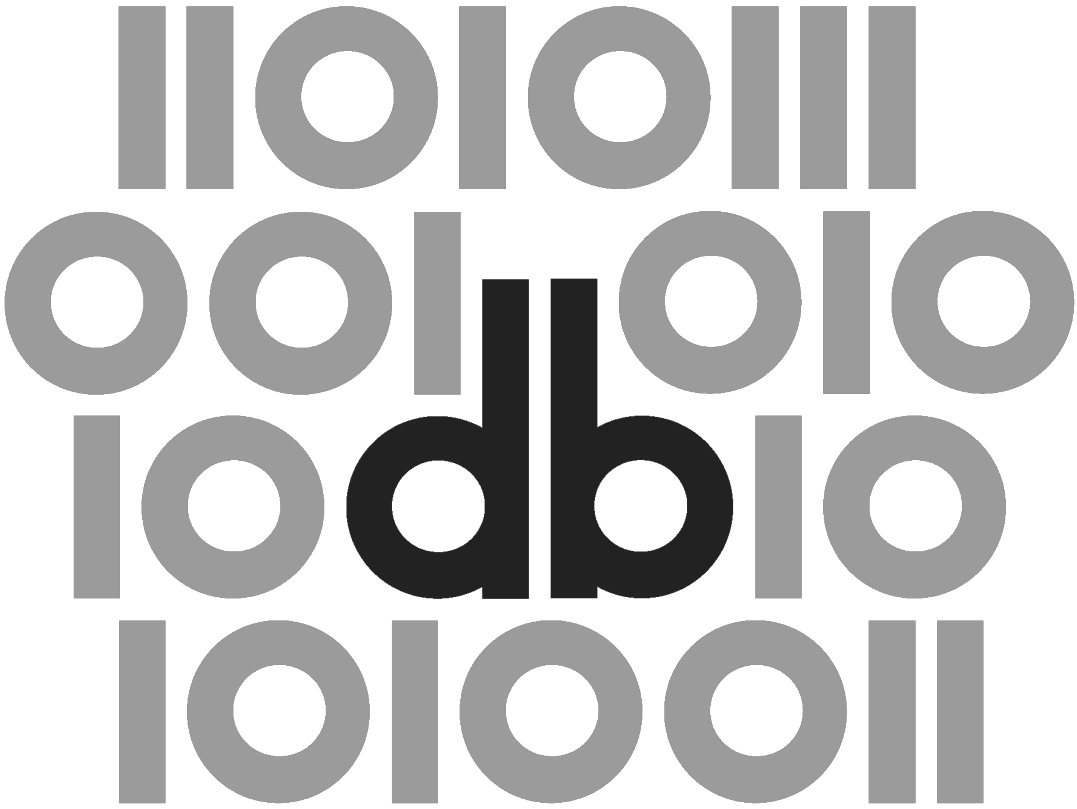
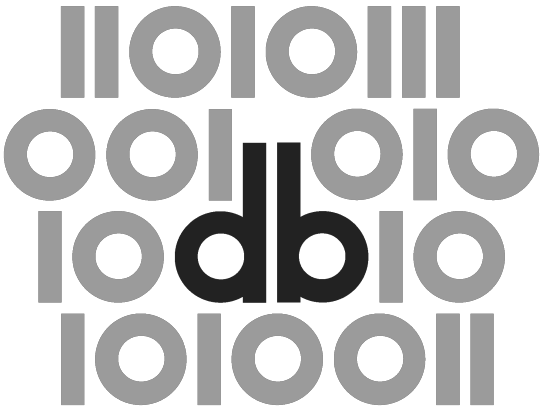


Digitális Bölcsészet
2021., negyedik szám

<DIGITÁLIS BÖLCSÉSZET>



4 (2021)

ISSN 2630-9696
DOI: 10.31400/dh-hun.2021.4

Kiadja a Bakonyi Géza Alapítvány és az ELTE BTK Régi Magyar Irodalom Tanszéke (1088 Budapest, Múzeum krt. 4/A).
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/digitalisbolcseszet`
Email cím: dbfolyoirat@gmail.com

Olvasószerkesztő: Bucsics Katalin
Tördelés: Hegedüs Béla
Grafika: Hegyi Gábor

<MŰHELY>

Péter Róbert 0000-0002-7972-4751

Szegedi Tudományegyetem

robert.peter@ieas-szeged.hu

Szántó Zsolt 0000-0002-8924-206X

Szegedi Tudományegyetem

szantozs@inf.u-szeged.hu

Bilicki Vilmos 0000-0002-7793-2661

Szegedi Tudományegyetem

bilickiv@inf.u-szeged.hu

Berend Gábor 0000-0002-3845-4978

Szegedi Tudományegyetem

berendg@inf.u-szeged.hu

Az AVOBMAT (Analysis and Visualization of Bibliographic Metadata and Texts) többnyelvű kutatási eszköz bemutatása

