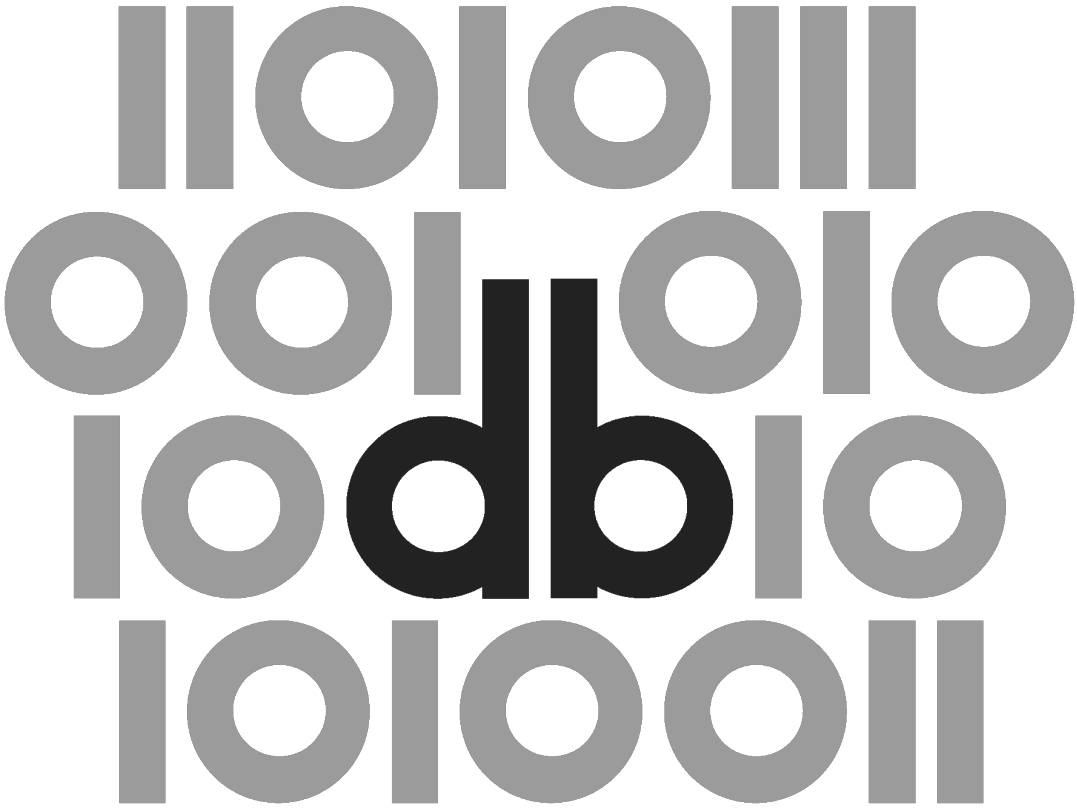
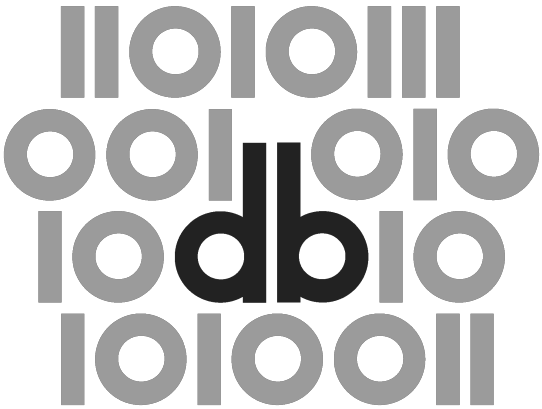
2 (2019) <DIGITÁLIS BÖLCSÉSZET>



2 (2019) </DIGITÁLIS BÖLCSÉSZET>

Digitális Bölcsészet
2019., második szám

<DIGITÁLIS BÖLCSÉSZET>



2 (2019)

ISSN 2630-9696
DOI 10.31400/dh-hun.2019.2
Kiadja az ELTE BTK Régi Magyar Irodalom Tanszéke (1088 Budapest, Múzeum krt. 4/A) és a Bakonyi Géza Alapítvány.
Felelős kiadó az ELTE BTK Régi Magyar Irodalom Tanszék vezetője.

Megjelenik az Open Journal Systems (OJS) v. 3. platformon, melynek működtetését az ELTE Egyetemi Könyvtár- és Levéltár biztosítja.

Honlap: `http://ojs.elte.hu/index.php/digitalisbolcseszet`
Email cím: dbfolyoirat@gmail.com

Tördelés: Hegedüs Béla
Grafika: Hegyi Gábor

<TANULMÁNYOK>

Király Péter

Göttingen eResearch Alliance,

Gesellschaft für wissenschaftliche Datenverarbeitung mbH, Göttingen

peter.kiraly@gwdg.de

Marco Büchler

Institute of Computer Science, Georg-August-Universität, Göttingen

mbuechler@etrap.eu

A teljesség minőségjelzőként való mérése az Europeanában[*]

Az Europeana – a kulturális örökség európai digitális platformja – több, mint 3200[1] adatszolgáltatótól beérkező metaadatrekord gyűjteménye a rekordok jellemzőit tekintve meglehetősen heterogén. A rekordok eredeti típusa és kontextusa eltérő. Ahhoz, hogy hatékony szolgáltatásokat építhessünk rájuk, ismernünk kell az adatok erősségeit és gyengeségeit, más szóval a minőségét. A tanulmány egy olyan módszert javasol (és nyílt forráskódú implementálását), ami az adatok meghatározott szerkezeti tulajdonságait méri (teljesség, többnyelvűség, egyediség, rekordmintázat), hogy ezáltal minőségi problémákra világítson rá.

Kulcsszavak:
big data-alkalmazások, adatelemzés, adatgyűjtemény, szolgáltatásminőség, minőségkezelés, metaadat, adatintegráció

Bevezetés

> Az utóbbi 24 órában sok időt pazaroltam el, mert olyan dolgokat feltételeztem a (meta)adatokról, melyek nem bizonyultak helyesnek. Hosszú időt töltöttem a kód hibaellenőrzésével, de a kód jó volt, pusztán azt nem találta meg, ami ott sem volt. A rossz feltételezések a legnehezebben elkapható programozási hibák.[2]

[*] Jelen tanulmány eredetileg angolul jelent meg: Péter Király and Marco Büchler, „Measuring completeness as metadata quality metric in Europeana," in 2018 IEEE International Conference on Big Data (Piscataway: IEEE, 2019), 2711–2720. https://doi.org/10.1109/BigData.2018.8622487

[1] A tanulmányban szereplő számok az Europeana 2018. októberi állapotára érvényesek.

[2] Felix Rau, német nyelvész a metaadat-problémák következményeiről. 2018. október 18., https://twitter.com/fxru/status/1052838758066868224

Az aggregált metaadat-gyűjtemények funkcionalitása nem független a metaadatrekordok minőségétől. A következőkben néhány, az Europeanából,[3] az európai kulturális örökség digitális platformjából vett példával világítjuk meg a metaadatok fontosságát:

(a) Az adatbázisban van néhány ezer olyan rekord, aminek a címe „Photo" (fénykép) – illetve ennek valamely szinonimája, nyelvi változata – minden további leírás nélkül. Hogyan találja meg a felhasználó azokat a objektumokat, amelyek egy bizonyos épületet ábrázolnak, ha a leírások egyáltalán nem, vagy csak pontatlanul állnak rendelkezésre?

(b) Több olyan adatszolgáltató (data provider) található az Europeana portál „intézmény" címkéjű keresési facettájában, aminek többféle névváltozata van (pl. „Cinecittà Luce S.p.A." [372 412 rekord], „Cinecittà Luce" [2405 rekord], „LUCE" [105 rekord]). Feltételezhetjük-e, hogy a felhasználó ki tudja választani az összes releváns névalakot, amikor az adott intézményhez tartozó rekordok között szeretne keresni? Ha nem, akkor a keresés nem lesz teljes.

(c) Ha az „év" facettában nem formalizált és egységes adatok vannak, akkor nem lehet interaktív időtartam-szűkítést sem végezni. Hogy interpretáljunk olyan értékeket, mint „13436" vagy „97500000", ha alapvetően valamiféle évszámot várnánk?

(d) Vannak olyan rekordok, amelyek kizárólag műszaki azonosítókból állnak, nincsenek bennük leíró jellegű mezők (cím, létrehozó, leírás, tárgyszavak stb.). Ezek a rekordok emberi szemmel egyáltalán nem értelmezhetőek. Ezen okból kifolyólag nem is támogatják az Europeana egyetlen alapfunkcióját sem.

(e) Többnyelvű környezetben a felhasználó azt várná, hogy ha egy ismert entitásra pl. Leonardo mesterművére, a Mona Lisára különféle nyelveken keres (akár a „La Gioconda", „La Joconde" kifejezésekkel), akkor ugyanazt a találati listát kapja. Ezzel szemben a különféle nyelvi változatokkal történő keresés eltérő találati listákat eredményez, ugyanis a nyelvi változatok nem tartoznak közös entitás alá.

