

Szemantikus adatgazdagítás példái a kulturális intézményekben

Examples for Semantic Enrichment in the GLAM Sector

Mátyás Melinda¹ 

¹ Monguz Információtechnológiai Kft.
ELTE Eötvös József Collegium
mmatyas@monguz.hu

Beérkezett: 2026. 05. 14. / Elfogadva: 2026. 06. 04. / Publikálva: 2026. 06. 30.

Tanulmányomban¹ a szemantikus adatgazdagításról és a módszereinek csoportosításáról szólok a közgyűjtemények világában. A szemantikus adatgazdagítás az adatgazdagításnak egy speciális formája, amely során megbízható helyekről származó linkekkel egészítik ki a meglevő forrásokat. Az Europeana meghatározása szerint a szemantikus adatgazdagítás az a folyamat, amely során kontrollált adatcsomagokból (például névterekből, online lexikonokból) vagy értékkészletszótárakból származó linkeket kapcsolnak a meglevő metaadatokhoz. Az egyes formátípusok bemutatásánál Marcia Lei Zeng, a téma egyik legkorábbi kutatójának csoportosítását használok (strukturált, félig strukturált és strukturálatlan adatok), valamint bemutatok néhány eszközt, valamint hazai és nemzetközi példákat. A tanulmány végén röviden arra is kitérek, hogy hogyan lehet felhasználni a mesterséges intelligenciát a meglevő adatgazdagítási eszközök mellett.

szemantikus adatgazdagítás, RDA, metaadatok, wikidata, kapcsolt adatok

My study¹ focuses on semantic enrichment within the GLAM sector. Semantic enrichment is a specialized form of data enrichment where data is enhanced with links to reliable external sources. According to the Europeana definition, semantic enrichment is a method in which data from controlled vocabularies (e.g., namespaces, online lexicons) is linked to existing metadata. In demonstrating these methods, I follow the framework established by Marcia Lei Zeng, one of the pioneers in this field. According to her, enrichment can be applied to structured, semi-structured, and unstructured data. I will present several tools and examples for these methods from both Hungarian and international GLAM institutions. Finally, I will briefly discuss how artificial intelligence can be integrated into semantic enrichment processes alongside existing tools.

semantic enrichment, linked data, metadata, wikidata

1. Bevezető: kulcsfogalmak és szakirodalom-vizsgálat

Tanulmányomban az adatgazdagítást tág értelemben vizsgálom. Azt a folyamatot értem adatgazdagítás alatt, amelynek során a kulturális intézmények a meglevő adataikhoz (akár leíró, akár audiovizuális) új, releváns adatokat, metaadatokat kapcsolnak kézzel és/vagy informatikai eszközökkel. Az adatgazdagításnak egy speciális formája a szemantikus adatgazdagítás, amikor megbízható helyekről származó linkekkel egészítik ki a meglevő forrásokat. A megbízható helyeket többféleképpen is meghatározhatjuk: informatikai szempontból nézve stabilan működő, hosszú távra tervezett szolgáltatást nyújtó helyek, szakmailag pedig ellenőrzött, általában kulturális intézmények által kezelt források, például helynevekről vagy személynevekről, amelyek rendelkeznek a kapcsolat kialakításához állandó linkekkel (URI). Az Europeana² meghatározása szerint³ a szemantikus adatgazdagítás az a folyamat, amelynek során kontrollált adatcsomagokból (például névterekből, online lexikonokból) vagy értékkészletszótárakból származó linkeket kapcsolnak a meglevő metaadatokhoz. A dolgozatom címében szereplő kulturális

intézmények alatt elsősorban a könyvtárakat, a múzeumokat és a levéltárakat értem (angolul a LAM Libraries, Archives, and Museums), és ilyen típusú intézményekből ismertetek példákat az adatgazdagítás tényleges megvalósítására.

Az adatgazdagítás nemcsak az adott közgyűjteménynek hasznos. A kulturális örökségi adatok a hitelességük kapcsán a digitális bölcsészeti kutatásokhoz is alapvető fontosságúak. Az értéknövelt adatok a kutatóknak is előnyösek, hiszen nő a visszakereshetőség, és nagyobb hangsúly kerül az adatok megosztására, hogy ezzel további kutatási lehetőségeket tegyenek lehetővé. A feladatra a gyűjteménytől és az erőforrásoktól függően különböző eszközöket használnak.

Az adatgazdagítás szemantikus formája szorosan kötődik a szemantikus webhez is, mivel a folyamat végeredményeként létrejövő metaadathoz kapcsolt linkek a szemantikus web alapját képezik. A szemantikus web Tim Berners-Lee meghatározása szerint a jelenlegi web kiterjesztése, amelyben az információt jól körülhatárolt leírásokkal látják el annak érdekében, hogy a gépek számára is értelmezhető legyen (Berners-Lee et al., 2001, p. 37). Az információ gép által értelmezhető egyik formája a LOD (Linked Open Data), amely strukturált és szabadon újrahazsnálható, géppel olvasható adat, s az információs rendszerek is fel tudják dolgozni, valamint felhasználni arra, hogy szemantikusan releváns kapcsolatokat hozzanak létre más LOD-adatcsomagokkal (Baca, Gill, 2015, p. 237). Az adatok közötti kapcsolat kialakításának egyik módja a Resource Description Framework (RDF) felhasználása, amely egy szabványos modell a webes adatcserére (RDF – Semantic Web Standards⁴). Az RDF egységes erőforrás-azonosítókat (Uniform Resource Identifier, URI) használ arra, hogy meghatározza az erőforrások nevét és helyét, valamint kifejezze az entitásokat és a közöttük levő kapcsolatokat ún. tripletek vagy hármasok formájában (alany, állítmány, tárgy kapcsolat), s ezzel építi fel az információ szemantikus hálózatát (Baca, Gill, 2015, p. 238). A könyvtári területen az International Federation of Library Associations and Institutions (IFLA) által 1998-ban létrehozott konceptuális modell a könyvtári adatbázisok új adatstruktúrájának kialakítására, a Functional Requirements for Bibliographic Records (FRBR) (OCLC: FRBR) is rendelkezik RDF-alapú kapcsolatleírással, ezért a könyvtáraktól, közgyűjteményektől sem áll távol az RDF alkalmazása. Ma ezt a modellt már felváltotta az IFLA LRM⁵ konceptuális modellje az adatok leírására. Az eredetileg FRBR-re és az FRAD-re (Functional Requirements for Authority Data) (Hamburger, 2011) épülő, digitális világ igényeihez igazított könyvtári szabvány az RDA (Resource Description and Access), amely egy rugalmas keretrendszer, a digitálisan és hagyományos úton keletkezett források leírására olyan módon, hogy a felhasználóközpontú kapcsolt adatokat állítja a középpontba. Megjegyzendő, hogy a 3R projekt (RDA Toolkit Restructure and Redesign, 2020) és az RDA-tartalom IFLA Library Reference Modell összehangolása (LRM) (Mering, 2019, p. 66) nagy lépést jelentett ahhoz, hogy ne teljesen különböző utakon fejlődjön a két meghatározó keretrendszer.

