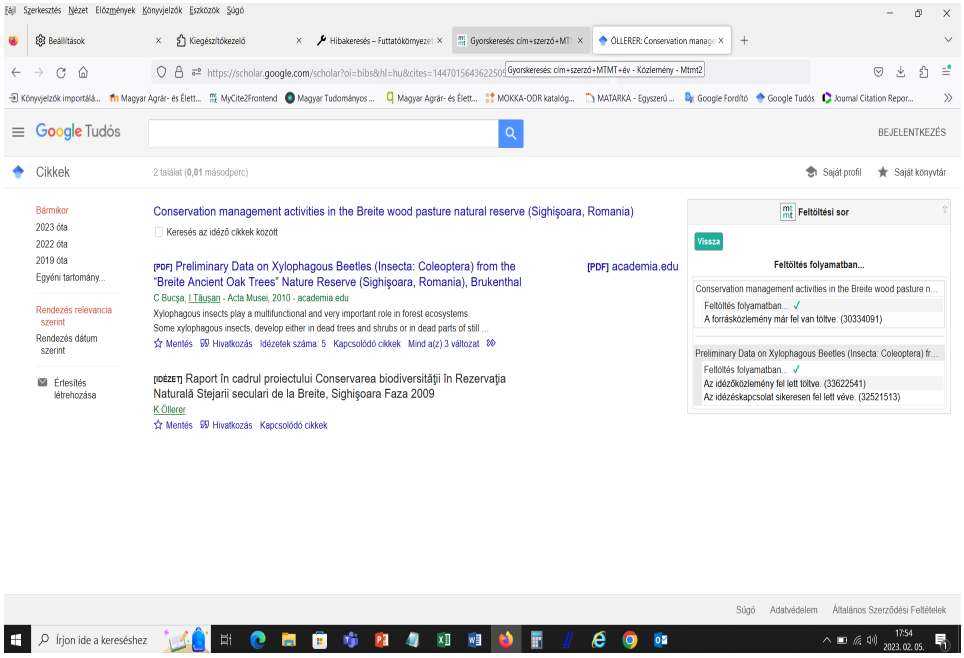
leverage this opportunity, we have developed a command-line–driven bulk citation processing script called Metaxa (Metadata Extractor), whose basic operating principles are presented here.

**Keywords:** Hungarian Science Bibliography, Scientometrics, Metadata extraction, Large Language Model, LLM

## Bevezetés

A Magyar Tudományos Művek Tára (MTMT) a hazai tudományos eredmények és eredményesség hiteles adatbázisa. Az MTMT 2.0 verzióját a SZTAKI Elosztott Rendszerek Osztálya tervezte és fejlesztette, valamint 2018 novembere óta a SZTAKI végzi a rendszer informatikai üzemeltetését is. A rendszerben a statisztikai, tudománymetriai elemek az elmúlt időszakban folyamatosan, az igények módosulása szerint változtak, például új folyóirat értékelési presztízsfaktorok jelentek meg. A kutatások eredményességének jelentős mutatója az idézettség, és ehhez az adatgyűjtés rengeteg emberi munkát és odafigyelést igényel. Az idézőközlemények beszámítását az egyes tudományos fórumok különböző feltételekhez kötik, melyek ellenőrzéséhez nem elég a szokásos alapadatok beszerzése az idézőkről. Számíthat az is, hogy szerepel-e az idézőközlemény a Web of Science-ben vagy a Scopusban, hogy hány oldal terjedelmű, hogy van-e ISSN vagy ISBN száma a kiadványnak, és így tovább. Nincs még általános megoldás a világban az idézéskapcsolatok keresésére, különböző adatbázisok különböző minőségben és lefedettséggel gyűjtik ezeket az adatokat. Az MTA mint fenntartó próbál központilag beszerezni és a rendszerbe betölteni hazai vonatkozású idézéskapcsolatokat, de az egyes tudományágak lefedettsége ezekben a nemzetközi adatbázisokban messze nem egyenletes, illetve a tudományos konferenciák feldolgozottsága is hiányos. Így a hiányok pótlása tudományterületenként változó arányban a szerzőkre és az MTMT adminisztrátoraira hárul, és ezt a feladatot általában kampányszerűen évente egyszer végzik el. Mivel akár egyetlen kutatónak is keletkezhet ezernél is több új hivatkozása egy év alatt, ezek összegyűjtése és feldolgozása óriási plusz terhet ró a kutatókra és a kutatást segítőkre.

E feladat megkönnyítésének lehetőségein már régóta gondolkozunk, és 2022-ben kifejlesztettük az MTMT böngésző plugint (1. ábra), amely a Google Scholar idézéslistái alapján képes volt a kiválasztott idézéseket az MTMT-be feltölteni. Ezzel a megoldással egy kattintással maximum 10 idézőközleményt lehetett feltölteni, de még ez is nagy segítséget jelentett a plugint használó adminisztrátoroknak. A böngészők bővítgetésének divatja azóta alábbhagyott, de más gond is volt ezzel a megoldással: nem sikerült a közlemények metaadatait elég jó minőségben összeállítani. Ez leginkább a ritkább típusok, mint például a disszertációk, szabadalmak esetén volt kritikus, de általánosan is előfordultak hibák.




*1. ábra. Az MTMT böngésző plugin munka közben*

A mesterséges intelligencia (MI) fejlődése hozta meg a lehetőségét annak, hogy teljesen új alapokra helyezzük az idézések feltöltését. A nagy nyelvi modellek strukturált kimenete, a parancssorban működő kódoló ágensek, illetve az ágensi működésmód adták azokat az építőelemeket, amelyre Metaxa (Metadata Extractor) nevű fejlesztésünket alapoztuk. A továbbiakban bemutatjuk a Metaxa műszaki megoldásait és működésének részleteit. Mindezek előtt azonban felvázoljuk a terület aktuális, fejlett megoldásait.

## Lehetőségek idézések gyűjtésére

Az egyetemek, kutatóintézetek általában előfizetnek kutatást segítő adatbázisokra és szolgáltatásokra, ahonnan a szerzőik idézeteit is le tudják keresni és menteni. A lementett idézetgyűjteményeket egy közleményhez egyszerre fel tudják tölteni az MTMT-be. Erre a legnépszerűbb szolgáltatások a Scopus és a Web of Science, és ehhez adódnak hozzá az intézetben művelt tudományágak saját preferált bibliográfiai adatbázisai. Ez a módszer még mindig sok emberi munkát igényel, mivel cikkenként és több helyről kell exportálni az adatokat, majd betölteni az MTMT-be.

Akadnak általános és ingyenesen hozzáférhető idézésgyűjtemények is, a legismertebb talán a Google Scholar. Bár a Scholar gyűjtési köre az ismertek közül talán a legszélesebb körű, ennek velejárójaként a minősége alacsonyabb, gyakran kerülnek be nem odavaló elemek vagy hiányos adatokkal szereplő közlemények. Több kezdeményezés fut teljesen ingyenesen letölthető, nyílt idézés adatbázisok létrehozására. Ilyen például az OpenCita-

tions[1], az OpenAlex[2] vagy a Semantic Scholar[3], de ezek még mindig jóval kevesebb idézést tartalmaznak a Google Scholarnál, és így fennáll a veszélye, hogy a szerzők lemaradnak egyes idézésekről.

Amennyiben megelégszünk a DOI-val rendelkező idézőkkel, akkor a CrossRef és a DataCite is nyújt ilyen keresési szolgáltatást, de ezek értelemszerűen az adott DOI-szolgáltatóra korlátozottak.

Összességében az idézések gyűjtését még mindig az üzleti megoldások uralják, és a nyílt tudomány mozgalom nem jutott még el egy széleskörűen használható nyílt idézettségi platform létrehozásáig.

### Tudományos közlemények import formátumai

Amennyiben hozzáférünk a megfelelő forrásokhoz, felmerül a kérdés, hogy milyen formátumban érdemes gyűjteni az idézéseket. Az ideális az lenne, hogy az idézett mű, idéző mű és az idézés adatai egyetlen jól definiált módon lennének tárolhatók, azonban ilyen megoldásról még nem tudunk, ami elterjedt lenne. Linked data technológiával és a megfelelő SPAR ontológiák használatával ez megoldható, és az OpenCitations használja is, de nem mondhatjuk egy széleskörűen használható technológiának. Ezért a problémát visszavezetjük az idéző közlemények metaadatainak exportálására, amire szinte minden szolgáltatónál van lehetőség.

Jelenleg a RIS-formátum az, amely a legtöbb helyen elérhető mint exportformátum, de kezelhetőség szempontjából sok hátrányos tulajdonsággal rendelkezik; egyrészt nem hierarchikus, ezért a komplex leírókat (pl. szerző neve, affiliációja, ORCID-ja) is egy szinten kell szerializálni, valamint nem egységes a formátum használata, az egyes mezők adattartalma szolgáltatónként eltérhet.

A BibTeX talán a második legelterjedtebb ilyen formátum, amely ráadásul fejlett ökoszisztémával rendelkezik a rekordok keresésétől kezdve a különféle referencialisták előállításáig. Azonban a BibTeX lassan vagy alig követi a tudományos publikálás fejlődését, pl. az ORCID és ROR azonosítók szerepeltetésére nincs egységes megoldás.

Figyelemre méltó e téren a Citation Style Language[4] (CSL), amely ugyan a különféle hivatkozási formátumok generálására jött létre, de létezik egy specifikációja a közlemények metaadatainak JSON ábrázolására is.

1 https://opencitations.net/

2 https://openalex.org/

3 https://www.semanticscholar.org

4 https://citationstyles.org/

A Zotero[5] egy referencia kezelő szolgáltatás, amelyben nagyon könnyen lehet hivatkozásnak való anyagot gyűjteni akár közösen is. A Zotero alapműködése a CSL-re épül, és a publikációs weboldalakból programozott úton nyeri ki a metaadatokat, amely azzal a hátránnyal jár, hogy a webes felület megváltozásával az adatkinyerés sérülhet. A Zoteron kívül több mint 60 szolgáltatás és szoftver támogatja a CSL-t, és a DOI-infrastruktúra is képes a tárolt metaadatokat CSL-re konvertálni[6]. Az összes DOI-szolgáltató által támogatott metaadat formátum között is számunkra a legígéretesebb a CSL-JSON.

### Közlemény metaadat kinyerés

A metaadatok automatikus kinyerésével számos kutatás foglalkozott, amint azt a hivatkozott két áttekintő cikk is bemutatja [2,3], de használható megoldások csak 1–2 éve jöttek létre. Kísérleti fejlesztésünk céljául a fentiek alapján azt tűztük ki, hogy egy közlemény webes megjelenéséből MI segítséggel CSL-JSON formában metaadatokat tudjunk jó minőségben és nagy pontossággal kinyerni.


```
{
   „title”: „RANSAC for Robotic Applications: A Survey”,
   „author”: [
     {
       „family”: „Martínez-Otzeta”,
       „given”: „José María”,
       „ORCID”: „https://orcid.org/0000-0001-5015-1315”,
       „affiliation”: [
         {
           „name”: „Department of Computer Science and Artificial
          Intelligence, University of the Basque Country”,
           „place”: „Donostia-San Sebastián, Spain”
         },
         {
           „name”: „Department of Languages and Information Sys-
          tems, University of the Basque Country”,
           „place”: „Donostia-San Sebastián, Spain”
         }
```


5 https://www.zotero.org/

6 https://citation.doi.org/docs.html

```
      ],
      „role”: „author”
    },
    ...
  ],
  „issued”: {
    „date-parts”: [
      [
        2022,
        12,
        28
      ]
    ]
  },
  „DOI”: „10.3390/s23010327”,
  „URL”: „https://www.mdpi.com/1424-8220/23/1/327”,
  „language”: „en”,
  „type”: „article-journal”,
  „container-title”: „Sensors”,
  „volume”: „23”,
  „issue”: „1”,
  „page”: „327-327”,
  „ISSN”: [
    „1424-8220”
  ],
  „publisher”: „Multidisciplinary Digital Publishing Institute”
}
```


*2. ábra. LLM által generált CSL JSON részlete*

Erre a feladatra a nagy nyelvi modellek strukturált kimenete kiváló támogatást nyújt. Mivel a CSL-hez elérhető hivatalos séma nem volt elég pontos, létre kellett hoznunk Zodban[7] az általunk preferált sémát, és ezt már meg lehet adni az LLM-nek mint elvárt output. A köz-

7 https://zod.dev/

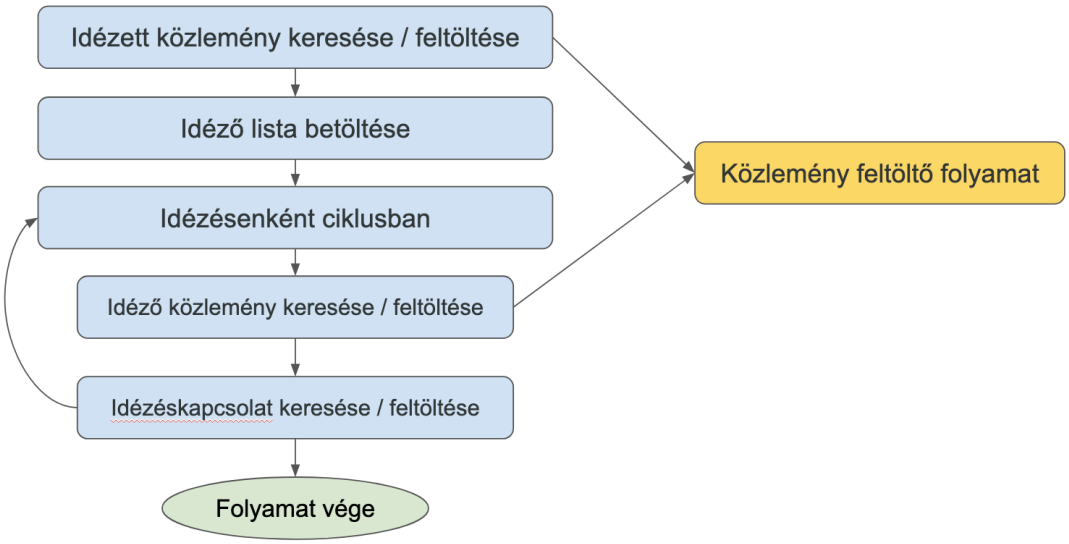
lemény CSL-t ily módon a közlemény HTML-ből jó minőségben elő tudtuk állítani kisebb modellekkel is, mint például a Qwen3.6-27b vagy a gpt-5-nano (2. ábra).

A CSL megszerzésének másik módja az, hogy lekérjük a DOI-hálózatból. Sajnos az így kapott CSL minősége szélsőséges végletek között mozog: a CrossRef metaadataiból készült CSL például tartalmazza a szerzők ORCID-ját és affiliációik ROR-azonosítóját is, míg repozitóriumokból származó szerzői változatok CSL-je esetén hiányosak, sőt hibásak is lehetnek a metaadatok. Az is meglepő volt, hogy a CSL-mezők elnevezéseiben is eltéréseket láttunk, főleg a CrossRef alkalmazott teljesen új mezőket (pl. „authenticated-orcid"), nem szabványos típusokat („article-journal" helyett „journal-article") vagy változtatott a mezőneveken ("event-title" helyett "event"). Az is változó a használatban, hogy egyes mezők értékkészlete mi lehet: karaktersorozat, tömb vagy rekord, ez a szerzői affiliációknál a legváltozatosabb.

Mindezek alapján szükséges egy utolsó ellenőrzés és hibajavítás a CSL-ben mielőtt az MTMT-be töltenénk, és ezt szintén LLM-mel végeztetjük. A kész CSL duplumellenőrzés után bekerül a rendszerbe, de az MTMT-ben megszokott módon ilyenkor még nem lesz nyilvános, szerzői vagy adminisztrátori jóváhagyás kell ahhoz, hogy a statisztikákban és nyilvános felületeken megjelenjen az új tétel. Ilyenkor még a hiányzó, kitöltetlen adatokat is pótolni kell (pl. jelleg és besorolás mezők), ezért a személyes ellenőrzése az ilyen módon bekerült új rekordoknak feltétlen szükséges.

## Az idézésfeltöltés munkafolyamata

Az idézések feltöltésének kiindulópontja egy lista, amely egyetlen idézett közleményt és az azt idéző közlemények felsorolását tartalmazza. Amikor az idézőközlemény már a rendszerben van, létre kell hozni az idézéskapcsolatot a két közlemény között (3. ábra). Ehhez a kapcsolathoz további információkat is lehet később csatolni, nevezetesen a függetlenség jelzését és az említések helyeit többszörös hivatkozás esetén, de ezek jelenleg nincsenek a Metaxában támogatva; a függetlenség csoportos gépi ellenőrzésére van külön menüpont az MTMT szerkesztői felületén.




*3. ábra. A tömeges idézésfeltöltés folyamata*

A Metaxa az indítás után folyamatosan jelzi a konzolon az előrehaladását és gyűjti a feldolgozási statisztikát, amelyet a végén táblázatos (CSV) formában elment. A táblázat alapján a feltöltést indító tájékozódhat, hogy melyik idéző milyen azonosítóval került be az MTMT-be, illetve ha nem sikerült a feltöltés, akkor mi volt a hiba.

A Metaxa Javascript nyelven készült, a LangChain, Puppeteer, Zod modulokra alapozva, és a felhasználó által beállított bármely korszerű, instruált LLM-et tudja használni a működéséhez.

## Összefoglalás

Az idézéseken alapuló tudományos teljesítménymméréshez szükség van az idézések részletes metaadataira. Ezek összegyűjtése és MTMT-be feltöltése óriási plusz terhet ró a kutatókra és a kutatást segítőkre. A friss MI-adta lehetőségek felhasználásával kifejlesztettünk egy megoldást ennek a feladatnak az automatizálására, amely a Metaxa (Metadata Extractor) nevet kapta. Eddigi kísérleteink alapján a közlemények metaadatainak CSL-formátumba való kinyerése a jelenlegi technológiával megbízhatóan működik, akár kisebb LLM-ekkel is. Háromféle folyamat támogatására folytatunk kísérleteket: (1) egyetlen közlemény automatikus feltöltése az MTMT-be; (2) egy közlemény új idézőinek és idézéskapcsolatainak feltöltése; (3) egy szerző új közleményeinek és idézéseinek feltöltése. A Metaxa alaposabb tesztelése után fogunk dönteni annak valós alkalmazási lehetőségeiről.

## Bibliográfia

[1] Peroni, S., Shotton, D. (2018). The SPAR Ontologies. In Proceedings of the 17th International Semantic Web Conference (ISWC 2018): 119–136. DOI: https://doi.org/10.1007/978-3-030-00668-6_8

[2] Nasar, Z., Jaffry, S.W. & Malik, M.K. (2018) Information extraction from scientific articles: a survey. Scientometrics 117, 1931–1990. https://doi.org/10.1007/s11192-018-2921-5

[3] Saxi Soni, Patel Ketankumar, Zeel Nakum. (2026). A Comprehensive Review of Methods, Frameworks, and Domains for Metadata Extraction from Scientific Texts. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 12(1), 141–146. https://doi.org/10.32628/CSEIT261218

DOI: 10.31915/NWS.2026.23



# AZ IDŐSZAKI KIADVÁNYOK MINT AGGREGÁTUMOK ÉS DIAKRONIKUS MŰVEK LEÍRÁSA AZ IFLA LRM ÉS AZ RDA ALAPJÁN

# DESCRIPTION OF PERIODICALS AS AGGREGATES AND DIACHRONIC WORKS BASED ON IFLA LRM AND RDA

*Vass Johanna*
*ELTE BTK KITI* (*óraadó oktató*)
*vass.johanna@btk.elte.hu*

## Absztrakt

Az időszaki kiadványok katalogizálása a könyvtári gyakorlat egyik állandóan visszatérő problématerülete. Sajátosságuk, hogy a katalógusban többnyire az első szám adatai alapján azonosított forrás időközben számos lényeges jellemzőjében megváltozhat; emellett bibliográfiai szempontból legalább három szint különíthető el esetükben: maga a folyóirat vagy másként az időszaki kiadvány mint egész; az egyes lapszámok; valamint az azokban megjelent önálló közlemények.

A tanulmány célja az időszaki kiadványokra vonatkozó új elméleti alapok, különösen az IFLA LRM és az RDA megközelítésének áttekintése. A dolgozat kitér az időszaki kiadvány fogalmi meghatározásának változásaira, az entitáshatárok kérdésére, valamint az összefoglaló és az analitikus leírás modellbeli eltéréseire. Emellett röviden érinti azokat a kapcsolódó modelleket és adatstruktúrákat is, amelyek az időszaki kiadványok részletesebb leírása szempontjából a jövőben jelentősek lehetnek.

**Kulcsszavak:** Resource Description and Access; RDA; serial works; időszaki kiadvány; bibliográfiai leírás; aggregátum

## Abstract

The cataloguing of serials has long been one of the most challenging areas of library practice. A serial usually enters the catalogue on the basis of data appearing in its first issue, yet many of its essential characteristics may change over time. Moreover, from a bibliographic point of view, at least three levels can be distinguished in the case of serials: the serial as a whole, the individual issues, and the articles or other contributions published within them.

The aim of this paper is to provide an overview of the new theoretical foundations for the description of serials, with particular attention to the approaches represented by IFLA LRM and RDA. The study discusses changes in the definition of serials, the question of entity boundaries, and the differences between summary and analytical description. It also briefly addresses some related models and data structures that may become increasingly relevant in the more detailed description of serial resources.

**Keywords**: Resource Description and Access; RDA; serials; bibliographic description; aggregate; diachronic work

## 1. Bevezetés

Az utóbbi évtizedek katalogizáláselméleti dokumentumai – elsősorban a Functional Requirements for Bibliographic Records (FRBR)[1] és az IFLA Library Reference Model (IFLA LRM)[2] –, valamint az ezek nyomán születő új leírási szabványok, különösen a Resource Description and Access (RDA)[3], nem csupán a bibliográfiai univerzum entitás-kapcsolat alapú szemléletét alakították át, hanem több hagyományos dokumentumtípus megközelítését is új alapokra helyezték. Különösen igaz ez az aggregátumokra és az időszaki kiadványokra.

Az RDA szabványhoz kapcsolódó magyar nyelvű szakirodalom az utóbbi években jelentősen gyarapodott[4], és a hazai könyvtári közösség is készül a szabvány bevezetésére[5]. Az előkészítő munka következő lépése lehet annak részletesebb vizsgálata, hogy az egyes forrástípusok miként írhatók le az IFLA könyvtári referenciamodellje és az RDA instrukciói alapján.

---

1 A bibliográfiai tételek funkcionális követelményei: zárójelentés. Ford. Berke Barnabásné. Budapest, Országos Széchényi Könyvtár, 2006. Forrás: https://www.ifla.org/wp-content/uploads/2019/05/assets/cataloguing/frbr/frbr-hu.pdf [Elérés: 2026.05.17.]

2 Riva, Pat – Boeuf, Patrick Le – Žumer, Maja: IFLA könyvtári referenciamodell. A bibliográfiai információk elméleti modellje. Ford. Gazdag Tiborné. Budapest, Országos Széchényi Könyvtár, 2018. Forrás: https://www.ifla.org/wp-content/uploads/2019/05/assets/cataloguing/frbr-lrm/ifla-lrm-august-2017_rev201712-hu.pdf [Elérés: 2026.05.17.]

3 A Resource Description and Access (RDA) szövege elérhető: RDA Toolkit. Forrás: https://www.rdatoolkit.org/

4 Legutóbb: Vass Johanna (2026): A forrásfeltárás korszerű módszerei. In: Kiszl, Péter (szerk.) Útmutató könyvtáralapításhoz 4. Budapest : ELTE Bölcsészettudományi Kar Könyvtár- és Információtudományi Intézet, ELTE Eötvös Kiadó, pp. 183-280. DOI: https://doi.org/10.21862/utm_kvta_4 [Elérés: 2026.05.17.]; Dancs Szabolcs (2025): RDA-alapok : Út az RDA megértéséhez. Budapest : Magyar Nemzeti Múzeum Közgyűjteményi Központ, Országos Széchényi Könyvtár, 81 p. Forrás: https://mek.oszk.hu/28400/28492 [Elérés: 2026.05.17.]
RDA-HU munkacsoport. Forrás: https://oszk.hu/rda-hu-munkacsoport [Elérés: 2026.05.17.]

5 Többek közt az MNMKK OSZK-ban működő RDA-HU munkacsoport tevékenysége, illetve az MNMKK OSZK Könyvtári Intézet által szervezett tanfolyamok révén. V.ö.: RDA-HU munkacsoport. Forrás: https://oszk.hu/rda-hu-munkacsoport [Elérés: 2026.05.17.] ; Forrásleírás és -hozzáférés a 21. században [tanfolyam]. MNMKK OSZK Könyvtári Intézet. Forrás: https://ki.oszk.hu/tanfolyamok/forrasleiras-es-hozzaferes-21-szazadban [Elérés: 2026.05.17.]

E tekintetben az időszaki kiadványok külön figyelmet érdemelnek. Bár a könyvtári gyakorlatban mindig is kiemelt szerepet játszottak, az újabb magyar szakirodalomban viszonylag kevés figyelem irányult arra, hogy e források miként értelmezhetők az LRM és az RDA fogalmi keretei között. Pedig az időszaki kiadványok éppen folyamatosan változó, mégis azonosítható egészként való fennállásuk miatt különösen próbára teszik a bibliográfiai modelleket.

A tanulmány célja ezért annak összefoglalása, hogy miként alakult az időszaki kiadvány fogalma, hogyan jelölhetők ki határai, és milyen módon kapcsolható össze az összefoglaló és az analitikus leírás logikája az LRM és az RDA alapján.

## 2. Az időszaki kiadvány meghatározása

Az időszaki kiadvány fogalma történetileg változó. Hosszú ideig alapvető kritériumnak az évenkéntinél gyakoribb megjelenés számított: az évi rendszerességű kiadványokat nem tekintették időszaki sajtónak. A magyar sajtóbibliográfiai hagyományban ennek jól ismert képviselője Dezsényi Béla, aki az időszaki sajtó körébe csak az egy évnél rövidebb időközökben megjelenő sajtótermékeket sorolta.[6]

A későbbi szabványokban a hangsúly fokozatosan eltolódott a periodicitás felől a nyitott végűség és a részegységekre tagoltság irányába. Az International Standard Bibliographical Description (ISBD) alapján kidolgozott meghatározások szerint az időszaki kiadvány olyan, előre meg nem határozott időtartamra tervezett kiadvány, amely egymást követő részegységekből áll, és ezeket számozás, keltezés vagy más megjelölés különbözteti meg egymástól.[7]

A 20. század végére azonban egyre több olyan forrástípus jelent meg – például cserelapos kiadványok vagy számozatlan sorozatok –, amelyek a dokumentumtípusok korábbi definícióiba nehezen voltak beilleszthetők. Ekkor vált hangsúlyossá a „publikációs státusz" szempontja, vagyis az a kérdés, hogy a közzététel lezárt vagy előre meg nem határozott ideig folytatódó folyamatként értelmezhető-e. Ez a szemlélet már kevésbé függ a formai ismérvektől, és elsősorban a forrás időbeli nyitottságára koncentrál, lényegét pedig egy újonnan bevezetett elnevezés, a „diachronic works" ragadja meg.[8]

Az FRBR az időszaki kiadványokkal még nem foglalkozott részletesen. Az IFLA LRM viszont már kitér rájuk, és két szempontból is lényeges előrelépést jelent. Egyrészt bevezeti az aggregáló mű fogalmát, amely lehetővé teszi az olyan források leírását, amelyek több önálló

6 Dezsényi Béla (1942): Sajtó és könyvészet. In: Magyar Könyvszemle, 66. évf. = 3. folyam, 2.sz., p. 126-152. Forrás: https://epa.oszk.hu/00000/00021/00209/pdf/ [Elérés: 2026.05.17.]

