Külföldi könyvtár- és információtudományi folyóiratokból (referátumok)

Hubay Miklós
Petőfi Irodalmi Múzeum
humán informatikai csoportvezető

Referenszontológia alkalmazása a szemantikai interoperabilitás előmozdítására

Patrício, H., Nogueira Ramos, P., Cordeiro, M. I. (2025) *Weaving the Threads of Bibliographic Ontologies - Application of a Reference Ontology to Advance Semantic Interoperability*, Information Technology and Libraries, 44(2), p. 1–35.
https://doi.org/10.5860/ital.v44i2.17289

A bibliográfiai ontológiák vagy szótárak a szemantikus web építésének elengedhetetlen eszközei: segítségével a közgyűjtemények elkészíthetik saját tudásgráfjaikat, azaz RDF-logikára, tripletek halmazára konvertálhatják hagyományos formában, pl. MARC-ban létező tudásvagyonukat. A könyvtári terület legismertebb szótárai, az LRM-, és az RDA-szótár, valamint a BIBFRAME azonban, adatmodelljük különbözősége miatt átjárhatósági problémákkal küzdenek, az egyik segítségével készült gráfok nehezen konvertálhatók át a másikra. A cikk szerzői ennek a problémának a megoldását kínálják az ún. Reference Ontology (referenszontológia, RO) megalkotásával. Az ún. mediációs megfeleltetésnek köszönhetően – szemben a transzformatív megoldásokkal, ahol minden egyes forrásontológiát mindegyik másikhoz hozzá kell igazítani – nem szükséges a szóban forgó ontológiák mindegyikét összepárosítani, elegendő csupán az RO-val kidolgozni a megfeleltetést, jelentős munka- és energiamegtakarítást elérve ezzel. A forrásontológiák között fennálló szemantikai kapcsolatok pedig automatikusan kikövetkeztethetők az explicit relációk segítségével.

A szerzők az egyes adatmodellek, továbbá magentitások bemutatása után az Os Lusiadas példáján mutatják be az átjárhatósági problémákat. Ezek egyike, hogy a bibliográfiai források négy legfontosabb entitása – a *mű*, *kifejezési forma*, *megjelenési forma*, illetve *példány* – egymással könnyen megfeleltethető módon, különálló osztályként található meg az LRM, illetve az RDA szótárában, a BIBFRAME azonban nem ennyire granuláris. A Kongresszusi Könyvtár által fejlesztett elemkészlet nem tartalmaz a *kifejezési forma* entitást leképező osztályt, hanem a bf:Work vonatkozik egyszerre a szellemi tartalomra, illetve az annak kifejezéséhez

felhasznált jelölésrendszerre is. Az így keletkező poliszémiát a szerzők úgy oldják meg, hogy a bf:Work alosztályait, amelyek voltaképpen kifejezési formákat írnak le (bf:Text, bf:Audio, bf:Cartography stb.), a referenszontológia orowl:Expressao (kifejezési forma) osztályának, míg a bf:Work-öt magát az orowl:Obra (mű) osztálynak feleltetik meg. Az RO ugyanis, hasonlóan az LRM- és RDA-szótárhoz, négy magentitással dolgozik.

A többértelműségi problémán kívül a cikk szerzői az egyes szótárelemek tulajdonságait elemezve megállapították, hogy az LRM, az RDA és a BIBFRAME szótárban a rész-egész viszonyokat leíró tulajdonságok, mint amilyen a bf:hasPart, nem tranzitív módon vannak definiálva, így például automatikusan nem következtethető ki, hogy egy többkötetes könyv első kötetének első fejezete nem csupán az első kötetnek, hanem a teljes műnek is része. A helyzetet a referenszontológia készítői úgy oldották meg, hogy egy parteDe (része) szupertulajdonságot definiáltak a forrásontológiák relációinak (így például a BIBFRAME-ben található bf:partOf) fölérendeltjeként, amelynek leírásából nem maradt ki a tranzitivitási paraméter sem.

Más problémák megoldására – a bf:language tulajdonság értékkészletének szűkítése, a parteDe tulajdonság non-reflexivitásának beállítása stb. – a szerzők SHACL alaki gráfok készítésével kerestek megoldást, amelyek az egyes tudásgráfok szerkezeti validálására szolgálnak.