A kérdés, hogy hogyan döntsük el, mely rekordokat kell javítani, és melyek elég jók. A „célnak való megfelelés" („fitness for purpose") a minőségbiztosítás jól ismert szlogenje, arra a koncepcióra épül, hogy a minőséget valamilyen üzleti cél kontextusában, annak megfelelően kell meghatározni. A metaadatok minőségét vizsgálva először azt kell tisztázni, hogy miért fontosak a metaadatok. Az Europeana esetében ez meglehetősen egyértelmű: digitális objektumokhoz nyújt hozzáférési pontokat. Ha a rekord – adatelemekben megnyilvánuló – tulajdonságai nem teszik lehetővé a metaadat megtalálását, a kívánt cél nem teljesül, a felhasználó nem fér hozzá a digitális objektumhoz és nem fogja azt használni. Következésképpen amellett lehet érvelni, hogy a rekord minősége rossz. Akad ugyanakkor egy fontos probléma: az összes rekord kézi ellenőrzését, a ráfordított idő és a szükséges szakértelem mennyisége miatt még egy közepes méretű gyűjtemény sem engedheti meg magának.

[3] `http://europana.eu`

 DOI 10.31400/dh-hun.2019.2.388

Jelen tanulmány egy olyan általános módszertant és skálázható szoftvercsomagot javasol megoldásként, amit mind az Europeanában, mind más, a kulturális örökséget érintő, kis vagy nagy adatmennyiséggel rendelkező gyűjteményben lehet alkalmazni.

Háttér és alapok

Az Europeana kulturális örökséggel kapcsolatos metaadatrekordokat gyűjt és szolgáltat. Az adatbázis a tanulmány megírásának idején több, mint 58 millió rekordból áll, melyek több, mint 3200 intézményből származnak.[4] A rekordokat az Europeana Adatmodell (Europeana Data Model, a továbbiakban EDM) metaadatsémának megfelelően tárolják. Az egyes intézmények EDM-ben, vagy más metaadatszabványban küldik az adataikat. Az eredeti adatformátumok, katalogizálási szabályok, nyelvek és szótárak változatosságának köszönhetően nagy eltérések vannak az egyes rekordok minőségét tekintve, ami komolyan befolyásolja az Europeana szolgáltatásainak egyes funkcióit.

2015-ben egy Europeana különbizottság megvizsgálta a metaadat-minőség problémáját és erről közre is adott egy jelentést,[5] a bizottságnak azonban – ahogy a jelentésben írják – „nem volt elég hatásköre […] a metaadat-minőség metrikáit […] vizsgálni […] .” 2016-ban alakult egy ennél szélesebb körű Adatminőségi Bizottság (Data Quality Committee, DQC).[6] Ebben különféle területekről (például a metaadatok elméleti vizsgálata, katalogizálás, tudományos kutatás, szoftverfejlesztés) érkező szakértők gyűltek össze azzal a céllal, hogy elemezzék és felülvizsgálják a metaadatsémát, megvitassák az adatnormalizálás lehetőségeit, funkcionális követelményelemzést végezzenek, és meghatározzák az egyes funkciók megvalósítását lehetővé tevő metaadatelemeket (megválaszolva afféle kérdéseket, mint „melyek az Europeana alapvető funkciói?” és „mely metaadatelemek támogatják ezeket?”). A bizottság ezenfelül egy „problémakatalógust” is épít, amely a gyakran ismétlődő, helytelen metaadat-minták gyűjteménye (egyebek mellett ilyenek a többszörösen rögzített értékek, a cím megismétlése a leírás mezőben, gépi feldolgozásra szánt értékek, pl. azonosítók elhelyezése emberi feldolgozásra szánt adatelemekben).[7] A többnyelvűség kérdései különleges hangsúlyt kaptak, lévén az Europeana természetéből adódóan törekszik a többnyelvű adatfeldolgozásra és szolgáltatásra.

Jelen kutatást a DQC-vel együttműködve, részben azon belül folytattuk. Azt tűztük ki célul, hogy módszereket és érvényes metrikákat találjunk az Europeana metaadat minőségének mérésére, és azt támogatandó kifejlesszünk egy nyílt forráskódú metaadatminőség-mérő keretrendszert (*Metadata Quality Assurance Framework*).[8] A javasolt eszköz szándékaink szerint általános célú metaadatminőség-mérő szoftver,

[4] A számokat azt Europeana kereső API-ja alapján közöljük.

[5] Marie-Claire Dangerfield et al., *Report and recommendations from the task force on metadata quality,* Technical report. (The Hague: Europeana Foundation, 2016.) `https://pro.europeana.eu/files/Europeana_Professional/Publications/Metadata%20Quality%20Report.pdf`

[6] `https://pro.europeana.eu/project/data-quality-committee`

[7] Timothy Hill and Hugo Manguinhas, *Internal DQC Problem Patterns,* Technical report (The Hague: Europeana Foundation, 2016). `http://bit.ly/2jIXQGU`

[8] Felhasználói felület elérhetősége a cikk írásának idején (hozzáférés: 2019.12.13): `http://rnd-2.eanadev.org/europeana-qa/`, forráskód és további háttérinformáció: `http://pkiraly.github.io`.

amely adaptálható különféle metaadatsémákra (a támogatni tervezett sémák többek között a MARC[9] és a Kódolt Levéltári Leírás – Encoded Archival Description, EAD[10]).

A szoftver skálázható, vagyis fel van készítve nagy tömegű adatok elemzésére, együttműködik továbbá az *Apache Hadoop*[11] elosztott fájlrendszerével, az általános, nagy mennyiségű adatok feldolgozására tervezett *Apache Spark*-kal[12] és az *Apache Cassandra*[13] adatbáziskezelővel. A megközelítésmód egyik legfontosabb jellemzője, hogy képes az adatkurátorok számára érthető jelentéseket készíteni, akiknek általában a szoftverfejlesztők, adattudósok és statisztikusok által használt szaknyelvi kifejezések nem sokat jelentenek. A jelentések azok számára készülnek, akik az ott tárolt információt cselekvési tervvé tudják formálni. A keretrendszer modulokból épül fel: egy sémafüggetlen magkönyvtár mellett sémaspecifikus kiegészítések találhatók (és építhetők). Azzal számolunk, hogy az eszközt a metaadatminőség-mérés egyfajta folyamatos integrált munkafolyamatában (*continuous integration*) lehet majd használni.[14]

A kutatás azt a kérdést teszi fel, hogy hogyan lehet a kulturális örökség metaadatainak a minőségét a leghatékonyabban mérni. Általános feltevés, hogy a minőség fogalma túl összetett, és lehetetlen az összes aspektusát mérni – egyrészt elméleti szempontból (mivel például a jelenleg rendelkezésünkre álló nyelvfelismerési módszerek nem működnek jól a metaadatokban tipikus módon megtalálható rövid szövegek esetében), másrészt gyakorlati okokból (tekintve például a kutatás során rendelkezésre álló erőforrások korlátos voltát). A metaadatrekordok számos szerkezeti jellemzője azonban mérhető, és ezen mérések eredménye a legtöbb esetben jó közelítést ad. Az ilyen eredményeket hívhatnánk „metaadatszagnak", hasonlóan, ahogy a szoftverfejlesztés *kódszagnak* nevezi „azon felszíni jelzéseket, melyek rendszerint a rendszer mélyebb problémáival vannak összefüggésben".[15] A közelítés a gyakorlatban azt jelenti, hogy az eredmények önmagukban nem perdöntőek, azok arra hívják fel a figyelmet, hogy a metaadat-szakértőknek ezeket a pontokat érdemes alaposabban ellenőrizniük. Ez ugyanakkor azzal is jár, hogy az eszköz nem tárja fel azokat a hibákat, melyek nem szerkezeti sajátosságokhoz kötődnek.

A kutatás legfőbb célja, hogy rávilágítson a javítandó metaadatrekordokra. Ha megtudjuk, merre vannak a hibák, és tudunk prioritizálni, a hibák kijavíthatóak lesznek, a javításokat pedig megfontoltan tudjuk tervezni a hibák fontossági rangsorának

[9] *MAchine Readable Cataloging*, `https://www.loc.gov/marc/`. A keretrendszerre épülve készül egy MARC-vizsgáló szoftver ami elérhető a `https://github.com/pkiraly/metadata-qa-marc` címen. Meg kell jegyeznünk, hogy a MARC sokkal összetettebb szabvány, mint az EDM, és a szigorúbb szabályrendszer megléte sokkal fontosabbá teszi az egyedi problémák kiszűrését a MARC esetében az Europeana rekordjainál, így ott a hangsúly a „pontosság" és a „követelményeknek való megfelelés" metrikákra esik.

[10] `http://www.loc.gov/ead/`

[11] `http://hadoop.apache.org/`

[12] `http://spark.apache.org/`

[13] `http://cassandra.apache.org/`

[14] Lásd `http://pkiraly.github.io/2016/07/02/making-general/` és Péter Király, „Towards an extensible measurement of metadata quality," in *Proceedings of the 2nd International Conference on Digital Access to Textual Cultural Heritage - DATeCH2017*, (New York: ACM Press, 2017) 111–115. `https://doi.org/10.1145/3078081.3078109`

[15] A fogalmat Kent Beck vezette be és Martin Fowler terjesztette el *Refactoring* című könyvében, lásd `https://martinfowler.com/bliki/CodeSmell.html`.