7 Többek közt a KSZ/3:2001 Bibliográfiai leírás. Időszaki kiadványok. Budapest : OSZK, 2001. 104 p.

8 Részletesebben összefoglalja Ilácsa Szabina (2024) Időszakosság a katalogizálásban, Central European Library and Information Science Review (CeLISR), 1(1). Forrás: https://doi.org/10.3311/celisr.37676 [Elérés: 2026.05.17.]

tartalmi egységet foglalnak össze. Másrészt az időszaki kiadványt a mű entitás sajátos eseteként értelmezi: ez a „serial work", amely egészében mint diakronikus mű ragadható meg, míg az egyes lapszámok az aggregátumok modellje alapján jellemezhetők. Fontos itt megjegyezni, hogy a fentiekből adódóan az LRM megjelenése előtti szakirodalom, amely még az FRBR által felvázolt struktúra – Work-Expression-Manifestation-Item – szerint próbálta modellezni az időszaki kiadványok leírását, nem használható teljes mértékben, legfeljebb egy-egy részmegállapítás vehető át az ott felvonultatott leírási gyakorlatból.[9]

Folyóiratok esetében az IFLA LRM elhagyja a standard FRBR-struktúrát, ahol egy mű több kifejezési formában is megvalósulhat, valamint ezek több megjelenési formában testesülhetnek meg. Az időszaki kiadványok viszont kihívást jelentenek ennek a struktúrának, elsősorban azért, mert az időszaki kiadvány határai szorosan kötődnek a címhez. Például egy fordítás címe az eredeti kiadvány címétől eltérő időpontban változhat meg, ami problémákat okozhat egy közös mű kifejezési formáinak összehangolásában. Ezenkívül az időszaki kiadványoknak nincs előre meghatározott végük, így az idők során a kifejezési formák eltávolodhatnak egymástól: a helyi kiadások kezdetben jobban hasonlíthatnak egymásra, majd egyre kevésbé; egy más nyelvű kiadás indulásakor hű lehet az eredetihez, majd kevésbé.[10] Az „időszaki kiadvány mű" praktikuma abban rejlik, hogy a leírandó forrás bármely változás – legyen az megváltozott cím; más nyelvű kiadás; eltérő hordozón megjelenő kiadás; egy lap helyi kiadásváltozatai stb. – mellett is önállóan azonosítható egységként jelenik meg. Ez viszont elvezet a WEM-lock problémájához: az előbbiekből következik, hogy a serial work esetében a műnek egy, és csakis egy kifejezési formája, illetve megjelenési formája lehetséges.

## 3. Az időszaki kiadvány határai

Az időszaki kiadványok bibliográfiai kezelésének egyik legnehezebb kérdése az, hogy hol húzódnak a forrás határai. Másként fogalmazva: mely változások férnek bele ugyanannak a műnek a történetébe, és mely ponton kell új entitásról beszélnünk.

A történeti gyakorlat e kérdésre is eltérő válaszokat adott, így elmondhatjuk, hogy a probléma eldöntése eleve konszenzuális természetű. Korábbi szabályok a közreadó nevének változását vagy a számozási struktúra módosulását tekintették döntőnek. A későbbi szab-

9 A teljesség igénye nélkül néhány példa a Mű-Kifejezési forma-Megjelenési forma felosztásra alapozó szakirodalomból: Bross, Valerie – Hawkins, Les – Nguyen, Hien (2013): CONSER serial RDA workflow. The Serials Librarian, 64(1–4), 211–215. https://doi.org/10.1080/0361526X.2013.760412 ; Hawkins, Les – Nguyen, Hien – Kharfen, Sarah (2015): Basic serials cataloging workshop. The Serials Librarian. https://doi.org/10.1080/0361526X.2015.1012473 ; Stachokas, George (2020): The role of the electronic resources librarian. In: Reimagining Technical Services. https://doi.org/10.1016/B978-0-08-102925-1.00007-6 [Elérés: 2026.05.17.]

10 V.ö.: Jones, Ed (2013): Description of serials, RDA, and the MARC 21 bibliographic format. The Serials Librarian, pp. 128., 135. Forrás: https://doi.org/10.1080/0361526X.2013.836465 [Elérés: 2026.05.17.]

ványosítási gyakorlatban viszont inkább a cím tartós és jelentős megváltozása vált meghatározóvá; generikus című kiadványok esetében ehhez társulhat a közreadó változása is. Ez a megoldás abból indul ki, hogy a számozási struktúra többször is módosulhat anélkül, hogy a kiadvány identitása megszakadna.[11]

Az LRM a kérdést más elméleti alapra helyezi. Mivel a művet önálló szellemi vagy művészi tartalomként értelmezi, az időszaki kiadványok elhatárolásában is a tartalmi és szerkesztői összetartozás szempontjai válnak fontossá. Ilyen lehet az azonos szerkesztői koncepció, az, hogy minden szám felismerhetően ugyanahhoz az egészhez tartozik, valamint a cím, a tipográfiai elrendezés vagy a rendszeres megjelenés.

Az RDA ugyanakkor világossá teszi, hogy az entitások határai részben helyi és kulturális konvenciókon alapulnak. Ennek megfelelően több olyan esetet is nevesít, amikor új mű létrehozását javasolja: ilyen lehet például a megjelenési gyakoriság jelentős változása, az ISSN változása, a hordozótípus módosulása vagy a preferált cím lényeges megváltozása. A címváltozások esetében sem pusztán formai eltérésről van szó, hanem arról, hogy a módosulás érinti-e a jelentést, a tárgyi irányultságot vagy a bibliográfiai azonosíthatóságot.

A gyakorlatban ezért az időszaki kiadvány határainak kijelölése továbbra sem tekinthető teljesen mechanikus műveletnek. A tisztán formai és a tisztán tartalmi megközelítés egyaránt bizonytalanságokat hagy maga után. Ez különösen akkor válik nyilvánvalóvá, amikor a kiadvány címe változatlan marad, de alcíme, illetve az abban jelzett tematikai irányultság jelentősen módosul.

Egy egyszerű, és talán sokak által ismert példával szemléltethetjük a fenti állítást. A 20. század utolsó harmadában népszerű, színvonalas közéleti, kulturális folyóirat, a Kritika (ISSN 0324-7775) 1972-ben indult „művelődéspolitikai és kritikai lap" alcímmel. Ez az alcím a 90/7. számig volt jelen a fejléc alatt, ezt követően egészen a 99/12. számig alcím nélkül jelent meg a kiadvány. A 2000/1. számon pedig új alcím jelezte – „társadalomelméleti és kulturális lap" alakban –, hogy a lap profilja kibővült a korábbiakhoz képest. (1. kép) A könyvtári gyakorlat 1972-től a folyóirat 2017-es megszűnéséig egyazon kiadványnak tekinti a fenti ISSN számmal azonosított periodikát. A nemzeti könyvtár katalógusa pedig számon tartja, hogy azonos címen, 1963 és 1971 között megjelent egy „másik" Kritika is, „a Magyar Tudományos Akadémia Irodalomtörténeti Intézete, a Magyar Irodalomtörténeti Társaság és a Magyar Írószövetség folyóirata", ezt a kiadványt a 0023-4818 ISSN-szám azonosítja. (2. kép) A legnagyobb hazai teljes szövegű digitális archívum, az Arcanum Újságok pedig mind a két lapot egy linken[12], kvázi egyetlen „entitásként" szolgáltatja…

11 Kiváló történeti összefoglalót ad az időszaki kiadványok kezeléséről Ed Jones előző lábjegyzetben hivatkozott írása.

12 Kritika. Forrás: https://adt.arcanum.com/hu/collection/MTA_Kritika/ [Elérés: 2026.05.17.]

## 4. Két lehetséges megközelítés: az összefoglaló és az analitikus leírás

Az időszaki kiadványok leírásában a könyvtári katalógusok hagyományosan az összefoglaló leírást részesítik előnyben. Ilyenkor a kiadvány mint folyamatosan megjelenő egész kap bibliográfiai leírást, rendszerint az első szám alapján. A nemzeti könyvtár aktuális gyakorlata ezt az eljárást kiegészíti azzal, hogy bizonyos adatok változását – például a megváltozott alcímet, közreadót, megjelenési adatokat, mellékleteket, valamint a periodicitást – megjegyzésként hozzáadja az eredeti leíráshoz. Ez a megoldás jól illeszkedik ahhoz az LRM-beli felfogáshoz, amely minden egyes időszaki kiadványt önálló műként kezel.

Számos esetben azonban nem elegendő a kiadvány egészének szintjén történő leírás. Tudományos közlemények követésekor, hivatkozásakor, illetve digitalizált történeti folyóiratok tartalmi feltárásakor az egyes lapszámok és a bennük megjelent közlemények analitikus leírására is szükség van. Ekkor az aggregátumok modellje kínál megfelelő keretet.

Az egyes lapszámok ugyanis rendszerint több önálló közleményt foglalnak magukba egy szerkesztői elgondolás alapján. Ilyen értelemben a lapszámok aggregáló műként értelmezhetők: az időszaki kiadvány egészének szintjén nem értelmezhető az a kapcsolódás, amikor részegységenként több különálló mű jelenik meg egyetlen kiadói és szerkesztői keretben. Ez a nézőpont lehetővé teszi, hogy az időszaki kiadvány egészét diakronikus műként, az egyes lapszámokat pedig aggregátumként kezeljük.

A bibliográfiai gyakorlat számára ebből az következik, hogy legalább három szint kapcsolatát kell tudatosan kezelni: a folyóirat mint egész, az egyes lapszámok és a bennük megjelent közlemények szintjét. A hagyományos katalógusok főként az első, a repertóriumok és cikkadatbázisok inkább a harmadik szint kifejezését hangsúlyozzák. A kapcsolt adatok környezetében viszont egyre inkább az a cél, hogy e szintek közötti viszonyok explicit módon, géppel is feldolgozható relációkként jelenjenek meg.

E ponton láthatóvá válik az LRM és az RDA egyik korlátja is. Bár fontos elméleti keretet adnak, nem dolgozzák ki részletesen az időszaki kiadványok közötti és az időszaki kiadványokon belüli összes lehetséges kapcsolatot. Ezért a helyi alkalmazásoknak vagy a részletesebb, időszaki kiadványokra „szabott” modelleknek maguknak kell kialakítaniuk azokat az elemkészleteket, amelyek pontosan kezelik például az előzmény–folytatás, a kiadásváltozat, a hordozóváltás vagy a részegységnek az egész kiadványhoz való kapcsolatait.

Ebben a tekintetben a szakirodalom[13,14] alapján a PRESSoo[15], illetve a BIBFRAME[16] felé fordulhatunk. (A PRESSoo az FRBRoo modell kiterjesztése, egy olyan formális ontológia, melynek célja kifejezetten az időszaki kiadványok leírására szolgáló bibliográfiai információk összegyűjtése és ábrázolása. A Library of Congress által fejlesztett BIBFRAME keretrendszer pedig a jövőbeni, Linked Data környezetben történő katalogizálást kívánja megalapozni.) Utóbbi olyan sajátos módon egyszerűsíti a helyzetet – miután a BIBFRAME modelljéből[17] eleve hiányzik az FRBR-modell Expression szintje –, hogy számos változatot eleve műként kezel. Bár ez nem mindenben felel meg az LRM differenciált logikájának, implementációs szempontból kedvezhet az időszaki kiadványok Linked Data környezetben történő reprezentációjának. A BIBFRAME HUB entitása[18] pedig arra is alkalmas lehet, hogy a változások ellenére összefogja egyazon lapcsalád különböző történeti alakzatait.

## 5. Következtetések

Az időszaki kiadványok fogalmi és bibliográfiai kezelése történetileg mindig is konvenciókon alapult. A különböző korszakok más és más ismérveket tekintettek meghatározónak: a periodicitást, a számozást, a címet, a közreadót vagy a tartalmi-szerkesztői azonosságot. Ebből következően az időszaki kiadványok határai ma sem jelölhetők ki teljesen automatikusan.

Az IFLA LRM mégis jelentős előrelépést képvisel, mert az időszaki kiadványt önálló elméleti problémaként kezeli. A diakronikus mű és az aggregáló mű fogalma, valamint a WEMlock felismerése olyan keretet kínál, amelyben az „időszaki kiadvány mű" mint egész és az egyes lapszámok közötti viszony következetesebben modellezhető. Az RDA ehhez gyakorlati instrukciókat is társít, de egyben azt is elismeri, hogy az entitáshatárok kijelölése részben helyi döntéseken múlik.

13 Jones, Ed i. m. p. 135–141.

14 Többek közt Senior, Alistair (2018): Bringing it all together: Mapping continuing resources vocabularies for linked data discovery. The Serials Librarian. https://doi.org/10.1080/0361526X.2018.1428463 ; Balster, Kevin – Rendall, Robert – Shrader, Tina (2018): Linked serial data: Mapping the CONSER standard record to BIBFRAME. Cataloging & Classification Quarterly, 56(2–3), 251–261. https://doi.org/10.1080/01639374.2017.1364316 ; El-Sherbini, Magda (2018): RDA implementation and the emergence of BIBFRAME. JLIS.it, 9(1). https://doi.org/10.4403/jlis.it-12443 [Elérés: 2026.05.17.]

15 Definition of PRESSOO : A conceptual model for Bibliographic Information Pertaining to Serials and Other Continuing Resources : Version 1.3. Ed by Patrick Le Bœuf. PRESSOO Review Group [...] IFLA Cataloguing Section, 2016 https://cidoc-crm.org/pressoo/fm_releases [Elérés: 2026.07.06.]

16 Bibliographic Framework Initiative / Library of Congress. [2011-2026] https://www.loc.gov/bibframe/ [Elérés: 2026.07.06.]

17 Overview of the BIBFRAME 2.0 Model. https://www.loc.gov/bibframe/docs/bibframe2-model.html [Elérés: 2026.07.06.]

18 A HUB egy olyan absztrakt entitás, mely az eredeti BIBFRAME-modellnek még nem volt része. Szerepe a különböző művek összegyűjtése, összekapcsolása Linked Data környezetben.

A részletesebb szabályok és kapcsolati elemek kidolgozása azonban még nagyrészt előttünk áll. A hazai gyakorlat számára különösen hasznos lehetne konkrét időszaki kiadványok példáin vizsgálni, hogy mely változások indokolnak új művet vagy új leírást, és miként kapcsolható össze egységes modellben a folyóirat, a lapszám és a cikk szintje. Az időszaki kiadványok pontosabb modellezése ezért nem lezárt kérdés, hanem a jövő katalogizálási és metaadat-fejlesztési munkájának egyik fontos területe.

## Irodalomjegyzék


Balster, Kevin – Rendall, Robert – Shrader, Tina (2018): Linked serial data: Mapping the CONSER standard record to BIBFRAME. Cataloging & Classification Quarterly, 56(2–3), 251–261. https://doi.org/10.1080/01639374.2017.1364316

A bibliográfiai tételek funkcionális követelményei: zárójelentés. Ford. Berke Barnabásné. Budapest, Országos Széchényi Könyvtár, 2006. Forrás: https://www.ifla.org/wp-content/uploads/2019/05/assets/cataloguing/frbr/frbr-hu.pdf

Bird, Nora J. (2025): Serials. In: Encyclopedia of Libraries, Librarianship, and Information Science. https://doi.org/10.1016/B978-0-323-95689-5.00052-3

Bross, Valerie – Hawkins, Les – Nguyen, Hien (2013): CONSER serial RDA workflow. The Serials Librarian, 64(1–4), 211–215. https://doi.org/10.1080/0361526X.2013.760412

Dezsényi Béla (1942): Sajtó és könyvészet. In: Magyar Könyvszemle, 66. évf. = 3. folyam, 2.sz., p. 126–152. Forrás: https://epa.oszk.hu/00000/00021/00209/pdf/

El-Sherbini, Magda (2018): RDA implementation and the emergence of BIBFRAME. JLIS.it, 9(1). https://doi.org/10.4403/jlis.it-12443

Hawkins, Les – Nguyen, Hien – Kharfen, Sarah (2015): Basic serials cataloging workshop. The Serials Librarian. https://doi.org/10.1080/0361526X.2015.1012473

Ilácsa Szabina (2024) Időszakosság a katalogizálásban, Central European Library and Information Science Review (CeLISR), 1(1). Forrás: https://doi.org/10.3311/celisr.37676

Jones, Ed (2013): Description of serials, RDA, and the MARC 21 bibliographic format. The Serials Librarian. https://doi.org/10.1080/0361526X.2013.836465

KSZ/3:2001 Bibliográfiai leírás. Időszaki kiadványok. Budapest : OSZK, 2001. 104 p.

Riva, Pat – Boeuf, Patrick Le – Žumer, Maja: IFLA könyvtári referenciamodell. A bibliográfiai információk elméleti modellje. Ford. Gazdag Tiborné. Budapest, Országos Széchényi Könyvtár, 2018. Forrás: https://www.ifla.org/wp-content/uploads/2019/05/assets/cataloguing/frbr-lrm/ifla-lrm-august-2017_rev201712-hu.pdf

Senior, Alistair (2018): Bringing it all together: Mapping continuing resources vocabularies for linked data discovery. The Serials Librarian. https://doi.org/10.1080/0361526X.2018.1428463

Stachokas, George (2020): The role of the electronic resources librarian. In: Reimagining Technical Services. https://doi.org/10.1016/B978-0-08-102925-1.00007-6



*1. kép A Kritika, ISSN 0324-7775 folyóirat [19]72/1. (február), illetve 30. évf. 1. sz. (2001. január) lapszámainak borítói (Forrás: Arcanum Újságok https://adt.arcanum.com/hu/collection/MTA_Kritika/)*


*2. kép A Kritika, ISSN 0023-4818, 1. évf. 1. sz. (1963. szeptember) lapszámainak borítója (Forrás: Arcanum Újságok https://adt.arcanum.com/hu/collection/MTA_Kritika/)*

DOI: 10.31915/NWS.2026.24



# A RAG RENDSZER LEHETŐSÉGEI A LEVÉLTÁRI KUTATÁSTÁMOGATÁSBAN

# THE APPLICABILITY OF RAG TECHNOLOGY IN THE HUNGARIAN NATIONAL ARCHIVES

*Varga Emese* (varga.emese@mnl.gov.hu)
*Havas Ádám* (havas@helion.hu)
*Bánki Zsolt* (bankizsolt@mnl.gov.hu)

**Absztrakt**

A mesterségesintelligencia-alapú technológiák egyre hangsúlyosabban jelennek meg a közgyűjteményi digitális szolgáltatásokban. A felhasználói igények már nem csupán a digitalizált források elérésére, hanem azok mélyebb, tartalmi feltárhatóságára, az eddigi kutatási módszertanokkal nehezen átlátható összefüggések felszínre hozatalára és célzott kereshetőségére irányulnak. Erre a kihívásra kínál új lehetőséget a RAG (Retrieval-Augmented Generation) technológián alapuló szolgáltatás.

Az alkalmazás egy felhasználói kérdésre levéltári források alapján szemantikus keresést futtat, majd egy nagy nyelvi modell (LLM) segítségével választ generál. A megbízható válasz előfeltétele az intézmény által biztosított jó minőségű forrás.

A tanulmány a Magyar Nemzeti Levéltár és a HELION Kft. együttműködésében fejlesztett alkalmazás levéltári integrálhatóságát mutatja be, tekintettel arra, hogy a RAG-alapú megoldás miként illeszkedik a meglévő online tartalomszolgáltatásokhoz. Fókuszba kerül a RAG fejlesztésének és levéltári implementálásának folyamata is.

E téren bemutatásra kerül, hogy a kétlépcsősen működő RAG-alkalmazás az első nyelvi modellhívás során megvizsgálja a kontextust, majd a felhasználói kérdésből kulcsszavakat és név entitásokat (NER) nyer ki, amelyek a dokumentumkeresés alapját képezik. FAISS (Facebook AI Similarity Search)-alapú szemantikus keresést kombinál Oracle adatbázis context indexre támaszkodó szűréssel. A keresés eredményeként kapott szövegrészek adják a második nyelvi modellhívás bemenetét. Mindkét híváshoz az OpenAI gpt-4.1 modelljét használja, a kérdésre adott végleges válasz streamelt formában érkezik.

**Kulcsszavak**: mesterséges intelligencia, RAG, szemantikus keresés, nagy nyelvi modellek, levéltári digitalizáció

## Abstract

Artificial intelligence-based technologies are becoming increasingly prominent in public collection digital services. User demands are no longer limited to accessing digitized resources, but also include deeper content exploration, the uncovering of connections that are difficult to discern using existing research methodologies, and targeted searchability. A service based on RAG (Retrieval-Augmented Generation) technology offers a new opportunity to meet this challenge.

The application runs a semantic search based on archival sources and then generates answers to user questions using a large language model (LLM). A prerequisite for reliable answers is the high quality of the sources provided by the institution.

The article will discuss the archival integration of the application developed in collaboration between the Hungarian National Archives and HELION Kft., with a focus on how the RAG-based solution fits into existing online content services. It will focus on the development of RAG, the process of its implementation in archives, and the experiences of its developers.

In this context, we will demonstrate that the two-stage RAG application examines the input during the first language model call, then extracts keywords and named entities (NER) from the user's question, which form the basis for document search. It combines FAISS (Facebook AI Similarity Search)-based semantic search with Oracle database context index-based filtering. The text fragments obtained as a result of the search provide the input for the second language model call. It uses the OpenAI gpt-4.1 model for both calls, and the final answer to the question is delivered in streamed form.

**Keywords:** artificial intelligence, RAG, semantic search, large language models, archival digitization

## Bevezetés

Napjainkban a közgyűjteményi digitális tartalomszolgáltatásokkal szemben már nem csupán a puszta elérhetőség és a nyílt hozzáférés az elvárás. A kombinált keresési metódusok – például a kulcsszavas és a fazettás keresés – mellett egyre nagyobb igény mutatkozik a szemantikus keresésre, az AI-alapú tartalomelemzési lehetőségekre és a kontextualizált válaszokat nyújtó rendszerekre is.

Ezek az AI-alapú eszközök a tudományos kutatás fontos eszközévé válnak, főként a szakirodalom-keresés és -elemzés, illetve a tartalomgenerálás terén, de a torzítások veszélye jelentős, emiatt a kutatóknak kritikus szemlélettel kell kezelniük az AI által generált ajánlásokat. Egy kutatástámogató szolgáltatásnál ezért fontos, hogy a válaszok ellenőrzött forrásokra épüljenek, és hivatkozásokkal legyenek alátámasztva. Erre kínál megoldást a RAG technológia.

A RAG (Retrieval Augmented Generation[1]) egy külső dokumentumgyűjteményből először visszakeresi a felhasználói kérdés szempontjából releváns információkat, majd egy nagy nyelvi modell (LLM[2]) segítségével választ generál a felhasználói kérdésekre. Nemzetközi példák mutatják, hogy a RAG rendszerek történeti dokumentumokra is alkalmazhatók. Különböző kísérletek vannak levéltári anyagokon és történeti újságcikkeken alapuló rendszerek kialakítására, illetve különböző típusú kulturális örökségi adatok RAG általi, egységes tudásbázisba való szervezhetőségére.

A technológia magyar levéltári használhatóságát az MNL és a Helion kft. által fejlesztett webalkalmazással demonstráljuk, amely Nagy Imre első kormányának minisztertanácsi jegyzőkönyveire[3] vonatkozó kérdések megválaszolását teszi lehetővé, és ezáltal egy új módszert kínál a levéltári anyagok kutatásához. Az alkalmazásba a tisztított, forráskiadványokhoz[4] készült PDF dokumentumokat használtuk. A technológia alkalmas arra, hogy nemcsak faktuális, hanem értelmező, definíciós, generatív kérdésekre is releváns választ adjon, így rámutathat összefüggésekre, lehetséges kapcsolatokra személyek, intézmények

1 A technológia működési elvét *visszakeresésre épülő válasz generálás*ként írhatjuk le magyarul a legpontosabban.

2 LLM: Large Language Model. Bővebben lásd: Naveed, Humza et.al.: A Comprehensive Overview of Large Language Models, *ACM Transactions on Intelligent Systems and Technology*, 2025/5, 1–72. https://doi.org/10.1145/3744746. [Megtekintve: 2026.07.08.]

3 Az eredeti forrásdokumentumok elérhetők az Adatbázisok Online felületén, lásd: Nagy Imre első kormánya 1953. július 4–1955. április 18., https://adatbazisokonline.mnl.gov.hu/adatbazis/minisztertanacsi-jegyzokonyvek-1944-1965/hierarchia. [Megtekintve: 2026.05.06.]

4 A forráskiadványok megjelenései: Nagy Imre első kormányának minisztertanácsi jegyzőkönyvei I. 1953. július 10. – 1954. január 15., szerk. Baráth Magdolna, Gecsényi Lajos, Nagy Imre Alapítvány, Magyar Nemzeti Levéltár, Budapest, 2018.; Nagy Imre első kormányának minisztertanácsi jegyzőkönyvei II. 1. rész, január 29. – 1954. július 2., szerk. Baráth Magdolna, Gecsényi Lajos, Nagy Imre Alapítvány, Magyar Nemzeti Levéltár, Budapest, 2022.; Nagy Imre első kormányának minisztertanácsi jegyzőkönyvei II. 2. rész, 1954. július 6. – 1954. október 26. Nagy Imre Alapítvány Magyar Nemzeti Levéltár, Budapest, 2022.

és események között, illetve időbeli és térbeli kontextusba helyezheti az információkat. Ugyanakkor nem helyettesíti a forráskritikát és a levéltári anyagok alapos kutatását, amelyek továbbra is a kutató feladatai maradnak.

## A RAG levéltári alkalmazhatósága

A levéltári anyagok AI-módszerekkel történő kutatását, vizsgálatát a tech cégek által kínált megvalósításokkal szemben olyan megoldással kell lehetővé tenni, amit az MNL képes kontrollálni. Az MNL igényeire szabott alkalmazás fejlesztése mellett sem magától értetődő, hogy annak miért kell a RAG technológiát használnia. Miért nem lehet a modellnek egyben átadni az egész dokumentumkorpuszt, és az alapján várni a válaszokat?

Napjaink nagy nyelvi modelljei elméletileg képesek hatalmas dokumentummennyiség fogadására egyetlen promptban, de ennek alkalmazása komoly problémákat vonna maga után:

- **Költség:** Az API[5]-k token[6]-alapú elszámolásúak. Sok millió token „beküldése" minden egyes kérdésnél rendkívül drága lenne, míg a RAG csak a releváns részeket (a visszakeresés eredményeit) küldi el, töredék áron.
- **Késleltetés:** Minél nagyobb a kontextus, annál lassabb a modell válaszideje. Például egy 1 millió tokenes kérés feldolgozása percekig is eltarthat (az alkalmazásunk által kezelt dokumentumok összes terjedelme kb. 7–8 millió token).
- **Figyelemvesztés:** Bár a modellek fejlődnek, még mindig megfigyelhető, hogy a hatalmas szövegek közepén elrejtett információkat kevésbé hatékonyan találják meg, mint a szöveg elején vagy végén lévőket.

A RAG működési elve egyszerű: a felhasználó által feltett kérdés alapján futtatni kell egy (vagy több) keresést, ezek eredményei alapján egy nagy nyelvi modell (LLM) választ ad. A megvalósítás több döntési pontot hozott:

- Milyen LLM-et használjunk? Helyben működőt vagy szolgáltatásként igénybevehetőt?
- Hogyan működjön a keresés?
- A több felhasználóval párhuzamosan futó kommunikációt az alkalmazás miképp különítse el úgy, hogy eközben mindenki számára megmaradjon a saját kontextusa?

5 API (Application Programming Interface, magyarul alkalmazásprogramozási felület)

6 A nagy nyelvi modellek nem szavakkal, hanem tokenekkel dolgoznak, amelyek szavak, szótöredékek vagy írásjelek lehetnek. Angol nyelven 1 token nagyjából 4 karakternek vagy 0,75 szónak felel meg.