A generatív mesterséges intelligencia és az információkeresés

Kalcsó Gyula
kalcso.gyula@oszk.hu
OSZK Digitális Bölcsészeti Központ, Digitális Filológiai és Webarchiválási Osztály
webarchiválási csoportvezető

Atwood, G. S., Swogger, S. E. (2025) *Generative AI and the Myth of Universal Search: Implications for Librarians*, Journal of Electronic Resources in Medical Libraries, 22(1–2), p. 9–12. https://doi.org/10.1080/15424065.2025.2496615

A cikk a generatív mesterséges intelligenciával kapcsolatos túlzott elvárások és félreértések hatását vizsgálja az információkeresésre és a könyvtárak jövőjére. Az utóbbi években a médiában olyan benyomás alakult ki, hogy a nagy nyelvi modellek „mindent meg tudnak keresni", mintha univerzális, határtalan hozzáférésük lenne a világ összes adatához. A szerzők rámutatnak, hogy ez a kép téves, és hasonló félelmeket idéz elő a könyvtárosokban, mint a Google 2004-es megjelenésekor.

A technológiai cégek marketingüzenetei – például az EXA Labs videója, amely a „világ összes adatát" kínálja a felhasználói számára – tovább erősítik a generatív MI körüli mítoszt. A valóság azonban az, hogy az MI-modellek nem férnek hozzá minden információhoz, különösen nem a fizetőfal mögötti, szerzői joggal védett vagy nem digitalizált tartalmakhoz. Ahogyan a Google esetében is kiderült, a keresőmotor korántsem „mindentudó": a tudományos irodalom nagy része nem elérhető számára, a találatok zajosak, nem átláthatók és nem reprodukálhatók – ez tette a Google-t alkalmatlanná a precíz, bizonyítékalapú keresési feladatokra, például a szisztematikus irodalomkutatásra.

A generatív MI esetében ugyanezek a korlátok ismét megjelennek. A tartalomtulajdonosok egy része (például a News Corp) partnerséget alakított ki az AI-cégekkel, mások (például az Elsevier) saját MI-eszközöket fejlesztenek, hogy fenntartsák az erőforrások feletti ellenőrzésüket. Eközben egyre több kiadó blokkolja az MI-modellek adathozzáférését, ami még jobban fragmentálja az információs környezetet. Mindez azt jelenti, hogy az „univerzális keresés" mítosza továbbra sem valósul meg: a felhasználók előreláthatólag továbbra is különböző „adatsilókban" keresnek majd, egymástól eltérő, saját MI-vel rendelkező rendszerekben.

A szerzők szerint ez a helyzet hosszú távon nem fenyegeti a könyvtárak és könyvtárosok szükségességét. Sőt, a szakemberek szerepe valószínűleg még fontosabbá válik. A gyűjteményszervezőknek továbbra is sokféle szolgáltató tartalmát és szerződéseit kell kezelniük, a könyvtárosoknak meg kell tanítaniuk a felhasználókat a megfelelő források használatára, az elektronikus forrásokért felelős könyvtárosok pedig továbbra is kulcsszerepet játszanak az eltérő rendszerek összekapcsolásában. A generatív MI-eszközök megjelenése tehát nem eltörli, hanem átalakítja ezeket a feladatokat.

A cikk felveti a lehetséges jövőbeli irányokat is. Például elképzelhető, hogy AI-ügynökök kapnak majd hozzáférést egyetemek vagy könyvtárak rendszereihez, és önállóan keresnek vagy ellenőrzik a hivatkozásokat. Ugyanakkor komoly kérdések merülnek fel: vajon a szolgáltatók engedélyeznék-e az ilyen hozzáférést? Mennyire lenne biztonságos? Javítaná-e a keresések minőségét, vagy inkább átláthatatlanná tenné őket? A szerzők szerint bár ezek izgalmas lehetőségek, egyelőre számos technikai, jogi és etikai akadály áll előttük.