E dolgozat célja, hogy bemutassa az AVOBMAT (Analysis and Visualization of Bibliographic Metadata and Texts) többnyelvű kutatási eszköz működéséhez kapcsolódó munkafolyamatot és a különböző elemzőfunkciókat. A webes alkalmazás segítségével nagy mennyiségű metaadatot és szöveget lehet feldolgozni és kritikusan elemezni adatvezérelt, mesterséges intelligenciával és természetesnyelv-feldolgozásos technológiákkal támogatott módszerekkel és eszközökkel. Az AVOBMAT szöveg- és adatbányászati eszköz újdonságai a következők: (i) számos nyelven képes előfeldolgozni, (szemantikusan) gazdagítani és elemezni metaadatokat és szövegeket; (ii) a beépített funkciók lehetőséget biztosítanak a szoros és távoli olvasásra egyaránt; (iii) egy felhasználóbarát, interaktív grafikus felületen integrál metaadat és szövegelemzéssel kapcsolatos kutatási eszközöket. A platformfüggetlen alkalmazás elsősorban olyan felhasználók számára lett kifejlesztve, akik nem rendelkeznek programozási ismeretekkel. Az egyszerűen használható felület interaktív paraméterbeállítást és vezérlést biztosít a normalizálást is támogató előfeldolgozástól az analitikai szakaszokig.

A felhasználók interaktív módon kísérletezhetnek az elemzések különböző beállításaival a munkafolyamat során. Ezáltal az AVOBMAT segít felismerni a számítógépes szöveg- és adatelemzés episztemológiai kihívásait, korlátait és erősségeit, valamint kritikus módon értelmezni az alkalmazott módszereket és eredményeket.

Kulcsszavak:
szöveg- és adatbányászat, természetesnyelv-feldolgozás, többnyelvű kutatási eszköz, metaadat, szemantikus adatgazdagítás

1. Bevezetés

Az elmúlt két évtizedben hatalmas mennyiségű nyomtatott forrást digitalizáltak és kódoltak eltérő minőségben és módokon. Ezek a digitális anyagok és hozzájuk tartozó bibliográfiai adatok nyílt hozzáférésű vagy előfizetést igénylő adatbázisokban korlátozott módon elérhetőek és kereshetőek. Temérdek – szövegek, metaadatok tárolására és keresésére használható – adatbázis jött létre (és tűnt el) az elmúlt időszakban,[1] de ezek az alkalmazások ritkán megfelelőek kutatási kérdések megválaszolására.[2] Emellett a főként angol, német és francia nyelvű dokumentumok vizsgálatára finomhangolt szoftverek is problémákat, nehézségeket okoznak a nem világnyelveken írott szövegekkel foglalkozó (digitális) bölcsészeti kutatások során. Ezeket a kihívásokat felismerve számos európai országban – jelentős részben a Digital Research Infrastructure for the Arts and Humanities (DARIAH) támogatásával – az elmúlt években kezdték kiépíteni azokat a digitális bölcsészeti kutatási infrastruktúrákat, amelyek lehetőséget biztosítanak a digitális tartalmak feltáró – kutatási problémák megoldását elősegítő – elemzéséhez, megosztásához és megőrzéséhez.[3] Többek között ilyen elképzelések

[1] Mivel a digitális bölcsészeti projektek jelentős része pályázati forrásból valósul meg, a pályázatok lezárását követően számos esetben nincs lehetőség a létrejött adatbázisok és szoftverek frissítésére, valamint ezek szervereken való működtetésére és tárolására. Ez az egyik fő oka annak, hogy számos digitális bölcsészeti eszköz és adatbázis csak korlátozott ideig elérhető. A problémát felismerve az elmúlt néhány évben kiírt európai uniós és hazai pályázatok különös hangsúlyt fektetnek a fenntarthatóságra és a kutatási infrastruktúrák kiépítésére.