A RAG technológia levéltári alkalmazhatósága szempontjából a válasz minőségét tartottuk a legfontosabbnak. A helyben működő LLM-mel szemben fontos érv volt a jelentős hardverköltség, valamint a hardver beszerzésének és üzembe helyezésének időigénye. Ezek alapján a szolgáltatásként igénybe vehető LLM mellett döntöttünk, ami API-hívások formájában használható. 2025-ben, a fejlesztés idején közzétett LLM összehasonlításokban nagyon jól szerepelt az OpenAI ***gpt-4.1*** modellje, ezért az alkalmazásba ezt építettük be.

## A forrásanyag előfeldolgozása

Kész megoldások egész sora[7] használható PDF fájlok RAG rendszerbe illesztésére. Ezek azonban csak a legegyszerűbb felépítésű PDF-eknél adnak elfogadható eredményt. Beépített algoritmusaik nem értelmezik a szövegeket és könnyen vezethetnek tartalmi szempontból rossz szerkezetű eredményhez.

A prototípusban használt közel 3000 oldal összterjedelmű PDF fájlokból úgy nyertük ki a szöveget, hogy annak tördelése, szerkezete és tartalma megfeleljen a szemantikus keresés és a nagy nyelvi modell számára. Az anyagot 10.487 szövegdarabra bontottuk: 380-tól 1300 karakterig terjedő méretűek, átlag 1113 karakteresek (kb. 400 token). Az átfedés kb. 150 karakter.

A szemantikus kereséshez mindegyik szövegdarabhoz egy-egy vektort rendeltünk. Igen magas (1536) dimenziószámú vektorokról van szó, amiket az OpenAI ***text-embedding-3-large***[8] modelljével állítottunk elő. Ezt a folyamatot szokás *beágyazás*nak nevezni, mert ezzel egy szöveget (szót, mondatot, dokumentumot) beemelünk egy matematikai térbe.

## A visszakeresés megvalósítása

A hasonló jelentésű szövegdarabok közel kerülnek egymáshoz ebben a matematikai térben, amire ráépíthető a szemantikus keresés: A felhasználói kérdés kulcsszavai és kifejezései alapján ugyanezzel a módszerrel vektort képezünk. Ezután megkeressük azokat a szövegdarabokat, amelyek vektorai közel állnak a kérdés vektoraihoz, vagyis tartalmilag a legrelevánsabbak. Az LLM ezekből a kiválasztott szövegrészekből generálja a választ. Az alkalmazásba a FAISS (Facebook AI Similarity Search[9]) keresőmotort építettük be, amelyet a keresés hatékonyságának és megbízhatóságának fokozása érdekében szemantikus kereséssel kombináltunk, az Oracle saját Text keresőmotorja segítségével.

---

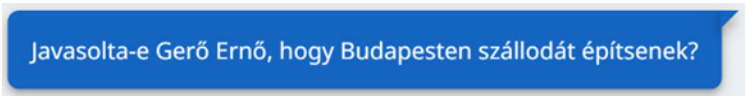
7 Például: pypdf, PyMuPDF, PyMuPDF4LLM, pdfminer.six, pdfplumber, pypdfium2, OCRmyPDF, docTR, Unstructured, Docling, stb.

8 Bővebben lásd: https://developers.openai.com/api/docs/models/text-embedding-3-large [Megtekintve: 2026.07.08.]

9 A FAISS (Facebook AI Similarity Search) a Meta által fejlesztett, nyílt forráskódú szoftverkönyvtár, amely lehetővé teszi nagy mennyiségű adathalmaz gyors átkutatását és hasonlósági keresését. Bővebben lásd: https://ai.meta.com/tools/faiss/ [Megtekintve: 2026.07.08.]

### Kulcsszavak kinyerése, NER (Named Entity Recognition)

A kombinált keresés hatékonyságához nagyon fontos, hogy a felhasználó által megfogalmazott kérdésből kiemeljük azokat a szavakat és különösen azokat a tulajdonneveket, amikre a keresést futtatni kell. A folyamat első lépése során (az 1. LLM hívás keretében) történik meg ez, ahogyan azt az alábbi egyszerű példa mutatja egy adott felhasználói kérdés esetében:

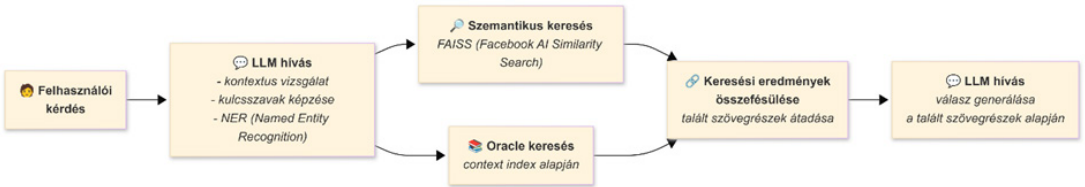



Az 1. LLM hívás eredménye (NE = named entity):

```
Kulcsszavak
├── 🧍 Gerő Ernő
│   └── Típus: személy (NE)
├── 📍 Budapest
│   └── Típus: hely (NE)
└── 🏨 szálloda építés
    └── Típus: egyéb (nem NE)
```

### A válasz generálásának folyamata

Az LLM-hívások folyamatát, a válasz generálásának logikáját az alábbi ábra mutatja be:




Fontos eleme a folyamatnak az LLM-hívások paraméterezése és a prompt, amit a felhasználói kérdéssel és a keresési eredményekkel egyidőben kap meg a modell.

### Előzmények (kontextus) figyelembevétele

Az alkalmazás minden egyes felhasználói munkamenethez (session-höz) két párhuzamos szálon követi az előzményeket.

1. A kontextusvizsgáló (és kulcsszókinyerő) LLM-hívás folyamatosan követi a saját előzményeit az alkalmazás használatának megkezdésétől a böngészőlap bezárásáig vagy kilépésig (illetve az F5 billentyű lenyomásáig).
2. A választ generáló LLM-hívás addig követi az előzményeket, amíg a kontextusvizsgálat alapján szükséges. Az előzményektől független, új kérdés hatására törli őket (ez természetesen nem azt jelenti, hogy a felületről eltűnnek).

Az előzmények figyelembevétele és szükség szerinti törlése két szempont miatt fontos:

- A párbeszéd folyamán a kérdések gyakran implicit módon hivatkoznak az előző kérdésre, ilyen esetekben fontos az előző kérdéshez kapott kulcsszavak továbbvitele.
- Ha viszont a kérdés újnak tekinthető, akkor éppen az a fontos, hogy előzmények nélkül, csak a szükséges kulcsszavak felhasználásával fusson le a keresés. Ennek költségvonzata is van, el kell kerülni a fölösleges tokenfelhasználást.

### A válaszok értékelése

A demó alkalmazás gyakorlati használhatóságát levéltáros és történész szakemberek bevonásával is teszteltük. A vizsgálat célja az volt, hogy különböző típusú tesztkérdések által kiderüljön, hogy a kutatók/felhasználók a chatbot-alapú szolgáltatást mennyire tekintik valódi segítségnek a kutatómunkájukban. Elégedettek-e a jelenlegi demóverzióval a publikus bevezetéshez? Relevánsak és pontosak-e a válaszok?

A tesztelők összesen 516 kérdést tettek fel a rendszernek, 1-től 5-ig adható pontszámmal értékelték a válaszokat, amelynek átlaga 4,3 lett.

Emellett az alábbi 7 kategória egyikébe is besoroltak minden egyes választ:

| **Válasz besorolása** | **Részarány (%)** |
|---|---|
| értékelhetetlen, hibás | 1,0 |
| téves, a forrásokban foglaltaknak nem felel meg | 1,6 |
| hiányos, nem tartalmazza a lényeget | 3,3 |
| hallucinációt tartalmaz | 1,6 |
| elfogadható, de nem elég tartalmas | 11,6 |
| jó, tartalmazza a lényeget | 42,8 |
| kitűnő, tartalmas és részletes | 38,2 |

Az eredmények arra utalnak, hogy a rendszer működése alapvetően megfelel a felhasználói elvárásoknak. A válaszok 81%-át jónak vagy kitűnőnek értékelték, a hallucinációk aránya alacsony volt, negatív értékelés főképp a válaszok hiányosságára vonatkozott.

A kifejezetten gyenge értékelések aránya mindössze 2,6% volt, ami arra utal, hogy a teljesen sikertelen válaszok ritkán fordultak elő.

A visszajelzések által kirajzolódtak a kutatói elvárások egy RAG-alapú szolgáltatással szemben. Fontos szempont volt a válasz strukturáltsága, pontossága, tartalmassága. A tesztelők pozitívan értékelték, ha a modell jelezte, hogy a források alapján nem tud teljeskörű választ adni a feltett kérdésre. Különösen fontos szempont volt még a forrásokhivatkozások pontossága, a forráskörnyezet láthatóvá tétele, és a teljes forrásanyag közvetlen elérhetősége a kutatók számára.

## A fejlesztés tapasztalatai

A fejlesztés bebizonyította, hogy a ma elérhető és használható RAG technológia alkalmas arra, hogy hatékonyan segítse a levéltári dokumentumok kutatását. Közel 3000 oldalnyi dokumentumból másodpercek alatt nyerhetünk ki olyan információkat, amelyekhez a hagyományos (digitális) keresési technikákkal hosszadalmas keresgélésre és a talált szövegrészek értelmezésére, feldolgozására lenne szükség.

Mindemellett fontos, hogy a felhasználók ismerjék e technika korlátait: nem várható mindig tökéletes és hiánytalan válasz. Bizonyos típusú kérdések eleve fel se tehetők a RAG technológia jellemzői miatt: pl. Hányszor szólalt fel Gerő Ernő? Ez a technika nem képes arra, hogy a teljes tartalmat végig vizsgálva megszámlálja a felszólalásokat, vagy bármilyen más szöveg előfordulásait.

A RAG technológia széleskörű használatbavételéhez a rendelkezésre álló dokumentumok előfeldolgozása jelentős munkaigénnyel fog járni, ugyanis a különböző PDF fájlok szerkezete teljesen heterogén, azokat nem úgy állították elő, hogy erre a célra módosítás nélkül alkalmasak legyenek. A lábjegyzetek – ha vannak – elkülönítése külön nehézséget jelent.

Fontos, hogy az alkalmazásba épített prompt tartalmán nagyon sok múlik. Kis módosítás gyakran jelentős következményeket von maga után, újra kell tesztelni sok-sok kérdéssel. A fejlesztés jelentős részben a modellparaméterek és a prompt variálásában merül ki, ezek kipróbálása, tesztelése időigényes. A fejlesztési folyamat nem mindig lineáris, néha inkább a „Trial and error" jellemzés illik rá. Az egyik nyelvi modellnél bevált paraméterek, promptok nem biztos, hogy jók egy másik modellváltozathoz is. Új modellre áttéréskor újra kell kezdeni a tesztelést, módosítani a promptot és a modellparaméterek beállítását.

Az alábbiakban összefoglaljuk mindazokat a szempontokat, amelyeket figyelembe kell venni egy RAG rendszer tervezésekor és fejlesztésekor. Jó eredményt csak akkor várhatunk, ha az alábbi feltételek mindegyike teljesül:

- A dokumentumkorpusz legyen megfelelően előfeldolgozott és feldarabolt. A szövegdarabok mérete és az átfedés nagysága is fontos (a teljesen automatizált, kész előfeldolgozó megoldások nem mindig adnak jó eredményt).

- Magas dimenziószámú vektorok generálása egy magyar nyelvet jól kezelő modellel.
- Gyors működést biztosító, lehetőleg kombinált (szemantikus és szöveges) keresés implementálása a tulajdonnevek kiemelt kezelésével (NER).
- Magyarul is kifogástalanul teljesítő, jó minőségű nagy nyelvi modell használata.
- Prompt kialakítása az alábbi szempontok alapján:
  - a használt nagy nyelvi modell tulajdonságai,
  - a dokumentumok tartalma,
  - a RAG rendszer célja.
- A modellparaméterek finomhangolása, amely a prompt fejlesztésével, finomításával és rengeteg teszteléssel egyszerre végzendő.

### Távlati célok

Az eddigi tapasztalatok és a kutatói visszajelzések alapján kirajzolódnak a fejlesztés további lépései. Elsőként a demóverzió élesítése a cél, egy jól definiált korszak iratainak közzétételével. Egyúttal folyamatosan újabb és újabb irategyütteseket teszünk a RAG-on keresztül kutathatóvá, hogy minél szélesebb körű tudásbázist alakítsunk ki. Jelen terveink alapján a minisztertanácsi iratokat középtávon az MTI[10] és az MSZMP[11] források követhetnék.

A különböző típusú források integrálása nemcsak a rendszer által feldolgozható információk mennyiségét növeli, hanem hozzájárul ahhoz is, hogy a felhasználók összetettebb történeti összefüggéseket tárhassanak fel. A forrásanyagok bővítésének egyik alapfeltétele a jó minőségű, OCR-rel[12] feldolgozott szövegállományok előállítása, hiszen a hibás vagy torzított szöveg rontja a visszakeresés pontosságát és a generált válaszok megbízhatóságát. Ehhez egyrészt szükség van a karakterfelismerési hibák csökkentésére, másrészt a szövegek strukturáltabb feldolgozására, például bekezdések, címsorok vagy egyéb tartalmi egységek pontosabb azonosítására.

Emellett fontos fejlesztési irány, hogy a rendszer által generált válaszokban szereplő hivatkozásokat közvetlenül az Adatbázisok Online[13] megfelelő rekordjaira mutató linkekkel kapcsoljuk össze. Ez nemcsak az átláthatóságot és az ellenőrizhetőséget növeli, hanem lehetőséget ad arra is, hogy a felhasználók a generált válaszokból kiindulva az eredeti forrásokban folytathassák a kutatást.

10 MTI (Magyar Távirati Iroda) forrásai, lásd: Adatbázisok Online: https://adatbazisokonline.mnl.gov.hu/adatbazis/magyar-tavirati-iroda. [Megtekintve: 2026.05.06.]

11 MSZMP (Magyar Szocialista Munkáspárt) jegyzőkönyvei.

12 Az OCR (Optikai Karakterfelismerés) technológia a képeken, beszkennelt dokumentumokon vagy PDF-fájlokban található gépelt, kézírott vagy nyomtatott szöveget géppel olvasható, szerkeszthető és kereshető szöveggé alakítja.

13 https://adatbazisokonline.mnl.gov.hu/

## Irodalomjegyzék


[1] Király Péter: Adat a könyvtárban. Hagyomány és újítás a 21. századi könyvtárban. Erdélyi évszázadok. A kolozsvári magyar történeti intézet évkönyve III. Bolyai Társaság – Kolozsvár, Egyetemi Műhely kiadó, 2018, 61.

[2] Hajdu Krisztián: A mesterséges intelligencia szerepe a kutatástámogatásban. Különleges Bánásmód Interdiszciplináris folyóirat, 2025/11,15. DOI https://doi.org/10.18458/KB.2025.2.7

[3] Lewis, Patrick et.al.: Retrieval Augmented Generation for Knowledge-Intensive NLP Tasks, ArXiv, 2025, 2.

[4] Lin, Claire et.al.: A Preliminary Study of RAG for Taiwanese Historical Archives. Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), 2025, 45–62.

[5] Tran, The Trung et.al.: Retrieval Augmented Generation for Historical Newspapers. JCDL '24: Proceedings of the 24th ACM/IEEE Joint Conference on Digital Libraries, 2025, 1–5. DOI: https://doi.org/10.1145/3677389.3702542

[6] Kelly, Paul – Schild, Jonathan – Jafari, Amir: FolkRAG: a retrieval-augmented generation system for cultural heritage materials. Neural Computing and Applications, 37., 2025/24, 20281–20297. DOI: https://doi.org/10.1007/s00521-025-11455-4

DOI: 10.31915/NWS.2026.25



# AZ ÖSSZETETT MEZŐK STRUKTURÁLT ADATAINAK SZÉTVÁLASZTÁSA AZ ÁLLAMI ANYAKÖNYVEKBEN

# SEPARATING STRUCTURED DATA IN THE COMPLEX FIELDS OF HUNGARIAN CIVIL REGISTERS

*Szűcs Kata Ágnes* szucs.kata.agnes@mnl.gov.hu
*Varga Emese* varga.emese@mnl.gov.hu
*Vadász Noémi* vadasz.noemi@mnl.gov.hu
*Záros Zsolt* zaros.zsolt@mnl.gov.hu
*Szatucsek Zoltán* szatucsek.zoltan@mnl.gov.hu

## Absztrakt

A kézírással készült adatok kinyerése szkennelt anyakönyvekből az összetett mezők miatt is kihívást jelent: egy mező többféle adattípust tartalmazhat, hiányozhat a fejléc és változhat a sorrend, miközben a HTR-kimenetek soralapúak. Tizenhét összetett mezőt azonosítottunk; a szabályalapú módszerek csak egyszerű esetekben működtek, ezért bevezettünk mezőre szabott LLM-megoldásokat prompt- és paraméterhangolással. Egyes modellek megbízható eredményeket adtak, másoknál adatvesztés, ID-kiesés, felesleges tokenek vagy instabilitás jelentkezett, amelyeket szigorú rendszerpromptokkal és prompt módosítással kezelünk. Reprezentatív tesztkorpuszt és mezőszintű mérőszámokat építettünk. Következtetés: nincs univerzális megoldás – néhány mező teljesen automatizálható, soknál maradnak bizonytalanságok; a legjobb eredmény a feladat részekre bontásából és a célzott LLM-ek szabályalapú ellenőrzésének kombinálásából adódott.

**Kulcsszavak:** állami anyakönyvek, kézírás-felismertetés (HTR), nagy nyelvi modellek (LLM), szemantikus szegmentálás, adatnormalizálás, Magyar Nemzeti Levéltár

## Abstract

Extracting handwritten data from scanned registry books is challenging due to complex fields that may contain multiple data types, missing headers, and variable ordering, while HTR outputs are row-based. We identified 17 complex field types; rule-based methods worked only for simple cases, so we introduced field-specific LLM solutions with prompt and parameter tuning. Some models were stable; others showed data loss, ID dropout, superfluous tokens or instability, which we have mitigated by strict system prompts and tuning. We built a representative test corpora and field-level metrics. Conclusion: there is

no universal solution – some fields can be fully automated, many remain ambiguous; best results arise from decomposing the task and combining targeted LLMs with rule-based validation.

**Keywords:** civil registers, handwritten text recognition (HTR), large language models (LLM), semantic segmentation, data normalization, National Archives of Hungary

## 1. Bevezetés

A Magyar Nemzeti Levéltár (MNL) egyik legjelentősebb digitális bölcsészeti vállalása az 1895 utáni állami anyakönyvi másolatok átfogó, automatizált és mesterséges intelligenciával támogatott feldolgozása. Az MNL őrizetében lévő, mintegy 22 millió oldalnyi születési, házassági és halálozási anyakönyv nem csupán hatalmas adatmennyiséget képvisel, hanem a történeti demográfiai kutatások megkerülhetetlen forrásbázisát is alkotja.

A projekt stratégiai célkitűzése a kézzel írt adatok kinyerése a strukturális információkkal együtt, lehetővé téve a rekordok névtérhez kapcsolását és az entitások összekapcsolását. Jelen tanulmány a nagy nyelvi modellek (Large Language Models, LLMs) alkalmazására tett kísérletek eredményeit foglalja össze az úgynevezett összetett mezők szemantikai szétválasztásában és tisztításában.

### 1.2 Technológiai környezet

A nagy nyelvi modellek (LLM) futtatása támasztotta számítástechnológiai igényeket az MNL más intézményekkel együttműködésben tudja kielégíteni. A számításigényes munkafolyamatokat a HUN-REN Wigner Kutatóközpont[1] által biztosított, dedikált felhőalapú kutatási infrastruktúra szolgálja ki. A tesztelési fázis során az alábbi hardveres és szoftveres beállítások álltak rendelkezésre:

- **Számítási teljesítmény:** 4 db NVIDIA A100 40GB GPU egység, amely elengedhetetlen a nagy paraméterszámú modellek futtatásához.
- **Adattárolás:** 520 GB SSD tárolókapacitás.
- **Biztonság:** Titkosított, távoli elérésű szerverkörnyezet, amely garantálja a levéltári adatok integritását a számítási folyamatok alatt.

## 2. Munkafolyamat és módszertan

A teljes anyakönyv-feldolgozási lánc kilenc egymásra épülő lépésben foglalható össze, amelyek a digitalizálástól a publikálásig kísérik végig a folyamatot. Ebben a folyamatban az összetett mezők esetében a nagy nyelvi modellek hidat képeznek a kézírás-felismertetés (Handwritten Text Recognition, HTR) nyers kimenete és a strukturált adatbázisrekordok között.

1 https://wignerdc.wigner.hu/home (Utolsó letöltés: 2026.04.07.)


**I. Előkészítési szakasz:**

1. **Képek előfeldolgozása:** Digitális tisztítás és orientáció javítása.
2. **Sablonkiválasztás:** A megfelelő oldalképformátum (layout) azonosítása.
3. **Szegmentálás:** A kitöltött rovatok geometriai elhatárolása a képen.

**II. HTR és utófeldolgozási szakasz**

4. **HTR (kézírás-felismerés):** Szöveghipotézisek generálása.
5. **Adatbázisba töltés:** Ebben a fázisban kerülnek rögzítésre a nyers HTR-adatok a későbbi szemantikai tisztítás előtt.
6. **Adatnormalizálás és javítás:**

   6.1 Szabályalapú utófeldolgozási eljárások: a) hipotézis kiválasztás a legmagasabb konfidencia érték alapján, b) normalizálás (pl. leány → nő), c) javítás (pl. rom. katólikus → római katolikus)

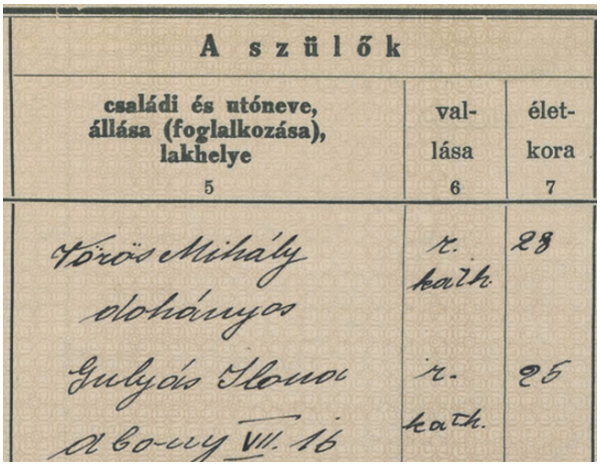
   *6.2 LLM-alapú adatszétválasztás és helyesírási korrekció.*

**III. Entitásösszekapcsolási és -publikálási szakasz**

7. **Adatok névtérhez kapcsolása:** Névtér azonosítók hozzárendelése a földrajzi nevekhez.
8. **Személyek összekapcsolása:** Rekord-összekapcsolás a későbbi családfa-rekonstrukcióhoz.
9. **Publikálás:** Az adatok indexelése és kereshetővé tétele a végfelhasználók számára.

### 2.1 Az összetett mezők kihívásai

Az összetett mezők kezelésének kérdése a teljes anyakönyv-feldolgozás komplex munkafolyamatának része. A „sima" és az „összetett" mezőket a cellában megjelenő adatok mennyisége mentén különböztetjük meg. Abban az esetben, amikor az anyakönyvvezetőnek több különböző típusú információt kellett beírnia egy-egy cellába, amelyek egy vagy több személyre is vonatkozhattak, összetett mezőkről beszélhetünk (pl. a szülők neve, állása, lakhelye vagy a menyasszony neve, foglalkozása, vallása, életkora és lakhelye).



| A szülők | | |
|---|---|---|
| családi és utóneve, állása (foglalkozása), lakhelye<br>5 | vallása<br>6 | életkora<br>7 |
| Kiss Mihály<br>dohányos<br>Gulyás Ilona<br>Arad-ucy VII. 16 | r. kath.<br>r. kath. | 28<br>25 |

*1. ábra: Szülők neve, állása, lakhelye – egy összetett mező adatai*

Míg az egyszerű – de akár egyes összetett mezők esetében is (pl. 1. ábra szülők vallása, szülők életkora) – szabályalapú utófeldolgozás is elegendő, a komplex eseteket nem tudjuk pusztán szabályalapon kezelni. A korábban külön lépésként lefutó hipotéziskiválasztás, adatbázisba töltés, illetve a normalizálás és szemantikai szétválasztás ezeknél egyetlen fázisban történik, még összetettebbé téve a feladatot.

Az összetett mezők (pl. „anyja neve, állása, lakhelye") feldolgozása az alábbi problémák miatt igényel fejlett nyelvi modelleket:

- **Határolójelek hiánya:** A különböző adatok (pl. személynév, foglalkozás) gyakran írásjelek nélkül követik egymást.
- **Szemantikai szegmentálás:** Az entitások jelentés szerinti elkülönítése az azonosításhoz (pl. szabó és Szabó).
- **Fejléc-következetlenség:** A cella tartalma gyakran eltér a fejlécben előírt adatoktól (pl. életkor helyett születési dátum megadása, családi állapot feltüntetése).
- **Szemantikai szegmentálás és entitássorrend inkonzisztenciája:** A kitöltő személy tetszőleges sorrendben rögzíthette az adatokat (pl. lakhely megelőzi a nevet, a családi állapot kerülhet a név elejére és a végére is, bizonyos adatokat kihagy, illetve nem ír le kétszer vagy épp megdupláz).

### 2.2 LLM-alapú feldolgozás: metodika

A kutatás során a modellek tesztelése három fázisban történt: (1) egyetlen komplex mezőn több modell mérése a „best-of" modell kiválasztásához; (2) a nyertes modell tesztelése több mezőtípuson; (3) modell-specifikus tesztelés dedikált mezőcsoportokon.

A modelleket finomhangolás nélkül és a reprodukálhatóság és visszamérhetőség érdekében rögzített paraméterekkel alkalmaztuk: seed=42 és hőmérséklet= 0,4.[2] Az első fázisban vizsgált modellek technikai jellemzőit az 1. táblázat foglalja össze.

Készítettünk egy benchmark korpuszt a tesztelt összetett mezőkre, amelyhez kétfajta átírást állítottunk elő, hogy kiderüljön, a modell mennyire távolodik el a hipotézisektől a normalizálás felé.

| Modell | Méret | Kontextus |
|---|---|---|
| **llama3.3:latest** | 43GB | 128K |
| **gemma3:27b** | 17GB | 128K |
| **qwen3:235b** | 142GB | 256K |
| **gpt-oss:120b** | 65GB | 128K |

*1. táblázat: Az anyja neve, állása, lakhelye mezőn tesztelt modellek*

A mérések kiterjedtek a **betűhív és szöveghű átírásra,** valamint a **sima** és a **transzformált** értékek számítására is. Utóbbi során az alábbi szabályokat érvényesítettük: kisbetűs konverzió, írásjelek és felesleges szóközök eltávolítása, valamint regex-alapú normalizálás (pl. cz → c mapping, kivéve a személynevekben, illetve a rövidítések feloldása: róm. kath. → római katolikus).

Az adatfeldolgozás során az LLM megkapja bementként a nyers HTR-hipotéziseket, majd ebből strukturált JSON formátumú kimenetet generál, melyben az összetett mezőket kulcs-érték párokra bontja a modell (pl. anya_nev: ”adat”, anya_foglalkozas: ”adat”, anya_lakhely: ”adat”).

2 Finomhangolás nélkül a modell csak a gyárilag előre betanított súlyokat használja; nem alkalmaztunk további feladatspecifikus finomhangolást (fine-tuning). A viselkedése így általánosabb, kisebb a pontosság, azonban az esély is a túlillesztésre. A seed bármely tetszőleges egész szám lehet; ugyanazzal a seed értékkel (és azonos környezettel) ugyanazt a pseudovéletlen sorozatot és így azonos generált kimenetet kapjuk. Tehát biztosítja a reprodukálhatóságot a generálás során előforduló véletlenszerű folyamatokhoz. A temperature (hőmérséklet) egy olyan kapcsoló, amely modell kockázatvállalását szabályozza a szöveggenerálás során. Az értéke 1 alatt kicsinek számít, a modell gyakran a legvalószínűbb szavakat választja, melynek következtében kevésbé változatos, stabil, „konzervatív” szöveget generál. 1-re állítva az alapviselkedése szerint működik, 1-nél magasabb érték esetén pedig több szó kap hasonló esélyt, tehát változatosabb, kreatívabb, de néha kevésbé koherens szöveget eredményezhet.