A tanulmány végkövetkeztetése szerint a generatív MI-eszközök továbbra is meghatározóak lesznek az információkeresésben, de nem valósítják meg az univerzális keresés álmát, és nem váltják ki a könyvtárosokat. A szakma feladata továbbra is a komplex, töredezett információs ökoszisztéma kezelése és az MI-integráció felelősségteljes irányítása marad.

Bubnó Katalin
bubno.katalin@oszk.hu
OSZK Könyvtári Intézet,
Könyvtártudományi Szakkönyvtár
osztályvezető

Relevancia nélküli discovery rétegek? Három közismert könyvtári keresőeszköz vizsgálata és bizonyítékokon alapuló értékelése

Szpunar, R., Bradley, E., Gabrielson, E., Pellegrino, P. (2025) *Irrelevant discovery layers*, Information Technology and Libraries, 44(2). https://doi.org/10.5860/ital.v44i2.17266

A tudományos könyvtárak az elmúlt másfél évtizedben széles körben bevezettek ún. discovery rétegeket (*discovery layers*) – magyar terminológia szerint egyablakos keresőeszközöket – annak érdekében, hogy megfeleljenek a modern felhasználók elvárásainak. Az ilyen rendszerek célja, hogy a könyvtár teljes gyűjteményére kiterjedő, Google-szerű keresési élményt nyújtsanak a felhasználóknak, lehetővé téve a hozzáférést a katalógushoz, az e-forrásokhoz és a különböző

adatbázisokhoz egyetlen felületről. Az ilyen típusú eszközök gyors elterjedése ellenére a közelmúltban kritikák is felmerültek a teljesítményüket és korlátaikat illetően, különös tekintettel a kapott találatok relevanciájára.

A Ruth Szpunar és szerzőtársai által végzett kutatás célja éppen e kritika alátámasztása vagy megcáfolása volt. A tanulmány bizonyítékokon alapuló értékelést hajtott végre, amely azt vizsgálta, hogy egy közkedveltnek számító discovery a legrelevánsabb találatokat produkálja-e a felhasználói kulcsszavas keresésekre, összehasonlítva két másik, széles körben hozzáférhető tudományos keresőeszközzel. A vizsgálat így közvetlen összehasonlítást nyújt a könyvtári rendszerek, a szakterületi adatbázisok és a szabadon hozzáférhető webes keresők között.

A kutatók a következő három keresőeszközt hasonlították össze a felhasználók által keresett tipikusnak mondható kulcsszavak segítségével:

1. WorldCat Discovery (WCD): Egy központi indexszel rendelkező discovery, amely a vizsgált intézmények könyvtári rendszerét reprezentálja. Ez az eszköz a könyvtárak által leginkább népszerűsített és ismert keresőfelület.
2. Academic Search Complete (ASC): Egy előfizetéses, interdiszciplináris indexelő és referáló adatbázis. Ez a típusú adatbázis hagyományosan a lektorált tudományos irodalom elsődleges forrása volt a discoveryk megjelenése előtt.
3. Google Scholar (GS): Egy szabadon hozzáférhető akadémiai webes keresőmotor.

A tanulmány kvalitatív módszert alkalmazott. A kutatók részletes értékelési szempontokat dolgoztak ki, amelyet több értékelő (a tanulmány szerzői) is alkalmazott a keresési eredmények relevancia szerinti megítélésére. A szempontok a következők voltak:

- Frissesség (*Currency*): A találat megjelenési dátuma, mennyire aktuális az adott téma szempontjából.
- Relevancia (*Relevancy*): A találat tartalmi kapcsolódása a keresési kulcsszavakhoz.
- Közelség (*Proximity*): A kulcsszavak elhelyezkedése a találatokban.
- Autoritás (*Authority*): A forrás hitelessége, a publikáció típusa és minősége.

Az értékelési folyamat során a fenti kritériumok megsértése esetén büntetőpontokat is adtak, így képződött egy pontszám az egyes keresőeszközök teljesítményének mutatójaként.

A kutatás megállapításai egyértelműen rávilágítottak a különböző keresőeszközök erősségeire és gyengeségeire, és megkérdőjelezték a discoveryk dominanciáját a tudományos keresésben.