[2] John Bradley, „Digital Tools in the Humanities: Some Fundamental Provocations?," *Digital Scholarship in the Humanities* 34, 1. sz. (2019): 13–20, https://doi.org/10.1093/llc/fqy033; Marijn Koolen, Jasmijn van Gorp and Jacco van Ossenbruggen, „Toward a Model for Digital Tool Criticism: Reflection as Integrative Practice," *Digital Scholarship in the Humanities* 34, 2. sz. (2019): 368–385, https://doi.org/10.1093/llc/fqy048; John Unsworth, „Scholarly Primitives: What Methods Do Humanities Researchers Have in Common, and How Might Our Tools Reflect This?" hozzáférés: 2021.12.15, https://johnunsworth.name/Kings.5–00/primitives.html.

[3] Maria Ågren, Claudine Moulin, Marko Tadic, Julianne Nyhan, Arianna Ciula, Margaret Kelleher, Elmar Mittler, Andrea Bozzi and Kristin Kuutma, *Science Policy Briefing: Research Infrastructures in the Digital Humanities* (Strasbourg: European Science Foundation, 2011), https://www.esf.org/fileadmin/user_upload/esf/RI_DigitalHumanities_B42_2011.pdf. Németországban, Finnországban és Lengyelországban 2021-ben indultak olyan projektek, melyek célja digitális bölcsészeti

 DOI 10.31400/dh-hun.2021.4.3530

motiválták az AVOBMAT (Analysis and Visualization of Bibliographic Metadata and Texts) kutatási eszköz – önmagában nem infrastruktúra – fejlesztését, amely vállalkozás 2017-ben kezdődött a Szegedi Tudományegyetemen.

A dolgozat célja, hogy bemutassa az AVOBMAT többnyelvű kutatási eszköz működéséhez kapcsolódó munkafolyamatot és a különböző elemző funkciókat.[4] A programozási ismereteket nem igénylő webes alkalmazás segítségével nagy mennyiségű metaadatot és szöveget lehet feldolgozni és kritikusan elemezni korszerű, adatvezérelt és természetesnyelv-feldolgozásos technológiákkal támogatott módszerekkel és eszközökkel.

Első lépésben ismertetjük és összehasonlítjuk az AVOBMAT-hoz funkcionalitásában hasonló alkalmazásokat és bemutatjuk az AVOBMAT újdonságait. A következő fejezet az elemezni kívánt dokumentumok előfeldolgozásáról és a különböző feltöltési opciókról ad áttekintést. Ezután szemléltetjük, milyen módokon kereshetünk a feltöltött dokumentumok metaadatai és szövegei között. A keresési funkciók segítségével szűkíthetjük a korpuszunkat és meghatározhatunk egy alkorpuszt, amelyen a különböző metaadat- és szövegelemzéseket elvégezhetjük. A dolgozat következő részében a metaadat-elemzési lehetőségek és ezekhez kapcsolódó vizualizációk kerülnek bemutatásra konkrét példákon keresztül. Majd a tanulmány leghosszabb fejezete betekintést nyújt a szövegek tartalmi vizsgálatára vonatkozó paraméterezhető elemzők (pl. témamodellezés, szóstatisztikai vizsgálatok, névelem-felismerés) működésébe. Az

kutatási infrastruktúrák létrehozása. Vannak olyan futó projektek is, melyek nemzetközi kutatási infrastruktúrát szeretnének létrehozni a digitális bölcsészet egy adott részterületén. Lásd például a Computational Literary Studies Infrastructure-t, hozzáférés: 2021.12.15, `https://clsinfra.io/`. Hazánkban jelenleg nincs digitális bölcsészeti kutatási infrastruktúra, bár 2021-ben nálunk is megfogalmazódtak ilyen irányú tervek (pl. Digital Humanities Platform: dHUpla) a Petőfi Irodalmi Múzeum Digitális Bölcsészeti Központjában, valamint a Kulturális Örökség Nemzeti Laboratóriumban, de konkrét, kutatók által is használható infrastruktúrák még nem állnak rendelkezésre. dHUpla, hozzáférés: 202112.15, `https://dhupla.hu/` és DH-LAB, hozzáférés: 2021.12.15, `https://dh--lab.hu/`.