## 3. Eredmények, kiértékelés

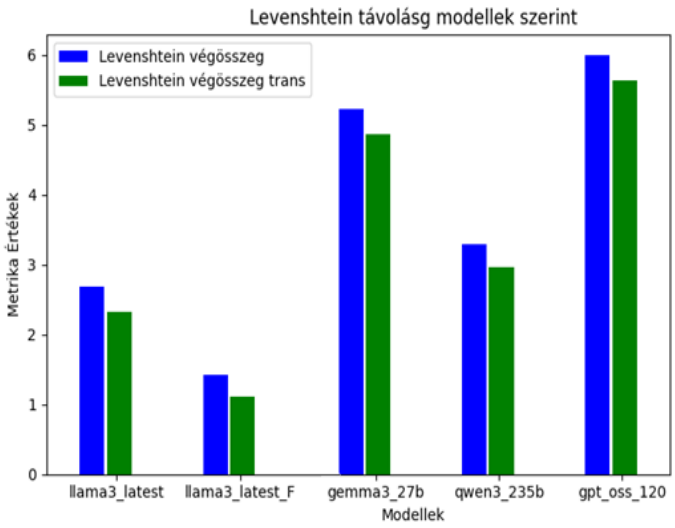
A teljesítmény értékelése mezőnkként egy 300 mintás véletlenszerű benchmark korpuszon alapult, Levenshtein-távolság[3] mérésével.

(1) A tesztelés első fázisában az anyja neve, állása, lakhelye mezőn tesztelt modellek közül – egy prompt finomhangolást követően (F) – a llama3.3 került ki legjobbként, átlagosan 1,4 értékű karaktereltéréssel. Az átlag Levenshtein-távolságok összege 2,7-ről 1,4-re csökkent, míg a transz értékek esetében 2,3-ról 1,1-re.

A 2. ábráról leolvasható, hogy a transzformált értékek minden esetben ugyanolyan mértékben alacsonyabbak, ami igazolja, hogy a normalizálási szabályok sikeresen távolítják el a stilisztikai zajt, amely nem tartalmi hiba.

Az eredmények további pontosítása végett a távolságmetrika mellett azt is vizsgáltuk, hogy mennyi volt a pontos egyezések, a hallucinációk és a kihagyások száma az összes értékhez viszonyítva. Ezeket a mutatókat mezőnkként generáljuk egy script segítségével.

Az anyja állása, lakhelye, neve mezőnél azt látjuk, hogy a 906 minta esetében 36 eltérés mutatkozik az elvárt kimenet és a modell által generált eredmény között. Ebből 6 esetben (0,6%) történt **hallucináció**, amikor a mérési alap (GT) üres volt, de az LLM adatot generált. Ezzel szemben 30 esetben (3,3%) fordult elő **kihagyás** (omission), ahol a modell nem adott választ a létező mérési alap mellett.




*2. ábra: Levenshtein-távolságok betűhív átírás esetén az anyja neve, állása, lakhelye mező esetében.*

3 A nyelvészetben és az informatika területén a Levenshtein-távolság egy karakterlánc-metrika, amely két sorozat közötti különbséget méri. Két szó közötti Levenshtein-távolság az a legkevesebb egykarakteres módosítás (beillesztés, törlés vagy csere), amely ahhoz szükséges, hogy az egyik szót a másikká alakítsuk.

Az eddigi pozitív eredmények mellett azonban figyelembe kell venni, hogy a pontos egyezések aránya (betűhív transzértékkel számolva) 39%. Ezen minták esetében tehát nincs szükség további beavatkozásra. Ehhez még a 3 alatti Levenshtein-távolsággal rendelkező mintákat is hozzátehetjük, ami további 26% -kal növeli az egyezések arányát, amely így 65%-ra egészül ki. A kimenetek 35%-a, tehát 4 és afölötti Levenshtein-távolságra van a benchmark korpuszban lejegyzettektől.

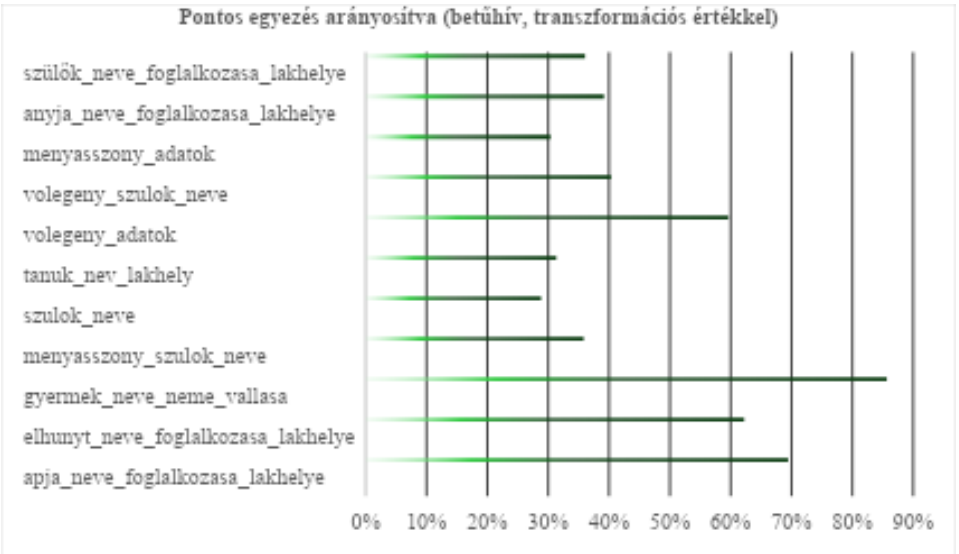
Ebből is látható, hogy egy-egy eredmény értelmezése nem annyira egyértelmű és sok számolást és gondolkodást igényel, amelyre azonban szükség van, ha a jövőbeni munkát és költségeket szeretnénk felmérni.

(2) A tesztelés második fázisban a többi mezőn is leteszteltük a llama 3.3 modellt. Ahogy az a 3. ábrán is látható, a pontos egyezések aránya széles kilengést mutatott.

Az eredmények tüzetes átvizsgálása során kiderült, hogy egyes mezőstruktúrák és adattípusok esetén sokat ront a modell, míg más esetekben könnyebben oldotta meg a feladatot.

A hibatípusok között a kihagyás több okból is előfordult. Az LLM rosszul kezelte a nagyon hosszú sorokat, egy név esetében – például, „néhai özvegy Bankó Jánosné született Halmos Sarolt," – a született utáni részt már lehagyta.

Előfordul olyan is, hogy az LLM rosszul szegmentálja a sztringeket – például egy hibás HTR-kimenet értelmezése esetén[4] – és nem ismeri fel, hogy az adott szövegrész nem a foglalkozáshoz, hanem a lakhelyhez tartozik. Ebből kétféle hiba is eredhet: egy kihagyás, és egy rosszul felcímkézett adat.




*3. ábra: Llama3.3 teljesítménye az összetett mezőkön*

4 Mivel a pipeline lépései egymásra épülnek, ezért minden korábbi folyamat teljesítménye kihatással van a végső eredményre.

Figyelembe véve az intézmény célkitűzését, egy anyakönyvi adatbázisban a legrosszabb hibatípus a hallucináció, tehát az, amikor LLM kitalál egy adatot ahelyett, hogy a megadott bemeneti hipotézisekből dolgozna.

Kiszámíthatatlanul működik a modell a helyesírási hibák javítása és az írott formák normalizálása esetében. Az alábbi prompt (utasítás)[5] mellett csak bizonyos esetekben modernizálta a helyesírást, több alkalommal őrizte meg az eredeti cz-s formákat.

(3) A harmadik fázisban csoportokba osztottuk a mezőket az alapján, hogy milyen típusú adatok fordulhatnak elő bennük (példák ld. Függelék). Kiemelt figyelmet kaptak a három különböző típusú adatot magukban foglalók, például az „anyja/apja/elhunyt neve, foglalkozása, lakhelye", illetve a több különböző adatot tartalmazók, mint a „menyasszony/vőlegény adatai", valamint a „tanúk neve és lakhelye" típusú komplex mezők. Ezekben ugyanis magas az előforduló személynevek aránya, amelyek még a lakcímben is előfordulhatnak.

Ebben a fázisban újabb modellekkel kísérleteztünk, annak reményében, hogy jobb eredményeket kapunk az összetettebb, illetve vegyes adatokat tartalmazó mezőkkel.

A qwen3.5:122b-a10b[6] indításakor olyan hibákba ütköztünk, amelyek hosszútávon megakadályozták a modellel való munkát: futtatás közben leállhat, vagy kihagyja a képekhez tartozó ID-kat, amelyek egy későbbi adatbázisba visszatöltés esetén elengedhetetlenek.

A qwen3-235b:latest[7] esetében problémát jelentett a folyamatosan jelenlévő „gondolkodás"[8], amelyre sem a think=False és a /set nothink kapcsolók, sem a system prompt módosítása sem volt hatással. Ez a gondolkodó viselkedés hosszútávon költség- és időigényes, a plusz generált tokenek miatt.

A system prompt egy speciális kezdeti utasítás vagy kontextus, amelyet a futtatási környezet rögzít a modellnek. Célja, hogy keretezze a modell viselkedését és szerepét a teljes beszélgetés során. Meghatározza a modell hangját, stílusát, tiltott/engedélyezett viselkedést és magas szintű szabályokat esetleg szerepmeghatározásokat rögzít. A modell számára látható, de általában a felhasználó nem látja közvetlenül. Az általunk használt utasítás a következő volt: „You are a specialized JSON generator. Output ONLY raw JSON. No conversational text, no markdown code blocks, no explanations."

Mivel nem találtunk olyan beállítást, amellyel sikerült volna lekorlátozni ezt a viselkedést ezért a feladat összetettségét csökkentettük. A kezdeti prompt szöveget úgy módosítottuk, hogy az eredeti elképzeléshez képest ne válasszon ki legjobb hipotézist, hanem az összes sort címkézze annak megfelelően, hogy milyen adatot tartalmaz. Például:

---

5 Végül javítsd az elgépeléseket a címkékben. A keresztneveket bátrabban, a vezetékneveket kevésbé. [prompt részlet]

6 https://ollama.com/library/qwen3.5:122b-a10b (Utolsó letöltés 2026.05.08)

7 https://ollama.com/library/qwen3:235b (Utolsó letöltés 2026.05.08)

8 Internális gondolat/elemzés tageket generál és ír ki, mielőtt a feladatot ellátná.

```
A feladatismertetés után kötegelten küldök mezőket. Minden mező
egy többsoros mező a szülők adataival.
Címkézd a mezők tartalmát.

A címkézés szabályai:
1. tanu_1_nev, tanu_1_lakhely, tanu_2_nev, tanu_2_lakhely címke
lehet.
2. nem biztos, hogy minden címkéhez van adat.
3. a sorokban lévő listák a sor különböző hipotézisei. Minden
sor összes hipotézistét tartsd meg a címkében, az eredeti for-
mában listázva, pl. „1. sor”: [„MátÉ”, „Máté”, „Mátén”, „Má-
tés”, „Mátó”], „2. sor”: [...]
4. a mezőben egy címke egy vagy több egymást követő sorból áll
össze. Az összetartozó sorokat a hipotézisekkel együtt tedd egy
címkébe.
5. ha előfordul a listában a „-” és a „¬” karakter, az elvá-
lasztást jelent, jelezve, hogy az adott sorban lévő hipotézisek
összetartoznak a következő sorban lévőkkel.
6. a személynevek elemei között logikai kapcsolat van: tanu1_
nev: első vezetéknév + első keresztnév, tanu2_nev: második ve-
zetéknév + második keresztnév.
6.1. A vezetéknevek előtt szerepelhetnek özv., özvegy, néh.,
néhai előtagok. Ezeket a névvel összefűzve őrizd meg minden
esetben.
7. A nem tipikus lakhelymegjelöléseket is vedd figyelembe, pl.
településnév + házszám vagy, településnév + kerületszám.
7.1. Ha a lakhelyet követő sorban számszerű adat van, akkor az
a hely adatokhoz tartozik.
7.2. A kerületszám római számmal van jelölve, pl. VII, III, V.

Válaszodat a JSON adatmezőjébe helyezd ilyen formában (és kizá-
rólag a JSON-el válaszolj, szöveges összefoglalást vagy egyéb
hozzáfűzést ne tegyél):

[     HU_MNL_...|...: [
      {
  „tanu_1_nev”: „sor: [hipotézisek]”,
  „tanu_1_lakhely”: „sor: [hipotézisek]”,
  „tanu_2_nev”: „sor: [hipotézisek]”,
  „tanu_2_lakhely”: „sor: [hipotézisek]”
}],
      ...
]
```
[9]

A kimenettel kapcsolatos problémákat nem szüntette meg teljes mértékben, azonban ezzel a megfontolással kezdtünk bele a qwen3:235b-instruct[10] utasításkövető verziójának tesztelésébe.

9 Vö.: Az első prompt variációban elvárt kimenet: {‚tanu_1_nev': ‚adat', ‚tanu_1_lakhely': ‚adat', ‚tanu_2_nev': ‚adat', ‚tanu_2_lakhely': ‚adat'}

10 https://ollama.com/library/qwen3:235b-instruct (Utolsó letöltés 2026.05.08)

Nem jelentkezett kihagyás sem az ID-kat, sem a batch egyes példáit illetően. Jellemző hiba maradt azonban a címkék keverése, ilyenkor például tévesen a lakhelyhez került a név. Ennek áthidalására egy konfigurációs Modelfile-ban megadtuk a futtatáshoz lefoglalt GPU-k számát (160) és a kontextusablak nagyságát tokenben (4096), amely a futtatási környezetben ideális GPU-felhasználást eredményezett. Ezen kívül a hőmérsékletet 0,1-re állítva, egyszerűsített system prompttal („Answer immediately and directly. Do NOT use <thought> tags. Start your response with the final answer.") futtatva a modell jól működött a címkézési feladatra és nem produkálta a korábbi modellek jellemző hibáit. A tanúk neve, lakhelye mező 182 példáján 84%-ban százalékban jól teljesített. Hibás címkézés olyan esetekben fordult elő, amikor csak egy lakhely volt megjelölve, vagy az utca neve egy személynév.

További javulást sikerült elérni a promptban meghatározott kimeneti példák egyértelműsítésével. A tanúk neve, lakhelye mező esetében pontosabb eredmények születtek:

```
{
„tanu_1_nev": {
    „1. sor": [„hipotézis"],
    „2. sor": [„hipotézis", „hipotézis"]
  },
„tanu_1_lakhely": {
    „3. sor": [„hipotézis"],
    „4. sor": [„hipotézis"]
  },
„tanu_2_nev": {
    „5. sor": [„hipotézis"],
    „6. sor": [„hipotézis", „hipotézis"]
  },
„tanu_2_lakhely": {
    „7. sor": [„hipotézis"],
    „8. sor": [„hipotézis"]
  }
}
```

## 4. Következtetések

Fontos megjegyezni, hogy ez egy folyamatban lévő kutatási és fejlesztési projekt, jelen tanulmány pedig az eddigi eredményeket és tapasztalatokat foglalja össze. Elmondható, hogy a modellek teljesítménye mezőtípusonként eltérő, ezért a promptok módosítása és a feladat egyszerűsítése (pl. csak szegmentálás kérése) elengedhetetlen a pontosság növeléséhez.

Vannak mezők, amelyeken a jelenlegi beállításokkal már jó eredményt érünk el, de egyik modell sem képes megbízhatóan kezelni az összetettebb szerkezetű mezőket, és könnyen lehet, hogy nem találunk egységes modellalapú megoldást minden mezőre.

Végső soron sikeresnek bizonyult a megközelítés, amely a feladatot részekre bontotta és egy lépésben kevesebbet várt el az LLM-től. Ez azonban további fejlesztést igényel, mivel modellek használatát szabályalapú eljárásokkal kell ötvözni a fennmaradó feladatok ellátására.

Bár egyes mezőkön már most is kiválóan teljesítenek a nagy nyelvi modellek, az összetettebb struktúrák továbbra is kihívást jelentenek, és elképzelhető, hogy nem találunk univerzális megoldást minden mezőre; az eddig tesztek során nyert tapasztalatok – a think=False hatékonysága a „gondolkodás" visszaszorításában, a rendszerpromptok és in-context példák szerepe az ellentmondások csökkentésében, valamint a qwen3:235b-instruct relatív stabilitása – azt mutatják, hogy a gyakorlatban a legjobb eredmény a feladat finom részekre bontásából és a nyelvi modellek célzott prompt- és paraméter-hangolásával kombinált szabályalapú ellenőrzésekből adódik.

## 5. Függelék

### 5.1 Anyja/apja/elhunyt neve, foglalkozása, lakhelye

```
JSON:
{
„HU_MNL_PVL_XXXIII_0001_a_0001_c_0005_0268|1”: [
    {
      „1. sor”: [
        „Tóth Imréné született Zeszenezki Te-”,
        „Tóth Imréné született Zeszenczki Te-”,
        „Tóth Imréné született Leszenczki Te-”,
        „Tóth Imréné született Zeszenczki Te¬”
      ]
    },
    {
      „2. sor”: [
        „véz, férje lakásn”,
        „véz, férje lakása”,
        „néz, férje lakása”,
        „vér, férje lakása”,
        „vér, férjé lakása”
      ]
    }
  ]
}
```

### 5.2 Menyasszony/vőlegény adatai

```
JSON:
{

  „HU_MNL_PVL_XXXIII_0001_a_0001_b_0006_0042|2”: [
    {
      „1. sor”: [
        „Klein Regina”
      ]
    },
    {
      „2. sor”: [
        „izraeltita”,
        „izraeltíta”
      ]
    },
    {
      „3. sor”: [
        „szül. 1884. június 2.”
      ]
    },
    {
      „4. sor”: [
        „lak. Abony I”,
        „loh. Abony I”
      ]
    },
    {
      „5. sor”: [
        „nelei u. 121.”
      ]
    }
  ]
}
```

### 5.3 Tanúk neve és lakhelye

```
JSON:
{
  „HU_MNL_PVL_XXXIII_0001_a_0001_b_0006_0042|2”: [
    {
      „1. sor”: [
        „Kagrenki”,
        „Kagvenki”,
        „Kugremki”,
        „Kugrenki”
      ]
    },
```

```
    {
      „2. sor”: [
        „Ligost”,
        „Ligosát”,
        „Lijosb”,
        „Lijosh”,
        „Lispát”
      ]
    },
    {
      „3. sor”: [
        „Abony II”
      ]
    },
    {
      „4. sor”: [
        „Andrássy u.”,
        „Andróssy u.”
      ]
    },
    {
      „5. sor”: [
        „Schwarcz”
      ]
    },
    {
      „6. sor”: [
        „Mór”
      ]
    },
    {
      „7. sor”: [
        „Abony II”
      ]
    },
    {
      „8. sor”: [
        „Andrincey úr.”,
        „Androssy u.”,
        „Androssy úr.”,
        „Andrászág u.”,
        „Andrásány u.”
      ]
    },
    {
      „9. sor”: [
        „23.”
      ]
    }
  ]
}
```

DOI: 10.31915/NWS.2026.26



# A KÖNYVTÁRI AUTOMATIZÁLÁS TÖRTÉNETE AZ MTA KÖNYVTÁR ÉS INFORMÁCIÓS KÖZPONTBAN: AZ INTEGRÁLT KÖNYVTÁRI RENDSZERTŐL A DISCOVERY-N ÁT A FELHŐALAPÚ SZOLGÁLTATÁSIG

# THE HISTORY OF LIBRARY AUTOMATION AT THE LIBRARY AND INFORMATION CENTRE OF THE HUNGARIAN ACADEMY OF SCIENCES: FROM THE INTEGRATED LIBRARY SYSTEM TO DISCOVERY AND CLOUD-BASED SERVICES

*Haász Antal*
*MTA KIK Szakinformatikai Osztály*
*haasz.antal@konyvtar.mta.hu*

**Absztrakt**

Az MTA Könyvtár és Információs Központban (MTA KIK) 1992-ben kezdődött a könyvtári gépesítés, az ALEPH integrált könyvtári rendszer bevezetésével.

Idővel a nem hagyományos dokumentumok kezeléséhez elektronikus forráskezelő- (SFX, 2003), digitális gyűjteménykezelő (ADAM, 2010), illetve repozitóriumi (REAL, 2009) szoftverekre volt szükség. Valamennyi erőforrástípus közös keresését tette lehetővé a Primo discovery-rendszer 2015-től. 2024-ben pedig sor került az egységes feldolgozást is biztosító könyvtári szolgáltatási platform, az ALMA bevezetésére.

Tanulmányomban ismertetem a korábbi rendszer-modell problematikáit, az új szoftver kiválasztásának szempontjait, valamint annak implementálását. Bemutatom, hogy az ALMA hogyan biztosítja a fizikai, elektronikus és digitális dokumentumok egységes forráskezelését. Végül szót ejtek a katalogizálást és forráskeresést támogató AI-alapú funkciók működéséről is.

**Kulcsszavak**: egységes könyvtári forráskezelés, könyvtári szolgáltatási platform (ALMA), discovery szolgáltatás (Primo), AI-technológia

## Abstract

Library automation at the Library and Information Centre of the Hungarian Academy of Sciences began in 1992 with the introduction of the ALEPH integrated library system.

Over time, the management of non-traditional materials required the implementation of an electronic resource management system (SFX, 2003), a digital asset management system (ADAM, 2010), and repository software (REAL, EPrints, 2009). The Primo discovery system was introduced to access all resource types in a single interface (2015). In 2024 a key strategic objective was realized with the acquisition of ALMA platform providing unified resource management.

This study discusses the challenges of the former system landscape, the key criteria of the software selection process, as well as the implementation of the new platform. It also provides an overview of how ALMA supports unified resource management for physical, electronic, and digital materials. Finally, it presents the AI-based functionalities supporting cataloguing and resource discovery.

**Keywords**: unified resource management, Library Services Platform (ALMA), Discovery Service (Primo), AI technologies

## Bevezetés

A 2000-es évek elejétől az elektronikus és digitális források iránti egyre erősödő felhasználói igények kielégítése céljából a magyarországi könyvtárak is megkezdték saját állományuk és szolgáltatási portfóliójuk ezirányú bővítését.

Azonban korán kiderült, hogy ezeknek az új típusú dokumentumoknak a menedzselése csak igen korlátozott módon valósítható meg az addig használt és a hagyományos (print) dokumentumok kezelésére optimalizált integrált könyvtári rendszerek (IKR) keretei között. A hatékonyság növelése érdekében szükségessé vált – az IKR-ek mellett – elektronikus forrás- és digitális gyűjteménykezelő rendszerek bevezetése a könyvtárakban. Viszont ezek az eszközök csak néhány, nagyobb költségvetéssel rendelkező könyvtár számára voltak elérhetőek – és nemcsak azok ára miatt, hiszen a használathoz szükséges szakértelmű könyvtárosok sem álltak mindenhol rendelkezésre. Ráadásul hamar szembesült a szakma a párhuzamos használatból adódó redundanciával és kompatibilitási problémákkal is.

A könyvtári szoftverek nemzetközi piacán a 2010-es évek első felében váltak elérhetővé olyan újgenerációs könyvtári rendszerek – más néven: *könyvtári szolgáltatási platformok* (*KSZP*)[1] –, amelyek lehetővé tették a hagyományos és nem hagyományos dokumentumok egységes keretrendszerben történő kezelését.

1 Szokás még könyvtári menedzsment rendszereknek (angolul Library Management System) is nevezni ezeket a szoftvereket. Azonban Dancs (2015) hiánypótló tanulmánya nyomán a magyar szakirodalomban inkább a *könyvtári szolgáltatási platform* változat terjedt el, ezért én is ezt használom.

Mindez felcsillantotta a reményt a magyarországi könyvtári szakemberek körében is, hogy egy ilyen rendszer bevezetése a felhasználói igények magasabb szintű kiszolgálása mellett a belső munkafolyamatok költség- és humánerőforrás hatékonyságának optimalizálását is lehetővé teszi majd.

Kijelenthető, hogy sajnos a folyamatok lassan zajlanak hazánkban: mostanáig csak néhány intézménynek sikerült megtalálnia a lehetőségeket a váltásra, és a könyvtárak döntő többsége továbbra is csak a tervezgetés fázisában van.

## Az átállás előzményei az MTA Könyvtár és Információs Központban

Az MTA KIK-ben a törzsállomány, illetve a különgyűjtemények[2] papír alapú állományának gépi feldolgozása 1992-ben kezdődött az ALEPH integrált könyvtári rendszer üzembe állításával. Az új felhasználói igényekre reagálva elindult az elektronikus és digitális források beszerzése, majd ezek kezelésére elektronikus forráskezelő (SFX, 2003), közös adatbáziskereső (MetaLib, 2003) majd később digitális gyűjteménykezelő (ADAM, 2010) szoftverek kerültek bevezetésre. Mindeközben létrejött a Könyvtár repozitóriuma (REAL, 2009) is.

Ebben a „több-rendszeres" modellben biztosítani lehetett az új típusú források menedzselését is, de a bevezetőben említett redundancia több területen is éreztette hatását. 2015-től a Primo discovery bevezetésével az egységes visszakeresés már ugyan biztosított volt, azonban a feldolgozási munkafolyamatok továbbra is „szétszabdaltan" folytak.

A könyvtár vezetése már 2016-ban kiemelt célkitűzéssé tette[3] egy könyvtári szolgáltatási platform bevezetését, de ehhez csak 2024-ben álltak rendelkezésre a feltételek.

A szoftver közbeszerzési eljárás keretében került kiválasztásra. A benyújtott műszaki dokumentáció legfontosabb pontjai a következők voltak: az új szoftver biztosítsa az egységes forráskezelést, a működése felhőtechnológián alapuljon, integráns része legyen egy globális tudásbázis és központi index, illeszthető legyen más, szabványosan működő rendszerekhez.

Az eljárás eredményeként az Ex Libris – Part of Clarivate cég *ALMA*[4] nevű szoftverét vásárolta meg az MTA KIK 2024. decemberében.

## Az ALMA implementációja az MTA KIK-ben

Az implementációs munkák 2025. februárjában kezdődtek meg könyvtárunkban. Első lépésként egy mátrix típusú belső projektszervezetet állítottunk fel, amelyben az egyes

2 Ezek a Keleti Gyűjtemény, illetve a Kézirattár és Régi Könyvek gyűjteménye

3 A 2016-ban kiadott szándéknyilatkozat az alábbi linken érhető el: https://konyvtar.mta.hu/index.php?name=v_1_5_szandeknyilatkozat

4 https://exlibrisgroup.com/products/alma-library-services-platform/

szakterületek (szerzeményezés, feldolgozás, olvasószolgálat, speciális különgyűjteményi feladatok) képviselői, az azok munkáját összefogó szakterületi felelősök, illetve az operatív feladatokat ellátó Szakinformatikai Osztály négy munkatársa vett részt. A projektcsapat elsődleges feladata az implementációs folyamat nyomon követése, a felmerülő problémák megvitatása és a szükséges döntések meghozatala volt.

Az operatív team feladatai közé tartozott többek között az oktatás szervezése, a konfigurációs dokumentáció kitöltése, az adatexport biztosítása a forrásrendszerekből, a tesztelések és a felmerülő javítási feladatok koordinálása, valamint az úgynevezett 3rd party integrációk menedzselése[5].

Az implementáció teljes idejében rendelkezésünkre állt egy tesztkörnyezet (Sandbox), ami nagyban segítette a tanulási folyamatot.

A *forrásrendszereinkből*[6] elsőként egy tesztelési célokat szolgáló adatbetöltésre került sor. Ez lehetővé tette számunkra az adatok ellenőrzését és a felfedezett hibák kezelését, biztosítva a második és egyben végleges adatmigráció sikerességét[7].

Ezután kerülhetett sor a beszerzési, katalogizáló és kölcsönzési munkafolyamatok megtervezésére, majd azok végleges kialakítására.