Az egyik könyvtári szférában elterjedt RDF-alapú adatmodell a BIBFRAME, amelyet a MARC-formátum (machine-readable cataloging) leváltására szántak, mivel a kapcsolt adatok elvein alapul, amelynek segítségével a bibliográfiai adatok jobban hasznosíthatók a könyvtáron belül és kívül is ("Librarianship studies & information technology: BIBFRAME (Bibliographic Framework)").

A szabályok és keretrendszerek ezen alkalmazásai is azt mutatják, hogy a kulturális intézmények törekednek arra, hogy kifejezzék adataik rejtett szemantikus potenciálját azáltal, hogy kapcsolatokat hoznak létre az adatbázisaikon belül és más hasonló kulturális intézmények adataival is oly módon, hogy az adataikat leírják szemantikus formában, és nyílt hozzáféréssel közzé is teszik (Daquino et al., 2017, 21:2). Az effajta közzététel abban is segít, hogy új lehetőségeket biztosít a kutatóknak az adatok felhasználására, valamint a felhasználóknak a böngészéshez és a kereséshez (Daquino et al., 2017, 21:2).

Az utóbbi évtizedekben a digitalizálás és online közzététel által nőtt ezen gyűjtemények hozzáférhetősége, de a közgyűjteményi adatok minőségének és megtalálhatóságának növelése érdekében a szemantikus gazdagítás is fontos stratégiává vált (Zeng, 2019, p. 2). A szemantikus webes átalakulás ösztönzésére már 2013-ban megjelent Mirna Willer és Gordon Dunsire könyve a bibliográfiai információk szervezéséről a szemantikus webes környezetben (Willer, Dunsire, 2013). A műben igyekeznek bemutatni a szemantikus webes technológiákat és hangsúlyozni, hogy az új technológiák alkalmazása miért válna a könyvtárak hasznára.

Magyarországon sajnos nem igazán terjedt el az adatgazdagítás a közgyűjteményekben. Véleményem szerint főként az intézmények informatikai területen mutatkozó erőforráshiánya, illetve az e területen jártas könyvtárosok alacsony száma miatt. Magyar nyelven egyelőre kevés gyakorlati alkalmazást találunk.

Az adatgazdagítással kapcsolatos joggyakorlatokkal a közgyűjtemények mellett a társadalom- és bölcsészettudományi kutatóközpontokban és kutatócsoportokban találkozhatunk, például a Bölcsészettudományi Kutatóközpont Irodalomtudományi

Intézetében (ma ELTE HTK Irodalomtudományi Kutatóintézet) és az ELTE Humanizmus Kelet-Közép-Európában Kutatócsoport munkásságában, amely műhelyekből szintén hozok példát.

A következőkben az adatgazdagításról és az adatgazdagítási módszerekről szólok általánosan, majd a szemantikus formájáról írok részletesebben.

2. Adatgazdagítási módszerek csoportosítása

Marcia Lei Zeng, a Kent University munkatársa az adatokat középpontba állítva határozta meg kategóriáit a szemantikus adatgazdagítás bemutatására (2019, 2):

- Strukturált adatok: tudásszervező rendszerek és szótárak, LOD használatával növelhető a metaadat minősége és megtalálhatósága.
- Félig strukturált adatok: átalakíthatók strukturálttá entitásalapú szemantikus elemzéssel és annotációval, ezzel is növelve a hozzáférési pontok számát.
- Strukturálatlan adatok: adatbányászattal és tudásbázisok létrehozásával emelhető a megtalálhatóságuk.

2.1. Kapcsolt linkekkel való kiegészítés, szemantikus adatgazdagítás

Névterekben, tudástárakban és egyéb adatsomagokban található linkek kapcsolása, akár gépi (pl. OpenRefine), akár kézi eszközökkel, a meglevő adatokhoz vagy tudásbázisok (pl. Getty Vocabularies) linkelése – ez a módszer főleg a strukturált adatok gazdagítására alkalmazható, mivel itt a legformalizáltabb a kiinduló adat, főként, ha authority adatokat⁶ egészítenek ki. A kapcsolt linkek funkciója, hogy elősegítsék a közgyűjteményi adatok megtalálhatóságát és növeljék az adatminőséget, továbbá, hogy több lehetőséget biztosítsanak a többnyelvű keresésre és több információt nyújtsanak egy-egy authority adatról (például személynév esetén a szerző további műveihez, életrajzához is, illetve akár egy kapcsolódó gyűjtemény tételéhez is eljuthassunk) (Zeng, 2019, p. 7).

Az Europeana a kulturális örökség objektumai többnyelvű módon való kereshetőségét szintén kontrollált szótárak segítségével teszi lehetővé (Stiller, 2024). A felhasználható szótáraknak van egy külön listája, az Europeana dereferenceable vocabularies (Vocabularies Supported by Europeana⁷), mely tartalmazza a jelenleg támogatott szótárakat, például a Getty szótárát, a német Gemeinsame Normdatei (GND), a Wikidata és a Geonames, valamint listázza az elfogadás alatt álló szótárakat. Itt elérhető például az Europeana Eagle8 (Vocabularies. EAGLE Portal⁸) anyag és objektum típus szótára, amely az ókori görög és latin feliratok leírásához biztosít kontrollált szótárakat. A közgyűjteményeknek a támogatott szótárak listája igen hasznos forrást nyújthat az adatgazdagítás tervezésében. Az Europeana azzal is támogat, hogy elérhetővé teszi a saját rendszeresen frissülő útmutatóját az automatikus szemantikus adatgazdagításokról (Europeana Pro: Europeana Semantic Enrichment⁹).

