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.2026.27

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