Végül az új rendszer éles indítására (go live) augusztus elején került sor.

## Az egységes forráskezelés modellje (inventory model) ALMA-ban

Az ALMA egy egymásnak megfeleltetett, hármas struktúrában „írja le" a különböző típusú erőforrásokat. Mindhárom (fizikai, elektronikus, digitális) forrástípus esetén a kiindulási pont a bibliográfiai rekord, amely az adott forrás leíró adatait tartalmazza.

A *fizikai forrásoknál* ehhez a deskriptív szinthez kapcsolódik az úgynevezett holding-rekord[8], amely az adott mű fizikai elhelyezésével kapcsolatos információk tárolására szolgál, majd ehhez kapcsolódik a példányrekord, amely a konkrét fizikai példány adatait tárolja.

---

5 Ennek részét képezte például egy LDAP-protokoll alapján működő olvasói nyilvántartó rendszer létrehozása. Ez az adatbázis az alapja a Primo-ban történő olvasói autentikációnak, illetve ez biztosítja a REST API-alapú adatszinkronizációt az ALMA felé.

6 A következő forrásrendszerekből történt meg az adatmigráció: a hagyományos dokumentumok rekordjait (bibliográfiai és authority) tartalmazó ALEPH integrált könyvtári rendszerből, ennek ADAM moduljából, amelyben a digitalizált dokumentumaink egy részét tároltuk, illetve az elektronikus forrásainkat tartalmazó elektronikus forráskezelő rendszerből, az SFX-ből.

7 Az adatmigrációt nagyban megkönnyítette, hogy katalogizálási gyakorlatunkban még 2016-ban átálltunk HUNMARC-ról MARC21 szabvány használatára, illetve, hogy a 2017-ben megvalósult WorldCat csatlakozásunk előkészítésekor további adattisztításokat végeztünk. Ezekről részletesebben lásd: Bilicsi (2018), valamint Gyuricza és Haász (2018).

8 Megjegyzendő, hogy a holding rekordoknak egyéb más funkciói is lehetnek az elhelyezési információk eltárolása mellet, mint például a példányszintű megjegyzések rögzítése. Mi egyelőre csak ezzel a szűkített funkcionalitással használjuk a holding-rekordokat.

Az *elektronikus dokumentumok* esetében a bibliográfia rekordhoz a 2. szinten az úgynevezett gyűjteményi (collection) adatok kapcsolódnak. A 3. szinten pedig a portfólió adatok tárolódnak, amely az elektronikus forrás mintegy példányosított változata, tartalmazva többek között az adott erőforrás lefedettségi és kapcsolati adatait.

A *digitális források* esetében a 2. szintet az úgynevezett reprezentáció képviseli, amely egy adott forrás digitális formátumát jeleníti meg. Ez alapján beszélhetünk – például állományvédelmi szempontból – master, illetve származékos másolat típusú reprezentációról. Míg az első esetben jellemzően nagyfelbontású 'tiff'-fájlok kapcsolódnak a reprezentációhoz, a második esetben „használatra szánt", például 'jpg'- vagy 'pdf'-állományok.

Ennek a hármas struktúrának a legfőbb előnye az, hogy az erőforrások közötti kapcsolatok létrehozásával biztosítani lehet a művek különböző kifejezési- és megjelenési formáinak együtt tartását, illetve együtt láttatását a felhasználóink számára[9], ahogy az az 1. ábrán is látható.


*1. ábra: A Tér és társadalom című folyóirat megjelenítése a Primo-ban*

## A tudásbázis és a központi index szerepe

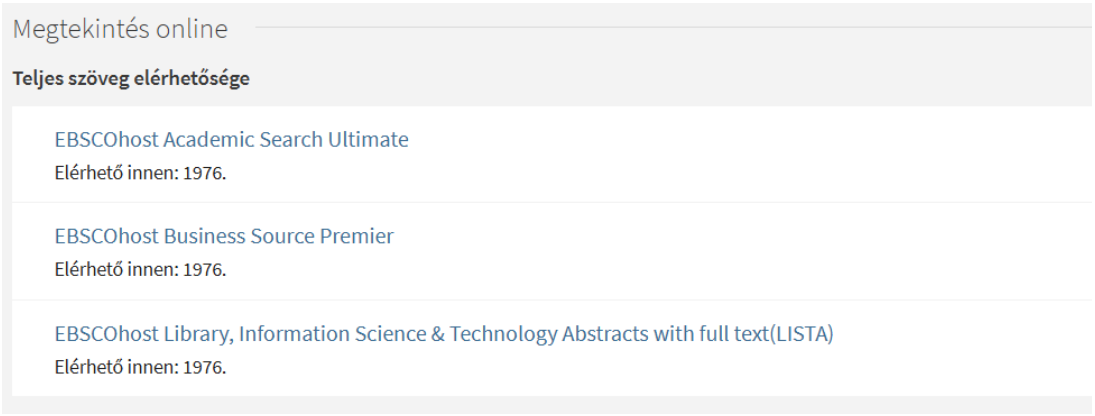
Az elektronikus források szolgáltatása esetében mindenképpen szót kell ejteni a tudásbázis (knowledge base), illetve a központi index (central index) szerepéről a KSZP-k működésében.

A *tudásbázis* tartalmazza a nemzetközi szolgáltatók e-kollekcióinak és open access gyűjtemények metaadatait, beleértve az egyes források bibliográfiai rekordjait, illetve a gyűjtemények lefedettségi és elérhetőségi adatait. A felhasználó könyvtárak a tudásbázisban tudják nyilvántartani és menedzselni saját előfizetéseiket és open access hozzáféréseiket.

ALMA esetében a tudásbázis (Central Knowledge Base - CKB) a CZ (Community Zone) része és mintegy 32 ezer kollekció és 71 millió portfólió adatait tartalmazza. Az adatok karbantartása központilag történik a tudásbázisban, azaz amennyiben az ALMA CZ-ben aktiválunk egy hozzáférést, a kollekcióban bekövetkező mindennemű változás automatikusan megjelenik.

9 Ezt a folyamatot erősíti Primo-ban az FRBR-alapokon működő úgynevezett „verziózás".

Szolgáltatási szempontból a CKB teszi lehetővé, hogy a találati listában egy adott elektronikus forrás minden olyan kollekcióból elérhető legyen, amelyhez a könyvtárnak hozzáférése van.




2. ábra: A KB alapján megjelenő alternatív hozzáférések egy adott elektronikus forráshoz

A *központi index* biztosítja a tudásbázisban tárolt elektronikus források cikk-, illetve könyvfejezet szintű feltárását és a linkfeloldó (link resolver) funkció révén a full-text-ek közvetlen elérését is.

Az ALMA központi indexe (Central Discovery Index - CDI) mintegy 5,2 milliárd rekordot tartalmaz. A CDI feltárja a rekordok közötti hivatkozási kapcsolatokat is, így például egy könyvismertetés kapcsolódik az általa bemutatott könyvhöz. A metaadatok mellett lehetőség van a full-text-ek tartalmának indexelésére is 65 ezer karakterig.

Összefoglalásként elmondható: a KSZP-k integráns részeiként, a tudásbázis és a központi index együttesen teszik lehetővé, hogy a könyvtárak saját gyűjteményeik mellett, előfizetett, illetve ingyenesen elérhető elektronikus tartalmaikat is feltárt módon, egy keresési felületen kínálva tegyék elérhetővé felhasználóiknak.

## Mesterséges intelligenciával támogatott katalogizálás az ALMA-ban

A szoftver működésébe folyamatosan kerülnek beépítésre mesterséges intelligenciára alapozott legújabb technológiai megoldások. Az egyik ilyen funkció az ALMA *katalogizáló asszisztense*, amely új bibliográfiai rekordok létrehozására, illetve meglévő leírások adatgazdagítására használható. A munkafolyamat első lépéseként a feldolgozni kívánt mű néhány[10] alapadatát kell rögzíteni a kapcsolódó űrlapon. A második lépés a mű releváns (például címoldal) részeiről készített fotók feltöltése. Ezek alapján a rendszer elkészíti a

10 Több adat is megadható, de csak egy, a címadat kötelező.

bibliográfiai rekordot, majd az automatikusan átkerül a bibliográfiai szerkesztő felületre (MDE).

| | | | |
|---|---|---|---|
| LDR | | | 00000nam#a2200000#i#4500 |
| 008 | | | 110000s2011####enk##########000#0#eng## |
| 020 | | | $$a 978-1-846-14338-0 |
| 041 | | | $$a eng |
| 100 | 1 | | $$a Davies, Norman, $$e author. |
| 245 | 1 | 0 | $$a Vanished Kingdoms : $$b The History of Half-Forgotten Europe / $$c Norman Davies. |
| 264 | | 1 | $$a London : $$b Allen Lane, an imprint of Penguin Books, $$c 2011. |
| 336 | | | $$a text $$b txt $$2 rdacontent |
| 337 | | | $$a unmediated $$b n $$2 rdamedia |
| 338 | | | $$a volume $$b nc $$2 rdacarrier |
| 520 | | | $$a Vanished Kingdoms explores the history of forgotten European states and kingdoms that once thrived but have since disappeared. Norman Davies delves into the stories of these lost realms, examining their rise, fall, and the cultural, political, and historical legacies they left behind. The book spans centuries, covering regions across Europe, and provides a detailed analysis of how these kingdoms shaped the continent's history. $$7 Generated by AI |

*3. ábra: Részlet egy katalogizáló asszisztenssel készített bibliográfiai rekordból*

Egy miniprojekt keretében megvizsgáltuk a katalogizáló asszisztens hatékonyságát. Ugyanazon öt mű került feldolgozásra hagyományos, „kézi" beviteli módszerrel, rekordátemeléssel, valamint AI-alapú katalogizálással.

Az új módszer előnyei között említhetjük: a rekordok generálásának gyorsaságát[11], a 008-as mező pontos kitöltését, releváns tárgyszavak megalkotását angol nyelven, kétnyelvű tartalmi összefoglaló elkészítését. Viszont néhány nehézséget is tapasztaltunk: időt és gondot kell fordítani a fotók előkészítésére és feltöltésére, zavaró, hogy nem állapítható meg az adatok forrása, bizonyos adatmezőket törölni vagy pótolni kellett[12].

Összeségében arra a megállapításra jutottunk, hogy a katalogizáló asszisztens számunkra elsősorban az adatgazdagítás (például tárgyszavazás), illetve speciális nyelvű dokumentumok feldolgozása során nyújt előnyöket.

## AI-alapú információkeresés a Primo-ban

A *Kutatássegéd* az olvasói felületen rendelkezésre álló AI-alapú keresőeszköz, aminek a használatával természetes nyelven fogalmazhatjuk meg a kutatási kérdéseinket. A keresés elindítása előtt kiválaszthatjuk a dokumentum típusát, a keresés időintervallumát és külön megjelölhető, hogy csak teljes szövegű forrásokat tartalmazzon a találati lista.

11 Amennyiben az előkészítési időt nem számoltuk (fotók elkészítése és feltöltése), az ötből 3 esetben az AI-katalogizálás, két esetben a rekordátemelés bizonyult a leggyorsabb módszernek.

12 Törölni kell az RDA-szabvány alapján bekerülő 336, 337 és 338-as mezőket, és minden esetben pótolni kell a 300-as mezőt.

A keresőkérdésre „válaszolva” – a CDI-ben elérhető forrásokból válogatva – a rendszer öt, az adott témában legfontosabbnak ítélt forrást jelenít meg. A források esetében lehetőség van egy rövid tartalmi kivonat megtekintésére, majd a teljes szöveg elérésére. Emellett egy rövid áttekintést is kapunk a témáról, megjelölve a felhasznált forrásokat. Ezenkívül egy kattintással elérjük a további forrásokat tartalmazó találati listát[13].


*4. ábra: Kutatássegéd a Primo-ban*

A Kutatássegéd az információgyűjtést további kapcsolódó kutatási kérdések felsorolásával is segíti. A kutatási kérdéseket a rendszer tárolja, így azok bármikor újra előhívhatók.

Úgy gondoljuk, hogy a funkció hasznos és hatékony kiegészítője a „hagyományos” kulcsszó alapú visszakeresésnek.

## Összegzés

Az „Akadémia Könyvtára” az 1990-es évek elejétől kezdődően folyamatos erőfeszítéseket tett egy olyan szoftverkörnyezet kialakításáért, amelyben egyszerre lehet optimalizálni a munkavégzés hatékonyságát, valamint a felhasználóknak nyújtott szolgáltatások minőségét. A 2025-ös év fordulópontot jelentett ebben a folyamatban, amikor is egy új típusú könyvtári rendszer implementálására került sor.

13 Ennek alapja az, hogy a rendszer a természetes nyelven megfogalmazott kérdést booelan-típusú keresőkifejezéssé alakítja, majd ez alapján végzi el a keresést.

Alig kilenc hónap elteltével már eredményekről is beszámolhatunk: a zökkenőmentes átállás mellett beállítottuk az automatikus adatszinkronizációt a MOKKA felé, betöltöttük előfizetett e-book gyűjteményeinket a Primo-ba. Az új lehetőségek birtokában terveink is nagyívűek: szeretnénk a magyar nyelvű open access irodalmat láthatóbbá tenni a discovery rendszerünkbe történő betöltés révén és tervezzük a linked data alapú lehetőségek minél szélesebb körű kihasználását az intézményi tudásvagyon globális tudáshálózatokba való integrálása érdekében.

## Köszönetnyilvánítás

Szeretném megköszönni Naszádos Edit és Sallai-Tóth László kollégámnak, hogy az implementáció küzdelmes útját együtt járhattuk végig. Szintén köszönöm Kasza Zsófiának a feldolgozási miniprojektben nyújtott segítségét, valamint Bilicsi Erikának, aki értékes tanácsaival segítette a tanulmány megírását.

## Irodalomjegyzék

Bilicsi, Erika, szerk.: A MARC21 szerinti katalogizálás bevezetése az MTA Könyvtár és Információs Központban. A Magyar Tudományos Akadémia Könyvtárának közleményei (39/114). MTA Könyvtár és Információs Központ, Budapest, 2018. (pdf) https://real-eod.mtak.hu/6918/

Dancs Szabolcs: Könyvtári szolgáltatási platformok, avagy ami az IKR-ek után következik. In: Könyvtári figyelő 61. évf. 3. sz. (2015.) pp. 359–371. http://epa.oszk.hu/00100/00143/00304/pdf/EPA00143_konyvtari_figyelo_2015_3_359-371.pdf

Gyuricza Andrea, Haász Antal: Az MTA Könyvtár és Információs Központ gyűjteménye a WorldCat nemzetközi katalógusban. In: Tick, József; Kokas, Károly; Holl, András (szerk.) NETWORKSHOP 2018 konferenciakötet. Budapest, Hungarnet (2018) pp. 51–58. https://real.mtak.hu/65520/

DOI: 10.31915/NWS.2026.27



# METAADATOK MINŐSÉGE A KÖNYVTÁRI DISCOVERY RENDSZEREKBEN

# METADATA QUALITY IN LIBRARY DISCOVERY SYSTEMS

*Balázs László Róbert*


## Absztrakt

A tanulmány a könyvtári discovery rendszerek technológiai hátterét és az előttük álló kihívásokat foglalja össze. A szerző bemutatja az átmenetet a hagyományos, adatbázis-specifikus keresőktől az egységes, index-alapú discovery szolgáltatásokig, amelyek a Google-szerű felhasználói élményt és az egypontos hozzáférést valósítják meg. A cikk részletesen elemzi a heterogén források használatából következő problémákat, különös tekintettel az analitikus rekordok (folyóiratcikkek) dominanciájára a monográfiákkal szemben, valamint a szerzői névalakok inkonzisztenciájára. A megoldási javaslatok között szerepel a mezőalapú rangsorolás (ranking), az ORCID azonosítók használata, valamint a nemzetközi névtárak (VIAF, ISNI) szisztematikusabb integrációja a rekordok előfeldolgozása során.


**Kulcsszavak:** Könyvtári discovery rendszerek, metaadat-minőség, névalak egységesítés, ranking, heterogén adatforrások, ORCID, VIAF, ISNI, analitikus rekordok


## Abstract

This paper explores the technological background of library discovery systems and the challenges they face regarding metadata quality. The author outlines the transition from traditional, database-specific search engines to unified, index-based discovery services designed to meet modern user expectations for Google-like interfaces and single-point access. The study provides a detailed analysis of issues arising from heterogeneous data sources, specifically focusing on the dominance of analytical records over monographs and the inconsistency of author name authorities. Proposed solutions include field-based ranking algorithms, the adoption of ORCID identifiers, and a more systematic integration of international authority files (VIAF, ISNI) during record pre-processing.


**Keywords:** Library discovery systems, metadata quality, name authority control, ranking, heterogeneous data sources, ORCID, VIAF, ISNI, analytical records

A könyvtárakban hagyományosan a keresők egy-egy adatbázishoz készültek, amelyekkel csak az adott adatbázisban lehetett keresni. Az adatbázisok rekordjai és az indexelés az adott tárgyhoz kapcsolódott. Igyekeztek a fejlesztők minden rendszert a lehető legjobban illeszteni az adatbázis tárgyához, és a tárgyhoz legjobban illeszkedő keresési stratégiákat támogatni. Később megjelentek a közös keresők, amelyekkel egyszerre több rendszerben kereshettünk, azonban a különálló rendszerek továbbra is eltérőek voltak.

A felhasználói igények fejlődése időközben oda vezetett, hogy ma már egy Google-szerű kereső a használói elvárás. Egy kereső felületen szeretnének minden adatbázisban keresni. Azért is, hogy ne kelljen megtanulni, hogy hol találom a megfelelő adatokat és azért is, hogy ne kelljen megtanulni a különböző rendszerek használatát. A különböző rendszerek használatát a mai felhasználók még akkor sem akarják megtanulni, ha az adott rendszerben jobb találatokat kaphatnának bizonyos kérdésekre. Így tehát erős a nyomás a rendszerek egységesítésére.

A Könyvtári discovery rendszerek technológiai háttere biztosíthatja ezeknek az elvárásoknak a teljesülését. Egyszerre és egyformán kereshetünk sok forrásban. Ehhez nem a korábbi keresők megoldásait használják, amikben elküldték a különálló rendszereknek a keresőkifejezést és összegyűjtötték a válaszokat, majd azokat rendezve megmutatták a felhasználónak. A discovery rendszerek a különböző rekordokat gyűjtik össze, előfeldolgozzák, indexelik, majd ezekben történik a keresés. Így kereshetünk olyan rekordokban is, amelyek nincsenek egy adatbázisba sem betöltve, csak szövegfájlban vannak meg, ilyenek lehetnek például a KBART-listák. A rendszerek másik alapvető előnye, hogy azonosan indexeljük a rekordokat, így azonos módon lehet keresni a betöltött rekordokban. A keresés ezekben a rendszerekben általában gyorsabb, mint az SQL vagy NoSQL adatbázisokban, a rekordfeldolgozás elkülönült rendszerekben történik, így nincs szükség tranzakciókezelésre.

Azonban hiába építünk gyors, egységes, jól használható rendszereket, a keresések sikeressége a metaadatok, vagyis a rekordok, minőségén múlik. A heterogén forrásokból érkező rekordok eltérő struktúrája, valamint az eltérő feldolgozási mélységek feszültségeket okoznak a rendszerben.

A cikkben két fő problémát fogok vizsgálni: az analitikus rekordok dominanciáját a találati listákon és a névalakok inkonzisztenciáját a facettákon.

A discovery rendszerek nem jelentenek újdonságot a könyvtáraknak, azonban a fejlesztésük egy véget nem érő feladat, folyamatosan kell fejlesztenünk a rendszereket, és különösen a metaadat-gazdagítás terén folyamatosan kell keresnünk az új lehetőségeket.

## A Discovery rendszerek előnyei

1. Az indexalapú keresés és a „federated search” közötti különbség

   A korábbi rendszerekben például a KözElKat[1]-ban a közös keresőfelületen megadott keresési kérdést a központi rendszer, a bróker továbbította a szakrendszereknek, a KözElKat esetén a különböző könyvtárak katalógusainak, amelyik lefolytatták a keresést, majd visszaküldték a találat halmazt. A találati halmazt a bróker rendezte (ranking), ezt követően megmutatta a felhasználónak. Ezzel a rendszerrel csak olyan keresőkifejezésekre érkezett használható találati lista, amelyekben minden rendszerben indexelt kritériumra kerestünk. Általában ilyen volt a cím és a szerző. Ha tárgyszóra kerestünk, az nem volt minden katalógusban, így, ahol ilyen nem volt, onnan nem érkezett találati lista. A KözElKat legnagyobb tanulsága az volt, hogy még azonos integrált könyvtári rendszert használó könyvtárak is annyira eltérően dolgozták fel a dokumentumaikat, és más indexelési stratégiát használtak, hogy korlátozottan használható a közös keresés ilyen adatbázisokra.

   Ezzel szemben a discovery rendszerekben a kereséseket megelőzően összegyűjtjük a rekordokat, előfeldolgozzuk, azonos elvek szerint indexeljük, és így egységesen tesszük kereshetővé.

   Ez az indexelés olyan rendszerekben történik, amelyekben az indexek felépítése lassabb, mint az SQL adatbázisokban, azonban a keresés lényegesen gyorsabb, nem alakulhatnak ki deadlock[2] események, a keresés sebessége nagyon tág terhelési tartományban közel azonos. A sebesség kérdése az utóbbi hónapokban nagyon fontos lett, elsősorban az LLM[3]-rendszerek tanítási lépései miatt. Egy-egy ilyen rendszer tanítása akár használhatatlanná is lassíthat egy keresőrendszert, még a discovery rendszerek esetében is kihívás a rengeteg címről érkező rengeteg kérés kiszolgálása.

2. Relevancia szerinti rendezés

   A discovery rendszerek technikai alapjai tartalmaznak egy ún. ranking algoritmust is. Ez az algoritmus rendezi sorba a találati rekordokat attól függően, hogy a keresőkérdés találati rekordokban hol és hányszor fordult elő, illetve a rekordok egyéb jellemzőit is figyelembe veheti. A ranking algoritmus folyamatos fejlesztése az olvasói keresési stratégiák elemzése alapján elvárás a könyvtárakkal és rendszerek fejlesztőivel szemben.

---

1 KözElKat – Közös Elektronikus Katalógus Magyar Közös kereső a IIF egyik tartalomszolgáltatási projektje 1996-ban. Ötletgazda: Kokas Károly, Rendszertervezés: Balázs László Róbert, Fejlesztés: Balázs László Róbert, Burgermeister Zsolt. https://web.archive.org/web/20081019002903/ http://www.kozelkat.iif.hu/

2 Olyan esemény, amely egy időre leállítja az adatokhoz történő hozzáférést

3 LLM – Large Language Modell Mesterséges Intelligencia rendszerek

## Az elszigetelt rendszerek használatának felszámolása

1. Egypontos hozzáférés

   A felhasználónak nem kell tudnia, hogy az információ a helyi katalógusban, az intézményi repozitóriumban, vagy egy előfizetett távoli adatbázisban van, mindegyik esetben szerepel majd a találati listában.

2. Láthatóság

   A felhasználó olyan információkra is rátalálhat, amelyekre korábban a rendszerek ismerete híján sosem talált volna meg.

## A felhasználói élmény fokozása és a vizualitás

A felhasználók – különösen a közkönyvtárak felhasználói – hozzászoktak a streaming szolgáltatók rendszereiben a grafikus tartalmakhoz, ezt mára minden rendszertől elvárják. A másik változás, hogy a felhasználók egyre kevésbé hajlandók megtanulni a rendszerek használatát. Ha nem látják át a működést pár másodpercen belül, akkor felhagynak a rendszer használatával és más forrásokat keresnek. Ezért az elvárások változását követve a könyvtárak a discovery rendszereikbe egyre több kiegészítő tartalmat töltenek be. A discovery rendszerek tehát nem csak egy új felületet jelentenek a könyvtárnak, hanem a könyvtári stratégia része, a bevezetésük célja, hogy a könyvtárak versenyképesek maradjanak a Google-al és az LLM-rendszerekkel. Fontos tehát tisztában lennünk az előnyeivel és azokkal a pontokkal is, amelyek fejlesztésre szorulnak.

1. A rekordok tartalmát gazdagítjuk

   A discovery rendszerek számítógépes interfészeken keresztül automatikusan kapcsolják a rekordokhoz a könyvek borítóképeit, a fülszövegeket, tartalomjegyzékeket, az idézettségi mutatókat, recenziókat, videós médiatartalmakat, mint például a booktrailer[4].

2. Keresési stratégia változása

   A discovery rendszerek alapvető jellemzője, hogy bár támogatják az összetett kereséseket is, nem ez az ajánlott keresési stratégia. Ezen rendszerek adekvát használata, egy egyszerű keresési kérdés és a találati halmaz szűkítése a rendszer által ajánlott szűkítési feltételekkel. Az összetett keresésben a szűkítési feltételek általánosak, vagyis akkor is kiválaszthatóak ha a találati halmazban nem is lesz értelme erre szűkíteni, hiszen nem tudhatjuk előre, minek lesz értelme. Ezzel szemben a Discovery rendszerek a találati halmazból építenek fel szűkítési feltételeket, ezeket nevezzük facettának. Ezek tehát csak akkor jelennek meg a

4 A Fővárosi Szabó Ervin Könyvtár innovációja a közösségi médiában megjelenő booktrailerek rekordokhoz kapcsolása.

találati lista mellett, ha van értelme kiválasztani azokat, ugyanis a találati lista feldolgozásával jönnek létre. Például csak olyan tárgyszavak jelennek meg a facettákban, amelyek előfordulnak a találati halmaz rekordjaiban. Ezek a facetták alkalmasak arra, hogy a nagy találati halmazokat gyorsan, egyszerűen szűkítsük áttekinthető méretűre.

## Kihívások a discovery rendszerek implementálása közben

1. Az analitikus rekordok túlsúlya

   Egy folyóirat cikkeinek száma általában jelentős, és ha sok folyóirat cikkeit töltjük be, az számban messze felülmúlhatja a könyvtár katalógusában lévő monográfiák rekordjainak számát. Ebből az következik, hogy bármire is keresünk, jórészt analitikus rekordokat kapunk. Ez különösen akkor zavaró, ha nem egy kutatást végzünk, hanem egy monográfiát szeretnénk megtalálni. A monográfiák eltűnnek az analitikus rekordok tengerében. Ez ellen a megfelelő rankigalgoritmus, a találati listák megfelelő sorba rendezési módszerének megtalálása segíthet.

2. Inkonzisztens névalakok

   A sok forrásból származó metaadatok sokféle szabvány szerint feldolgozott adatbázisból származnak. A sokféle szabvány a névalakok írásmódját is befolyásolja. A kiadói listákon lehet, hogy a szerző keresztneve csak egy betűvel szerepel, elképzelhető, hogy az intézményi repozitóriumban is más a nevek feldolgozásának módja, mint a könyvtári katalógusban. Magyarországon az intézményi repozitóriumokban nincs, a könyvtári katalógusokban általában van authority kontroll. Ez a sokféleség azt jelenti, hogy a nevek az összetöltött adatbázisban nem lesznek egységesek. Egy szerzőnek több formában is szerepelhet a neve az összetöltött adatbázisban. pl: Kovács János, J. Kovács, Kovács, János.

   Ez a sokféleség a facetták hatékonyságát jelentősen rontja, nem lehet leszűkíteni egy szerzőre egy szűréssel a találati listát, és a felhasználónak kell tudnia, hogy melyik névalak azonos a másikkal.

3. Heterogén adatforrások

   A discovery rendszerek több forrást töltenek össze, ezekben a forrásokban más és más szabvány szerint történik az adatfeldolgozás. Vannak MARC21-források, általában ilyen a katalógus, vannak KBART-listák, a kiadói egyszerűsített e-folyóirat listák, vannak a repozitóriumból származó Dublin Core rekordok. Ez a feldolgozási sokféleség jelentősen megnehezíti a rekordok egységes kezelését. Hiába dolgozzuk fel ezeket betöltéskor, a metaadatok teljessége és minősége is változó, ami közvetlenül kihat a rankingalgoritmusra, a rangsorolás minőségére.