Az Academic Search Complete (ASC) kapta a legmagasabb összesített pontszámot. Ez azt jelzi, hogy a szakterületi indexelő és absztrakt adatbázis a vizsgált négy kritérium (frissesség, relevancia, közelség, autoritás) tekintetében összességében a legjobb minőségű találatokat biztosította.

A WorldCat Discovery (WCD) kapta a legalacsonyabb összesített pontszámot. Ez a legfontosabb megállapítás megerősíti a discoverykkel kapcsolatos kritikát: a Google-szerű keresési élmény ígérete ellenére a WCD kevésbé releváns és minőségi eredményeket produkált, mint a hagyományos adatbázis vagy a nyílt akadémiai kereső.

A részletes kritériumok szerinti bontás további árnyalatokat mutatott. Az Academic Search Complete (ASC) nyújtotta a legfrissebb és a leghitelesebb forrásokat. Ez megerősíti a szakterületi indexelő adatbázisok értékét, mivel azok szigorúbb kritériumok szerint válogatják a tartalmat, biztosítva a magasabb tudományos minőséget. Ezzel szemben a Google Scholar (GS) nyújtotta a legre-

levánsabb forrásokat. Ez a megállapítás rámutat a Google Scholar algoritmikus erősségére, amely gyakran jobban értelmezi a felhasználó szándékát a kulcsszavakból, és tágabb merítésének köszönhetően gyakran a legpontosabb tartalmi egyezést találja meg.

A tanulmány eredményei fontos tanulságokkal szolgálnak az akadémiai könyvtárak számára. A discoveryk, bár kényelmes, *egylépéses* keresést kínálnak, nem feltétlenül biztosítják a legmagasabb minőségű vagy leghitelesebb találatokat, és összességében alulmaradnak mind a hagyományos szakterületi adatbázisokkal, mind a Google Scholarral szemben.

A szerzők következtetése szerint a könyvtáraknak a jövőre nézve újra kell gondolniuk a discovery eszközök szerepét, figyelembe véve azok előnyeit és költségeit. A tiszta felhasználói felület és az egyszerű hozzáférés, mint előny, nem feltétlenül ellensúlyozza a gyengébb relevanciát és autoritást a professzionális tudományos keresés szempontjából.

A tanulmány utal arra, hogy a könyvtáraknak érdemes lehet továbbra is hangsúlyt fektetniük a szakterületi indexelő s referáló adatbázisok használatának oktatására, mivel ezek a rendszerek bizonyítottan jobb minőségi mutatókkal rendelkeznek a tudományos eredmények szempontjából. Ezen felül érdemes megvizsgálni, hogyan lehetne a Google Scholar *relevancia* erejét a könyvtári rendszerekbe integrálni, vagy a felhasználókat ösztönözni a Google Scholar és a könyvtári linkelési szolgáltatások együttes használatára.

Összességében a kutatás egyértelmű bizonyítékokon alapuló kritikát fogalmaz meg a discoverykkel szemben, és arra ösztönzi a könyvtárosokat, hogy stratégiai döntéseik meghozatalakor tartsák szem előtt a keresőeszközök valós, objektív teljesítményét a felhasználóknak nyújtott tudományos érték szempontjából.

Kalcsó Gyula
kalcso.gyula@oszk.hu
OSZK Digitális Bölcsészeti Központ, Digitális Filológiai és Webarchiválási Osztály
webarchiválási csoportvezető

Az adatlekéréssel kiegészített generálás (RAG) integrálása az akadémiai könyvtári kereső- és lekérdezőrendszerekbe

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

A tanulmány azt vizsgálja, hogy miként integrálható a „Retrieval Augmented Generation" (RAG), azaz adatlekéréssel kiegészített generálás a tudományos könyvtári kereső- és lekérdezőrendszerekbe, és milyen hatással lehet ez a folyamatokra. A szerzők abból indulnak ki, hogy a tudományos könyvtárak hagyományos keresőrendszerei (kulcsszavas keresés, metaadat-alapú lekérdezés) ma már nem mindig felelnek meg a kutatók és diákok igényeinek, különösen interdiszciplináris vizsgálódások, komplex igények esetén.