A félig strukturált adatok szemantikus gazdagítására is van lehetőség, ehhez először szükséges a nem formalizált módon tárolt adatok elemzése, hogy a bennük található elemek beazonosításával egyértelművé váljon melyek alakíthatók hozzáférési ponttá. Ilyen félig strukturált források lehetnek a könyvtári leírásokban bizonyos MARC-mezők, egy fotógyűjtemény leírása vagy a levéltári segédanyagok szövege (Zeng, 2019, p. 13). Az adatok azonosítása különböző módszerekkel történhet. Az egyik lehetőség a szemantikus analízis, amelyhez több informatikai megoldás is a rendelkezésünkre áll. Használhatjuk többek között az OpenCalais¹⁰ vagy a Cogito Intelligence API¹¹ alkalmazásokat, amelyek többféle szakterületi tudás szervezési rendszerét és taxonómiáját használják a gépi tanulás és a mesterséges intelligencia vívmányaival kiegészítve (Zeng, 2019, p. 19).

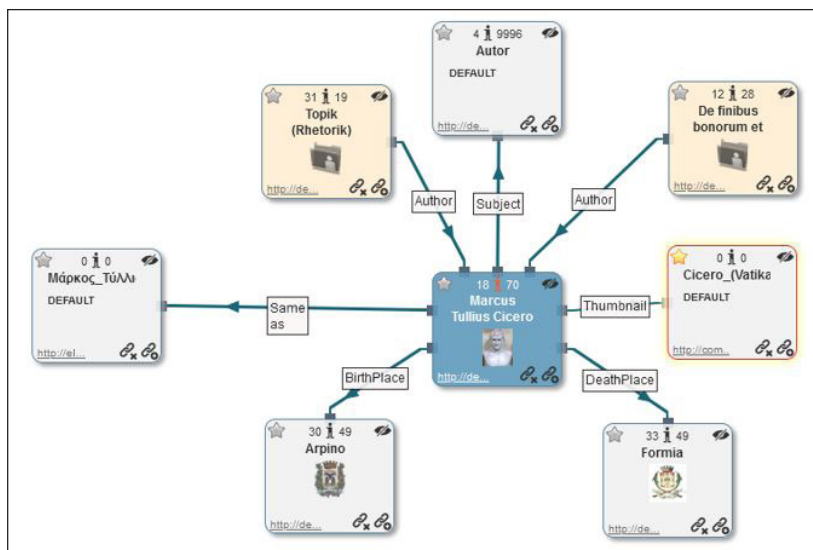
A szolgáltatások működési elve, hogy beazonosítják a szövegben többek között a tulajdonneveket (Named Entity Recognition, rövidítve NER) a természetes nyelvi feldolgozás (Natural Language Processing, rövidítve NLP) segítségével, bizonyos algoritmusok felhasználásával. A feldolgozás során ezeket az azonosított elemeket általában ontológiákba szervezik (Hinze et al., 2019, p. 311). A közgyűjtemény munkatársai a beazonosított kapcsolatokat vissza tudják építeni linkeként az eredeti adatbázisukba.

Az OpenRefine¹² program is hasonló eszköz, amelynek segítségével például megoldható, hogy a megnyitott táblázatos fájlban a kijelölt oszlopokon lefuttassuk az ún. Reconciliation (egyeztetés keresése) műveletet, amellyel az oszlop adataihoz többek között Wikidata-forrást keres, majd ellenőrzés és jóváhagyás után ki is egészíti a Wikidata-linkkel a táblázatunkat.

A strukturálatlan adatok esetében megkülönböztetjük a szövegalapú strukturálatlan adatokat. Ezeknél az adatgazdagításhoz a címkézés, szemantikus analízis, kivonatolás és adatbányászat lehet hatékony megoldás (Zeng, 2019, p. 20). A nem szövegalapú, strukturálatlan adatok adatgazdagításában a szemantikus annotáció és az ontológia alapú tudásbázisok felhasználása vezethet eredményre. Ezen eszközökre példák a Pundit¹³, a Mirador¹⁴, a Brat Rapid Annotation Tool¹⁵, a Recogito¹⁶ és a MapHub¹⁷ (Zeng, 2019, p. 20). Az eszközök kiválasztása attól függ, hogy milyen típusú strukturálatlan adatot szeretnénk gazdagítani: például képeket annotálni, videó- vagy hangfelvételeket ellátni kulcsszavakkal vagy tartalomjegyzékkel vagy akár szkennelt térképeket bővíteni szemantikusan kapcsolt adatokkal.

2.1.1. A DBpedia bemutatása és egy példa adatvizualizációs eszköz ismertetése

A szemantikus adatgazdagítás illusztrálására jó példa a DBpedia¹⁸ tudásgráf, amelyben a Wikipédia tartalmát jelenítik meg (rendszeresen frissítve) szemantikus kapcsolatokkal ellátva. A tudásgráfban 2015-ben 37 millió RDF-link mutatott több mint 30 különböző adatforrásra (Lehmann et al., 2015). A DBpedia szemantikus hálójára építve olyan adatvizualizációs megoldásokat lehet készíteni, mint például a HUN-REN Számítástechnikai és Automatizálási Kutatóintézet LODmilla¹⁹ eszköze, amelyen egy csomópontot (node) kiválasztva vizualizálhatók a hozzá vezető vagy tőle kifelé mutató kapcsolatok. Példámban Marcus Tullius Cicerót és néhány kapcsolatát tettem fel az adatmegjelenítő felületre (1. ábra). Az elkészített vizualizáció lementhető, megosztható. Az eredeti DBpedia-adatokat a 2. ábra szemlélteti, amely tartalmazza a DBpedia saját metaadatelemléteit is, amely a vizualizáción a csomópontokat összekötő élek (edge) elnevezéseiként látható. A LODmilla közel 10 éve készült, ami látszik is a felületen – első ránézésre nem tűnik modernnek –, viszont véleményem szerint a tudásgráf szemléltetésére napjainkban is egy jól használható eszköz (Micsik, Turbucz, Tóth, 2015). A bevezetőben említett linked data-alkalmazásokat összefoglaló cikk is bemutatja a LODmillát a csomópont – kapcsolatok vizualizációs eszközök között. A szerzők kiemelik, hogy azáltal, hogy az adatokat nem jeleníti meg egyből (a felhasználóknak kell rátenni azokat a felületre), megkönnyíti az információk felfedezését és a mélyebben levő kapcsolatok felderítését (Bernasconi et al., 2023, p. 11).