[4] Az AVOBMAT korábbi verzióját az alábbi publikációban és poszteren mutattuk be: Róbert Péter, Zsolt Szántó, József Seres, Vilmos Bilicki and Gábor Berend, „AVOBMAT: A Digital Toolkit for Analysing and Visualizing Bibliographic Metadata and Texts," in Berend Gábor, Gosztolya Gábor és Vincze Veronika, szerk., *XVI. Magyar Számítógépes Nyelvészeti Konferencia,* 43–55 (Szeged: Szegedi Tudományegyetem, Informatikai Intézet, 2020), `http://acta.bibl.u-szeged.hu/67682/`; Zsolt Szántó, József Seres, Vilmos Bilicki, Bendegúz M. Bendicsek, Gábor Berend and Róbert Péter, „Introducing the AVOBMAT (Analysis and Visualization of Bibliographic Metadata and Texts) Multilingual Research Tool," in *DARIAH: Virtual Annual Event 2020. Poster Exhibition,* hozzáférés: 2021.12.15, `https://www.virtualdariah2020.dariah.eu/posters/#lightbox--gallery-1/7/`. Az AVOBMAT fejlesztését részben az EFOP-3.6.1-16-2016-00008 és az EFOP-3.6.3-VEKOP-16-2017-0002 azonosítószámú pályázatok támogatták. Az előbbi pályázat keretében jött létre a TANIT *(Text ANalysIs Tools)* morfológiai elemző (`http://dighum.bibl.u-szeged.hu/tanit/`). Labádi Gergely, Farkas Richárd, Nagy Roland és Péter Róbert, „TANIT: Magyar nyelvű szövegeket elemző eszköz összehasonlító digitális bölcsészeti feladatokhoz," in Vincze Veronika, szerk., *XIV. Magyar Számítógépes Nyelvészeti Konferencia*, 450–455 (Szeged: JATEPress, 2018), `http://real.mtak.hu/86149/1/teljesB5-460-465.pdf`. Köszönettel tartozunk Ficand Tamás, Simon Gábor, Dér Gergely, Seres József és Bendicsek M. Bendegúz hallgatóknak, akik részt vettek az AVOBMAT fejlesztésében, valamint Kokas Károly, Sándor Ákos, Nagy Gyula és Erdődi Zoltán (SZTE Klebelsberg Könyvtár) informatikus könyvtárosoknak, akik biztosították a technikai hátteret és számos adatbázist az említett szoftverek teszteléséhez.

összegzés és a fejlesztési tervek felvázolása előtt röviden szólunk a jogosultságkezelésről, valamint a felhasználók által nem látható adminisztrátori felület által kínált lehetőségekről.

2. Hasonló alkalmazások és újdonságok

Az olyan digitális bölcsészethez köthető kereskedelmi termékekben, mint a *Gale Digital Scholar Lab*[5] (az ára miatt csak a leggazdagabb egyetemeken és kutatóintézetekben érhető el), az egyéni vagy könyvtári adatbázisok, a(z előre betanított) nyelvi modellek és szótárak nem tölthetők fel. A szintén magas előfizetési áron megrendelhető *ProQuest Text* and *Data Mining Studio*[6] alkalmazás használatához Python és R programozási ismeretek szükségesek. A legtöbb jelenleg elérhető webes szövegelemző eszköz, mint például az angol nyelvű dokumentumok feldolgozására fókuszáló *Voyant Tools*[7] vagy a *Topics Explorer,*[8] nem tud megbirkózni nagy szövegkorpuszokkal. A szöveg- és adatelemzésre alkalmazható Python- és R-könyvtárak konfigurációs beállításaihoz képest a viszonylag kevés böngészőalapú digitális bölcsészettudományi alkalmazásban csak korlátozott számú konfigurációs paramétert lehet beállítani az egyes elemzők esetében. A *Paper Machine*[9] (a *Zotero* hivatkozáskezelő szoftver bővítménye) egykor ötvözte az alapvető bibliográfiai metaadatokat (dátum, cím és kiadási hely) és témamodellezési elemzőket, de ez már nem kompatibilis a *Zotero* jelenlegi, 5-ös verziójával. Az *Interactive Text Mining Suite*[10] webes alkalmazás előfeldolgozza a TXT és PDF formátumú szövegeket, amelyeken klaszter-, téma- és gyakoriságelemzéseket végez, de nagyon kevés metaadatmezőt (szerző, év, cím és kategória) tud kezelni, valamint a metaadatokat nem képes önállóan elemezni. A böngészőalapú alkalmazások közül a *Lexos*[11] kínálja a legtöbb eszközt a szövegek (TXT, HTML és XML) előfeldolgozására és szegmentálására. A *Lexos* tokenizálja a szövegeket, azonosítja az n-gramokat, statisztikai összefoglalókat készít, vizualizálja az eredményeket különböző típusú szófelhők, dendrogramok és konszenzusfák segítségével. A szövegek összehasonlítása mellett, a MALLET[12] által generált adatokon alapuló „témafelhőt" is lehet a *Lexos*szal készíteni,

[5] *Gale Digital Scholar Lab,* hozzáférés: 2021.12.15, https://www.gale.com/primary-sources/digital-scholar-lab.