## Megoldási javaslatok

A felvázolt problémák megoldása nem elképzelhetetlen, de ehhez együtt kell dolgoznia a kiadóknak, a szoftverfejlesztőknek és a könyvtáraknak.

1. Intelligens rangsorolás

   A találati lista megfelelő sorba rendezése sokat segíthet a monográfiák láthatóságán.

   - Szükség van mezőalapú súlyozásra, vagyis a találati listában szerepeljen előbb, ha egy rekordban valamely fontosabb mezőben találjuk meg a keresett kifejezést, pl. cím, szerző.
   - A keresés helyszíne és a dokumentum elérhetősége is súlyozási tényező lehet, így a helyben elérhető monográfiák előrébb kerülhetnek.
   - Lehetséges a dokumentumtípusokat is figyelembe venni a súlyozásnál, ha azt szeretnénk.

2. A szerzői névalakok egységesítése

   Betöltéskor feldolgozzuk a rekordokat. Ez a feldolgozás általában formátumkonverziót jelent, de ha javítani akarunk a rekordok minőségén, a feldolgozásnak ki kell terjednie a nevek konszolidációjára is.

   - A kiadókkal az ORCID azonosítókat használva tudunk együttműködni. Ez egy kutatói azonosító, amivel a kutató általában egyértelműen azonosítható. Sajnos nem minden esetben, mert egy kutatónak több azonosítója is lehet. Az általában nem fordul elő, hogy egy azonosítóhoz több kutató tartozzon.
   - A kutatói nevek eltérő írásmódjainak összekapcsolására a VIAF és az ISNI rekordok lehetnének a megfelelő kiindulópont. Ezekben a rekordokban felsorolható az összes névalak, amit az bibliográfiai rekordok előzetes feldolgozásakor felhasználhatunk az egységesítésre.

## Problémák az egységesítés közben

Az előfeldolgozás közbeni egységesítéshez valamilyen adatbázisban szerepelnie kell a megfelelő névalakoknak, a megfelelő ORCID azonosítóknak, ráadásul olyan API-nak kellene szolgáltatnia, ami megfelelő sebességgel képes kiszolgálni a betöltéshez kapcsolódó előfeldolgozás közben akár több millió kérést könyvtáranként. A VIAF API rendszerét a közelmúltban alakították át, az új rendszer jobban megfelel ennek az elvárásnak. A VIAF-ban a nemzeti könyvtárak tartják karban a besorolási rekordokat. A magyar nyelvi rekordokat 2015 előtt töltötte be az OSZK és azóta nem aktualizálta. Sajnos a VIAF rekordok nem tartalmazzák az ORCID azonosítókat. Lehetséges lenne, hogy a névalakokat és azonosítókat a Nemzeti Névtérben találjuk meg, azonban annak a karbantartása sem folyamatos és az API rendszerét nem ilyen alkalmazásra tervezték.

Jelenleg tehát az egyetlen megoldás az, ha minden könyvtár a saját authority rekordjaiban sorolja fel az általa elképzelhetőnek tartott összes névalakot és azonosítót, és azt folyamatosan karban is tartja. Ez nagyon nagy munka, amit felesleges lenne minden könyvtárban folyamatosan végezni.

## Összefoglalás

A discovery rendszerek térnyerése nem csupán egy technológiai váltás a könyvtárak életében, hanem egy alapvető szemléletmódváltás a tudományos információhoz való hozzáférésben. Láttuk, hogy az út nem mentes a nehézségektől: az adatmennyiség robbanásszerű növekedése és a források sokfélesége olyan kihívások elé állítja a rendszereket, amelyekre a hagyományos katalógusoknak még nem kellett válaszolniuk, illetve a szűkebb rendszerekre kidolgozott válaszaik kiterjesztése a discovery rendszerekre nem egyszerű.

Láttuk, hogy a discovery rendszerekbe töltött metaadatok minősége javítható, de ehhez megoldásokat munkafolyamatokat kell kidolgoznunk, és azokat folyamatosan végre is kell hajtanunk.

A könyvtárosok és fejlesztők közös feladata most az, hogy a technológiai lehetőségeket a minőségi adatokkal ötvözve biztosítsák: a tudás ne csak elérhető, hanem valóban felfedezhető is maradjon a digitális korszakban.

## Irodalomjegyzék


Holl András – Bilicsi Erika (2017) Orcid – egy újabb szerző-azonosító tudományos közleményekhez [kézirat] Könyvtári Figyelő 2017(3) https://real.mtak.hu/65517/1/ORCID_KF.pdf

Carolyn Caizzi, Ami Deschenes (2026) From Card Catalogs to Semantic Search – Building a Human-Centered Discovery Platform Powered By AI technology Information Technology and Libraries 2026(3) https://doi.org/10.5860/ital.v45i1.17511

Balázs László (2010) Virtuális közös lekérdezés, vagy valós központi adatbázis Tudományos Műszaki Tájékoztatás 2010(2) https://www.pp.bme.hu/tmt/article/view/33121/18836

Burgermeister Zsolt, Balázs László (1997) KözElKat: A közös Elektronikus Katalógus NIIF-es és Tempuszos élete.

Sujan Saha, Sukumar Mandal (2022) Application and Integration of VIAF and OCLC FAST, with KOHA, VuFIND, and Google CSE Journal of Library Metadata (2022)3–4 https://doi.org/10.1080/19386389.2022.2103346

Monica Moore (2016) But is My Resource Included? How to Manage, Develop, and Think about the Content in Your Discovery Tool, The Serials Librarian, 70:1–4, 149–157, https://doi.org/10.1080/0361526X.2016.1147877

Patricia M. Dragon, Janet L. Mayo, An Carol Stocks, Rebecca Tatterson (2025) Enhancing library discovery: An approach to understanding user access to electronic resources The journal of Academic Librarianship (2025) https://doi.org/10.1016/j.acalib.2025.103064

DOI: 10.31915/NWS.2026.28



# RAG MINT KÖNYVTÁRI TUDÁSRENDSZER

# A MESTERSÉGES INTELLIGENCIA HANGJA AZ IR VS. DR VITÁBAN

# RAG AS A LIBRARY KNOWLEDGE SYSTEM

# THE VOICE OF ARTIFICIAL INTELLIGENCE IN THE IR VS. DR DEBATE

*Fülöp Endre*
*endre.fulop@qulto.eu*
*Qulto – Monguz, rendszertervező*
*ELTE BTK Könyvtár- és Információtudományi Intézet, doktorandusz*

**Absztrakt**

Miután Calvin Mooers – bizonyos értelemben Paul Otlet monografikus elvét követve – 1950-ben megalkotja az információ-visszakeresés (Information Retrieval, IR) terminust, a könyvtártudomány szinte azonnal átveszi és magáévá teszi a fogalmat és a mögötte álló elvet. Ezzel lényegében azt az elvárást támasztja a könyvtári tudásrendszerrel szemben, hogy amikor a felhasználó információt keres, akkor a könyvtári rendszer ne pusztán a kérdés szempontjából releváns dokumentumokhoz legyen képes őt irányítani, hanem adjon közvetlenül és azonnal tényszerű választ a megfogalmazott kérdésre. A könyvtári tudásrendszer IR-rendszerként való felfogásának gondolatát nem fogadta egyöntetű lelkesedés. Számos elméletalkotó (többek között Robert Fairthorne, Bernd Frohmann, Birger Hjørland és Rafael Capurro) különböző okokból és nézőpontokból, de egyaránt bírálja a paradigmaváltást, és amellett érvel, hogy a könyvtári tudásrendszer fő feladatának mégis inkább a dokumentum-visszakeresést (Document Retrieval, DR) kellene tekinteni. Tanulmányomban arra a kérdésre keresek választ, hogy a nagy nyelvi modellek (LLM) és különösen a Retrieval Augmented Generation (RAG) megjelenését tekinthetjük-e fordulópontnak ebben az IR vs. DR vitában. Szolgáltat-e a RAG érveket akár az egyik, akár a másik oldal számára? Mit kell, mit lehet, mit célszerű elvárnunk egy könyvtári tudásrendszerként működő RAG-alkalmazástól?

**Kulcsszavak** információfogalom, információfilozófia, információ-visszakeresés, dokumentum-visszakeresés, mesterséges intelligencia, LLM, RAG, metainformáció

**Abstract**

After Calvin Mooers coined the term Information Retrieval (IR) in 1950 – in a sense following Paul Otlet's monographic principle – library science almost immediately adopted and integrated both the concept and its underlying principle. This essentially imposed an expectation on library knowledge systems: when a user searches for information, the system should not merely direct them to documents relevant to their query, but should provide a direct and immediate factual answer to the formulated question. The notion of conceiving the library knowledge system as an IR system was not met with unanimous enthusiasm. Numerous theorists (including Robert Fairthorne, Bernd Frohmann, Birger Hjørland, and Rafael Capurro), for various reasons and from different perspectives, criticized this paradigm shift, arguing that the primary task of library knowledge systems should remain Document Retrieval (DR). In my paper, I seek to answer whether the emergence of Large Language Models (LLM) and, in particular, Retrieval-Augmented Generation (RAG) can be considered a turning point in this IR vs. DR debate. Does RAG provide arguments for either side? What should, what can, and what is it practical to expect from a RAG application functioning as a library knowledge system?

**Keywords**: concept of information, philosophy of information, information retrieval, document retrieval, artificial intelligence, LLM, RAG, meta-information

## 1. Információ-visszakeresés (IR) és dokumentum-visszakeresés (DR)

Ha könyvtártudományi tanulmányokat folytatva vagy könyvtártudományi kutatásokat végezve azzal a kérdéssel találjuk szembe magunkat, hogy *mikor és milyen körülmények között született meg az információ-visszakeresés terminus*, s könyvtári tudásrendszerhez fordulunk segítségért, kétféle válaszra számíthatunk. Kaphatunk egyfelől közvetlen, azonnali választ, például az alábbi formában: „*Az információ-visszakeresés kifejezést Calvin Mooers amerikai informatikus alkotta meg 1950-ben.*" A könyvtári tudásrendszer által adott másik fajta válasz megbízható, tudományos igényű publikációk listája lehet az információtudomány vagy információtudomány-történet területéről, amelyekben a kifejezés megszületésének körülményei is feldolgozásra vagy legalábbis említésre kerülnek. Egy lehetséges példa ilyen listára:

- Wolfgang G. Stock, Mechtild Stock: *Handbook of Information Science*[1]
- Csík Tibor: *Az információtudomány létrejötte és Horváth Tibor információtudományi eszméinek gyökerei*[2]

1 Stock – Stock, 2013:93

2 Csík, 2017:59

- Mark Bowles: *The Information Wars: Two Cultures and the Conflict in Information Retrieval, 1945-1999*[3]
- Ronald R. Kline: *What Is Information Theory a Theory Of?*[4]

Arra, hogy a kétféle válasz közül melyik hasznosabb, melyik elégíti ki jobban a kérdező igényét, általános érvénnyel nem lehet felelni. Bizonyos használati esetekben egyik, másfajta használati esetekben a másik nyújt több segítséget. Az is kétségtelen, hogy elképzelhetőek másfajta válaszok is a fenti két típus valamiféle kombinációjának formájában. A két fenti választípus mibenlétének megértése, különbségeik számbavétele azonban nem haszontalan és hiábavaló. Ideáltípusokról lévén szó tanulmányozásuk hozzásegíthet a könyvtári tudásrendszerrel szemben támasztható igények, követelmények tisztázásához.

A továbbiakban arra, amit az első esetben végez a könyvtári tudásrendszer, az *információ-visszakeresés* (*Information Retrieval*), arra pedig, amit a második esetben végez, a *dokumentum-visszakeresés* (*Document Retrieval*) kifejezést fogom használni. Ezek az elnevezések természetesen vitathatóak. Az információ-visszakeresés kifejezést sokféle értelemben használják, nem ritkán beleértik ebbe a fogalomba a dokumentum-visszakeresést is.[5] Mások – talán éppen ezért – az *adat-* vagy *tényvisszakeresés* kifejezést részesítik előnyben az első típus vonatkozásában.[6] Az elnevezések maguk azonban talán kevésbé lényegesek, ha megfelelően tisztázott az, hogy a kifejezések használója melyik alatt mit ért pontosan. *Információ-visszakeresés* alatt a továbbiakban azt értem, ha a könyvtári tudásrendszer nem pusztán a felhasználó kérdése szempontjából releváns dokumentumokhoz irányítja, hanem közvetlenül és azonnal tényszerű választ ad a megfogalmazott kérdésre. *Dokumentum-visszakeresés* alatt pedig azt, ha a tudásrendszer kitüntetett feladatának a felhasználói kérdés szempontjából legrelevánsabb dokumentumok listájának összeállítását tekintjük.

## 2. Az otlet-i vízió és a kritikája

Bár – ahogy ez már a fent kiderült – az *információtudomány* terminus Calvin Mooers nevéhez fűződik, a kifejezés mögött álló gondolat legnagyobb hatású képviselőjének Paul Otlet belga bibliográfust tekinthetjük. Ő volt az, aki az eddig bevett gyakorlattal szembehelyezkedve nem a dokumentumok, hanem „a dokumentumokban található egyedi információk”[7] számbavételét jelölte meg a könyvtárak legfontosabb feladata gyanánt. 1934-ben írt *Traité de documentation* címet viselő főművében a következőképpen fogalmazza ezt az célkitűzést: „Maguk az adatok jól elkülöníthetők azoktól a dokumentumoktól, amelyek-

3 Bowles, 1999:158

4 Kline, 2004:20

5 van Rijsbergen, 1979:1

6 Capurro–Hjørland, 2003:384

7 Pogányné, 2007:66

ben rögzítették őket. Ezen tények és adatok halmazainak rendszerezett megszervezéséről van szó."[8] Az Otlet által *monografikus elv*nek is nevezett vízió szerint egy könyvtári tudásrendszernek nem a létező, nemritkán feleslegesen bőbeszédű és gyakran csupán közismert információkat ismételgető dokumentumok katalogizálásán és osztályozásán kell alapulnia. Az értékes információkat – Otlet szóhasználatával a tényeket és adatokat – elemi formában és a maguk identitásában[9] ki kell nyerni, ki kell szabadítani a dokumentumokból, és ezekből az önmagában értelmezhető, önmagában teljes információelemekből kell a könyvtári tudásrendszert felépíteni, amely így inkább lesz hasonlatos egy univerzális enciklopédiához, mint könyvtári katalógushoz.

Ehhez a vízióhoz csatlakozott Calvin Mooers a múlt század 50-es éveiben új terminus hozzáadásával. „Az információ-visszakeresés kifejezést alig nyolc évvel ezelőtt volt szerencsém megalkotni egy másik számítástechnikai konferencia alkalmával. A név azóta hosszú utat tett meg. […] Jelenleg az információ-visszakeresés már többről szól, mint a dokumentumok puszta megtalálása és rendelkezésre bocsátása; most már az információ felfedezésével és átadásával foglalkozunk, teljesen függetlenül annak dokumentumformájától."[10] Mooers terminusát a könyvtártudomány szinte azonnal átveszi, mind a mai napig használja.

A mögötte álló otlet-i víziót, azaz könyvtári tudásrendszer IR-rendszerként való felfogásának gondolatát azonban nem fogadta egyöntetű lelkesedés.[11] Boyd Rayward, ausztrál kutató például markáns kritikával illeti ezt az elképzelést. Úgy látja, hogy a dokumentumok Otlet szemében csak erősen korlátozott értékkel bírnak, mert ismétlésektől hemzsegnek, megfogalmazásuk zavaros, és egyszerre vannak tele igaz és téves állításokkal. „Számára a dokumentumok tartalmának azon aspektusa, amellyel foglalkoznunk kell, a tények. Szinte mindenütt tényekről beszél. Egyértelmű, hogy az általa kifejlesztendőnek vélt új típusú rendszerekkel szemben az elsődleges követelmény az, hogy képesek legyenek felszabadítani az értékes információkat – amit ő tényeknek nevez – az azokat fogva tartó és elrejtő dokumentumokból."[12] Ez az elképzelés azonban Rayward szerint „tekintélyelvű, redukcionista, pozitivista, leegyszerűsítő – és optimista."[13] Arra a nehezen igazolható hitre épül, hogy ami tényszerű és igaz, könnyen azonosítható egy dokumentumban, hiszen ez csupán tartalmának elemzésére és rendszerezésére szolgáló folyamatok szabályozásának és intézményesítésének kérdése.

8 Otlet, 2021:13

9 Rayward, 1990:1

10 Mooers, 1960:229

11 Ld. pl. Hjørland, 2026

12 Rayward, 1994:247

13 Rayward, 1994:247

Birger Hjørland és Rafael Capurro is hasonló álláspontot képvisel, mint Rayward. Szerintük az az elképzelés, amely a könyvtári tudásrendszerek információ-visszakereső rendszerré alakítása mögött áll, idejétmúlt, „meghaladott pozitivizmus."[14] Robert Fairthorne pedig ironikusan *az információ flogiszton-elméletéről* ír ennek kapcsán, arra utalva ezzel, hogy a dokumentumokban lévő információk visszakeresését propagáló elméletek ugyanúgy tévesen – lényegében egy nem létező dolog feltételezésével – értelmeznek egy jelenséget, ahogy a 17–18. században a flogiszton-elmélet az égést.[15] Ha szöveget olvasunk, nem az történik, hogy valami, ami a dokumentumban van (az információ-flogiszton), „átfolyik" s bekerül a fejünkbe. „Abból, hogy egy rendszer reagál egy ingerre, nem következhetünk valamilyen megmaradó mennyiség létezésére, amely átáramlik rajta. Amikor az ujj meghúzza a ravaszt, nem áramlik semmi az ujjból a golyóba."[16]

Bernd Frohmann Jacques Derrida posztstrukturalista filozófiája mentén fogalmaz meg kritikát Otlet víziójával szemben. Szerinte elemi információk dokumentumokból való kiemelése és kontextusfüggetlen leírása illúzió. A leírás mindig jelekkel (szavakkal) történik, s ezek a jelek nem rendelkeznek azzal a stabilitással, amelyet az otlet-i koncepció feltételez és elvár tőlük. A jelek képlékenyek, mást jelenthetnek helytől és időtől függően. Derrida szerint egy jel éppen attól jel, hogy végtelen különböző kontextusba helyezhető. Így, Frohmann szerint, „derridai probléma merül fel Otlet azon elképzelésével kapcsolatban, miszerint a tények teljes jelenlétük világosságában leírhatóak a tudat számára egy olyan dokumentációs fegyelem révén, amely a tények rögzítését és a jeleik közötti kapcsolatok stabilizálását szolgálja."[17]

## 3. Keresőmotorok, chatbotok és RAG

Ezekben a kritikákban minden bizonnyal sok igazság rejlik, azt állítani azonban, hogy Paul Otlet víziója nem valóban létező, máig jelen lévő felhasználói (kutatói, olvasói) igényre próbálna megoldást találni, túlzás, sőt – megkockáztathatjuk – tévedés volna. Ha másból nem, a nagy nyelvi modellekre épülő generatív mesterségesintelligencia-alapú chatbotok gyors elterjedéséből és rendkívüli népszerűségéből következtethetünk erre. Észre kell ugyanis vennünk, hogy MI chatbotokat (pl. ChatGPT, Gemini) – egyebek mellett – az különbözteti meg a „hagyományos" webes keresőmotoroktól (pl. Google, Bing), hogy míg az utóbbiak dokumentum-visszakeresést végeznek, addig az előbbiek lényegében információ-visszakeresést hajtanak végre. Ha egy webes keresőmotorhoz fordulunk azzal a kérdéssel, hogy „*Milyen hangszerek szólnak Gluck Orfeusz és Euridiké című operájában?*"

14 Capurro–Hjørland, 2003:385

15 Fairthorne, 1967:711

16 Fairthorne, 1967:711

17 Frohmann, 2017:85

akkor egy webes keresőmotor online enciklopédiák szócikkeihez, komolyzenei portálok előadás- és lemezkritikáihoz, digitalizált zenetörténeti tankönyvekhez vagy tanulmányokhoz, operaházak webes műsorfüzeteihez és online elérhető operapartitúrákhoz vezető linkek listáját kínálja, azaz olyan dokumentumokhoz vezet el, amelyek relevánsak a kérdésem szempontjából, és amelyekből jó eséllyel választ kapok arra, amit tudni szeretnék. Ez pedig nem más, mint amit fent dokumentum-visszakeresésként definiáltunk. Ha viszont ugyanezt a kérdést egy MI chatbotnak teszem fel, akkor azonnal és közvetlenül megkapom azoknak a hangszereknek a listáját, amelyek Gluck a nevezett operában használt. Ez pedig az információ-visszakeresés fenti definíciójának felel meg. A nagy nyelvi modellre épülő MI chatbotok egyre terjedő használatából tehát joggal következtethetünk arra, hogy az információ-visszakeresés létező felhasználói igény.

Fontos azonban azt is látni, hogy a nagy népszerűségnek örvendő chatbotok nem mindenben felelnek meg az otlet-i víziónak. A monografikus elvnek ugyanis fontos részét képezi az is, hogy az információ-visszakereséshez használt elemi információk köre miből és hogyan épül fel. Nem pusztán dokumentumokból kiszabadított információknak, hanem *gondosan válogatott dokumentumokból körültekintően kinyert s szakszerűen leírt információk*nak kell a visszakeresés alapjául szolgálniuk. Ennek a kritériumnak a népszerű chatbotok nem felelnek meg, hiszen a nagy nyelvi modellek építéséhez kis túlzással válogatás nélkül használtak fel mindenféle szöveges dokumentumot. Hamar ki is derült, hogy naprakész pontosságot és megbízhatóságot igénylő szakértői tudásrendszerként ezek az MI chatbotok nem tudnak működni.[18] Ennek a korlátnak a felismerése hívta életre a RAG-rendszereket, amelyek éppen ebből adódóan az tanulmány első felében tárgyalt vita szempontjából is meghatározó jelentőséggel bírnak.

A RAG (Retrieval Augmented Generation) olyan architektúrát jelent, amely a generatív, nagy nyelvi modellre épülő MI-rendszerbe egy új komponenst (vektoradatbázist) és egy új fázist (retrieval-fázis) illeszt. A RAG-rendszerek építési (indexelési) fázisában ki kell választani azokat a dokumentumokat, amelyeket a rendszer tervezői megbízható információk forrásának tekintenek. Az indexelés során ezeknek a kiválasztott dokumentumoknak a kis darabjait (chunks) numerikus reprezentációkká, ún. vektorokká alakítják, s ezekből a vektorokból speciális vektoradatbázist építenek. Amikor a felhasználó kérdést tesz fel a rendszernek, először a fent említett új (retrieval) fázis indul el, amelynek során rendszer gyors, szemantikus keresést végez az indexelés során felépített vektoradatbázisban. A szemantikus keresés eredményeként a RAG-rendszer megtalálja a megbízhatóként megjelölt, indexelésre kiválasztott dokumentumok kérdés szempontjából legrelevánsabb szövegrészleteit. Az LLM-komponens feladata pedig már „csak" annyi, hogy ezekből a szövegrészekből, információelemekből természetes nyelven legenerálja a pontos és tényszerű választ.

18 Gao et al., 2023:1

A RAG-rendszerek tehát már sokkal jobban megközelítik Otlet vízióját, hiszen itt az információ-visszakeresés gondosan és körültekintően kiválasztott dokumentumok szövegrészletei alapján történik. A RAG-rendszerek lehetőségeinek és korlátainak számba vétele így az otlet-i koncepció tanulmányozására és értékelésére is alkalmasnak látszik.

## 4. Információ a dokumentumban és információ a dokumentumról

Mind Otlet, mind a RAG-rendszerek koncepciójának alapmozzanata, hogy az információkat, amelyekre az információ-visszakeresést építeni kívánják, a kiválasztott dokumentumokból nyerik ki. Az információ-visszakeresés tehát mindkét esetben *a dokumentumban lévő információ*ra épül. Természetesen Otlet tisztában volt azzal, hogy a dokumentumok nem kizárólag információkból állnak, de minden bizonnyal nem tévedünk nagyot, ha azt állítjuk, hogy úgy gondolta, hogy az általa megálmodott információ-visszakeresésnek csak az információkkal van tennivalója. De vajon igaz-e, hogy egy könyvtári tudásrendszer számára csak a dokumentumokban lévő információk relevánsak?

Nézzünk két használati esetet! Az elsőben azért fordul a felhasználó könyvtári tudásrendszerhez, mert gyermekek környezeti neveléséhez a búbos banka hangjáról gyűjt anyagot, a másodikban pedig spanyolországi utazás előtt állva szeretne az Alhambráról tájékozódni. Az első használati eset kapcsán hasonlítsuk össze e két szövegrészletet:

($A_1$) „[A búbosbanka] nászhangja öblös, tompa, többször ismételt »up-up-up«."[19]

($B_1$) „… ebbe a csendbe lágy tavaszi kiáltással dobol bele a búbos banka.

– Hu-pu-pu, hu-pu-pu…

Ruhája ékes, fején gyönyörű bóbita, de mást nem tud. – Hu-pu-pu, hu-pu-pu…

Ez a kiáltás mindig közel van, és mindig távol. Egyszerre szól a földön és a levegőben, rászáll a csendre, és csenddé lesz, nincs benne semmi, mégis virágos szomorúság marad utána, mert mindenki tudja, hogyha elhallgat: ledobja szűzi koszorúját a Tavasz, elnehezedik, megbarnul, és tarka, nagy köténye alatt dús anyaságát takargatja a Nyár."[20]

19 Svensson – Grant, 2020:467

20 Fekete, 1971:94

Az első idézet egy terepi madárhatározóból, a második Fekete István *Kele* című regényéből származik. A második használati eset vonatkozásában is vessünk össze két szövegrészletet, az előbbit egy útikönyvből, az utóbbit Száraz Miklós György *¡Ó, Santo Domingo!* című esszéregényéből idézem[21].

($A_2$) „V. Muhammad remekművét 124 kecses oszlopon nyugvó boltív szegélyezi. A középső kutat 12 zömök kőoroszlán tartja."[22]

($B_2$) „Az Alhambra pedig nem építészet, hanem fény. Az ablakok, az áttört díszek, a kagyló- és cseppkőstukkók mind-mind vezetik, terelik, szűrik a fényt, átengedik és megmozgatják, láthatóvá teszik a levegőt. […] Egyszerre üdítő és lenyűgöző. Csak addig hat, amíg benne vagy. Akkor viszont uralkodik rajtad, nemcsak körülvesz, hanem beléd is költözik."[23]

Talán egyetérthetünk abban, hogy az *A* állításokat joggal tekinthetjük információnak, mind az $A_1$, mind az $A_2$ állítás belefér az információ legszigorúbb meghatározásába is. Megkockáztathatjuk, hogy Otlet monografikus elve szerinti információkinyeréshez is megfelelő alapanyagul szolgálhatnak. Ha egy RAG-rendszer retrieval-fázisa e szövegrészeket keresi elő a vektoradatbázisból, akkor az LLM-fázis megfelelő természetes nyelvű választ fog generálni – amely nem lesz ugyan teljesen azonos a kapott szövegrészlettel, de ennek az eltérésnek (az átfogalmazásnak) a válasz relevanciája és pontossága szempontjából semmi jelentősége nincs.