1. ábra: Cicero - SZTAKI LODmilla

 DBpedia Browse using • Formats • Faceted Browser Spinal Endpoint	
 Following Caesar's death, Cicero became an enemy of Mark Antony in the ensuing power struggle, attacking him in a series of speeches. He was proscribed as an enemy of the state by the Second Triumvirate and consequently executed by soldiers operating on their behalf in 43 BC having been intercepted during an attempted flight from the Italian peninsula. His severed hands and head were then, as a final revenge of Mark Antony, displayed on the Rostra. Petrarch's rediscovery of Cicero's letters is often credited for initiating the 14th-century Renaissance in public affairs, humanism, and classical Roman culture. According to Polish historian Józef Zieliński, "the Renaissance was above all things a revival of Cicero, and only after him and through him of the rest of Classical antiquity." The peak of Cicero's authority and prestige came during the 18th-century Enlightenment, and his impact on leading Enlightenment thinkers and political theorists such as John Locke, David Hume, Montesquieu and Edmund Burke was substantial. His works rank among the most influential in global culture, and today still constitute one of the most important bodies of primary material for the writing and revision of Roman history, especially the last days of the Roman Republic. (en)	
 <code>db:birthDate</code>	 -106-01-03 (xsd:date)
 <code>db:birthPlace</code>	 - db:Roman_Italy - db:Roman_Republic - db:Arpino
 <code>db:birthYear</code>	 -106-01-01 (xsd:gYear)
 <code>db:deathDate</code>	 -043-12-07 (xsd:date)

2. ábra: Cicero - DBpedia

2.1.2. Szemantikus adatgazdagítás az Europeanán

A 2008-ban elérhetővé tett Europeana-gyűjteményben már 2011-től elkezdődött egy hároméves adatgazdagítási projekt azzal a céllal, hogy a természettudományi anyagok szélesebb köréhez biztosítson hozzáférést, és összekapcsolja azokat a kulturális örökség területével (OpenUp! Network: Opening up the Natural History Heritage for Europeana, 2024; Koch, Koch 2017, p. 172). A program során a Pentaho Kettle²⁰ szolgáltatást használták a különböző forrásokból érkező heterogén adatok egységesítésére és az ETL (Extract, Transform, Load = kivonatolás, átalakítás és betöltés) feladatokra. Az átalakítást követő ellenőrzés után OAI-PMH-val importálták az adatokat a felhasználáshoz. Az Europeanán így olyan köznevekkel, címkékkel gazdagított kollekciók jöttek létre a természettudományos forrásokból, amelyek a szélesebb közönség (például turisták, diákok) számára is hozzáférhetőbbé tették az adatokat (Koch, Koch 2017, p. 172–73).

Napjainkra az OpenUp Hálózat az Europeana természettudományi aggregátorává vált, és folyamatosan bővíti az adatok körét, illetve várja az újabb, természettudományhoz kapcsolódó intézmények csatlakozását. Jelenleg többnyelvű szakterületi szótár elemeivel, biodiverzitási irodalommal, georeferált adatokkal és a laikusok számára is jobb megértést biztosító kulcsszavakkal gazdagítják a hálózat gyűjteményeit. Ez utóbbihoz kapcsolódóan létezik egy crowdsourcing felület is, ahol önkéntesek bevonásával²¹ kiegészítő információk hozzáadásával (például kapcsolódó Europeana-tartalmak linkelésével) az aktuális természettudományi projektek adatait egy igen látványos és felhasználóbarát keretrendszerben jelenítik meg (De Smedt, S. et al., 2024, p. 32.).

Az Europeana korábban hivatkozott szemantikus adatgazdagítási keretrendszere leírja az általános adatgazdagítás részleteit (Europeana Semantic Enrichment Framework, 2021). Ebből fontos kiemelni, hogy a gazdagítás a platform összes tartalmára kiterjed, amilyen mértékben az lehetséges. Az adatgazdagítás sikeressége ugyanis nagymértékben függ a meglévő aggregált adatok minőségétől. A gazdagításra használt források köre jól meghatározott és folyamatosan revideált.²² Az adatok megfeleltetése szabályrendszer alapján történik. A gazdagításra használt forrásokban kis- és nagybetű-érzékeny módon keresik le gépi eszközökkel a beazonosítandó helyek, fogalmak, szereplők, időszakok és szervezetek szöveges címkéjét. A folyamat eredményeképpen SKOS²³ címkékhez rendelik hozzá a beazonosított entitásokat, és csatolják a forrásból a megfelelő URI-t az Europeana adatához. Az adatgazdagításhoz többféle gépi eszközt használnak: az Annocultor egy továbbfejlesztett változatát REST API-val,²⁴ a Metist és MongoDB-t a tároláshoz. A gazdagítás eredményét értékelik is, ebben a *Többnyelvű és szemantikus gazdagítási stratégia* csoportnak²⁵ a legnagyobb szerepe.

Ha megnézzük a MuseuMap²⁶, a hazai múzeumi ágazati szektor aggregátorának a tartalmait az Europeanán, megfigyelhetjük az adatgazdagítás hozadékát: látunk példát arra, hogy az alkotóhoz Wikipédia-azonosító került („Fodor Ilda: Tenger növényei - Meduza”), és arra is, hogy a készítés helye ki lett egészítve Geonames-URI-val („Könyv: kézi aranyozás”) (3. ábra).

Tenger növényei - Meduza

Plasztikus, különféle eszközökkel áttört, strukturált felületű organikus alakzat. Felületét vékony, fekete szabálytalan mintázat hálózta be. A kúpos oldalán nyolc csigavonalas nyúlvány található.

Ezt az tárgyat az **Jósa András Múzeum** biztosítja és karbantartja
 Megtekintés a szolgáltató intézmény honlapján

Jó tudni **Minden metaadat** hely

Alkotó	https://hu.wikipedia.org/wiki/Fodor_Ilda
Tárgy	égetett ; festett
Az tárgy típusa	szobor ; képzőművészet ; sculpture (visual works) ; szobrászat ; Szobrászat
Közepes	samott
Szolgáltató intézmény	Jósa András Múzeum
Aggregátor	MuseuMap

3. ábra: MuseuMapból aggregált leírás adatgazdagítással

Az még valószínűleg további fejlesztendő a felületen, hogy a beszúrt linkek tényleges linkként is működjenek, illetve nem szerencsés, hogy Fodor Ilda esetében a meglevő nevet felülírta a Wikipédia-link, a készítés helye esetében az érték a meglevő mellé került (4. ábra).