[6] *ProQuest Text and Data Mining Studio,* hozzáférés: 2021.12.15, https://about.proquest.com/en/products-services/TDM-Studio/.

[7] Stéfan Sinclair and Geoffrey Rockwell, *Voyant Tools*, hozzáférés: 2021.12.15, http://voyant-tools.org/.

[8] *TopicsExplorer*, hozzáférés: 2021.12.15, https://dariah-de.github.io/TopicsExplorer/.

[9] *Paper Machines*, hozzáférés: 2021.12.15, http://papermachines.org/.

[10] *Interactive Text Mining Suite,* hozzáférés: 2021.12.15, https://languagevariationsuite.wordpress.com/2016/03/18/interactive-text-mining-suite-itms/; Olga Scrivner and Jefferson Davis, „Interactive Text Mining Suite: Data Visualization for Literary Studies," in Thierry Declerck and Sandra Kübler, eds., *Proceedings of the Workshop on Corpora in the Digital Humanities*, 29–38 (Bloomington, 2017), http://ceur-ws.org/Vol-1786/scrivner.pdf.

[11] *Lexos*, hozzáférés: 2021.12.15, http://lexos.wheatoncollege.edu/upload; Scott Kleinman, Mark D. LeBlanc, Michael D. C. Drout, and Weiqi Feng, *Lexos. v4.0*, hozzáférés: 2021.12.15, https://github.com/WheatonCS/Lexos/.

[12] Andrew Kachites McCallum, „MALLET: A Machine Learning for Language Toolkit," hozzáférés: 2021.12.15, http://mallet.cs.umass.edu.

 DOI 10.31400/dh-hun.2021.4.3530

de hagyományos (pl. Latent Dirichlet Allocation alapú) témamodellezést nem lehet ezzel végezni. Az eddig említett ingyenesen hozzáférhető eszközök egyike sem képes bibliográfiai adatok és szövegek (szemantikus) gazdagítására és elemzésére, valamint a szövegelemzéshez feltöltött digitális gyűjtemények szűrésére metaadatok vagy (teljes szövegű) kulcsszavas keresések segítségével, beleértve a közelítő *(fuzzy),* szószomszédsági *(proximity)* és parancssori lekérdezéseket. Az eleddig felsorolt szövegelemzési funkciók jelentős része (és számos egyéb, szövegek összehasonlítására alkalmas elemző) megtalálható az *impresso: Media Monitoring of the Past,*[13] újságok elemzésére specializálódott kutatási eszközben. Ennek a korszerű szövegbányászati eszköznek az a hátránya, hogy csak fix, főleg svájci újságkorpuszokon működik. Az ismertetett alkalmazásokból merítettünk ötleteket és számos elemző funkciót beépítettünk az AVOBMAT-ba is.

Az AVOBMAT kutatási eszköz újdonságai a következők: (i) számos nyelven képes előfeldolgozni, (szemantikusan) gazdagítani és elemezni metaadatokat és szövegeket; (ii) a beépített funkciók lehetőséget biztosítanak a szoros *(close)* és távoli *(distant)* olvasásra egyaránt; (iii) egy felhasználóbarát, programozási tudást nem igénylő grafikus felületen integrál metaadat- és szövegelemzéssel kapcsolatos kutatási eszközöket. Az interaktív felületen a legtöbb esetben ki-be kapcsolhatóak a megjelenített metaadatmezők, valamint módosíthatóak az elemzési paraméterek és ezután újrafuttathatók az elemzések. A távoli és szoros olvasási megközelítések elemzési keretrendszerünkben történő kombinálásával a felhasználók új perspektívákat azonosíthatnak a bibliográfiai adatok és a szövegelemzés kapcsán, valamint eleddig ismeretlen összefüggéseket fedezhetnek fel a digitális gyűjteményekben. Az AVOBMAT lehetővé teszi, hogy a felhasználó által tetszőlegesen konfigurálható előfeldolgozás után feltöltött adatbázisokat különböző típusú metaadatok és teljes szöveges keresések alapján szűrjék, és a szűrt alkorpuszon bibliográfiai, hálózati, valamint természetesnyelv-feldolgozással kapcsolatos elemzéseket végezzenek. Az eddigi digitális módszereket használó kutatások nem igazán aknázták ki a bibliográfiai (meta)adatok elemzésében rejlő lehetőségeket. A legújabb kutatások igazolják, hogy a szövegbányászathoz hasonlóan a bibliográfiai adatok (úgyis mint *big data*) kritikus vizsgálata is számos új felismerést nyújthat, eddig figyelmen kívül hagyott mintákat és trendeket tárhat fel, új típusú bizonyítékokkal és eredményekkel szolgálhat, valamint megkérdőjelezhet, finomíthat régi hipotéziseket a bölcsészettudományok területén.[14] Például a 18. századi tanulmá-