A RAG olyan hibrid modell, amely a nagy nyelvi modellek (LLM-ek) generálási képességeit kombinálja strukturált és gyakran vektor- vagy beágyazásalapú visszakereséssel, így képes lehet a felhasználói kérdés (természetes nyelvű lekérdezés) és a könyvtári tudásbázis közötti szakadék áthidalására.

A tanulmány részletesen bemutatja, mi is a RAG: olyan eljárás, amelynél a nagy nyelvi modell a lekérdezés után először releváns dokumentum- vagy tudáselemeket gyűjt (retrieval), majd ezekre alapozva generál választ vagy összegzést (generation). A nyelvi modellek önmagukban csak a tréningadatukban szereplő tudásra hagyatkoznak, és így elavultak vagy pontatlanok lehetnek újabb témák esetén. Az RAG segít bepótolni ezt a hiányt azáltal, hogy aktuális, a lekérdezés szempontjából releváns dokumentumokat kapcsol be a folyamatba.

A szerzők hangsúlyozzák, hogy a tudományos könyvtárak kiváló alapot jelentenek egy RAG rendszer számára, mert strukturált metaadatokkal rendelkeznek (pl. MARC-, RDA-szabványok) amelyek segítik a dokumentumok indexelését és felhasználását a RAG-ban. Fontos továbbá, hogy a tartalmak hitelesek, lektoráltak, így a visszakeresett információ minősége magasabb, mint az internetes keresési találatoké. Ez azt jelenti, hogy ha egy tudományos könyvtár megfelelően integrálja a RAG-ot, akkor javulhat a keresési találatok relevanciája, pontossága és a felhasználói élmény.

A szerzők külön fejezetben foglalkoznak azzal, hogy milyen technikai komponensek kellenek a RAG-rendszer bevezetéséhez egy könyvtári környezetben: szólnak a dokumentumbeágyazásról, a vektoradatbázisokról, a köztes szoftverek (middleware) és API-k integrációjáról, az adatvédelmi és szerzői jogi szempontokról. A tanulmány rámutat arra, hogy a technológiai infrastruktúra mellett fontos a skálázhatóság, karbantarthatóság és költséghatékonyság kérdése is – mivel sok könyvtár korlátozott erőforrásokkal rendelkezik.

A tanulmány áttekinti, hogy miként változhat meg a felhasználói keresési élmény a RAG bevezetése által. A felhasználók nemcsak kulcsszavakkal kereshetnek, hanem teljes, természetes nyelvű kérdéseket is feltehetnek. A válaszok kontextusérzékenyek és személyre szabottak, a modell képes figyelembe venni a felhasználó korábbi interakcióit, a tématerületet, s így relevánsabb találatokat adni.

A szerzők következtetése, hogy a RAG új lehetőségeket kínál az akadémiai könyvtári keresés és lekérdezés fejlődésében – lehetővé teszi a szemantikus, kontextusérzékeny keresést, a természetes nyelvű interakciót, és egy magasabb szintű felhasználói élményt. Ugyanakkor a sikeres bevezetéshez meg kell tervezni az infrastruktúrát, gondoskodni kell az adatvédelmi és szerzői jogi megfelelésről, valamint a felhasználói bizalom kiépítéséről. A jövőbeli kutatási irányok közé tartozik a skálázhatóság vizsgálata, költséghatékony megoldások kidolgozása, valamint az etikai és átláthatósági kérdések mélyebb feltárása.

Szőnyegi Zsanett
szonyegi.zsanett@oszk.hu
OSZK Könyvtári Intézet, Könyvtári Szak- és Továbbképzési Osztály
oktatásszervező

Generatív mesterséges intelligencia egyetemisták számára: közösen fejlesztett online tananyag a generatív MI-ről az egyetemi képzésben

Willenborg, A., Withorn, T. (2025) *Generative AI for College Students: A Collaboratively Developed Online Microcourse on GenAI in the College Classroom*, Communications in Information Literacy, 19(1), p. 113–130. https://doi.org/10.15760/comminfolit.2025.19.1.7