Könyv	
Ezt az tárgyat az Iparművészeti Múzeum biztosítja és karbantartja	
Megtekintés a szolgáltató intézmény honlapján ↗	
Jó tudni	Minden metaadat hely
Tárgy	aranyozott ; bőrkötés ; vaknyomással díszített
Az tárgy típusa	könyv ; iparművészet ; books ; könyv ; Könyv ; Iparművészet
Közepes	bőr ; papírtábla ; papír ; Papír
Szolgáltató intézmény	Iparművészeti Múzeum
Aggregátor	MuseuMap
Az ebben a tárgyban szereplő adathordozó licence (hacsak másképp nincs meghatározva).	http://creativecommons.org/licenses/by-nc-sa/4.0/
Létrehozás dátuma	1716 : 1716
Helyek	http://www.geonames.org/3057124 ; Tyrnau : Nagyszombat ; Szlovákia
Azonosító	oai-aggregated-bib9569779 ; imm-object-35192 ; 19882. ; imm:35192 ; imm:35192
Terjedelem	magasság : 15.5 cm ; szélesség : 10 cm ; vastagság : 3.2 cm ;

4. ábra: MuseuMapból aggregált tartalom - Geonames-linkkel kiegészítve

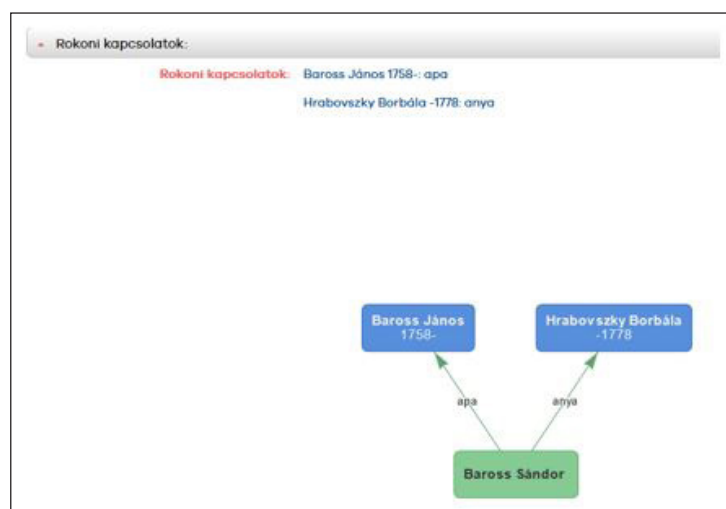
2.1.3. Félig strukturált levéltári és régészeti adatok gazdagítása

A Kanti Egyetem munkatársai többféle szemantikus annotáló eszközt használva gazdagították 16 különböző intézmény levéltári segédleteinek leírását²⁷ (Zeng 2019, p. 15–17). A felismert entitások közé tartoztak a személynevek, az intézménynevek, a földrajzi nevek és az események; kevésbé felismertekhez pedig a közösségi címkék (social tags), a termékek és az iparági kifejezések (Zeng, 2019, p. 16).

Egy másik projektben CIDOC-CRM-alapú²⁸ ontológiát és szemantikus annotáló eszközöket használva adtak értelmes plusz információkat régészeti jelentéseket tartalmazó digitális könyvtári állományhoz (Vlachidis et al., 2013). A vállalkozás eredményeit bemutató Androkinos portál sajnos 2024 júniusa óta nem elérhető, de szerencsére a Wayback Machine²⁹ segítségével még visszakereshető.

2.1.4. Adatgazdagítás a Petőfi Irodalmi Múzeum személynévterében

A Petőfi Irodalmi Múzeum (PIM)³⁰ szemantikus webbel kapcsolatos törekvéseit Hubay Miklós mutatta be. Tanulmányából két megvalósítást emelek ki: a hálóban ábrázolt családi kapcsolatokat és a névterekre vagy külső forrásokra mutató linkekkel való gazdagítást (Hubay, 2020b, p. 119). Az előbbinél a rokoni kapcsolatok a szöveges leírason kívül megjelennek egy kapcsolati hálóban is, amelyen az élek a rokoni kapcsolatot mutatják, a csomópontok pedig az egyes személyek neveit (5. ábra). Az utóbbi során Wikipédia, Wikidata, ISNI (International Standard Name Identifier) és VIAF (Virtual International Authority File) linkekkel egészültek ki a személynévrekordok, ami úttörő és példaértékű a hazai közgyűjtemények katalógusában (6. ábra).



5. ábra: PIM Névtér - rokoni kapcsolatok

József Attila

Adatbázis: Magyar életrajzi index (Név); Magyar életrajzi index (Nekrológ); Magyar írók sírjai; Díjazottak; Budapest topográfia/Lakók

Egységesített név: József Attila

Születés/Halálozás: 1905-1937

Névvarians: J. A.; -JA-

Foglalkozás: költő (foglalkozás); műfordító (foglalkozás); lapszerkesztő (foglalkozás); újságíró (foglalkozás)

Születési hely: Budapest

Pontos születési dátum: 1905. IV. 11.

Halálozási hely: Balatonszárszó

Pontos halálozási dátum: 1937. XII. 3.

Temető: Kerepesi

Parcella: 35-2-59

Tagság: Magyar Újságírók Szövetsége

Halálozási adat forrása: BH 1937. dec. 5

Rokoni kapcsolatok:

Díj:

Megjegyzés:

Adatforrások:

Wiki: https://hu.wikipedia.org/wiki/J%C3%B3zsef_Attila

Külső névterek:

ISNI: <https://isni.org/xslt/DB-12/CMD?ACT=SRCH&IKT=8006&TRM=ISN%3A0000000083739388>

Wikidata: <http://www.wikidata.org/entity/Q313359>

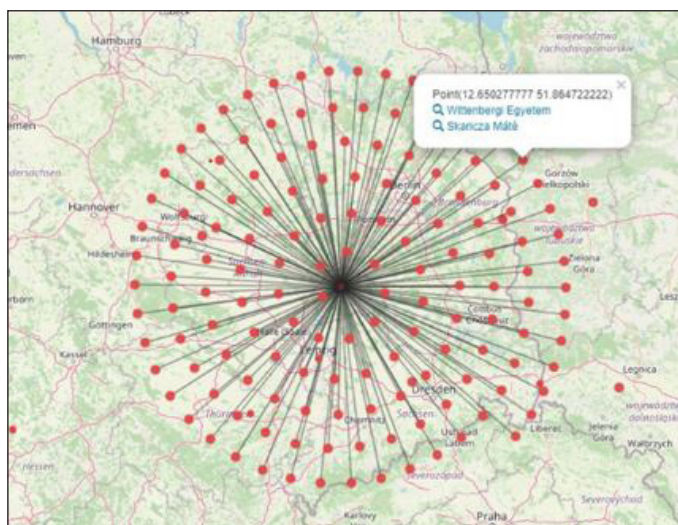
VIAF: <https://viaf.org/viaf/39379044>

6. ábra: PIM Névtér - névtér linkek

Hubay Miklós 2024-ben történt szóbeli közlése alapján a képernyőképek készítése óta a külső szolgáltatásokra mutató linkek köre tovább bővült: bekerültek a PIM névtérbe³¹ a Magyar Nemzeti Levéltár (MNL) online adatbázisaiban található közel 200 irodalmi vonatkozású személy oldalára mutató linkek, illetve az MNL oldaláról³² is visszamutatnak ezeknek a személyeknek a rekordjai a PIM Névtérre külső hivatkozásként.