[13] impresso. Media Monitoring of the Past, hozzáférés: 2021.12.15, `https://impresso-project.ch`; Matteo Romanello, Maud Ehrmann, Simon Clematide and Daniele Guido, „The Impresso System Architecture in a Nutshell," *Technical Report, EuropeanaTech Insights*, 16 (2020), `https://pro.europeana.eu/page/issue-16-newspapers#the-impresso-system-architecture-in-a--nutshell`.

[14] Iraklis Varlamis and George Tsatsaronis, „Visualizing Bibliographic Databases as Graphs and Mining Potential Research Synergies," in Randall Bilof, ed., *2011 International Conference on Advances in Social Networks Analysis and Mining*, 53–60 (Piscatawy, NJ: The Institute of Electrical and Electronics Engineers, 2011), `https://doi.org/10.1109/ASONAM.2011.52`; Róbert Péter, „Researching (British Digital) Press Archives with New Quantitative Methods," *Hungarian Journal for English and American Studies* 17, 2. sz. (2011): 283–300, https://www.jstor.org/stable/43487818; Katrina Fenlon, Miles Efron and Peter Organisciak, „Tooling the Aggregator's Workbench: Metadata Visualization through Statistical Text Analysis," *Proceedings of the American Society for Information Science and Technology* 49, 1. sz. (2012): 1–10, `https://doi.org/10.1002/meet.14504901161`; Franco Mo-

nyok kapcsán ezekre kiváló példákat találunk többek között Mikko Tolonen, Simon Burrows, Dan Edelstein, Mark Towsey és Alicia Montoya vezette kutatócsoportok publikációiban.[15]

3. A konfigurálható előfeldolgozás és feltöltés

Az előfeldolgozási fázisban a felhasználó konfigurálhatja az egyes elemzőket, valamint a metaadat- és szemantikus gazdagítást végző eszközöket. A munkafolyamat első lépése az a feltöltendő szövegekre épülő automatikus nyelvdetekció, melynek eredményét az elemzések során figyelembe veszi a program. Ehhez a művelethez a *lang-*

retti, *Distant Reading* (London; New York: Verso, 2013); Andrew Prescott, „Bibliographic Records as Humanities Big Data," in Xiaohua Tony Hu et al., eds., *2013 IEEE International Conference on Big Data,* 55–58 (Piscatawy, NJ, 2013), https://doi.org/10.1109/BigData.2013.6691670; Matthew L. Jockers, *Macroanalysis: Digital Methods and Literary History* (Champaign, IL: University of Illinois Press, 2013); Andrew Prescott, *Big Data in the Arts and Humanities: Some Arts and Humanities Research Council Projects* ([Glasgow]: University of Glasgow, 2015); Shawn Graham, Ian Milligan and Scott Weingart, *Exploring Big Historical Data: the Historian's Macroscope* (London: Imperial College Press, 2016), https://doi.org/10.1142/p981; Jean-Philippe Moreux, „Innovative Approaches of Historical Newspapers: Data Mining, Data Visualization, Semantic Enrichment: Facilitating Access for various Profiles of Users," in *IFLA News Media Section, Lexington, August 2016, At Lexington, USA,* 1–16 (Lexington: IFLA, IFLA, 2016), https://hal-bnf.archives-ouvertes.fr/hal-01389455/document; Giovanni Schiuma and Daniela Carlucci, *Big Data in the Arts and Humanities: Theory and Practice* (Boca Raton, FL: Taylor and Francis, 2018); DARIAH Bibliographical Data Working Group, „An Analysis of the Current Bibliographical Data Landscape in the Humanities. The Joint Bibliodata Agendas of Public Stakeholders", (2022), https://doi.org/10.5281/zenodo.6559857.