 DOI 10.31400/dh-hun.2019.2.388

megfelelően. Mivel az Europeana egy adataggregátor, a javításokat az információ forrásánál, az adott adatszolgáltató adatbázisán belül kell elvégezni. A minőségileg jobb adatok megbízhatóbb funkciókat támogatnak, így a gyenge minőségű rekordok kijavításával az Europeana erősebb szolgáltatásokat képes építeni. A tipikus hibák megtalálása másrészről az alapul szolgáló metaadatséma és annak dokumentációja fejlesztéséhez is elvezethet (bizonyos hibák feltehetően a séma dokumentációjában előforduló nyelvhasználat nem egyértelmű megfogalmazásaiból fakadnak), továbbá a mérés során olyan rekordokat lehet találni, melyek illusztrálják egyes metaadatelemek helyes vagy helytelen használatát. Végezetül a kiemelkedő minőségű metaadatrekordok használhatók a „követendő metaadat-gyakorlatok" elterjesztésére, vagy új szolgáltatások prototípusainak elkészítésekor.

Kutatási helyzetkép

Az elmúlt évtizedben a metaaadatok minőségmérésének informatikai alapú módszerei megjelentek a kulturális örökség területén.[16] Legutóbb Palavitsinis értékelte a téma releváns eredményeit.[17] A kulturális örökség területével némileg átfedő kapcsolt adatok (Linked Data) területén alkalmazott metrikákat Amrapali Zaveri és szerzőtársai összegezték.[18] Az ezekben hivatkozott tanulmányok meghatározták a minőség metrikáit, és számítási módszereket is javasoltak. Többnyire azonban kisebb rekordhalmazokat és az EDM-nél egyszerűbb metaadatsémákat elemeztek, továbbá általában homogénebb adathalmazokra alkalmaztak módszereket (jelentősebb kivételek a 7 millió rekordot elemző Newman és munkatársai,[19] valamint a 25 millió rekordot elemző Harper). Jelen kutatás újdonsága az, hogy megnöveli az elemzett rekordok számát, új adatvizualizációs megoldásokat és minőségjelentéseket vezet be, és más gyűjteményekben is újrahasznosítható nyílt forráskódú implementációt kínál.

A kulturális örökség metaadat értékeléséről szóló bibliográfiát lásd a Zotero hivatkozáskezelő rendszer „Metadata Assessment" (metaadat-értékelés) nevű könyvtárá-

[16] Thomas R. Bruce and Diane I. Hillmann, „The continuum of metadata quality: Defining, expressing, exploiting," in D. Hillman and E. Westbrooks, eds. *Metadata in practice* (ALA Editions, 2004) 238–256., Besiki Stvilia, Les Gasser, Michael B. Twidale and Linda C. Smith, „A framework for information quality assessment," *Journal of the American Society for Information Science and Technology* 58, 12. sz. (2007): 1720–1733. https://doi.org/10.1002/asi.20652, Xavier Ochoa and Erik Duval, „Automatic evaluation of metadata quality in digital repositories," *International Journal on Digital Libraries* 10, 2. sz. (2009): 67–91. https://doi.org/10.1007/s00799-009--0054-4, Corey Harper, „Metadata Analytics, Visualization, and Optimization: Experiments in statistical analysis of the Digital Public Library of America (DPLA)," *The Code4Lib Journal* 33. sz. (2016) http://journal.code4lib.org/articles/11752

[17] Nikos Palavitsinis, *Metadata Quality Issues in Learning Repositories*. PhD thesis, (Alcala de Henares, 2014) https://www.researchgate.net/publication/260424499_Metadata-_Quality_Issues_in_Learning_Repositories

[18] Amrapali Zaveri, Anisa Rula, Andrea Maurino, Ricardo Pietrobon, Jens Lehmann and Sören Auer, „Quality assessment for linked data: A survey," *Semantic Web* 7, 1. sz. (2015): 63–93. https://doi.org/10.3233/SW-150175

[19] David Newman, Kat Hagedorn, Chaitanya Chemudugunta and Padhraic Smyth, „Subject metadata enrichment using statistical topic models," in *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL '07*, (New York: ACM, 2007) 366–375. https://doi.org/10.1145/1255175.1255248

ban,[20] amelyet az amerikai Digitális Könyvtári Szövetség (Digital Library Federation) Metaadat-értékelés csoportja[21] és a DQC tagjai, köztük jelen dolgozat első szerzője állított össze.

Módszertan

Az EDM-séma

Az EDM-rekord[22] több entitásból tevődik össze. A rekord magja az *adatszolgáltató proxyja* (*provider proxy*), ami azokat az adatokat tartalmazza, amit az egyes szervezetek (*adatszolgáltatók*) az Europeanába beküldtek. Az adatok eredeti formátuma lehet EDM vagy számos egyéb, a kulturális örökség területén használatban lévő metaadatséma (például *Dublin Core*, EAD, MARC stb.). Ez utóbbi esetben az adatszolgáltatók vagy az Europeana átalakítja ezeket EDM-re. A rekord további lényeges részei a *kontextuális entitások* (*contextual entities*): résztvevők (*agents*), fogalmak, helyek és idősávok (*timespans*) – azon entitások (személyek, helynevek stb.) leírását tartalmazzák, melyek valamiféle kapcsolatban állnak a rekord tárgyával. Ezen kontextuális entitásoknak két fontos tulajdonságuk van:

(1) A forrásuk valamilyen többnyelvű szótár, így példányaik a nevüket több nyelven rögzítik.
(2) Amennyiben lehetséges, az entitások kapcsolódnak más entitásokhoz (a kapcsolati típusokat a Simple Knowledge Organization System (SKOS) ontológia[23] definiálja).

Az utolsó itt tárgyalt entitás neve *Europeana proxy*. Szerkezetileg megegyezik az adatszolgáltató proxyjával, de ez csak az adatszolgáltató proxy elemeit kontextuális entitásokkal összekötő linkeket tartalmaz, melyeket egy automatikus szemantikus gazdagító eljárás alakít ki.

Minden adatelem egy vagy több, az adatra épülő funkcionalitást vagy szolgáltatást tesz lehetővé. Az Adatminőségi Bizottság elemzi a funkcionális követelményeket, aminek során a tipikus felhasználói forgatókönyvek alapján (t.i. hogyan lépnek kapcsolatba a gyűjteménnyel) meghatározza a legfontosabb funkciókat, és elemzi, hogy mely metaadatelemek támogatják ezeket.[24] Vegyük például a többnyelvű visszakeresést (*cross-language recall*). A bizottság által megállapított felhasználói forgatókönyv a következő: „Felhasználóként az Europeana gyűjteményeiben az általam leginkább ismert nyelven szeretnék keresni, ugyanakkor szeretnék biztos lenni abban, hogy a dokumentumok nyelvétől függetlenül a legrelevánsabb találatokat kapom." Az említett

[20] `https://zotero.org/groups/metadata_assessment`
[21] `https://dlfmetadataassessment.github.io/`
[22] Az EDM-dokumentáció, útmutatók és más anyagok megtalálhatóak a `https://pro.europeana.eu/page/edm-documentation` címen.
[23] `https://www.w3.org/2004/02/skos/`
[24] Timothy Hill, Valentine Charles and Antoine Isaac, *Discovery – User Scenarios and their Metadata Requirements – v.3,* Technical report (The Hague: Europeana, 2015) `https://pro.europeana.eu/files/Europeana_Professional/EuropeanaTech/EuropeanaTech_WG/DataQualityCommittee/DQC_DiscoveryUserScenarios_v3.pdf`

 DOI 10.31400/dh-hun.2019.2.388

kontextuális elemek jórészt többnyelvűek. A funkcionalitást „lehetővé tevő" (*enabling*) elemekre vonatkozó követelmény: „minden literál értékadást[25] támogató EDM elemet nyelvi címkével *kell* ellátni, ezen felül *javasolt* az EDM elemek többnyelvű kontextuális entitáshoz való kapcsolása."[26]

Mivel ezen lehetővé tevő elemek meghatározása egyelőre nincs összhangban a mérés céljával és a meglevő rekordok tulajdonságaival, egy *aldimenziók*nak nevezett egyszerűbb modellel kezdtünk dolgozni. Ennek a modellnek az alapja az összetettebb felhasználói forgatókönyvek helyett a bizottság két tagja, Valentine Charles és Cecile Devarenne által alkotott mátrix, ami az általános funkciókat (aldimenziókat) kapcsolja össze az ezeket lehetővé tevő elemekkel. Az aldimenziók a következők:

- *Kötelező elemek* – azon mezők, melyeknek minden rekordban jelen kell lenniük. A modell kezelni tudja az olyan mezőcsoportokat, melyekből legalább az egyiküknek jelen kell lenni, pl. a tárgyszó típusú elemekből (dc:type, dc:subject, dc:coverage, dcterms:temporal, dcterms:spatial) legalább egynek;
- *leírhatóság* (*descriptiveness*) – mennyi információt hordoz a metaadat ahhoz, hogy leírja azt a tárgyat, amiről szól;
- *kereshetőség* (*searchability*) – a keresés során leggyakrabban használt mezők;
- *kontextusba helyezhetőség* (*contextualization*) – annak alapja, hogy csatolt entitásokat (személyek, helyek, időpontok stb.) találjunk a rekordban;
- *beazonosítás* (*identification*) – az objektum egyértelmű beazonosítását segítő mezők;
- *böngészés* (*browsing*) – az Europeana-portál böngészési jellemzői;
- *megtekintés* (*viewing*) – a portálon való megjelenítésben segítő mezők;
- *újrahasznosíthatóság* (*re-usability*) – a metaadatrekordok más rendszerekben való felhasználását lehetővé tevő mezők;
- *többnyelvűség* (*multilinguality*) – a többnyelvűség szempontjai, hogy a rekordok minden európai polgár számára érthetőek legyenek.