2.1.5. Wiki-alapú hazai példák

Az ITIdata³³ és az ELTEdata³⁴ szemantikus adatbázis felületének alapja az eleve szemantikus kapcsolatokra épülő, nyílt forráskódú Wikibase szoftver. Az adatszerkezetük részben követi a Wikidata struktúráját (Dobás, Fellegi, Palkó, 2023, p. 192). Mindkét adatbázisban a szülő intézményhez (a Bölcsészettudományi Kutatóközpont Irodalomtudományi Intézetéhez és az Eötvös Loránd Tudományegyetemhez) kapcsolódó kutatási projekteknél keletkezett adathalmazokat tesznek közzé RDF-állítások formájában. Ez lehetőséget biztosít az adatok legkülönbözőbb lekérdezésére SPARQL-formátumban, illetve különböző vizualizációk készítésére. A HECE-lexikonból³⁵ például vizualizált módon ábrázolták a humanisták peregrinációit (7. ábra). Az ITIdata egyik nagy projektje a Magyar Irodalomtörténeti Bibliográfia migrálása relációs adatbázisból a Wikibase-be, amelyhez 2025-ben készült egy felhasználóbarát keresőfelület, amelyről Dobás Kata és Szekrényes István számol be. (Dobás, Szekrényes, 2025 p. 60–61)



7. ábra: HECE – vizualizáció

3. Összefoglalás és kitekintés a jövőbeli lehetőségekre

A téma sokrétűsége és szerteágazósága kapcsán tanulmányomban igyekeztem egy általános képet adni a szemantikus adatgazdagítás részéről. Az adatgazdagítással a hazai kulturális intézményeknek is érdemes lenne többet foglalkozniuk, mivel ez alapját képezhetné a későbbi szemantikus adatkezelésnek, illetve az új szabványoknak való megfelelésnek is.

A jövőre tekintve igen izgalmas kérdés, hogy az adatgazdagítás eszköztárát hogyan fogja befolyásolni a mesterséges intelligencia (MI). Kiváltja-e a mostani számos szemantikus annotáló eszközt vagy az emberi validálás feladatait az eredmények ellenőrzésében. A többször idézett Marcia Lei Zengnek és szerzőtársának már született egy tanulmánya a kulturális örökség adatainak kezelése és a mesterséges intelligencia kapcsolatáról, amely egyelőre csak kínai nyelven (angol absztrakttal) érhető el (Fan, Zeng, 2024). A szerzők hangsúlyozzák, hogy az emberi értelem és felügyelet továbbra is meghatározó szerepet kell kapjon az adatmodellezésben és az adatok értelmezésében, valamint a mesterséges intelligencia által generált adatok fölött is, azért, hogy biztosítsák az eredményül létrejött új adatok megbízhatóságát, hitelességét és értelmezhetőségét.

Valószínűsíthető, hogy a mesterséges intelligencia használatának elterjedésével háttérbe szorulnak a közösségi közreműködési alkalmazások, mivel sok feladatot át tudnak majd venni az MI-alapú rendszerek. A kapcsolt adatokkal való kiegészítés és a könyvtárak elmozdulása az új könyvtári szabványok felé várhatóan megerősödik, mivel az ezen folyamatok során hozzáadott értékek a mesterséges intelligencia számára is hasznos forrásul szolgálhatnak majd. 2024-ben európai uniós támogatással létrejött egy kapacitásbővítésre szánt platform³⁶ azzal a céllal, hogy a kulturális örökségi intézmények minél hatékonyabban tudják kihasználni a mesterséges intelligenciát.

Jegyzetek

¹ A tanulmány a következő szakdolgozat egy részének szerkesztett változata: Mátyás, M. (2024) *Adatgazdagítás a kulturális intézményekben: hazai és nemzetközi példák*. Eötvös Loránd Tudományegyetem, Budapest.

² Az Europeana az Európai Unió által létrehozott gyűjtőportál, amely digitalizált kulturálisörökség-tartalmakat gyűjt ágazati aggregátorokon keresztül elsősorban európai országokból. A gyűjtőportál funkció mellett innovatív digitális megoldásokkal törekszik a metaadatok egységes kezelésére és adatgazdagítására is.

³ Europeana Semantic Enrichment Framework. 2021. Elérhető: <https://docs.google.com/document/d/1JvjrWMTpMIH7WnuieNqCt0zpJAXUPo6x4uMBj1pEx0Y/> (Utolsó elérés: 2026. 06. 01.)

⁴ RDF – Semantic Web Standards. Elérhető: <https://www.w3.org/RDF/> (Utolsó elérés: 2026. 06. 01.)

⁵ IFLA Library Reference Model: A Conceptual Model for Bibliographic Information. Elérhető: <https://repository.ifla.org/handle/20.500.14598/40.2>; a magyar fordítása itt érhető el: <https://repository.ifla.org/handle/20.500.14598/46> (Utolsó elérés: 2026. 06. 01.)

⁶ Az authority – besorolási rekord elnevezéssel Ungváry Rudolf foglalkozik (l. Ungváry 2019, p. 333), de az Országos Könyvtári Szabványosítási Bizottság ajánlása alapján írok authorityt.

⁷ A tanulmány írásakor 40 tétel szerepelt a listán: Vocabularies Supported by Europeana. Europeana PRO. Elérhető: https://docs.google.com/spreadsheets/u/1/d/1BoDNolkcp_qfvVShdOZYGcf61XslcwKF2MdGcgYs20 (Utolsó elérés: 2026. 06. 01.)

⁸ Vocabularies. EAGLE Portal. Elérhető: <https://www.eagle-network.eu/resources/vocabularies/> (Utolsó elérés: 2026. 06. 01.)

⁹ Europeana Pro: Europeana Semantic Enrichment. Elérhető: <https://pro.europeana.eu/page/europeana-semantic-enrichment> (Utolsó elérés: 2026. 06. 03.)