Az egyetemi könyvtárak eltérő módon reagálnak a generatív mesterséges intelligencia (továbbiakban: GenAI) felsőoktatásban történő megjelenésére. Sok intézményben nincs egységes irányelv, az oktatók kevés szakmai útmutatást kapnak arra vonatkozóan, hogyan alkalmazzák a GenAI-t a tanórákon. A könyvtárosok, az egyetemi íráskészség-fejlesztő központok munkatársai és a digitális médiával foglalkozó munkatársak segítségével olyan innovatív, hatrészes mikrotanfolyamot dolgoztak ki, amely a GenAI működését, korlátait, valamint etikus és hatékony felhasználását mutatja be. A mikrotanulás lehetőséget ad arra, hogy több szakértő tudását integrálják a tananyagba, miközben a tanfolyam rugalmasan illeszkedik az oktatók időkereteihez és könnyedén beilleszthető bármelyik egyetemen oktatott kurzusba.

A projekt a Louisville-i Egyetemen valósult meg, több részlegének – az egyetemi könyvtár (Ekstrom Library), az íráskészség-fejlesztő központ (the University Writing Center) és a digitális médialabor (the Digital Media Suite) – együttes közreműködésével. A tanfolyam rövid, könnyen feldolgozható tanegységekre épül, melyeket aszinkron módon, online formában sajátíthatnak el a hallgatók. Az egymástól függetlenül is elvégezhető leckék kb. 15–20 percesek, és a témák magunkban foglalják a GenAI működését és korlátait. A projekt kidolgozása előtt a kollégák részt vettek az ACRL (Association of College and Research Libraries – Amerikai Felsőoktatási és Tudományos Könyvtárak Egyesülete) szakmai továbbképzésén –, ez iránymutatást adott a folyamat kidolgozásához. Az alábbi hat lecke készült el:

1. lecke: Mit jelent a GenAI és hogyan működik?
2. lecke: Milyen kihívásokkal kell szembenéznie a GenAI-nek?
3. lecke: Hogyan használhatjuk fel a GenAI által nyújtott lehetőségeket?
4. lecke: GenAI és az egyetemi kutatás
5. lecke: GenAI és a tudományos írás
6. lecke: GenAI és a felsőoktatás etikája a Louisville-i egyetemen

A Blackboard Ultra rendszerben több részből álló tanulási modulokat hoznak létre az oktatók, amely több tanulási elemet is tartalmaz, például: teszteket, feladatokat, űrlapokat, hivatkozásokat. Mindegyik lecke négy részből áll: előzetes tudásmérő (a hallgatók előzetes ismereteit térképezik fel); videó (videó formájá-

ban hallgathatnak előadást az adott témáról); tevékenység (feladatok elvégzése a tanultak alapján); reflexió (a tanulási folyamat értékelése). A mikrokurzus indulása előtt a projektben részt vevő partnerek elkészítették a marketinganyagokat, valamint a tanszékek tájékoztatására egy LibGuide-ot hoztak létre, amely bemutatta a kurzus célját, tartalmát és elérhetőségét.

A mikrokurzuson résztvevő oktatóknak és hallgatóknak így lehetősége nyílik kritikai gondolkodásuk fejlesztésére, miközben elsajátíthatják a mesterséges intelligencia használatával kapcsolatos alapvető ismereteket.

Egy digitális megőrzési (pilot) program indítása

Rényi Mátyás
renyi.matyas@oszk.hu
OSZK, Gyűjteménymegőrzési Főosztály, Digitális Megőrzési Osztály
osztályvezető

Hoffman, K. (2025) *Starting up a Digital Preservation (Pilot) Program*, Information Technology and Libraries, 44(2). https://doi.org/10.5860/ital.v44i2.17452

A tanulmány a New York állambeli Hamilton College könyvtára 2024-ben elindított digitális megőrzési pilotprogramjának tapasztalatait mutatja be. A szerző Kim Hoffman, az intézmény első digitális megőrzési feladatokat ellátó könyvtárosa, így a nulláról kényszerült felépíteni egy rendszert, amely alkalmas a hosszú távú digitális megőrzés ellátására. A fő cél egy olyan, aktív megőrzést biztosító eszköz bevezetése volt, amely túlmutat a korábban alkalmazott, pusztán passzív tárolási gyakorlaton.

