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

NETWORKSHOP 2026
35. Országos Informatikai Konferencia

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

Szerkesztette: Tick József, Kokas Károly, Holl András

HUNGARNET Egyesület
Budapest, 2026

HUN-REN
Magyar Kutatási Hálózat

NETWORKSHOP

Szerkesztette: Tick József, Kokas Károly, Holl András

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

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

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

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

Borítókép: freepik.com

DOI: 10.31915/NWS.2026.22

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

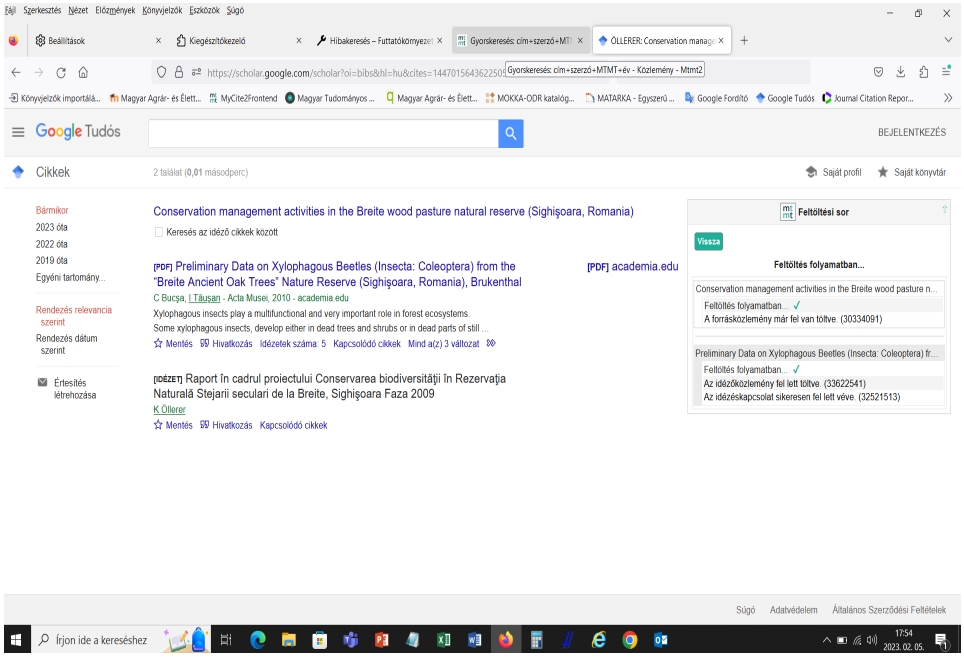
leverage this opportunity, we have developed a command-line–driven bulk citation processing script called Metaxa (Metadata Extractor), whose basic operating principles are presented here.

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/

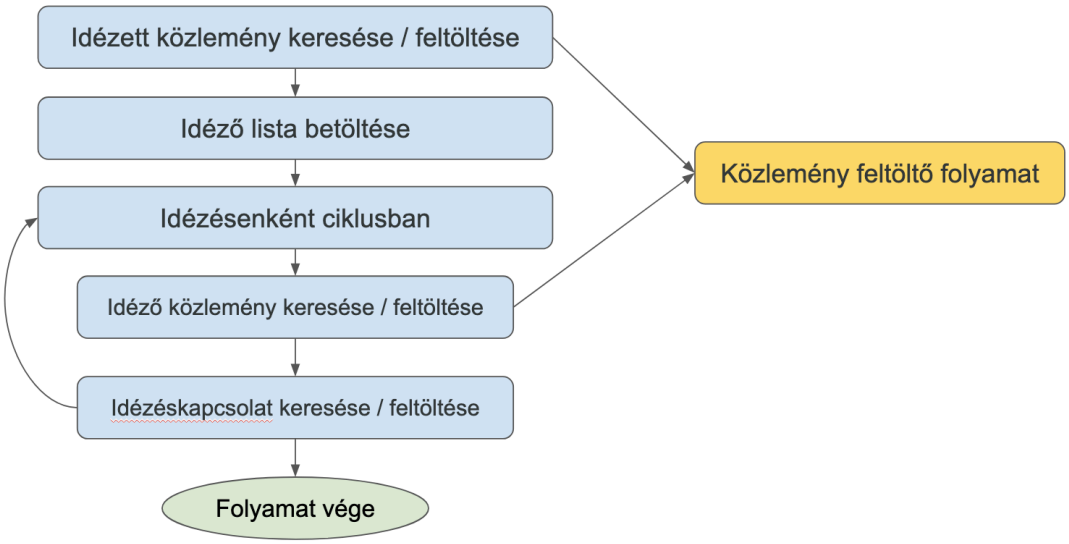
lemény CSL-t ily módon a közlemény HTML-ből jó minőségben elő tudtuk állítani kisebb modellekkel is, mint például a Qwen3.6-27b vagy a gpt-5-nano (2. ábra).

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