¹⁰ Open Calais. Elérhető: <https://weadapt.org/organisation/open-calais/> (Utolsó elérés: 2026. 06. 03.) Kiegészítő információ: Az OpenCalais eredetileg a Thomson Reuters terméke volt, de napjainkra LSEG Data & Analytics (korábban Refinitiv) vállalkozáshoz tartozik. Lsd. https://developers.lseg.com/content/devportal/en_us/api-catalog/open-perm-id/intelligent-tagging-restful-api.html (Utolsó elérés: 2026. 06. 03.)

¹¹ Egy leírás és képernyőkép elérhető róla itt, de már nem működik sajnos az eredeti weboldala és a dokumentációja, archívumokban még lehetséges, hogy megtalálható: <https://www.outsystems.com/forge/component-overview/3410/expert-system-cogito-api-o11> (Utolsó elérés: 2026. 06. 03.)

¹² Az OpenRefine elérhetősége: <https://openrefine.org/> (Utolsó elérés: 2026. 06. 03.)

¹³ A Pundit elérhetősége: <https://thepund.it/> (Utolsó elérés: 2026. 06. 03.)

¹⁴ A Mirador képnézegető annotációs pluginja: <https://github.com/ProjectMirador/mirador-annotations> (Utolsó elérés: 2026. 06. 03.)

¹⁵ A Brat Rapid Annotation Tool elérhető itt, bár nem a legfrissebb, de részletes dokumentációt és útmutatót is tartalmaz az oldal: <https://brat.nlplab.org/index.html> (Utolsó elérés: 2026. 06. 03.)

¹⁶ A Recogito elérhetősége: <https://recogito.pelagios.org/> (Utolsó elérés: 2026. 06. 03.)

¹⁷ A MapHub elérhetősége: <https://maphub.net/> (Utolsó elérés: 2026. 06. 03.)

¹⁸ A DBpedia elérhetősége: <https://dbpedia.org/fct/> (Utolsó elérés: 2026. 06. 03.)

¹⁹ A LODMilla elérhető itt: <https://lodmilla.sztaki.hu/lodmilla/> (Utolsó elérés: 2026. 06. 03.)

²⁰ A Pentaho program elérhetősége: <https://pentaho.com/products/pentaho-data-integration/> (Utolsó elérés: 2026. 06. 03.)

²¹ A LinBi adatgazdagító felület linkje: <https://linbi.ait.co.at/linbi-enrichment-platform> (Utolsó elérés: 2026. 06. 03.)

²² Lásd a tanulmány 2.1. alfejezetének általános részét.

²³ SKOS Simple Knowledge Organization System. Elérhető: <https://www.w3.org/2004/02/skos/> (Utolsó elérés: 2026. 06. 03.)

²⁴ Enrichment Framework API & Instructions. Elérhető: <https://docs.google.com/document/d/1fjVG-F4K7xl4oHuv1COGOVP-1H6PhM2Fd6hMSwQdvDYY/> (Utolsó elérés: 2026. 06. 03.)

²⁵ Europeana Pro: Multilingual and Semantic Enrichment Strategy. Elérhető: <https://pro.europeana.eu/project/multilingual-and-semantic-enrichment-strategy> (Utolsó elérés: 2026. 06. 01.)

²⁶ A MuseuMap elérhetősége: <https://www.museumap.hu/> (Utolsó elérés: 2026. 06. 01.)

²⁷ A levéltári segédletek általában szöveges dokumentumok, amelyekben útmutatást adnak a kutatóknak az adott gyűjtemény használatához.

- ²⁸ A CIDOC CRM modellről bővebben: <https://cidoc-crm.org/> (Utolsó elérés: 2026. 06. 01.)
- ²⁹ Wayback Machine: Androkinos: Digital Humanities Semantic Annotation of Archaeological Grey Literature. Elérhető: <http://web.archive.org/web/20240614180901/http://www.andronikos.co.uk/> (Utolsó elérés: 2026. 06. 03.)
- ³⁰ Petőfi Irodalmi Múzeum. Elérhető: <https://pim.hu/> (Utolsó elérés: 2026. 06. 03.)
- ³¹ PIM Névtér: Arany János. Elérhető: <https://resolver.pim.hu/auth/PIM41404> (Utolsó elérés: 2026. 06. 03.)
- ³² Petőfi közössége. MNL AdatbázisokOnline: Arany János. Elérhető: <https://adatbazisokonline.mnl.gov.hu/adatbazis/petofi-kozossege/adatlap/103482> (Utolsó elérés: 2026. 06. 03.)
- ³³ DigiPhil: ITIdata. Elérhető: https://itidata.abtk.hu/wiki/Main_Page (Utolsó elérés: 2026. 06. 03.)
- ³⁴ elte.dh: ELTEdata. Elérhető: https://eltdata.elte-dh.hu/wiki/Main_Page (Utolsó elérés: 2026. 06. 03.)
- ³⁵ MTA–ELTE Humanizmus Kelet-Közép-Európában Lendület Kutatócsoport, Történeti Értelmisségi Hálózatok - HECEdata. Elérhető: <https://eltdata.elte-dh.hu/wiki/HECE> (Utolsó elérés: 2026. 06. 03.)
- ³⁶ AI4Culture. Elérhető: <https://ai4culture.eu/project/mission> (Utolsó elérés: 2026. 06. 03.)