A Hamilton College digitális gyűjteményei gyorsan növekednek, és már nem csak papíralapú digitalizált dokumentumokat tartalmaznak, hanem egyedi, sérülékeny, illetve avuló hordozókon érkező digitális állományokat is. Emiatt sürgetővé vált a megfelelő eszközök, szabályzatok (policy) és munkafolyamatok kialakítása. A választásuk az Archivematica nevű, nyílt forráskódú digitális megőrzési platformra esett, mivel a könyvtár technikai háttere képes támogatni egy ilyen rendszer működtetését.

Bár a megőrzési területtel kapcsolatos irodalom bőséges, a megvalósításhoz hasznos sorvezetők és implementációk bemutatására alig akad példa. A szerző hat lépésben foglalja össze, hogyan érdemes egy ilyen pilot projektet megszervezni:

Eszközválasztás: A lehetőségek, igények, költségvetés és intézményi kompetenciák felmérése nélkülözhetetlen. A Hamilton esetében az Archivematica volt az egyetlen komolyan vizsgált opció.

Érdekelt felek bevonása: A projekt sikeréhez szükség volt informatikai szakemberekre, költségvetési döntéshozókra, gyűjteményi szakértőkre és magára a megőrzési könyvtárosra. Fontosnak bizonyult az elköteleződés biztosítása és a felelősségek rögzítése.

Konkrét feladatok kidolgozása: A csapat meghatározta, milyen célokra szeretnék használni a rendszert:

- sérült vagy hiányzó fájl visszaállítása,
- formátummigráció,
- fájlok integritásellenőrzése (checksum),
- metaadatok feltöltése és kezelése,
- korlátozott hozzáférés biztosítása az embargós dokumentumokhoz.

Visszakérdezés időszaka: A pilot során felmerülő kérdések megfogalmazása és tisztázásnak szakasza:

- Összekeverhetünk nem nyilvános és nyilvános archívumok részét képező dokumentumokat?
- Minden dokumentum az Archivematicába kerül?
- Mi történik, ha úgy döntenek, mást rendszert használnak, a fájlok ingeritásellenőrzése kompatibilis lesz más rendszerekkel?
- Ki férhet hozzá az Archivematicához? Jogosultsági kérdések.
- Tárolási architektúrával kapcsolatos kérdések megfogalmazása.

Munkafolyamatok megtervezése: A szerző folyamatábrákat készített, hogy vizuálisan láthatóvá tegye a digitális dokumentumok útját, honnan érkeznek a fájlok, hogyan kerülnek előkészítésre és ingestálásra, és hol lesznek a Megőrzési Információs Csomagok (AIP-ok) eltárolva.

Elvárási küszöb meghatározása: A csapatnak fel kellett ismernie, hogy a pilot sikeréhez nem feltétlenül szükséges az eredetileg tervezett rendszerek közötti integráció. A legfontosabb az, hogy bevezethető legyen egy működő digitális megőrzési rendszer.

A projekt ugyan nem a tervezett módon ment végbe, de Hoffman szerint így is komoly előrelépés történt. A pilot tapasztalatai egyértelműsítették, hogy az Archivematica megfelel az intézmény igényeinek, és megkezdődhet az átállás a hosszú távú megőrzésre szánt digitális anyagok feldolgozására. A szerző visszatekintve úgy látja, hogy a strukturáltabb, jobban dokumentált projektirányítás meggyorsíthatta volna a folyamatot, de az így is értékes tanulságokkal szolgálhat más intézmények számára.