A tanulmány írása idején a modell csak a mezők meglétét vizsgálja, nem ellenőrzi, hogy tartalmuk megfelel-e az elvárásoknak – ezt a feladatot a kutatás egy későbbi pontján oldjuk meg.

Mérés

Minden rekord esetében számos olyan jellemzőt mérünk, melyek kapcsolatosak a rekord minőségével. A főbb tulajdonságcsoportok a következők:

- *egyszerű teljesség* (*completeness*) – a rekordban meglevő mezők aránya a sémában definiáltakhoz képest;
- *az aldimenziók teljessége* – adott funkciót támogató mezőcsoportok, lásd fentebb;

[25] Közvetlenül megadott érték, pl. szám, karaktersorozat, szemben a referenciális értékekkel, mint amilyen az EDM-ben az URL.

[26] Uo., 9–10.

- *mezők megléte és számossága* – mely mezők vannak jelen a rekordban és hányszor;
- *problémakatalógus* – ismert metaadat-problémák jelenléte;[27]
- *a leíró mezők egyedisége* (cím, egyéb cím, leírás);
- *többnyelvűség*;[28]
- *rekordminták* – mely mezők alkotják a „tipikus rekordokat".

A mérés három szinten történik: az egyes rekordok esetében, a gyűjtemény részhalmazaiban (pl. egy adatszolgáltató összes rekordja), végül a teljes adathalmazon.

Az első szinten az eszköz végigjár minden metaadatrekordot. Ezeket elemzi, az egyes rekordok mérési eredményeit pedig egy vesszővel határolt (*comma separated values*) fájl soraiba menti. Összességében minden rekord esetében több, mint ezer mérési eredményt (pontszámot) vagy egyéb jellemzőt nyerünk ki, melyek mindegyike egy-egy mező, mezőcsoport vagy a teljes rekord valamilyen minőséggel összefüggő tulajdonságát jelenti. Az eredményeket különféle pontozási algoritmusok számolják ki.

A második szint a részhalmazoké. Jelenleg a következő részhalmazokkal számolunk: az Europeana adathalmazokban (datasets) azok a rekordok találhatóak, melyek ugyanabban az adatgyűjtési (*data ingestion*) folyamatban kerültek be a gyűjteménybe (ezek a rekordok általában ugyanazon az átalakítási folyamaton mennek keresztül, amikor az Europeana letölti őket az adatszolgáltatóktól); az ugyanazon adatszolgáltatóktól származó rekordok; ugyanazon köztes szolgáltatóktól származó rekordok (az Europeana és az adatszolgáltatók között túlnyomó esetben van egy köztes réteg, egy olyan szolgáltató, ami tematikus vagy regionális alapon koordinál egy adatszolgáltatói csoportot); azonos nyelvű rekordok; azonos országból érkező rekordok. Az első három esetben leképezzük az alhalmazok metszetét is, amelybe azok a rekordok kerülnek, amelyek esetében mind a három vagy kettő tulajdonság közös (pl. ugyanabból a gyűjteményből és köztes szolgáltatótól származó rekordok). A jövőben a DQC kibővítheti

[27] Ez a mérés az Europeana kontextusában kísérleti fázisban van. A teljes problémakatalógust formálisan a Shapes Constraint Language (SHACL) szabvánnyal tervezzük leírni, lásd Holger Knublauch and Dimitris Kontokostas, *Shapes constraint language (SHACL),* W3C recommendation (W3C, 2017. júl. 20.) `https://www.w3.org/TR/2017/REC-shacl-20170720/`.

[28] Lásd Juliane Stiller and Péter Király, „Multilinguality of metadata. Measuring the Multilingual Degree of Europeana's Metadata," in M. Gäde et al. eds., *Everything Changes, Everything Stays the Same? Understanding Information Spaces.* Proceedings of the 15th International Symposium of Information Science (ISI 2017) Schriften zur Informationswissenschaft, (Glückstadt: Werner Hülsbusch, 2017) 164–176. `https://www.researchgate.net/publication/314879735_Multilinguality_of_Metadata_Measuring_the_Multilingual_Degree_of_Europeana's_Metadata`, Valentine Charles, Juliane Stiller, Péter Király, Werner Bailer and Nuno Freire, „Evaluating data quality in europeana: Metrics for multilinguality," in A. Caputo et al. eds., *Joint Proceedings of the 1st Workshop on Temporal Dynamics in Digital Libraries, the (Meta)-Data Quality Workshop and the Workshop on Modeling Societal Future co-located with 21st International Conference on Theory and Practice of Digital Libraries (TPLD 2017)* (Aachen: CEUR, 2017). `http://ceur-ws.org/Vol-2038/paper6.pdf`, valamint Péter Király, Juliane Stiller, Valentine Charles, Werner Bailer and Nuno Freire. "Evaluating Data Quality in Europeana: Metrics for Multilinguality," in *Metadata and Semantic Research 2018* 12th International Conference, MTSR 2018, Limassol, Cyprus, October 23–26, 2018, Revised Selected Papers (Communications in Computer and Information Science, volume 846) (Cham: Springer, 2019) 199–211. `https://doi.org/10.1007/978-3-030-14401-2_19`

 DOI 10.31400/dh-hun.2019.2.388

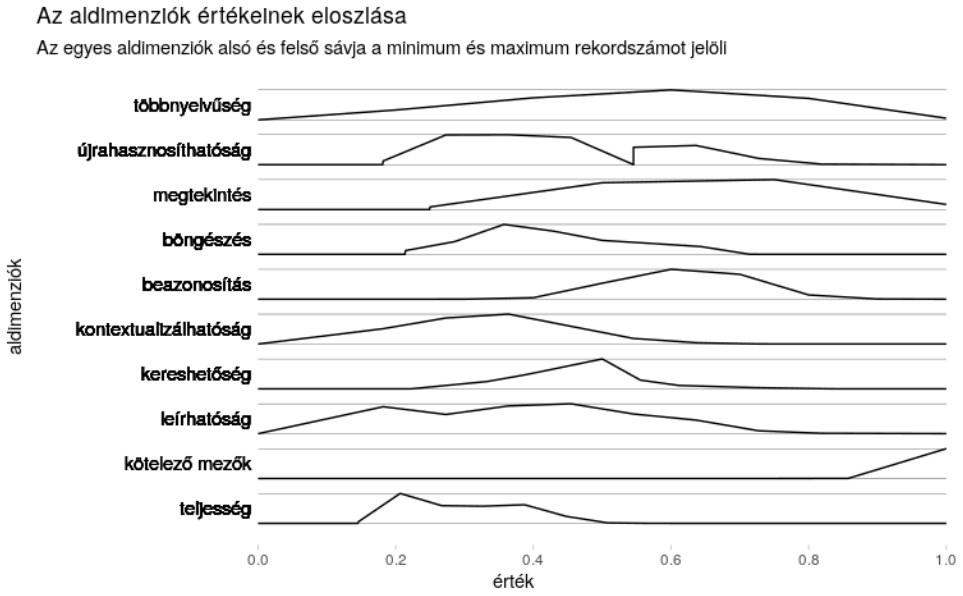
a részhalmazok körét, és olyan csoportokat is képez, melyek valamilyen másik, a metaadatsémában meghatározott tulajdonságon osztoznak.

A második és harmadik szinten az aggregált metrikákat számoljuk ki, statisztikailag összegezve az egyes rekordok esetében kiszámolt eredményeket.

A teljesség kiszámolásakor két eltérő súlyozási modellt alkalmazunk. Az első esetben a súlyok az aldimenziókat veszik figyelembe. A naiv, „egyszerű teljesség" (ahol minden adatelem ugyanazzal a súllyal szerepel) 5 súlyponttal szerepel, a kötelező elemek teljességének súlya 3, az aldimenziók súlya pedig 2. A számítási képlet (1) a következő:

$$c_{subdimensions} = \frac{\sum_{i=1}^{d} score_i \times w_i}{\sum_{i=1}^{d} w_i} \tag{1}$$

ahol d az aldimenziók száma, $score_i$ az adott aldimenzióhoz tartozó mezők közül a rekordban meglevők aránya (0-tól 1-ig terjedő skálán), a w_i pedig az aldimenzió súlya. Vagyis minden aldimenzió esetében megnézzük, hogy az ott szereplő mezők hány százalékban vannak jelen az adott rekordban, majd ezt az aldimenzió súlyának megfelelően vesszük figyelembe (ha egy aldimenzió például teljes, de a súlya alacsony, a végső pontszámnál kevesebbet fog számítani, mint egy gyengébb, de nagyobb súlyú társa). A végeredmény minimális értéke 0, maximális értéke pedig 1.