Felhasznált és ajánlott források

- Baca, M., Gill, M. (2015) *Encoding Multilingual Knowledge Systems in the Digital Age: The Getty Vocabularies*. Knowledge Organization, 42(4), p. 232–243. <https://doi.org/10.5771/0943-7444-2015-4-232>
- Bernasconi, E. et al. (2023) *Linked Data Interfaces: A Survey*. Information, 14(9), art. 483. <https://doi.org/10.3390/info14090483>
- Berners-Lee, T., Hendler, J., Lassila, O. (2001) *The semantic web*. Scientific American, 284(5), p. 34–43. <https://doi.org/10.1038/scientificamerican0501-34>
- Daquino, M. et al. (2017) *Enhancing Semantic Expressivity in the Cultural Heritage Domain: Exposing the Zeri Photo Archive as Linked Open Data*. Journal on Computing and Cultural Heritage, 10(4), art. 21. <https://doi.org/10.1145/3051487>
- De Smedt, S. et al. (2024) *Ten lessons learned from the mass digitisation of a herbarium collection*. PhytoKeys, 244, p. 23–37. <https://doi.org/10.3897/phytokeys.244.120112>
- Dobás, K., Fellegi, Zs., Palkó, G. (2023) *A kis gömböc meséje – az ITIdata irodalomtudományos adatbázis fejlesztése 2022–2023-ban*. In: Hungarnet Egyesület (szerk.) *Új technológiákkal, új tartalmakkal a jövő digitális transzformációja felé*: 32. Networkshop: országos konferencia, Veszprém, április 12–14. p. 192–198. Hungarnet Egyesület, Budapest. <https://doi.org/10.31915/NWS.2023.28>
- Dobás, K., Szekrényes, I. (2025) *Az ITIdata szemantikus adatbázis új keresőfelületei*. In: Oktatási, kutatási és közgyűjteményi infrastruktúrák és tartalmak: digitális transzformáció felsőfokon: Workshop 2025: 34. Országos Informatikai Konferencia, Győr, május 13–15. p. 58–65. Hungarnet Egyesület, Budapest. ISBN: 978-615-6792-15-0. <https://doi.org/10.31915/NWS.2025.7>
- Fan, W., Zeng, M. L. (2024) *Exploring the Generation Path of Smart Data for the Activation and Utilization of Cultural Heritage in the New Era of AI*. Journal of Library Science in China, 50(2), p. 4–29. <https://doi.org/10.13530/j.cnki.jlis.2024010>
- Gracy, K. F., Zeng, M. L., Skirvin, L. (2013) *Exploring Methods to Improve Access to Music Resources by Aligning Library Data with Linked Data: A Report of Methodologies and Preliminary Findings*. Journal of the American Society for Information Science and Technology, 64(10), p. 2078–2099. <https://doi.org/10.1002/asi.22914>
- Hamburger, S. (2011) *Functional Requirements for Authority Data: A Conceptual Model*. Library Collections, Acquisitions, & Technical Services, 35(1), p. 41. <https://doi.org/10.1080/14649055.2011.10766293>

- Hinze, A. et al. (2019) *Capisco: Low-Cost Concept-Based Access to Digital Libraries*. International Journal on Digital Libraries, 20(4), p. 307–334. <https://doi.org/10.1007/s00799-018-0232-3>
- Hubay, M. (2019) *Tények, mítoszok és lehetőségek a szemantikus web világában*. Könyvtári Figyelő, 29(2), p. 245–253. Elérhető: <https://epa.oszk.hu/00100/00143/00356/pdf/> (Utolsó elérés: 2026. 06. 01.)
- Hubay, M. (2020a) *Adatgazdagítás, adatszolgáltatás és discovery hagyományos és szemantikus metaadat-környezetben*. Tudományos és Műszaki Tájékoztatás, 67(5), p. 291–300. Elérhető: <https://journals.bme.hu/tmt/article/view/35421> (Utolsó elérés: 2026. 06. 01.)
- Hubay, M. (2020b) *Rekordkapcsolattól a szemantikus kapcsolatig: A Petőfi Irodalmi Múzeum személynévtére és az értelemmel teli világháló*. In: Hagyományok és kihívások, p. 115–123. https://doi.org/10.21862/HagyKihiv_2019.115
- Koch, W., Koch, G. (2017) *Aggregation and Management of Metadata in the Context of Europeana*. VOEB-Mitteilungen, 70(2), p. 170–178. <https://doi.org/10.31263/voebm.v70i2.1776>
- Lehmann, J. et al. (2015) *DBpedia – A Large-Scale, Multilingual Knowledge Base Extracted from Wikipedia*. Semantic Web, 6(2), p. 167–195. <https://doi.org/10.3233/SW-140134>
- Mering, M. (2019) *IFLA Library Reference Model, RDA, and Serials in a Nutshell*. Serials Review, 45(1–2), p. 66–68. <https://doi.org/10.1080/00987913.2019.1627691>
- Micsik, A., Turbucz, S., Tóth, Z. (2015) *Exploring Publication Metadata Graphs with the LODmilla Browser and Editor*. International Journal on Digital Libraries, 16(1), p. 15–24. <https://doi.org/10.1007/s00799-014-0130-2>
- Neuböck, V. G. (2019) *Crowdsourcing an der oberösterreichischen Landesbibliothek*. VOEB-Mitteilungen, 72(2), p. 297–309. <https://doi.org/10.31263/voebm.v72i2.2834>
- Stiller, J. (2024) *Multilingual Developments in Digital Cultural Heritage - Problem Space and Solutions*. EuropeanaTech Insight, 2024(17). Elérhető: <https://pro.europeana.eu/page/issue-17-multilinguality> (Utolsó elérés: 2026. 06. 04.)
- Subirats, I., Zeng, M. L. (2020) *Linked Open Data Enabled Bibliographical Data (LODE-BD) 3.0*. FAO, Rome. Elérhető: <https://openknowledge.fao.org/handle/20.500.14283/cb2209en> (Utolsó elérés: 2026. 06. 01.)
- Ungváry, R. (2019) *Besorolási, szabványosított, normatív vagy »autorizált«*. Tudományos és Műszaki Tájékoztatás, 66(6), p. 328–342. Elérhető: <https://journals.bme.hu/tmt/article/view/35873> (Utolsó elérés: 2026. 06. 01.)
- Vlachidis, A. et al. (2013) *Automatic Metadata Generation in an Archaeological Digital Library: Semantic Annotation of Grey Literature*. In: Przepiórkowski, A. et al. (szerk.) Computational Linguistics: Applications. p. 187–202. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-34399-5_10
- Weitz, J. et al. (2016) *Mining MARC's Hidden Treasures: Initial Investigations Into How Notes of the Past Might Shape Our Future*. Journal of Library Metadata, 16(3–4), p. 166–180. <https://doi.org/10.1080/19386389.2016.1262653>
- Willer, M., Dunsire, G. (2013) *Bibliographic Information Organization in the Semantic Web*. Chandos Publishing, Oxford. <https://doi.org/10.1533/9781780633978>
- Zeng, M. L., Chan, L. M. (2004) *Trends and Issues in Establishing Interoperability among Knowledge Organization Systems*. Journal of the American Society for Information Science and Technology, 55(5), p. 377–395. <https://doi.org/10.1002/asi.10387>
- Zeng, M. L. (2019) *Semantic Enrichment for Enhancing LAM Data and Supporting Digital Humanities*. Profesional de La Información, 28(1). <https://doi.org/10.3145/epi.2019.ene.03>
- Zumer, M., Zeng, M. L., Mitchell, J. S. (2012) *FRBRizing KOS Relationships: Applying the FRBR Model to Versions of the DDC*. In: Nee-lameghan, A., Raghavan, K. S. (szerk.) Categories, Contexts and Relations in Knowledge Organization. p. 190–194. Ergon-Verlag GmbH, Würzburg. Elérhető: <https://www.webofscience.com/wos/woscc/full-record/WOS:000342292200029> (Utolsó elérés: 2026. 06. 01.)