Kalcsó Gyula
kalcso.gyula@oszk.hu
OSZK Digitális Bölcsészeti Központ, Digitális Filológiai és Webarchiválási Osztály
webarchiválási csoportvezető

Webarhívumok metaadatainak automatikus generálása a GPT-4o generatív nyelvi modellel

Nair, A., Goh, Z. R., Liu, T., Huang, A. Y. (2025) *Web Archives Metadata Generation with GPT-4o: Challenges and Insights*, Information Technology and Libraries, 44(2).
https://doi.org/10.5860/ital.v44i2.17305

A tanulmány szerzői azt vizsgálják, hogy a GPT-4o generatív nyelvi modell hogyan alkalmazható webarhívumok metaadatainak automatikus generálására. Fő motivációjuk az, hogy a webarchívumok (a WARC-fájlok) hatalmas mennyiségű, strukturálatlan adatot tartalmaznak, és a metaadatok kézi előállítása rendkívül munkaigényes, idő- és erőforrás-igényes folyamat.
A cikk egy prototípusrendszert mutat be, amely GPT-4o modellt használ arra, hogy a webarchívum tartalmát értelmezve releváns metaadatokat generáljon. A módszer lényege, hogy a weboldalakat HTML vagy szöveges formátumban betáplálják a modellbe, amely ebből olyan metaadatmezőket állít elő, mint például a cím,

leírás, témakörök vagy a kulcsszavak. A kutatás során a szerzők tesztelték a modell pontosságát, hatékonyságát és skálázhatóságát – azaz megnézték, hogyan teljesít nagyobb archívumokon, és milyen erőforrásigénye van a rendszernek.

A GPT-4o jól teljesít a metaadatok generálásában: releváns és jól használható adatmezőket képes előállítani. Ugyanakkor a modell nem hibátlan: előfordulnak pontatlan vagy félreértelmezett metaadatok is, különösen, ha az eredeti weboldal bonyolult szerkezetű vagy sok rajta a dinamikus tartalom.

Az automatizált metaadat-generálás jelentősen csökkentheti az emberi munkaigényt az archívumoknál, ami költséghatékonyabb működést tesz lehetővé. Ugyanakkor maga a generálási folyamat számításigényes: a modell futtatása jelentős erőforrásokat igényel, és az infrastruktúrát megfelelően kell beállítani, ha nagyobb mennyiségű adathalmazt kell feldolgozni.

A szerzők kiemelik az olyan etikai és minőségbiztosítási kérdéseket, mint például a generált metaadatok hitelessége és megbízhatósága. Ha a modell téves információkat generál, az ronthatja az archívum használhatóságát, az automatizáció ellenére is szükség lehet emberi felülvizsgálatra bizonyos metaadatmezők esetében.

A kutatás rávilágít arra, hogy a GPT-4o és hasonló generatív modellek komoly potenciált jelenthetnek a digitális archiválás területén. Az automatizált metaadat-generálás felgyorsíthatja a webarchívumok katalogizálását, és segíthet abban, hogy ezek az archívumok könnyebben kutathatóak és visszakereshetőek legyenek. Ez különösen fontos lehet olyan intézmények számára, amelyek nagy mennyiségű webes tartalmat archiválnak (pl. nemzeti könyvtárak, kutatóközpontok), de korlátozott emberi erőforrással dolgoznak.

A tanulmány szerzői több problémára is rámutatnak, és javasolják további vizsgálatukat. Szükséges lehet a modell által generált metaadatok felülvizsgálata és javítása emberi munkával, minőségbiztosítás céljából. Az infrastruktúra skálázhatóságának optimalizálása fontos kérdés, hogy a rendszer nagy archívumok esetén is fenntartható maradjon. Felhívják a figyelmet a generált metaadatok etikai és jogi dimenzióinak vizsgálatára is, különösen tekintettel a szerzői jogi kérdésekre vagy adatvédelmi szempontokra. Kiemelik, hogy szükséges lehet alternatív modellek és beállítások tesztelése is, amelyek olcsóbbak lehetnek, de mégis elfogadható pontosságot nyújtanak.

A *Web Archives Metadata Generation with GPT-4o* című cikk ígéretes irányt mutat a webarchívumok automatikus feldolgozása szempontjából: a GPT-4o segítségével automatizálhatóvá válhat a metaadatok előállítása, ami jelentősen felgyorsítja és olcsóbbá teheti a webarchívumok katalogizálását. Ugyanakkor a megközelítés még nem teljesen kiforrott – a pontosság, skálázhatóság és etikai megbízhatóság terén is vannak még megoldandó feladatok. A tanulmány ezért nemcsak egy működő prototípust mutat be, hanem alapot szolgáltat további kutatásokhoz és fejlesztésekhez is.