1. ábra. Az aldimenziók és az „egyszerű teljesség" értékeinek eloszlása

A második megközelítésben az elsőrendű faktor a számosság, vagyis hányszor fordul elő a mező az adott rekordban. A szélsőséges értékek torzító hatásának csökkentése érdekében nem közvetlenül ezt a számot, hanem ennek normalizált verzióját vettük alapul, amely inkább a számosság nagyságrendjét jelzi. Ilyen szélsőséges eset, amikor

egy-egy mező (pl. tárgyszó) több száz feletti példányban van jelen egy rekordban, aminek túlságosan nagy súlya lenne, s így túlságosan befolyásolná a pontszámot. A normalizálást az 1. táblázat tartalmazza.

mezőpéldányok száma	0	1	2–4	5–10	11–
normalizált pontszám	0.0	0.25	0.50	0.75	1.0

1. táblázat. A számosság normalizálása

A számosságon alapuló súlyozás egyszerű: minden mező súlya 1, kivéve az egyes entitásokat azonosító rdf:about mezőket, melyek 10 pontot kapnak, így a súlyozást főként az entitások száma és kevésbé azok „kitöltöttsége" tükrözi. Az egyenlet (2):

$$c_{cardinality} = \frac{\sum_{i=1}^{d} norm(cardinality_i) \times w_i}{\sum_{i=1}^{d} w_i} \tag{2}$$

ahol d a mezők száma, $cardinality_i$ a mezők kardinalitása, a $norm()$ a normalizálási funkció (lásd 1. táblázat) és a w_i a mező súlya. Minden mező esetében megszámoljuk, hogy hány példányuk érhető el a rekordban, a számot a táblázatnak megfelelően normalizáljuk és súlyozzuk. Eredményül – mint korábban is – egy 0-tól 1-ig terjedő számot kapunk.

A végső egyenlet a két megközelítés kombinációja, ahol az első, aldimenziós, vagyis a mezők fontosságán alapuló megközelítésmód súlya (és fontossága) két és félszer nagyobb, mint a második, az entitások és mezők számosságán alapuló megközelítésmód:

$$c = \frac{(c_{subdimensions} \times 2.5) + c_{cardinality}}{2.5} \tag{3}$$

Az alapul vett súlyszámot némileg szubjektívan állapítottuk meg, különböző mérések alapján úgy találtuk, hogy az Europeana céljainak ez az arány felel meg.

Implementáció

Az adatfeldolgozó munkafolyamatnak négy fázisa van. Az adatok forrása egy *MongoDB*-adatbázis, amiből az adatokat sororientált JSON-fájlokba exportáljuk (ahol minden sor egy külön rekord), amit Linux fájlrendszerben vagy *Apache Hadoop* fájlrendszerben tárolunk (a kutatás során rendelkezésre álló erőforrások esetében a kettő között nincs jelentős különbség, de egy több számítógépből álló klaszter esetében a *Hadoop* fájlrendszer jobb választás lehet). A rekord szintű elemzést egy, az *Apache Spark* API-t kihasználó *Java* nyelvű szoftver végzi.[29] Mivel a *Spark* automatikusan és

[29] A könyvtár magja: `https://github.com/pkiraly/metadata-qa-api`, Europeana-specifikus kiterjesztés: `https://github.com/pkiraly/europeana-qa-api`, *Spark*-felület: `https://github.com/pkiraly/europeana-qa-spark`. Az API-k (és a MARC elemzőeszköz) lefordított Java könyvtárként is elérhető a Maven központi repozitóriumban: `https://mvnrepository.com/a`

 DOI 10.31400/dh-hun.2019.2.388

konfigurálható módon támogatja a többszálú programfuttatást, az eszköz hatékonyan tudja kihasználni a futtatókörnyezet rendelkezésre álló erőforrásait (akár egyetlen, többmagos processzorú számítógépen, akár nagy kapacitású, több gépből álló számítási klaszteren dolgozunk). A számítások eredménye néhány CSV-fájl, melyeket *Apache Solr* keresőgéppel indexelünk a későbbi visszakeresés céljából – ez az eszköz kijelző felületét („műszerfal") fogja segíteni, ahol a jelentések mögött keresési találati listák húzódnak.

A harmadik fázis a rekordszintű mérési eredmények statisztikai elemzése. Ezek a szoftverek R,[30] illetve (kihasználva a Spark adatelemző API-ját) *Scala* nyelven [31] készültek. Az elemzés az előző fázisban készült CSV-fájlokat olvassa be. A kimenetet a nyers statisztikákat tartalmazó CSV- és JSON-fájlok, illetve a központi tendenciákat vagy az adatok egyéb statisztikai jellegzetességeit tükröző adatvizualizációkat tartalmazó képfájlok alkotják. Az R-nek van azonban egy gyenge pontja: kizárólag a memóriában dolgozik, így a memória mérete meghatározza a feldogozható adathalmaz méretét is. A teljes Europeana adathalmaz statisztikai elemzéséhez az általunk hozzáférhető memória kevésnek bizonyult, ezért kénytelenek voltunk a *Spark* API *Scala* nyelvű megvalósításra áttérni, és mivel a Scala statisztikai eszköztára jóval kisebb, mint a kifejezetten statisztikai elemzésekre tervezett R, ezen a ponton egyelőre kompromisszumokra kényszerültünk.

Az utolsó fázis egy online statisztikai „műszerfal", egy pehelysúlyú PHP és JavaScript alapú weboldal, ami az előző fázisok eredményeit mutatja be.[32] Az összes fázis egyetlen, közepes teljesítményű számítógépen fut (Intel Core i7-4770 Quad-Core processzor, 32 GB DDR3 RAM, Ubuntu 16.04 operációs rendszer), amit párhuzamosan más kutatás-fejlesztési projektek is használtak, ezért az, hogy a számítások erőforrás-kímélőek legyenek a szoftvertervezés során, fontos szemponttá vált.

A számítás adatforrását az Europeana-adatokról készült mentések képezik. Az első ilyen mentés 2015 végén készült az Europeana OAI-PMH szolgáltatása segítségével, amely 46 millió rekordot, 1747 adathalmazt és 3550 adatszolgáltatót tartalmaz.[33] A kutatás időtartama alatt további mentések készültek, a legutolsó 2018 augusztusában (62 millió rekord, összesen 1.27 TB-nyi fájl, az adatforrás ezúttal az Europeana *MongoDB* adatbázisának másolata volt).[34] A DQC célja, hogy havi frissítési ciklust vezessen be, vagyis az Europeana élő adatbázisa és az adatminőséget jelentő weboldal frissítése között ne legyen egy hónapnál hosszabb különbség.

`rtifact/de.gwdg.metadataqa`, így ezek más által írt Java vagy Scala szoftvercsomagokban is felhasználhatóak.

[30] Forráskód: `https://github.com/pkiraly/europeana-qa-r`

[31] `https://github.com/pkiraly/europeana-qa-spark/tree/master/scala`

[32] Forráskód: `https://github.com/pkiraly/europeana-qa-web`

[33] Az adatszolgáltatók neve nincs normalizálva, így előfordul, hogy ugyanaz az intézmény több különböző néven is szerepel.

[34] A kutatás reprodukálhatóságának kedvéért három teljes mentés elérhető a `http://hdl.handle.net/21.11101/0000-0001-781F-7` címről. Az elsőt hosszú távra archiváltuk a göttingeni Humanities Data Centre-ben: `https://hdl.handle.net/21.11101/EAEA0-826A-2D06-1569-0`. A mentések formátuma JSON, soronként egy rekorddal.

Eredmények

Teljesség

Az aldimenziókon (ahol a mezők fontossága számít) és a mezők számosságán (ahol a mező-előfordulások száma a döntő) alapuló megközelítések pontszámainak összehasonlítása rávilágít az eltérő eredményekre. Pearson-féle korrelációs koefficienssel kifejezve 0.59 a korrelációjuk, azonban eloszlásuk alakja és elhelyezkedése különbözik. A számítás módja miatt az összetett pontszám az első megközelítésmódhoz áll közelebb, a számosságon alapulónak kisebb hatása van a végső pontszámra. Az aldimenzión alapuló pontszám alsó és felső határa 0.22 és 0.92 (0–1 közötti skálán), míg a számosságon alapulóé 0.05 és 0.48. Az eloszlás részletei a 2. táblázatban és a 2. ábrán láthatók.

metrika	átlag	szórás	min.	max.
aldimenzión alapuló	0.50	0.07	0.22	0.93
számosságon alapuló	0.12	0.05	0.05	0.48
összetett	0.39	0.06	0.17	0.78

2. táblázat. A teljesség-számítás alapstatisztikái

2. ábra. A teljességszámítások eredményeinek eloszlása

Vannak olyan adatszolgáltatók, melyek összes (esetenként tízezernél is több) rekordjának ugyanaz a pontszáma; ez arra utal, hogy a rekordok struktúrája teljesen egyforma, mivel egyetlen számmal nem, csak mezőszintű elemzéssel lehet igazolni, hogy ezek a

 DOI 10.31400/dh-hun.2019.2.388

rekordok tényleg ugyanabból a (Dublin Core alapú) mezőhalmazból állnak. A másik végponton vannak azok a gyűjtemények, melyekben a pontszám nagy változatosságot mutat. Például aldimenziók tekintetében egy adatszolgáltatónak öt, 0.4-től 0.8-ig terjedő, szinte tökéletesen egyenlő eloszlású pontszáma van, míg az – ugyanezen számítás tekintetében – egyik legjobb gyűjtemény majdhogynem teljesen homogén: a rekordok 99.7%-ának pontszáma 0.9 (és még a maradék 0.3%-nak is 0.8). Ez azt jelenti, hogy az érintett mezők[35] általában nincsenek jelen az első adathalmaz esetében, de szinte mindig jelen vannak a második esetben. Az eszköz különféle ábrákat és táblázatokat kínál a pontszámok megoszlásának vizualizációjára.