[15] Mikko Tolonen, Leo Lahti and Niko Ilomäki, „A Quantitative Study of History in the English Short-Title Catalogue (ESTC), 1470-1800," *Liber Quarterly* 25, 2. sz. (2015): 87–116, https://doi.org/10.18352/lq.10112; Péter Róbert, „Digitális és módszertani fordulat a sajtókutatásban: A 17–18. századi magyar vonatkozású angol újságcikkek »távolságtartó olvasása«," *Aetas* 29, 1. sz. (2015), 5–30, http://acta.bibl.u-szeged.hu/35222/; Dan Edelstein, Paula Findlen, Giovanna Ceserani, Caroline Winterer and Nicole Coleman, „Historical Research in a Digital Age: Reflections from the Mapping the Republic of Letters Project," *American Historical Review* 122, 2. sz. (2017): 401–424, https://academic.oup.com/ahr/article/122/2/400/3096208, https://doi.org/10.1093/ahr/122.2.400; Simon Burrows, *The French Book Trade in Enlightenment Europe II: Enlightenment Bestsellers* (London: Bloomsbury Academic, 2018); Mark Towsey, „Book Use and Sociability in Lost Libraries of the Eighteenth Century: Towards a Union Catalogue," in Flavis Bruni and Andrew Pettegree, eds., *Lost Books: Reconstructing the Print World of Pre-Industrial Europe*, 414–438 (Leiden: Brill, 2016), https://doi.org/10.1163/9789004311824_021; Leo Lahti, Jani Marjanen, Hege Roivainen and Mikko Tolonen, „Bibliographic Data Science and the History of the Book (c. 1500–1800)," *Cataloging & Classification Quarterly* 57, 1. sz. (2019): 5–23, https://doi.org/10.1080/01639374.2018.1543747; Mark J. Hill, Ville Vaara, Tanja Säily, Leo Lahti and Mikko Tolonen, „Reconstructing Intellectual Networks: From the ESTC's Bibliographic Metadata to Historical Material," in *Proceedings of the Digital Humanities in the Nordic Countries,* 201–219 (Copenhagen: CEUR-WS.org, 2019), http://ceur-ws.org/Vol-2364/19_paper.pdf; Alicia C. Montoya, „Enlightenment? What Enlightenment? Reflections on Half a Million Books (British, French and Dutch Private Libraries, 1665 – 1830)," *Eighteenth-Century Studies* 54, 2. sz. (2021): 909–934, https://doi.org/10.1353/ecs.2021.0097; Simon Burrows and Terhi Nurmikko-Fuller, „Charting Cultural History Through Historical Bibliometric Research: Methods, Concepts, Challenges, Results," in Kristen Schuster and Stuart Dunn, eds., *Routledge International Handbook of Research Methods in Digital Humanities,* 109–124 (New York: Routledge, 2020), https://doi.org/10.4324/9780429777028-9.

 DOI 10.31400/dh-hun.2021.4.3530

detect nyelvfelismerő programot használjuk.[16] Ha egynyelvű korpuszt elemzünk, az automatikus nyelvfelismerés helyett megadhatjuk az elemezni kívánt szövegek nyelvét: 61 nyelv közül tudunk választani egy gördülősáv segítségével. Az automatikus nyelvdetekció gazdagítja és pontosíthatja az analóg módon megadott, dokumentumok nyelvére vonatkozó metaadatokat.[17]

Ha nem szeretnénk a teljes szövegállományt elemezni, akkor a kontextusszűrő *(Context filter)* funkció segítségével megadhatunk kulcsszavakat, valamint a kulcsszavaktól balra és jobbra (külön-külön) található szavak számát. A későbbi elemzések során az AVOBMAT csak az így definiált szövegdobozokban található szavakat elemzi, a többi szöveget eltávolítja a dokumentumokból. Ez a funkció hasznos lehet például kisebb szövegrészek elkülönítésére, így cikkek szerint nem szegmentált újságkorpuszok kezdeti feldolgozására is.

A helyettesítés *(replace)* funkció segítségével az optikai szövegfelismerésből (Optical Character Recognition: OCR) adódó hibákat javíthatunk, összevonhatunk szinonimákat, vagy modernizálhatjuk a régi, nem standardizált helyesírást használó szövegeinket. A cserepárok megadása során reguláris kifejezéseket is használhatunk. Így például lehetőségünk van különleges karakterek törlésére, rövidítések feloldására, az elválasztott szavak összevonására, melyek fontos lépések a szövegtisztítás és normalizálás során.