A *B* állításokról ugyanez viszont nem mondható el. Ha egy RAG-rendszert a búbosbanka hangjáról kérdezve azt a választ kapnánk, hogy „*a búbosbanka hangja egyszerre szól a levegőben és a földön, egyszerre van közel és távol, semmi sincs benne, de szomorúságot okoz*", akkor joggal lennénk elégedetlenek a rendszer válaszával. S ugyanígy meghökkentőnek és érthetetlennek találnánk, ha egy RAG-rendszer LLM-fázisa a $B_2$ szövegrészlet alapján azt a választ generálná az Alhambrával kapcsolatos kérdésünkre, hogy az „*Alhambra nem építészeti alkotás, hanem fényjelenség, amely beléd költözik, ha benne vagy*". Ebből arra – az intuíciónknak egyébként tökéleseten megfelelő – következtetésre juthatunk, hogy *B* állítások nem tekinthetőek információnak abban az értelemben, amelyben az *A* állítások azok. Kétségkívül van ezeknek is (pozitív) hozadéka a búbosbanka hangjáról, illetve az Alhambráról szóló ismereteink szempontjából, de olvasásuk nem azt a hatást váltja ki

21 Hogy a B1 és a B2 szövegrészletek szépirodalmi művekből származnak, nem véletlen. Tanulmányomban arra is fel szeretném hívni a figyelmet, hogy a tudásátadás szempontjából azoknak a könyveknek is jelentős szerepe lehet, amelyekről hajlamosak vagyunk elfeledkezni, amikor (a szűken vett) információközlésről gondolkodunk. Kétségkívül igaza van ugyanis Wittgensteinnek abban, „hogy a költemény, bár az információ nyelvén fogalmazták meg, nem az információközlés nyelvjátékában használatos" [Zettel §160, idézi: Floridi, 2011:83], de ehhez azt is hozzá kell tennünk, hogy jóllehet a szépirodalom nem az információközlés nyelvi játékának része, a tudásközvetítésben gyakran játszik fontos szerepet.

22 Sághi – Dávid, 2006:467

23 Száraz, 2003:162-3, 159

belőlünk, amelyet az *A* állításokhoz hasonló információközlés általában kivált, azt ti., hogy *olvastam valamit, amit eddig nem tudtam, de most már tudom.* Fairthorne nyomán úgy is fogalmazhatnánk, hogy az *A* állítások „információ-flogisztonként" „átfolynak" belénk, a *B* állítások „csak" befolyásolják a gondolkodásunkat, nézőpontunkat, valóságérzékelésünket, változást indukálnak, hozzáárulnak a jelenségek jobb, teljesebb, pontosabb megértésében. Megkülönbözteti a *B* részleteket az *A* állításoktól az is, hogy átfogalmazásukkal elveszik a lényegük, forrásmegjelölés nélkül felhasználói szempontból nehezen értelmezhetőek és értékelhetőek – ami azt is jelenti, hogy a Frohmann által említett derridai probléma ezeknél sokkal határozottabb és egyértelműbben jelentkezik, mint az *A* típusú szövegrészleteknél.

De vajon abból a tényből, hogy a *B* szövegrészletek nem információk, azt a következtetést kell-e levonnunk, hogy érdektelenek egy könyvtári tudásrendszer számára? Semmiképpen sem. Az a tudásszolgáltatás, amelyre a *B* állítások jelentenek példát, legalább olyan fontos, mint az *A* állításokhoz hasonló, szigorúan tényszerű információszolgáltatás. A gyermekek környezeti nevelése szempontjából a *Kele*ből vett idézet legalább annyira releváns és hasznos, mint a madárhatározó szikár tényközlése, s bizonyára sokan vannak olyanok is, akik az utazásra való felkészülés tekintetében értékesebbnek találják Száraz Miklós György szubjektív útleírását, mint az útikönyv egzakt információit.

Az *A* típusú szövegrészleteket használva *a dokumentumban lévő információ*ból megfelelő választ tud generálni egy információ-visszakereső rendszer. *B* típusú szövegrészletek kapcsán azonban nem *a dokumentumban lévő információ*ra, hanem *a dokumentumról szóló információ*kra, azaz *metainformáció*ra építve kell a választ kigenerálnia[24]:

($M_1$) Fekete István *Kele* című művében lírai leírást találunk a búbosbanka hangjáról.

($M_2$) Száraz Miklós György *¡Ó, Santo Domingo!* című esszéregényében az Alhambra Oroszlános udvaráról is olvashatunk szubjektív útleírást.

Az ilyen válaszok azonban már nem egy információ-visszakereső rendszer válaszai, hanem egy dokumentum-visszakereső rendszeré, hiszen releváns dokumentumokhoz irányítanak, s nem közvetlen, tényszerű választ közölnek a feltett kérdésre. Ami másképpen fogalmazva azt jelenti, hogy annak, hogy *A* típusú dokumentumokból információ-visszakereső rendszert építsünk, van létjogosultsága, *B* típusú dokumentumokból azonban csak dokumentum-visszakereső rendszert lehet és szabad építeni. Egy könyvtári tudásrendszernek – legyen az akár hagyományos, akár MI- vagy RAG-rendszer – különbséget kell tudni tenni *A* és *B* típusú dokumentumok, dokumentumrészletek között, és ennek függvényében kell információ- vagy dokumentum-visszakereső rendszerként működni.

Minden bizonnyal kiválóan lehet amellett érvelni, hogy az *A* típusú információszolgáltatás előnyére válhat egy könyvtári tudásrendszernek, de joggal gondolhatja azt is bárki, hogy

24 idézi: Hjørland, 2026:4

az adott szakterület (ornitológia, turizmus) domain-specifikus tudásrendszere legalább annyira alkalmas eszköz lenne erre. Az azonban, hogy az *M* típusú metainformáció-szolgáltatás számára aligha találhatunk megfelelőbb, autentikus helyet egy könyvtári tudásrendszernél, azt gondolom, nehezen vitatható.

## 5. Konklúzió

A RAG-rendszerek több fontos tekintetben rokonságot mutatnak az otlet–mooersi-i vízió információ-visszakereső rendszereivel, hiszen

- gondosan kiválasztott dokumentumokból dolgoznak
- a *dokumentumokban lévő információ*ra alapozva, s
- közvetlen, tényszerű választ adnak
- a dokumentumok szövegének átfogalmazásával.

Nem véletlen, hogy a könyvtári informatika egyik legmeghatározóbb trendje napjainkban a RAG-technológia integrálása, s hogy a RAG-technológiára épülő könyvtári tudásrendszer kérdése a kutatási és fejlesztési törekvések fókuszába áll.[25] A RAG-rendszerek működésének – akár csak gondolatkísérlet révén történő tanulmányozása segítséget jelenthet a könyvtári információ-visszakeresés lehetőségeinek, korlátainak jobb megértéséhez is.

## Felhasznált források


Bevara, R. V. K. – Lund, B. D. – Mannuru, N. R. – Karedla, S. P. – Mohammed, Y. – Kolapudi, S. T. – Mannuru, A. (2025). Prospects of Retrieval Augmented Generation (RAG) for Academic Library Search and Retrieval. *Information Technology and Libraries*, *44*(2). https://doi.org/10.5860/ital.v44i2.17361

Bowles, M. (1999). The Information Wars: Two Cultures and the Conflict in Information Retrieval, 1945-1999. In: Bowden, M. E. – Hahn, T.B. – Williams, R.V. (eds.). *History and Heritage of Science Information Systems*. Medford: Information Today, 155-166.

Capurro, R. – Hjørland, B. (2003). The Concept of Information. *Annual Review of Information Science and Technology.* 37(1), 343–411. https://doi.org/10.1002/aris.1440370109

Csík T. (2017). Az információtudomány létrejötte és Horváth Tibor információtudományi eszméinek gyökerei. *Könyvtári Figyelő*. 27(1), 56–67.

Fairthorne, R. A. (1967). Morphology of „Information Flow". *Journal of the Association for Computing Machinery*. 14(4), 710–719. https://doi.org/10.1145/321420.321430

Fekete I. (1971). *Kele*. Budapest: Móra.

Floridi, L. (2011). The Philosophy of Information. Oxford, Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199232383.001.0001


25 Ld. pl. Bevara–Lund–Mannuru–Karedla–Mohammed–Kolapudi–Mannuru, 2025

Frohmann, B. (2017). The Role of Facts in Paul Otlet's Modernist Project of Documentation. In: Rayward, W. Boyd (ed). *European Modernism and the Information Society.* Aldershot: Routledge, 75-88. https://doi.org/10.4324/9781315580951-5

Gao, Y. – Xiong, Y. – Gao, X. – Jia, K. – Pan, J. – Bi, Y. – Dai, Y. – Sun, J. – Guo, Q. – Wang, M. – Wang, H. (2023). Retrieval-Augmented Generation for Large Language Models: A Survey. *ArXiv*, https://doi.org/abs/2312.10997

Hjørland, B. (2026). Information Retrieval or Document Retrieval? *Journal of Association for Information Science and Technology. 7*(5), 714-726. https://doi.org/10.1002/asi.70057

Kline, R. (2004). What Is Information Theory a Theory of? In: Rayward, W. B. – Bowden, M. E. (eds). *The History and Heritage of Scientific and Technical Information Systems*. Medford: Information Today, 15–28.

Mooers, C. N. (1960). The Next Twenty Years in Information Retrieval. *American Documentation*. 11(3), 229–236. https://doi.org/10.1002/asi.5090110306

Otlet, P. (2021). Traité de documentation. La Plaine-Saint-Denis: Éditions des maisons des sciences de l'homme associées. https://doi.org/10.4000/books.emsha.482.

Pogányné R. G. (2007). Törekvések a világ tudásának egyetemes számbavételére Paul Otlet és Henri La Fontaine munkásságában. *Könyvtári Figyelő*. 53(1), 55–73.

Rayward, W. B. (1990). Introduction. In: Otlet, Paul (1990), *International Organisation and Dissemination of Knowledge: Selected Essays of Paul Otlet*. Amsterdam and New York: Elsevier, 1–10.

Rayward, W. B. (1994). Visions of Xanadu: Paul Otlet (1868–1944) and Hypertext. *JASIST*. 45(4), 235–250. https://doi.org/10.1002/(SICI)1097-4571(199405)45:4<235::AID-ASI2>3.0.CO;2-Y

Sághi B. – Dávid K. (2006). *Spanyolország*. Budapest: Panemex-Grafó.

Stock, W. G. – Stock, M. (2013). *Handbook of Information Science*. Berlin: De Gruyter Saur. https://doi.org/10.1515/9783110235005

Svensson, L. – Grant, P. J. (2005). Madárhatározó. *Európa és Magyarország legátfogóbb terepi határozója*. Budapest: Park.

Száraz M. Gy. (2003). !Ó, Santo Domingo! Budapest: Magyar Könyvklub.

van Rijsbergen, C. J. (1979). *Information Retrieval*. London: Butterworths. Forrás: http://www.dcs.gla.ac.uMKeith/Preface.html

Zins, Ch. (2007). Conceptual Approaches for Defining Data, Information, and Knowledge: Research Articles. *Journal of the American Society for Information Science and Technology*. 58(4), 479–493. https://doi.org/10.1002/asi.20508

DOI: 10.31915/NWS.2026.29



# OAIS A GYAKORLATBAN

# OAIS IN PRACTICE

*Görög Dániel*
*MNMKK OSZK*
*gorog.daniel.aron@oszk.hu*

## Absztrakt

Az OAIS (Open Archival Information System) a hosszú távú megőrzés nemzetközileg elfogadott szabványosítási kerete. A 2012-es kiadás magyar fordítása révén hazai szakmai körökben is ismert, legalábbis az elméleti része. Az MNMKK OSZK a digitális dokumentumok megőrzése kapcsán használja az OAIS szabványt.

A tanulmány bemutatja – részben a Digital Preservation Capability Maturity Model (DPCMM) szempontrendszer segítségével – hogy az OAIS szabvány hogyan és mennyiben valósul meg az OSZK digitális megőrzési gyakorlatában. A vizsgálat reflektál arra, hogy a már meghonosított vagy a megőrzési rendszerünk folyamatban levő cseréje révén bevezetendő gyakorlatok mennyiben OAIS-komptatibilisek, illetve milyen területen vannak még elmaradások a szabványhoz képest. A tanulmány a másik oldalról az OAIS szabvány megvalósíthatóságát is bemutatja valós megőrzési munkán keresztül.

A tanulmány kitér az OAIS szabvány 2024. decemberében megjelent kiadásában található újdonságokra, melyek a magyar közösség számára még kevéssé ismertek, főként a Preservation objectives fogalmára, illetve a Preservation Description Information és a Content Data kapcsolatára.

**Kulcsszavak:** OAIS, digitális megőrzés, hosszú távú megőrzés, OSZK

## Abstract

The OAIS (Open Archival Information System) is the internationally accepted standard in the long-term preservation field. Its 2012 edition is already translated into Hungarian and is reasonably known, at least as a theory. The National Széchényi Library (NSZL) applies the OAIS standard in line with its digital preservation duties.

The paper presents, in part using the Digital Preservation Maturity Model (DCPMM), how and to what extent the digital preservation practices of NSZL are OAIS compliant. The paper reflects on whether the already implemented, or planned and yet to be implemented practices are OAIS compliant; and in which fields do we have shortfalls in

relation to the standard. From another viewpoint, the paper presents the applicability of the OAIS standard through a very real preservation practice.

The paper touches upon the new features in the 2024 december edition of the OAIS standard, that have not been previously discussed in Hungarian: especially the concept of Preservation Objectives; and the relation between Preservation Description Information és a Content Data.

**Keywords:** OAIS, digital preservation, long-term preservation, NSZL

A nyílt archívumi információs rendszer (angol betűszóval: OAIS) a hosszú távú megőrzés nemzetközileg elfogadott szabványa. Maga a szabvány valamennyire már ismert Magyarországon, fordítása is létezik, a Networkshop közössége előtt Bódog András ismertette 2022-ben. Az eddigi feldolgozások azonban jellemzően a szabványra mint elméletre vonatkoztak. Az Országos Széchényi Könyvtár azon kevés hazai intézmény egyike, amelyek tudatosan követi az OAIS szabványt a digitális megőrzés terén. A tanulmány célja tehát, hogy az OSZK példáján keresztül az OAIS szabvány gyakorlati alkalmazását, alkalmazhatóságát mutassa be.

Az OAIS szabvány esszenciáját Bódog András a következő mondatban idézi a szabványból: „Az információnak a célközönség által függetlenül is értelmezhető formában való fenntartása hosszú távon, a hitelességet alátámasztó bizonyítékkal." Az idézett mondatnak minden eleme befolyásolja egy, a szabvány megvalósítására törekvő intézmény működését.

Az első lényegi kifejezés, az információ a megőrzés tárgyát jelöli, az információ megőrzése a cél, nem a hordozóé vagy mindegyik másolaté. A legnyilvánvalóbb gyakorlati következmény, hogy azonos fizikai példányról készült digitális másolatokból egyet őrzünk meg, hiszen további másolatok nem jelentenek többletinformációt. Ha meghatározható, akkor a jobb minőségű másolatot őrizzük meg. Az információ hiánya viszont információ, ezért az eredeti dokumentum részét képező üres lapokat, margókat megőrizzük, így bizonyítható, hogy nem vesztünk el információt.

A második lényegi kifejezés a „célközönség által függetlenül is értelmezhető"-ség. Az extrém példa a phaisztoszi korong, lineáris A írással van írva, amit egyetlen ma élő ember sem tud elolvasni. Itt tehát hiába jól kivehetőek a jelek, a megőrzés sikertelen, hiszen az értelmezhetőség elveszett. De közelebbi példa a János vitézből vett részlet, és akkor gondolkodjunk, hogy mit kell mellé rakni a megőrzési csomagba, hogy érthető legyen. Ha tíz embert megállítunk az utcán, hogy mit csinál, milyen állattal foglalkozik a juhászbojtár? Lehet, hogy a 21. században ez már magyarázatra szorul. Más kultúrában esetleg magyarázni kellhet, hogy a szerelemmel a hőérzetet asszociáljuk, és akkor ezt is be kell tenni a szöveg mellé a megőrzésbe. És akkor ad absurdum, ha a magyar nyelv elfelejtődik, akkor a nyelvtant is be kell tenni a János vitéz szövege mellé, hogy érthető legyen. Ilyen típusú kér-

déseket kellene egy OAIS-alapú megőrzési rendszernek monitoroznia, és ha a célközönség tudásbázisa megváltozik, a megőrzött dokumentumhoz csatolnia. Értelmezést segítő információ utólagos csatolásával az OSZK-ban eddig még nem foglalkoztunk, két okból, egyrészt mert szabványos digitális megőrzéssel négy éve foglalkozunk és nem számítunk rá, hogy a célközönség tudásbázisa ezalatt megváltozott; másrészt bízunk abban, hogy ha a nemzeti könyvtár teljes állományát sikerül digitális megőrzési rendszerbe adni, az önfeltáróként fog működni: teljes nyelvet, metaforakészletet rekonstruálni lehet belőle.

A harmadik kifejezés a hosszú táv, amit a szabvány úgy ír le, hogy a megőrzésben alkalmazott technológiák elavulását figyelembe kell venni. A megőrzésben alkalmazott rendszereknek és fájlformátumoknak tehát olyannak kell lenniük, amelyeket időtállónak gondolunk, vagy más esetben felkészültnek kell lenni a migrálásra. Mivel az OSZK még nem teljes négy éve rendelkezik OAIS szabványt követő digitális megőrzési programmal, ennyi idő alatt a már megőrzésben levő anyagok migrálására eddig még nem volt szükség.

A negyedik kitétel a hitelességet alátámasztó bizonyíték megléte. Ha a hitelességre nincs bizonyíték, nem tudjuk igazolni, hogy a megőrzési rendszerből előhívott dokumentum információtartalma megegyezik az eredetivel. Az állományi dokumentumokat ezért a jelzettel együtt digitalizáljuk és őrizzük meg, míg az állományba nem tartozó könyveket – például a Mikes programból származó könyveket – az egyedi példány azonosítását lehetővé tevő slejfnivel. A digitális megőrzésre előkészítés során mindegyik fájl kap egy checksumot, ami a fájl ujjlenyomataként viselkedik, viszonylag könnyen újra lekérhető, és egyetlen bit megváltozása esetén már jelzi a fájl megváltozását. A cheksum azonossága tehát a fájl azonosságát és így a hitelességet jelenti.

Az OAIS működési elvei közül szeretném említeni a csomagok elkülönítését: a létrehozó felőli, a megőrzési és a szolgáltatási csomagot külön egységként kell tartani, illetve a hozzáférhetőséget kezelni. Az OSZK gyakorlatában a csomag szintű kezelés még nem valósul meg, helyette a bundle vagy köteg rendszert használjuk, tehát az egyes fájlok köteg jelzőt kapnak (például: eredeti, mester stb.), így lehet külön jogokat adni azonos tételhez tartozó fájloknak.

Az eljárásainkat és az eszközeinket folyamatosan fejlesztjük, a következő nagy lépés a megőrzési rendszerként használt Dspace frissítése lesz az 5.5-ös verzióról a 8.0-s verzióra. Az átállással a munkafolyamataink is át fognak alakulni, OAIS-kompatibilis AIP-okat és DIP-eket fogunk gyártani; illetve a folyamatban korábban társítjuk majd a megőrzendő tartalmakhoz a hitelességet bizonyító azonosítókat, így jobban tudjuk majd bizonyítani, hogy maga a megőrzésbe kerülés folyamata nem változtatta meg a tartalmat. Az átállás hátulütője, hogy a szolgálatási képek csak a megőrzési folyamat végén lesznek elérhetők, míg a mostani munkamenetben általában gyorsan, a megőrzés előtt át tudjuk adni a szolgáltatás részére a szükséges tartalmakat.

A digitális megőrzési elveknek megfelelést a Digital Preservation Capability Maturity Model (DPCMM) segítségével lehet mérni. A modell 15 szempontot vizsgál infrastruktúra és szolgáltatások terén. Végül az öt szint valamelyikébe sorolható be az intézmény. Az első szint a névleges szintű digitális megőrzési képességet jelenti, amely nem követ külső standardokat, és nincs külön szervezeti egység megbízva a digitális megőrzési feladatokkal, így a megőrzés sikeressége nem garantált, csak névleges. Az ötös szinthez – egyebek mellett – intézményi fókusz és a technológiai változások proaktív követése szükséges. Az OSZK digitális megőrzési gyakorlata még nem esett át DPCMM-megfeleltetésen, hiszen az elérni kívánt gyakorlat még nem teljesen épült ki. Természetesen DPCMM szerinti értékelést nem végezhet az intézmény képviselője saját intézményén, így itt csak informálisan és részeiben néhány pontot szeretnék kiemelni. Egyes területeken, mint amilyen a fájlformátumok szabványos használata vagy a bevételezés, az OSZK gyakorlata jól dokumentált és támogatott, megfelelhet a modell hármas vagy akár négyes szintjének, azonban olyan területek is vannak, ahol még nem épült ki a kellő szinten az OSZK metódusa: ilyen például a georedundáns tárolás vagy a formátumok megújításának gyakorlata. A DPCMM követése az intézményeknek önellenőrzésre ad lehetőséget, hogy felmérjék, milyen területen szükséges még fejleszteniük digitális megőrzési gyakorlatukat.

Végül szeretnék kitérni arra, hogy az OAIS-nak 2024-ben új kiadása jelent meg, amit tudtommal magyarul még nem mutattak be nyilvánosan. Mennydörgő újdonság nincs a 2012-es kiadáshoz képest, inkább pontosítások és néhány kérdés részletesebb kifejtése történt meg. Ilyen a Preservation Objective (Megőrzési cél) fogalmának bevezetése és a függetlenül érthetőség fogalmának mérhetővé tétele. Az új kiadás specifikálja, hogy a megőrzött információnak nehezen hozzáférhető források bevonása nélkül kell értelmezhetőnek maradnia, és szempontokat ad az érthetőség teszteléséhez: meghatározottság, végrehajthatóság, mérhetőség. Ha tényleg megkérdezek tíz embert az utcán, hogy tudják-e, mivel foglalkozik egy juhászbojtár, az ilyen teszt lenne az érthetőség tekintetében, mert vagy tudják, mit csinál, vagy nem. További kifejtést közöl a 2024-es szabvány az archívumok közötti adatátadásokról, míg a korábbi szabvány migrálás néven említette ezt a kapcsolatot, az új modell inkább handover-nek (átadás-nak) nevezi a jelenséget, és megkülönbözteti a primary illetve supporting (elsődleges és kisegítő) archívum fogalmát.

## Felhasznált irodalom

Bódog A., A nyílt archívumi információs rendszer (OAIS) szabványának honosítása, Valós térben – az online térért: Networkshop 31: országos konferencia, szerkesztette Tick J, Kokas K, Holl A., HUNGARNET Egyesület, Budapest, 2022,

Görög D., Megőrzés és szolgáltatás a nemzeti könyvtár digitális megőrzési gyakorlatában, CeLISR, 2025. https://journals.bme.hu/celisr/article/view/40552/23886 (Utolsó elérés: 2026. 03. 27.)

Kómár É., Bánki Zs. (szerk.) (2019) Fehér Könyv: Módszertani útmutató a közgyűjteményi kulturális örökség digitalizálásához és közzétételéhez, EMMI, Bp.,. Elérhető: https://pulszky.hu/wp-content/uploads/2019/12/feher_konyv.pdf (Utolsó elérés: 2026. 03. 27.)

Reference Model for an Open Archival Information System(OAIS): Recommended Practice (2012), CCSDS, Washington DC.

Reference Model for an Open Archival Information System(OAIS): Recommended Practice (2024), CCSDS, Washington DC.

Rényi M. A. (2023) Repository or preservation system?: The change of perspective of the National Széchényi Library in the field of digital preservation, ppt előadásvázlat.

Űradat- és információközvetítő rendszerek. Nyílt archívumi információs rendszer (OAIS). Referenciamodell (2022), MSZ ISO 14721:2022.

DOI: 10.31915/NWS.2026.30



# SZABVÁNYALAPÚ DIGITÁLIS ARCHÍVUM ÉPÍTÉSE NYÍLT FORRÁSKÓDÚ ESZKÖZÖKKEL – BETEKINTÉS EGY FOLKLÓRARCHÍVUMI IMPLEMENTÁCIÓ FOLYAMATÁBA

# BUILDING A STANDARDS-BASED DIGITAL ARCHIVE WITH OPEN-SOURCE TOOLS: INSIGHTS INTO A FOLKLORE ARCHIVE IMPLEMENTATION PROCESS

*Zagyva Natália*
*Hagyományok Háza*
*zagyva.natalia@hagyomanyokhaza.hu*

**Absztrakt**

A Hagyományok Háza Folklór Archívuma a magyar népzene-, néptánc- és szövegfolklór-kutatás egyik legjelentősebb audiovizuális gyűjteménye, amely számára kritikus fontosságú a rábízott nemzeti kincs hiteles és hosszú távú megőrzése. A tanulmány betekintést nyújt abba a digitális transzformációs folyamatba, amely során az archívum a töredezett, Excel-alapú nyilvántartásoktól eljut egy szabványalapú, integrált és automatizált munkafolyamatokra épülő informatikai rendszerig.

**Kulcsszavak:** kulturális örökség, digitális hosszú távú megőrzés, nyílt forráskódú szoftverek, Archivematica, AtoM, Wikibase, Heratio

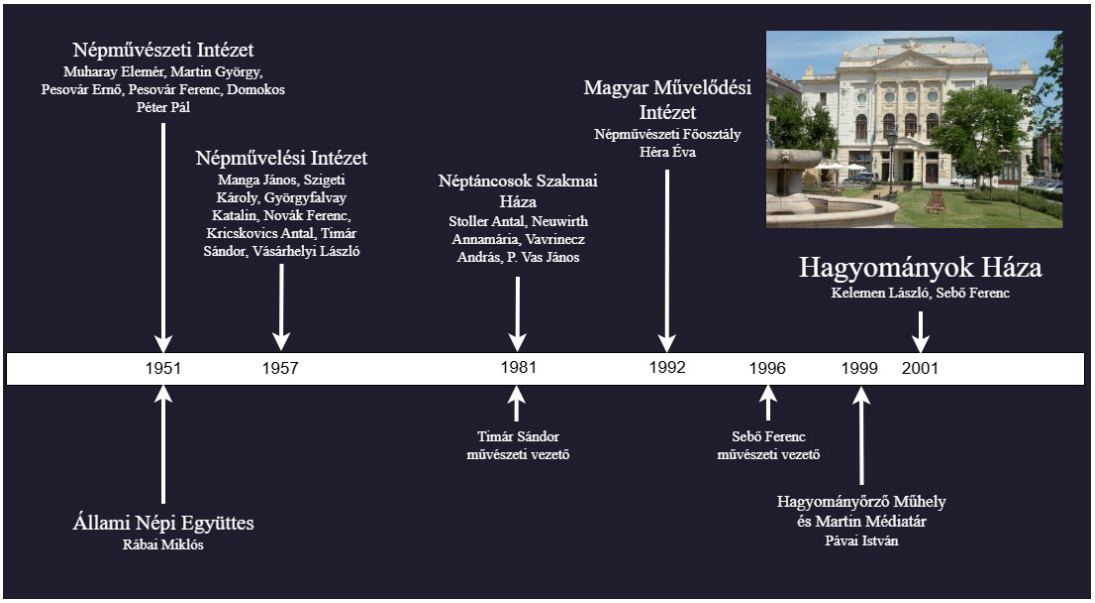
**Abstract**

The Folklore Archive of the Hungarian Heritage House is one of the most significant audiovisual collections supporting research on Hungarian folk music, folk dance, and verbal folklore. Ensuring the reliable long-term preservation and accessibility of this entrusted cultural heritage is therefore of critical importance. This paper provides insight into the digital transformation process through which the archive is transitioning from fragmented Excel-based records to a standards-based integrated information system built on automated and workflow-oriented processes.

**Keywords:** cultural heritage, long-term digital preservation, open-source software, Archivematica, AtoM, Wikibase, Heratio

## Bevezetés

A Hagyományok Háza küldetése „a hagyományos kultúra értékeinek széles körű társadalmi tudatosítása és hozzáférhetővé tétele”[1]. Tevékenységében a művészeti, a közművelődési és a közgyűjteményi terület egyaránt fontos szerepet kap. Közgyűjteményi feladatait a Magyar Népi Iparművészeti Múzeum, a Martin György Szakkönyvtár és a Folklór Archívum keretében valósítja meg.