A mezők eloszlásának vizsgálatából levonható első következtetés az, hogy sok rekordban nincsenek kontextuális entitások, és csak néhány adatszolgáltató rekordjaiban van 100%-os lefedettség (vagyis, ahol minden rekordban található valamelyik kontextuális entitás). A rekordok 6%-ban van *agent*, 28%-ban *place*, 32%-ban *timespan* és 40%-ban *concept* entitás. Kizárólag a kötelező technikai jellegű adatelemek érhetőek el minden rekordban. Vannak olyan mezők, melyeket ugyan definiál az adatséma, de a rekordokban egyáltalán nem szerepelnek. Vannak ugyanakkor „túlhasznált" mezők, például a dc:description-t gyakorta használják valamilyen specifikusabb mező (tartalomjegyzék, tárgyszó, egyéb cím) helyett.

A felhasználók a „műszerfalon" tudják ellenőrizni a teljes Europeana, egy adott gyűjtemény vagy egy-egy rekord minőségének jellemzőit. Az adatszolgáltatók világos képet kaphatnak az adatokról, és erre az elemzésre alapozva tervezhetik az adattisztító vagy adatjavító lépéseket.

Többnyelvűség

A DQC a közelmúltban publikálta a többnyelvűség számításának részleteit és eredményeit,[36] így ez a rész csak az eredmények rövid összefoglalása. Az EDM az RDF nyelvi annotációs modelljét követi, így az adatok létrehozói meg tudják jelölni, hogy egy adott sztring egy bizonyos nyelven íródott (például *"Brandenburg Gate"@en*, ahol a „Brandenburg Gate" a mezőérték, míg az „en" az angol nyelvet jelöli). A szerkezet neve „felcímkézett érték" (*tagged literal*). A DQC négy releváns rekord szintű metrikát azonosított.

- a felcímkézett értékek száma
- az egyedi nyelvi címkék száma
- a felcímkézett értékek száma nyelvenként
- a nyelvek átlagos száma mezőnként (azokban az esetekben, ahol legalább egy felcímkézett érték szerepel)

Ezeket az értékeket kiszámoltuk az adatszolgáltató proxyjára (vagyis az intézmények szolgáltatta eredeti adatokra), az Europeana proxyjára (ami a tipikusan többnyelvű

[35] Az adatszolgáltatói proxy dc:title, dcterms:alternative, dc:description, dc:type, dc:identifier, dcterms:created, dc:date és dcterms:issued mezői, illetve az aggregálás entitás edm:provider és edm:dataProvider mezői.

[36] Charles et al. 2018, Király et al., 2019.

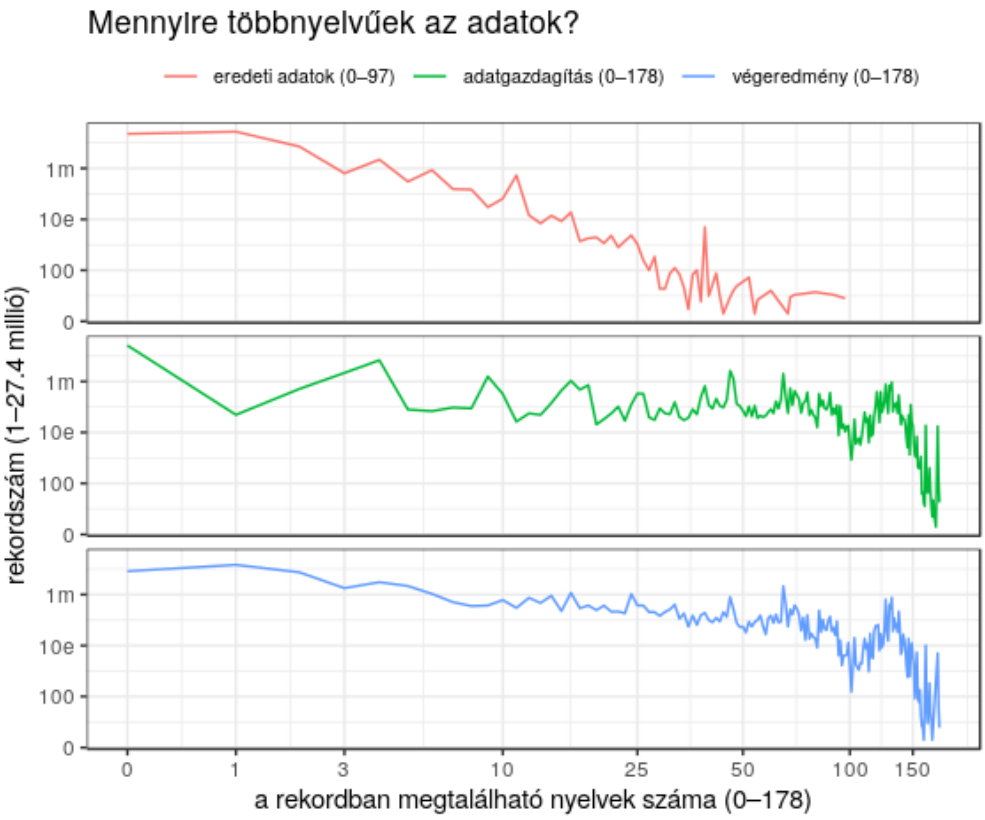
szótárakból származó adatgazdagítást tartalmazza), végül a teljes objektumra. Az eredményt a 3. és 4. táblázat és a 3. ábra tartalmazza.

metrika	szolgáltató	Europeana	teljes
a felcímkézett értékek száma	5.44	64.34	69.79
az egyedi nyelvi címkék száma	1.67	37.92	38.79
a felcímkézett értékek száma nyelvenként	2.64	0.95	2.17
a nyelvek átlagos száma mezőnként azokban az esetekben, ahol legalább egy felcímkézett érték szerepel	1.10	28.10	20.21

3. táblázat. A többnyelvűség metrikái (átlagok)

entitás	0	1	2 vagy több
szolgáltató	22.4M (36.2%)	27.3M (44.1%)	12.1M (19.6%)
Europeana	25.8M (41.7%)	49K (0.07%)	36.1M (58.2%)
teljes	8.2M (13.3%)	14.6M (23.7%)	39.1M (63.0%)

4. táblázat. A nyelvek átlagos számának eloszlása a rekordokban



3. ábra. Többnyelvűség

DOI 10.31400/dh-hun.2019.2.388

A 4. táblázat azt tükrözi, hogy csak a rekordok 20%-ában van két vagy több nyelven elérhető mező az adatszolgáltató proxyjában. Az adatgazdagítási eljárás miatt – ami többnyelvű adatforrásokból, például a DBpediából származó külső kontextuális információkat (a rekorddal kapcsolatos szereplők, fogalmak, helynevek és időpontok adatait) ad az Europeana rekordjaihoz – a többnyelvűség átfogó pontszáma magasabbá vált. Nemcsak hogy növekedett a két vagy többnyelvű mezők száma, de emellett csökkent a nyelvi annotáció nélküli rekordok száma is.

További felismerés, hogy a nyelvi címkék nem mindig szabványosak. Különféle adatszolgáltatók különféle szabványokat követnek, sőt, alkalmasint ad hoc címkéket is használnak. Az egész adathalmazban összesen több, mint 400 különféle nyelvi címke található, melyek közül több is ugyanazt a nyelvet jelöli („en", „eng", „Eng" stb. például az angolt). További kutatásokat kell a normalizált nyelvi címkékkel ellátott rekordok elemzésének szentelni, hogy helyes képet kapjunk a nyelvhasználatról.

Egyediség

A tanulmány elején említettük a hasonló címek példáját, amikor több rekordnak ugyanaz a címe. Ahhoz, hogy megtaláljuk ezeket a rekordokat, ki kell számolnunk az értékek egyediségét (*uniqueness*). Az egyediség azokban a mezőkben pozitív tulajdonság, amelyek az objektum egyedi tulajdonságait írják le, viszont kevésbé pozitív vagy egyenesen negatív azokban, amelyek a rekordokat kontextuális információkhoz kapcsolják, és ahol az értékek szükségképpen valamilyen ellenőrzött szótárból származnak, és így (ideális esetben) több rekord is ugyanazt az értéket használja. Azért, hogy hatékonyan állapíthassuk meg egy érték egyediségét, olyan keresőmotort használtunk, amelyben a mező értékeit teljes kifejezésként, az értéket elmentve indexeltünk. Mivel egy ilyen index felépítése a teljes adathalmaz összes mezőjére több erőforrást igényelt volna, mint amivel rendelkeztünk, a három, ebből a szempontból legfontosabb mezőt indexeltük: a címet (*title*), az egyéb címet (*alternative title*) és a leírást (*description*). A pontszám kiszámítására Solr relevanciaszámításának egy módosított változatát használtuk:

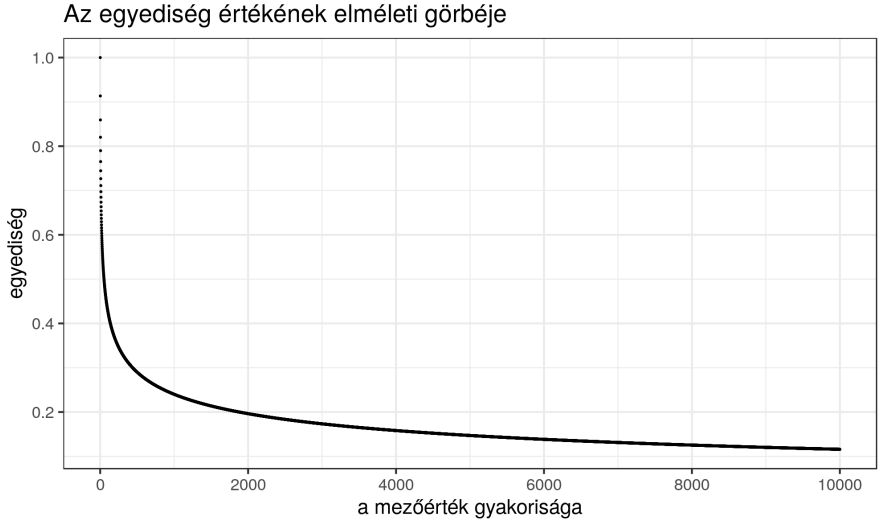
$$score(records_f,\ terms_f) = log\left(1 + \frac{records_f - terms_f + 0.5}{terms_f + 0.5}\right) \tag{4}$$

$$uniqueness_f = \left(\frac{score(records_f,\ terms_f)}{score(records_f,\ 1.0)}\right)^3 \tag{5}$$

ahol $records_f$ azon rekordok száma, amiben f mező elérhető, $terms_f$ az érték gyakorisága. A számítási mód azoknak az értékeknek ad magas pontszámot, melyek egyediek vagy nagyon kevés rekordban fordulnak elő. Minél többször fordul elő egy adott érték, az „egyedisége" annál alacsonyabb. Az eredmény minden esetben 0 és 1 közé eső szám (1 az egyedi mezőértékek pontszáma).

Ahogy a 4. ábrán látszik, a gyakoriság növekedésével a pontszám progresszíven csökken. A felhasználói felületen a következő kategorizálást vezettük be: az egyedi

értékek jelzésén túl további öt, csillaggal jelölt kategória található. A 5. táblázat jelöli a három mező kategóriahatárait.



4. ábra. Az egyediség pontszámának elméleti görbéje. Ahogy a gyakoriság nő, az egyediség pontszáma radikálisan csökken a nulla felé.

mező	*****	****	***	**	*
cím	2-	8-	37-	293-	5226-
egyéb cím	2-	6-	23-	132-	1514-
leírás	2-	7-	34-	252-	4128-

5. táblázat. A gyakoriságon alapuló egyediségkategóriák

A kategorizálás eredménye a 6. táblázatban látható. A rekordok abszolút többsége mindhárom mező esetében egyedi értékeket tartalmaz, azonban milliónyi rekordnak van egy vagy több mező esetében is alacsony pontszáma, sőt legalább tízezer rekordban a három mező közül egyik sem fordul elő. Amikor közösen vizsgáljuk a három mezőt, kiszámítva az eredmények átlagát (lásd a táblázat utolsó sorát) azt találjuk, hogy 25 millió rekord esetében mindhárom mezőnek egyedi értékei vannak, másfelől pedig a rekordoknak csak 3.62%-a tartozik a legalacsonyabb kategóriába. Ez azt jelenti, hogy bár vannak alacsony értékek, a legtöbb esetben van legalább egy mező, aminek kevésbé alacsony az értéke, vagyis nagyobb az esély rá, hogy a rekordot valamilyen keresőkifejezéssel meg lehet találni.

mező	egyedi	*****	****	***	**	*
cím	59.4	9.5	8.3	8.7	7.1	6.6
egyéb cím	62.4	11.2	7.1	3.6	2.7	12.7
leírás	54.6	9.0	7.3	10.2	6.7	11.9
közösen	45.4	10.8	15.6	18.2	6.3	3.62

6. táblázat. Mennyire egyediek az Europeana rekordjai? (%)

 DOI 10.31400/dh-hun.2019.2.388

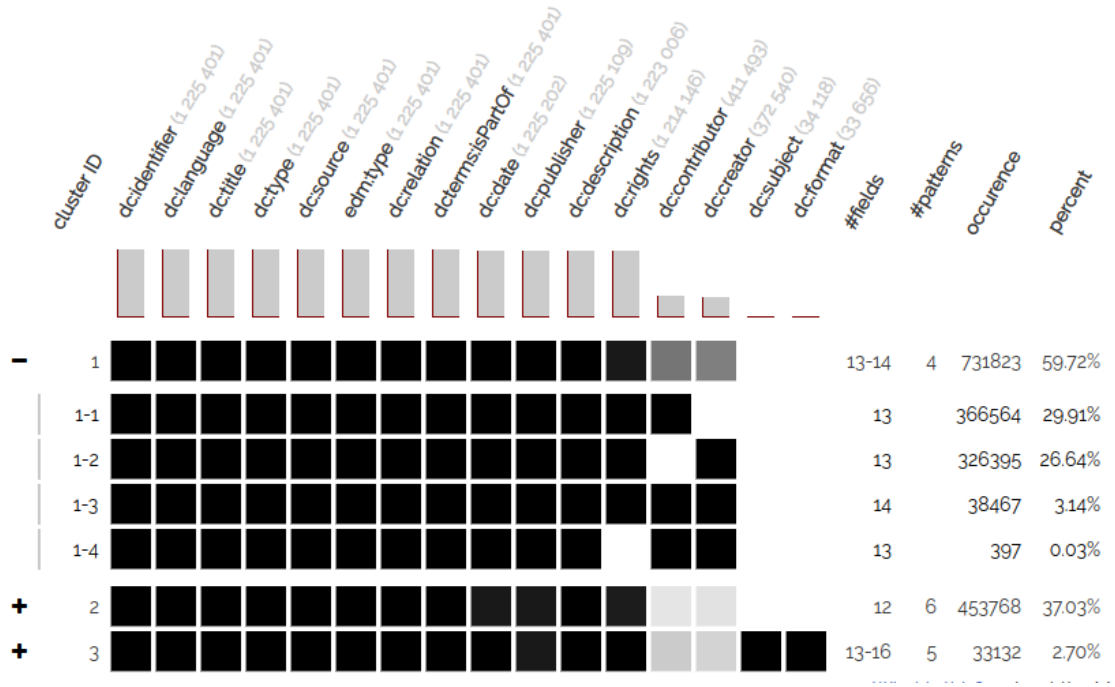
A Solr indexből ki lehet nyerni a leggyakoribb kifejezéseket. A fent említett „photograph" kifejezésen túl számos hasonló található a cím mezőben. Ezek többek között hiányzó információt (például „Unbekannt", „Onbekend" vagy „+++EMPTY+++"), gyűjtemény-, folyóirat- vagy intézménynevet („Journal des débats politiques et littéraires", „ROMAN COIN") vagy valamilyen általános leíró kifejezést („Porträtt", „Château", „Plakat", „Rijksmonument") takarnak. További vizsgálatot igényel azon gyakran előforduló kifejezések kiszűrése, melyek olyan rekordokban szerepelnek, melyekben a többi leíró mező sem rendelkezik a szükséges egyediséggel. Az eszköz megbízható kiindulópontot nyújt egy ilyen vizsgálathoz.

Rekordmintázatok

Milyen mezőkből áll a tipikus rekord? Másként fogalmazva: mely mezőket használják az adatszolgáltatók? A rekordmintázatok a rendszeresen együttálló mezők. Mivel a teljességmérés kinyeri az összes mező jelenlétét, ebből egy MapReduce algoritmus alapú elemzéssel ki lehet nyerni az együttállási mintázatokat. Itt a leképező „mapping" funkció létrehozza a mintázatot (ami a rekordban elérhető mezők rendezett listája), a redukáló „reduce" funkció pedig megszámolja azokat. Az algoritmus első megvalósításakor kiderült, hogy túl sok hasonló minta van, amelyeket érdemes lenne csoportosítani, hogy hatékonyan elemezhessük, ezért egy hasonlóan működő csoportosító algoritmust is alkalmaztunk. Ebben minden mintát először nullákból és egyesekből álló sztringgé alakítottunk. Egy-egy gyűjteményben előforduló összes mezőt szabott sorrendbe raktunk. A mezőket a következő három osztályba soroltuk: kötelező mezők, fontos mezők (melyek előfordulnak valamelyik aldimenzióban) és nem kitüntetett mezők. Ha a mező előfordul egy mintában, akkor azt egy vagy több egyes szám jelöli, máskülönben egy vagy több nulla. A kötelező mezők három számot kapnak, a fontosak kettőt, a maradék pedig egyet. Ekkor azok a minták, melyekben ugyanazok a fontos mezők, de különböznek, a nem fontos mezők közelebb kerülnek egymáshoz, mint azok, melyekben a nem fontos mezők egyeznek meg. A hasonlóságot a Jaro–Winkler algoritmussal[37] számoltuk. Megjelenítéskor (lásd 5. ábra) alapértelmezésben a csoportok jelennek meg, de a felhasználó kattintására megjelennek a csoportban szereplő egyes minták. A táblázat a csoporthoz/mintához tartozó rekordok száma alapján van rendezve, így a legtipikusabb rekordok kerültek legfelülre. Ha a mező nem érhető el minden rekordban, az azt reprezentáló négyzet szürke színben jelenik meg (a színárnyalat a rekordok számával arányos). Alapértelmezésben csak azok a csoportok jelennek meg, melyek a rekordok legalább 1%-át reprezentálják.