*1. ábra: A Hagyományok Háza elődintézményei alapításuk sorrendjében, valamint a Folklór Archívum gyűjteményének alakulására meghatározó hatást gyakorló, emblematikus szakembereik*[2]

A Folklór Archívum gyűjtőkörét elsősorban az elődintézmények (1. ábra) és a hetvenes években kibontakozó táncházmozgalom hatásai, igényei és gyűjteményei határozzák meg (Sebő, 2007). A jelenlegi gyűjtemény túlnyomó része népzenét és néptáncot rögzítő audio- és videofelvétel[3]. Bár az archívum elsődleges feladata a folklórdokumentumok megőrzése, katalogizálása és hozzáférhetővé tétele, gyűjtőköre egyre inkább kiterjed a folklorizmustörténeti

1 A Hagyományok Háza víziója: https://hagyomanyokhaza.hu/hu/hagyomanyok-haza (2026.05.09.)

2 Megjegyzések az 1. ábrához: Az 1950-ben szerveződött, 1951-ben bemutatkozó együttest a korabeli híradások már akkor is szórványosan Magyar Állami Népi Együttesnek nevezték, de az intézmény létrehozásáról szóló 79/1951. (III. 31.) MT rendelet szerint a hivatalos megnevezés Állami Népi Együttes (ÁNE). A Folklór Archívum közvetlen előzményének tekinthető Hagyományőrző Műhely és Martin Médiatár 1999-ben Sebő Ferenc kezdeményezésére az ÁNE keretében jött létre. Az ábrán a kiemelkedő hatással bíró, emblematikus munkatársak felsorolása nem teljes körű.

3 A Folklór Archívum forrásállománya 2026 áprilisában: 11.500 audiohordozó, 8350 videohordozó, 73.800 fényképhordozó, 96 folyóméter irat, 17 TB born digital anyag.

dokumentumokra is. A már feldolgozott és részletesen adatolt népzene-, néptánc- és szövegfolklór-gyűjtések válogatott részei a Folklóradatbázisban[4] szabadon elérhetők.

A Hagyományok Háza stratégiai céljainak megfelelően az elmúlt években megkezdődött a Folklór Archívum gyűjteménykezelési és digitális megőrzési folyamatainak átfogó megújítása[5]. A transzformáció célkitűzéseit, a szabványalapú megközelítés elméleti hátterét, a nyílt forráskódú opciók preferálásának indoklását és a szoftverválasztás módszerét egy nemrég megjelent tanulmányban részletesen bemutattam (Zagyva, 2026). Jelen írás a tervezett rendszer gyakorlati megvalósításának előkészítésére fókuszál, különös tekintettel a rendszerarchitektúra kialakítására, az adatmigráció módszertani kérdéseire, valamint a bevezetés során felmerülő főbb technikai és szervezeti kihívásokra. A cikk a jelenleg is zajló fejlesztési folyamatba nyújt betekintést.

## Megoldandó problémák

Bár a Folklóradatbázis létrehozása korát megelőző és nagy jelentőségű lépés volt, mely a gyűjtemények egy részének kiemelkedő részletességű, egységes struktúrában történő metaadatolását és közreadását tette lehetővé, a teljes archívum gyűjteménykezelési gyakorlata decentralizált és heterogén maradt. A gyűjteményi egységek nyilvántartása több ezer, egymástól független Excel-táblában történik, amelyek részben eltérő szerkezetűek, redundáns adatokat tartalmaznak és nem biztosítják az egységes kereshetőséget. Az Excel-táblákban és a Folklóradatbázisban történő párhuzamos adatkezelés növeli a hibalehetőségeket és megnehezíti az adatok konzisztens karbantartását. A megőrzési és szolgáltatási célú digitális állományok előállítása manuális munkafolyamatokra épül, egységes verziókezelés és dokumentált eljárásrend nélkül. A hosszú távú megőrzést szolgáló állományokat – melyeknek összmérete meghaladja a 350 TB-ot – offline adathordozókon tároljuk, ami növeli az adatvesztés kockázatát. A rendszer nem támogat szabványos protokollokon alapuló gépi adatcserét és nem teszi lehetővé külső szolgáltatások integrációját.

## Rendszerterv és -implementáció

Az archívum digitális állományait kezdetektől egy olyan hierarchikus fájlstruktúrában tároljuk, amelynek felső szintjét a forrásarchívumok alkotják és a fájlneveket is ez a hierarchia határozza meg. A gyűjteményi egységek eltérő, különböző mélységű szerkezete, a provenienca elvén nyugvó rendszerezés vezettek ahhoz, hogy az archívumi rekordok nyilvántartásához a rugalmas levéltári rendszerek és szabványok alkalmazhatóságát kezdtük vizsgálni.

4 A Folklóradatbázis weblapja: https://folkloradatbazis.hu (2026.05.09.) 2026 áprilisában 6878 folklóresemény dokumentációja érhető el, 1478 óra videofelvétel és 5041 óra hangfelvétel formájában.

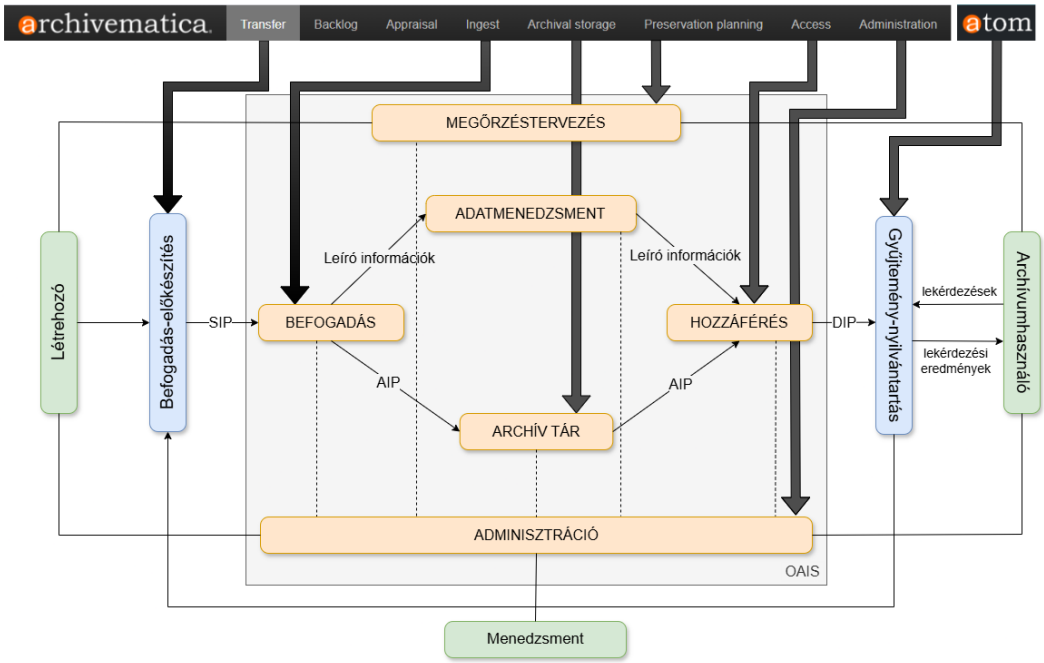
5 A Hagyományok Háza egészének modernizációs terveiről bővebben a Both Miklós főigazgatóval kinevezése után készített interjúban olvashatunk (Csinta 2021).

Mivel a legmodernebb levéltári modellhez, a RiC (Records in Contexts) szabványhoz (ICA 2023) még nem áll rendelkezésre széles körben használható szoftveres implementáció, ezért az ISAD(G) metaadat-szabványra (ICA 2000a) épülő megoldások felé fordultunk. Az alkalmazás során ugyanakkor figyelembe kell venni, hogy a szabvány terminológiája – különösen a magyar fordítás (ICA 2000b) – elsősorban papíralapú iratanyagok leírására készült, ezért az audiovizuális folklórgyűjtemény sajátosságaihoz több ponton terminológiai adaptációra van szükség[6,7].

A megújítás során olyan, az OAIS referenciamodellnek (OAIS, 2024) megfelelő rendszerarchitektúra kialakítása volt a cél, amely lefedi a gyűjteménykezelési, a megőrzési és a hozzáférési feladatokat ellátó funkciókat. A nemzetközi ajánlások és jógyakorlatok áttekintése és a felmerülő nyílt forráskódú megoldások összehasonlítása után a leíró metaadatok kezelésére és a felhasználói hozzáférés biztosítására az AtoM (Access to Memory) rendszert (AtoM 2026), míg a digitális objektumok hosszú távú megőrzésére az Archivematica platformot (Archivematica 2026) választottuk. A két webalapú, nyílt forráskódú rendszer kombinált alkalmazása lehetővé teszi, hogy a megőrzési folyamatok és a publikációs felületek egymástól függetlenül, mégis integrált módon működjenek. A 2. ábra a 2022-ben honosított, magyar szabványnak megfelelő terminológiát (MSZT, 2022) alkalmazva mutatja be az Archivematica és az AtoM együttműködése által támogatott digitális megőrzési és hozzáférési munkafolyamat OAIS kompatibilitását.

6 Példák olyan kifejezésekre, amelyeknél az angol verzió még beleilleszthető egy audiovizuális archívum szóhasználatába, de a magyar verzió általánosan nem alkalmazható: record – irat, item – iratdarab, creator – iratképző, archival description – levéltári leírás, file – ügyirat. A végleges terminológiáról levéltáros szakemberekkel történő egyeztetést követően fogunk dönteni.

7 Bár ötleteink már vannak, az alkalmazandó terminológiáról levéltáros szakemberekkel történő egyeztetést követően fogunk dönteni.




*2. ábra: OAIS-alapú digitális megőrzési és hozzáférési munkafolyamat az Archivematica–AtoM-ökoszisztémában (Zagyva, 2026. 4. ábra)*

A *befogadás-előkészítés* Archivematica által nyújtott transzferfunkcióit kiegészítjük egy egyszerű webes beadófelülettel, hogy a most készülő *born digital* anyagok esetében a *létrehozó* ne csak az Archivematica kezelőfelületéhez hozzáférő és értő archívumi munkatárs lehessen, hanem a felvételek készítője is. A 2. ábrán látható *archívumhasználó* lehet ember (kutató vagy egyszerű érdeklődő), de szoftver is. A későbbiekben például a Folklóradatbázis is innen kapja majd a médiafájlokat és a hozzájuk tartozó alapadatokat.

## Az adatmigráció folyamata

Az új rendszer bevezetésének egyik legkritikusabb eleme a meglévő nyilvántartások migrációja, amely nem csupán technikai feladatként, hanem komplex adatminőségi és adatmodellezési problémaként jelentkezik. A tanulmány készítésekor, 2026 májusában még a migrációs folyamat részleteinek kidolgozása előtt állunk, ezért az alábbiakban csak nagy vonalakban vázolom a terveinket.

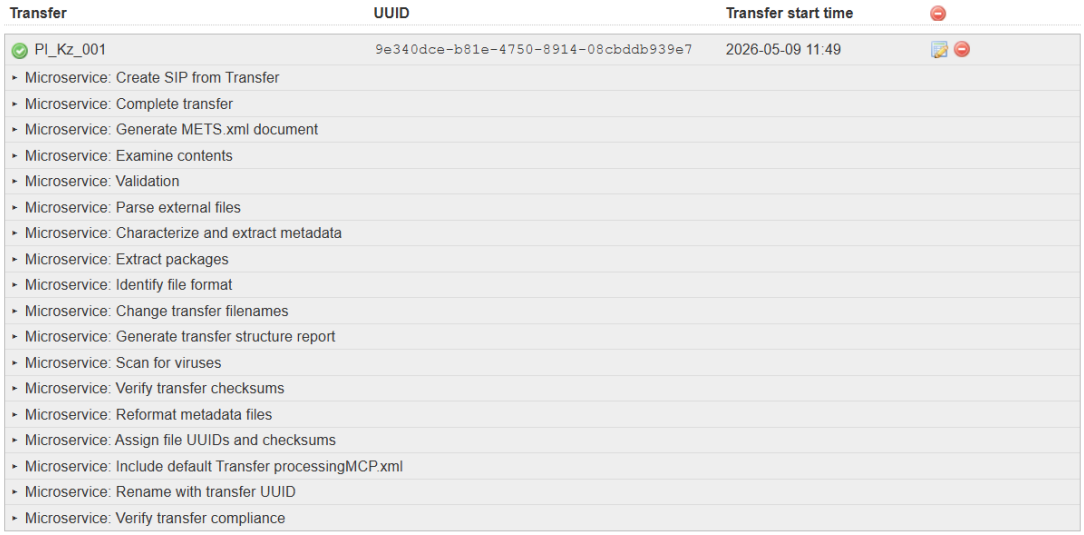
A Folklór Archívumban felhalmozott adatok több évtizedes gyűjtő- és feldolgozómunka eredményeként jöttek létre, különböző célok és munkafolyamatok során, ezért szerkezetük és részletességük jelentős eltéréseket mutat. A heterogén forrásadatok miatt a migráció nem valósítható meg egyetlen lépésben anélkül, hogy jelentős adatvesztés vagy torzulás ne következne be. Ennek kezelésére kétlépcsős migrációs stratégiát alakítottunk ki. A kétlép-

csős megközelítés előnye, hogy csökkenti a migrációs hibák kockázatát és lehetőséget ad az adatok minőségének fokozatos javítására.

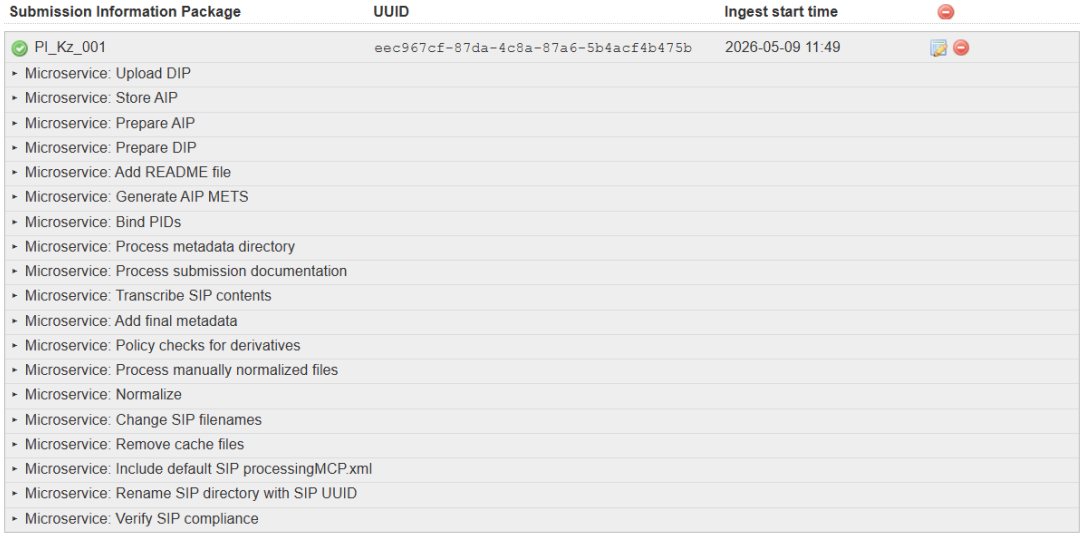
Az első lépés az úgynevezett alapimport, amelynek célja a teljes gyűjteményt lefedő, minimális, de konzisztens adatkészlet betöltése az AtoM rendszerbe. Ebben a fázisban létrehozzuk a teljes gyűjteményhez a hierarchikus archívumi struktúra felső szintjeit (jellemzően fond, állag, sorozat) a megbízhatóan rendelkezésre álló leíró adatokkal. Az alapimport elsődleges célja nem a teljes körű leírás biztosítása, hanem egy stabil, egységes adatstruktúra létrehozása. Ebben a fázisban még nem csatolunk digitális objektumokat az egyes rekordokhoz.

A második lépés az adatgazdagítás, amely során a rekordok adatait gyűjteményi egységenként, ellenőrzött módon egészítjük ki és pontosítjuk. Ez a folyamat már nem automatizált tömegműveletként, hanem szakértői validálással zajlik, lehetővé téve az Excel-táblákból hiányzó adatok pótlását, az inkonzisztenciák feloldását és az egységes metaadat-szerkezet szerinti katalogizálást. Az adatgazdagítás mélysége nem minden gyűjteményi egység esetében azonos: a feldolgozás sorrendjét és részletességét a rendelkezésre álló adatok minősége, a gyűjteményi jelentőség, valamint a megőrzési és szolgáltatási szempontok alapján határozzuk meg. Ebben a lépésben kerül sor az egyes gyűjteményi egységekhez tartozó digitális állományok (médiaállományok, nyilvántartások, megállapodások és egyéb adminisztratív dokumentumok) áttekintésére, rendszerezésére is. Az archívum teljes anyagának feldolgozása a szűkös gyűjteménykezelői kapacitás miatt sok éves munka lesz, de igyekszünk a prioritásokat úgy meghatározni, hogy a hosszú távú megőrzés biztonságának és a szolgáltatási igényeknek a lehető legjobban megfeleljünk.

A digitális objektumok feldolgozása az Archivematica *transfer* folyamatával kezdődik. Miután előkészítjük az adott egységhez tartozó megőrzendő fájlokat és a metaadatokat tartalmazó csv fájlokat, elindítjuk az aktuális igényeinknek megfelelően konfigurált transzferfolyamatot. A 3. ábrán láthatóak azok a mikroszolgáltatások, amelyek lefutása után létrejön az átadási információs csomag (SIP). Az *ingest* (befogadás) folyamat biztosítja a fájlok normalizálását, a technikai és megőrzési metaadatok előállítását, valamint a megőrzési (AIP) és disszeminációs (DIP) információs csomag létrehozását (4. ábra).

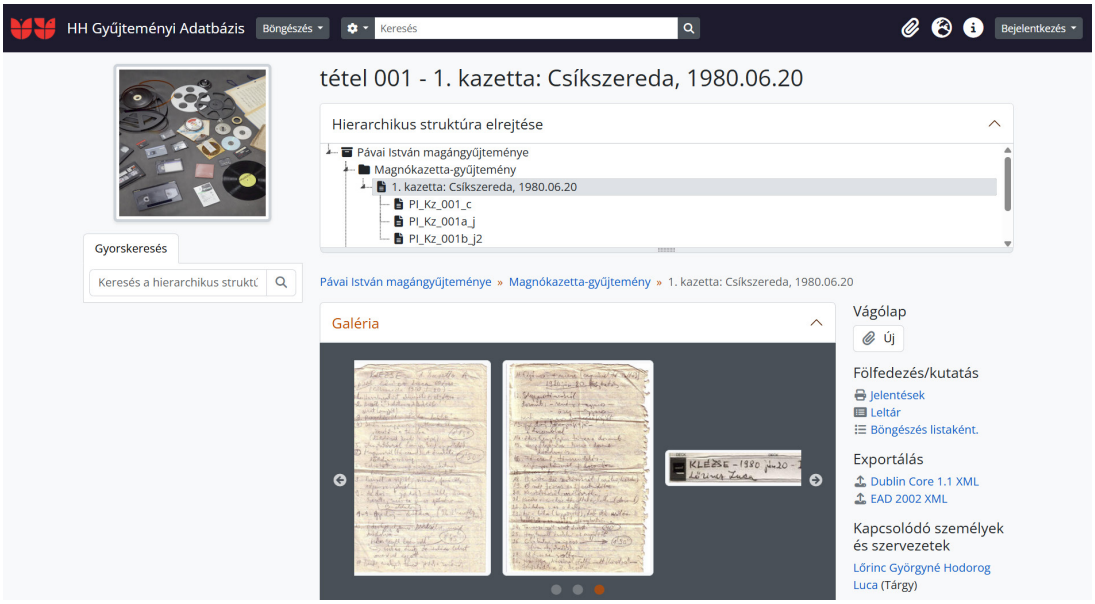



*3. ábra: Az Archivematica transfer folyamatának mikroszolgáltatásai (alulról felfelé haladó időrendben)*




*4. ábra: Az Archivematica ingest folyamatának mikroszolgáltatásai (alulról felfelé haladó időrendben)*

A létrejövő megőrzési információs csomag az archív tárba kerül hosszú távú megőrzésre. A szolgáltatási célú reprezentációk az AtoM rendszerben az előre előkészített hierarchikus struktúra megadott pontján válnak elérhetővé (5. ábra). A két rendszer közötti kapcsolat biztosítja, hogy a felhasználói felületen megjelenő digitális állományok mögött ellenőrzött és szabványos megőrzési folyamat álljon.




*5. ábra: Egy tétel az AtoM felületén az Archivematica-ból átadott és közzétett digitális objektumokkal. A két rendszer közötti kapcsolat biztosítja, hogy a felhasználói felületen megjelenő digitális állományok mögött ellenőrzött és szabványos megőrzési folyamat álljon.*

## További tervek

Az Excel-táblákban a személy- és helynevek egységesítés nélkül, szabad szöveges formában szerepelnek, ami megnehezíti a következetes keresést és az adatok külső forrásokkal (pl. VIAF, Wikidata) való összekapcsolását. A fejlesztés következő mérföldköve a Wikibase szoftver (Wikimedia Deutschland 2026) integrálása lesz. A Wales-i Nemzeti Könyvtár példáját (National Library of Wales 2026) követve létrehozandó Wikibase-alapú, az archívum által kontrollált személyi és földrajzi névtér lehetővé teszi a szemantikus kapcsolódást, kiemeli az archívumot az elszigetelt adatbázisok köréből és bekapcsolja a globális tudáshálózatba.

2026 januárjában az AtoM levelezési listán[8] megjelent egy új, Laravel alapú nyílt forráskódú keretrendszer híre. A Heratio (AHG 2026) az AtoM adatbázis-sémájára és leíró szabványaira épül, de már a teljes GLAM-szektor (galériák, könyvtárak, levéltárak, múzeumok) igényeit célozza. A rendszer megőrzi az AtoM-ökoszisztémával való kompatibilitást, ugyanakkor rugalmasabb architektúrát és korszerűbb adatkezelési megoldásokat kínál. A Heratio egyik legfontosabb újítása a bővíthető, plugin-alapú felépítés, valamint a legmodernebb levéltári szabvány, a RiC támogatása, amely lehetővé teszi a gráfos megjelenítést és a szemantikus kapcsolatokon alapuló feltárást. A rendszer emellett mesterséges intelligenciát alkalmazó funkciókkal támogatja az automatikus entitásfelismerést, a leiratkészítést és a metaadat-kinyerést, valamint kiemelt figyelmet fordít az adatvédelmi megfelelés-

8 Az AtoM levelezőlista: https://groups.google.com/g/ica-atom-users (2026.05.09.)

re is. A Heratio keretrendszer fejlesztését figyelemmel kísérjük és pilot jelleggel vizsgáljuk. Kedvező tapasztalatok esetén a későbbiekben megfontoljuk a bevezetését.

## Összefoglalás

A bemutatott fejlesztési folyamat rávilágít arra, hogy egy szabványalapú digitális archívumi infrastruktúra kialakítása nem kizárólag technológiai feladat, hanem az adatkezelési gyakorlatok és a munkafolyamatok átfogó újragondolását is igényli. Az Archivematica és az AtoM rendszerére épülő nyílt forráskódú ökoszisztéma lehetőséget biztosít a hosszú távon fenntartható, szabványos digitális megőrzési és szolgáltatási környezet kialakítására, ugyanakkor jelentős technikai és szervezeti felkészültséget követel meg.

A fejlesztés egyik legnagyobb kihívását a heterogén, részben strukturálatlan nyilvántartások migrációja jelenti. Az automatizált feldolgozás csak korlátozottan alkalmazható, ezért az adattisztítás és validálás jelentős manuális munkát igényel. Emellett a rendszer bevezetése a munkafolyamatok formalizálását és új kompetenciák[9] kialakítását is szükségessé teszi. A bevezetés szervezeti feltételeinek – így a szükséges munkaerő-kapacitásnak, az átállás ütemezésének és a kollégák továbbképzésének – részletes bemutatása egy következő tanulmány tárgya lehet.

A projekt eddigi eredményei ugyanakkor jól mutatják, hogy a nyílt forráskódú, levéltári szabványokra épülő megközelítés az audiovizuális archívumi környezetben is életképes alternatívát jelenthet. A tervezett további fejlesztések – különösen a szemantikus kapcsolatokra épülő névterek és az új generációs archívumi keretrendszer bevezetése – hosszabb távon egy rugalmasabb és jobban integrálható digitális infrastruktúra kialakítását teszik lehetővé.

## Irodalomjegyzék

AHG (2026) AtoM Heratio – Pure Laravel replacement for AtoM archival software. The Archive and Heritage Group. [online]. Elérhető: https://github.com/ArchiveHeritageGroup/heratio (Utolsó elérés: 2026. 05. 09.)

Archivematica (2026) Open-source Digital Preservation System [online]. Elérhető: https://www.archivematica.org/ (Utolsó elérés: 2026. 05. 09.)

AtoM (2026) Open Source Archival Description Software [online]. Elérhető: https://www.accesstomemory.org/ (Utolsó elérés: 2026. 05. 09.)

The Archive and Heritage Group (2026) AtoM Heratio – Pure Laravel replacement for AtoM archival software [online]. Elérhető: https://github.com/ArchiveHeritageGroup/heratio (Utolsó elérés: 2026. 05. 09.)

9 Az új munkafolyamatok bevezetése megköveteli a levéltári szabványok, a digitális megőrzési modellek, az Archivematica és az AtoM adminisztrációs eszközeinek ismeretét.

Csinta, S. (2021) „Célom a mozgalomból intézményesült rendszer korszerűsítése". Beszélgetés Both Miklóssal, a Hagyományok Háza új főigazgatójával, Folkmagazin, 2021(6). https://lapozo.folkmagazin.hu/mag21_6/?page=34 (Utolsó elérés: 2026. 05. 09.)

ICA (2000a) ISAD(G): General international standard archival description. 2. kiadás. Ottawa: International Council on Archives. Elérhető: https://www.ica.org/app/uploads/2024/01/CBPS_2000_Guidelines_ISADG_Second-edition_EN.pdf (Utolsó elérés: 2026. 05. 09.)

ICA (2000b): ISAD(G): Az Általános Nemzetközi Levéltári Leírás Szabványa. 2. kiadás. Magyar fordítás. Elérhető: https://archivportal.hu/wp-content/uploads/2022/06/ISAD-G.pdf (Utolsó elérés: 2026. 05. 09.)

ICA (2023) Records in Contexts: A Conceptual Model for Archival Description. Version 1.0. International Council on Archives. [online] Elérhető: https://www.ica.org/ica-network/expert-groups/egad/records-in-contexts-ric/ (Utolsó elérés: 2026. 05. 09.)

MSZ ISO 14721:2022 szabvány adatai (2022) [online]. Elérhető: https://ugyintezes.mszt.hu/webaruhaz/szabvany-adatok?standard=140845 (Utolsó elérés: 2026. 05. 09.)

National Library of Wales (2026) Semantic Name Authority Repository Cymru. Linked database of authority records and related data for people and places in Welsh archives. [online] Elérhető: https://snarc-llgc.wikibase.cloud/wiki/Main_Page (Utolsó elérés: 2026. 05. 09.)

OAIS (2024) Reference Model for an Open Archival Information System (OAIS). CCSDS Secretariat. Elérhető: https://ccsds.org/Pubs/650x0m3.pdf (Utolsó elérés: 2026. 05. 09.)

Sebő, F. (2007) A Népművészeti Intézettől a Hagyományok Házáig, Szín, 12(6). https://epa.oszk.hu/01300/01306/00094/pdf/EPA01306_Szin_2007_12_06_december_19-21.pdf (Utolsó elérés: 2026. 05. 09.)

Wikimedia Deutschland (2026) Wikibase. [online] Elérhető: https://wikiba.se/ (Utolsó elérés: 2026. 05. 09.)

Zagyva, N. (2026) A Hagyományok Háza Folklór Archívumának digitális stratégiája: Szabványalapú gyűjteménykezelés és hosszú távú megőrzés nyílt forráskódú szoftverekkel, Central European Library and Information Science Review (CeLISR), 3(1). https://doi.org/10.3311/celisr.43246 (Utolsó elérés: 2026. 05. 09.)