A rekordminták segítségével eddig kétfajta minőségi problémát tártunk fel. Az első azon rekordokra vonatkozik, melyek kevés számú mezőt tartalmaznak. Több, mint 150 000 olyan rekord van, melynek szolgáltatói proxyjában csak a következő négy mező szerepel: dc:title, dc:type, dc:rights, edm:type. Ezek közül csak az első kettő tartalmaz leíró jellegű információkat az objektumról. Nyilvánvaló, hogy a felhasználók nagy eséllyel nem lesznek képesek ezen rekordokhoz hozzáférni a facettákon keresztül, mivel hiányoznak az ehhez szükséges információk. A második probléma a szerkezeti homogeneitás: bizonyos gyűjtemények minden rekordja ugyanazokat a

[37] `https://en.wikipedia.org/wiki/Jaro%E2%80%93Winkler_distance`



5. ábra. A rekordmintázatok csoportosítása. Az első sor a hasonló minták csoportját jelenti. A következő négy sor a csoportba tartozó egyes mintákat. A felső szürke sáv jelenti a mezők gyakoriságát az adott gyűjteményen belül.

mezőket tartalmazza. Az Europeanában összesen 906 ilyen adatszolgáltató van, de szerencsére többségük viszonylag kis gyűjtemény, csak huszonhatuknak van ezernél több rekordja. Ugyanakkor a legnagyobb homogén gyűjtemény (több, mint 500 000 rekorddal) csak 5 mezőt tartalmaz, melyből 3 leíró jellegű. Ezekkel a rekordokkal az a probléma, hogy speciális mezők helyett általános mezőket tartalmaznak (például nem tesznek különbséget a fogalmi, térbeli és időbeli tárgyszavak, osztályozások között, és eltérő típusú kontextuális információkat egyaránt a dc:type vagy a dc:subject mezőben tárolnak).

További kutatási tervek

Az Europeana a Metis[38] nevű új begyűjtési rendszer bevezetésén dolgozik, mely a tervek szerint integrálni fogja a jelen kutatás során fejlesztett eszközt. Amikor egy új rekordhalmaz importjára kerül sor, a mérés automatikusan elindul, a folyamat koordinálásával megbízott munkatárs ellenőrizheti a minőségjelentést, és ennek kimenetét, valamint saját következtetéseit megoszthatja az adatszolgáltatóval, aki reagálhat erre akár úgy, hogy megváltoztatja az adatátalakítás szabályait, akár úgy, hogy – ha lehetséges – javítják a hibákat.

A tárgyalt számítási modelleken felül számos olyan metrika van, amit kiszámítani tervezünk a közeljövőben (pontosság, információtartalom, naprakészség, ismert metaadat anti-patternek felismerése). A releváns szakirodalom azt ajánlja, hogy alakítsunk

[38] `https://github.com/europeana/metis-framework`

 DOI 10.31400/dh-hun.2019.2.388

ki egy olyan csúcs szintű pontszámot, ami az összes metrikát egy számban összegzi, így egymagában jellemzi a metaadatrekord minőségét. Ezt a metrikák súlyozásával, illetve olyan dimenziócsökkentő gépi tanuló algoritmusokkal lehet elérni, mint az elsődleges komponenselemzés (*Principal Component Analysis*).[39] Korábban említettük, hogy a jelenlegi teljességszámítási megközelítések a mező jelenlétét elemzik. Ezen a fronton a következő lépés a modell kiterjesztése a releváns mezők tartalmi értékelésének irányába, a felhasználói forgatókönyvek elemzésének megfelelően.[40]

A DQC-n belül tervezzük az itt ismertetett eredmények szakértői értékeléssel és a (naplófájlokon alapuló) használati adatokkal való összevetését. Corey Harper ismertette az Europeanához céljában és metaadatsémájában is hasonló Amerikai Digitális Közkönyvtár (Digital Public Library of America) adatain lefuttatott tesztet, amelyben azt kísérelte feltárni, hogy van-e összefüggés az objektum használata (a portálon és az API-n mérhető használati gyakoriság) és a minőségmérés során számított pontok között. A teszt sajnos sikertelen volt részben azért, mert a kutatás időpontjában még nem állt rendelkezésre statisztikai következtetések levonására alkalmas mennyiségű adat; a javasolt módszer azonban ígéretes, és ha az Europeanának elérhetőek a naplófájljai, érdemes lenne lefuttatni a kísérletet.

Tervezzük továbbá a problémakatalógus elemeinek szabatos meghatározását a W3C Shapes Constraint Language (alaki követelmények nyelve) segítségével. További tervünk, hogy az eredményeket az Adatminőségi Ontológiának megfelelő kapcsolt adatként publikáljuk.[41]

Az itt javasolt módszert más metaadatsémákra is alkalmazni lehet, például többek között MARC alapú könyvtári katalógusokra,[42] EAD alapú levéltári gyűjteményekre.[43]

Összegzés

A kutatás során (a DQC-vel együttműködésben) újragondoltuk a funkcionalitás és a metaadatséma viszonyát, és implementáltunk egy keretrendszert, amellyel sikerrel tudtuk mérni a metaadatproblémákkal korreláló szerkezeti jellegzetességeket. A keretrendszer felhasználója képes kiválogatni alacsony és magas minőségű rekordokat. Kutatási hipotézisunk szerint az olyan szerkezeti jellemzők, mint a mezők jelenléte és számossága korrelálnak a metaadat minőségével, ami igaznak bizonyult. A kutatás azáltal, hogy a korábbi szakirodalomban nem említett big data eszközöket vezetett be, kiterjesztette az elemzett rekordok mennyiségét.

A kutatás egy bizonyos adathalmazt és metaadatsémát fogott át, azonban az alkalmazott módszer általános algoritmusokon alapszik, így az más adatsémára is alkalmazható. Számos digitális bölcsészeti kutatás (néhány példa: KOLIMO [Corpus of

[39] Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani, *An Introduction to Statistical Learning*, (New York: Springer, 2013) `https://doi.org/10.1007/978-1-4614-7138-7`

[40] Hill–Charles–Isaac, *Discovery - User scenarios.*

[41] Ricardo Albertoni and Antoine Isaac, *„Data on the Web Best Practices: Data Quality Vocabulary,"* W3C note, (W3C, 2016) `https://www.w3.org/TR/2016/NOTE-vocab-dqv-20161215/`

[42] Mivel a MARC-ban számos adattartalomra vonatkozó szigorú szabály van, az EDM-ben pedig kevés, számottevő eltérés mutatkozik a két mérési módszer között.

[43] A legnagyobb európai levéltári gyűjtemény, az Archives Portal Europe (`http://www.archivesportaleurope.net/`) az adatait egy REST API segítségével CC0 licensszel publikálta.

Literary Modernism],[44] Omniart,[45] Schmidt[46]) alapul valamilyen kulturális adatbázist definiáló sémán. Ezeket a kutatási folyamatokat javítani lehet az adatforrások gyenge pontjainak feltárásával, és – Felix Raunak a dolgozat elején idézett tweetjére reagálva – a forrásokra vonatkozó pontosabb feltételezésekkel; így a következtetések is megbízhatóbbak lesznek.

Köszönetnyilvánítás

Az első szerző szeretne köszönetet mondani az Europeana Adatminőségi Bizottság régebbi és jelenlegi tagjainak; doktori kutatása témavezetőinek, Gerhard Lauernek és Ramin Yahyapournak; Jakob Voßnak, Juliane Stillernek, Mark Philippsnek a visszajezésekért; Christina Harlownak és Zaveri Amrapalinak az inspirációért; Felix Raunak a mottóért, a GWDG-nek pedig a kutatás támogatásáért.

Measuring completeness as metadata quality metric in Europeana

Europeana, the European digital platform for cultural heritage, has a heterogeneous collection of metadata records ingested from more than 3200 data providers. The original nature and context of these records were different. In order to create effective services upon them we should know the strength and weakness or in other words the quality of these data. This paper proposes a method and an open source implementation to measure some structural features of these data, such as completeness, multilinguality, uniqueness, record patterns, to reveal quality issues.

Keywords:
big data applications, data analysis, data collection, quality of service, quality management, metadata, data integration

[44] `https://kolimo.uni-goettingen.de/`

[45] Gjorgji Strezoski and Marcel Worring, „OmniArt: Multi-task Deep Learning for Artistic Data Analysis," *CoRR*, 2017. `http://arxiv.org/abs/1708.00684`

[46] Benjamin Schmidt, „Stable Random Projection: Standardized Universal Dimensionality Reduction for Library-Scale Data," in R. Lewis et al. eds., *Digital Humanities 2017. Conference Abstracts* (Montréal: McGill University–Université de Montréal, 2017) 340–342. `https://dh2017.adho.org/abstracts/497/497.pdf`

 DOI 10.31400/dh-hun.2019.2.388