A metaadat-gazdagítás magában foglalja a szövegek nyelvének detektálását, valamint a szerzők nemének automatikus azonosítását. Az utóbbi célra a *gender-guesser* nevű Python-csomag általunk továbbfejlesztett verzióját használjuk. Az androgün keresztnevek létezése miatt a nemek azonosítása nem mindig kivitelezhető egyértelműen. A dokumentumok szerzőit férfi, női vagy ismeretlen (például ha csak a keresztnév rövidítése adott) kategóriákba soroljuk. Külön metainformációként kezeljük, ha egy dokumentumnak egyáltalán nincs szerzője. Annak érdekében, hogy csökkentsük azon esetek számát, amikor nem tudjuk megállapítani a szerző nemét, a dokumentum nyelvét is bevonjuk a döntéshozatali folyamatba. Azt az egyszerűsítő feltételezést vesszük alapul, hogy a szerzői nevek ugyanazon a nyelven szerepelnek, mint maguk a dokumentumok. E feltételezés alapján tudjuk kezelni az apofóniát, például Kovács Imréné magyar szerző esetén következtethetünk arra, hogy az Imréné név az Imre névből származik, tehát női szerzőséget rendelünk hozzá. A nyelvi információk felhasználásával csökkenthetjük azt a bizonytalanságot is, amely abból ered, hogy ugyanaz a keresztnév különböző nyelvekben különböző nemű személyekre utalhat. A programba beépített női és férfi keresztnév-adatbázisok (*gender-guesser* csomag) tartalmát a felhasználó bővítheti saját női és férfi névlistáival is. A magyar nyelv esetén

[16] *Langdetect*, hozzáférés: 2021.12.15, `https://pypi.org/project/langdetect/`.

[17] A Szegedi Tudományegyetem *Egyetemi Kiadványok* nevű repozitóriumában található Kurdy Fehér János *Oleskeluharjoituksia* című finn versét magyar nyelvű versként tünteti fel a katalógus. Kurdy Fehér János, „Oleskeluharjoituksia," *Gondolat-jel* (1993), 1–2. sz., 22, `http://acta.bibl.u-szeged.hu/11488/`. Az AVOBMAT számos – a katalógusban a dokumentumok nyelvére vonatkozó metaadatmezőben nem szereplő – nyelvű dokumentumot azonosított. Itt meg kell jegyezni, hogy az automatikus nyelvdetekció tévedhet rövid (pl. képaláírások) vagy rosszul OCR-ezett dokumentumok esetében. Az utóbbira jó példa a *Délmagyarország* 1945. február 21-i száma, melynek OCR-ezett első oldala nem összefüggő szavakat, hanem ezek szóközökkel elválasztott betűsorát tartalmazza. *Délmagyarország*, 1945. febr. 21., 1, `http://dmarchiv.bibl.u-szeged.hu/10369/`.

például az ELKH Nyelvtudományi Kutatóközpont által anyakönyvi bejegyzésre alkalmasnak minősített utónevek jegyzékét[18] is használhatjuk erre a célra, de tetszőleges – a szoftveres elemzést pontosító – névlistákat is megadhatunk. A kiegészítésképpen megadott férfi és női nevek felülírják a program által valószínűsített nemi kategóriákat.[19]

A szövegelemző funkciókhoz köthető előfeldolgozási műveletek minden elemző esetén egyedileg konfigurálhatók. Opcionálisan beállítható a következő hat paraméter: (i) lemmatizálás (24 nyelven);[20] (ii) kisbetűsítés; (iii) számok; (iv) nem alfanumerikus karakterek; (v) írásjelek és (vi) stopszavak eltávolítása. A stopszavas és a punktuációs előfeldolgozás során a felhasználó is kiegészítheti az AVOBMAT-ba épített spaCy nyelvmodulokban található listákat. A stopszavak kiszűrésénél és a szótövesítés során a program figyelembe veszi az adott dokumentum automatikusan azonosított vagy manuálisan megadott nyelvét. Továbbá az n-gram elemző esetében a felhasználó megadhatja az azonosítandó szógramok hosszát (1–5). A lexikaidiverzitás-elemző segítségével nyolc metrika szerint tudjuk kiszámoltatni az egyes szövegek lexikai gazdagságát: Type-token ratio (TTR), Guiraud (Root TTR), Herdan (Log TTR), Mass TTR, Mean Segmental TTR (MSTTR), Moving Average TTR (MATTR), Measure of Textual Lexical Diversity (MTLD) és Hypergeometric Distribution Diversity (HDD). Az MSTTR és az MATTR esetében – a korábban említett előfeldolgozási paraméterek mellett – beállíthatjuk a „szövegablak" méretét (szószám) is. Itt fontos megjegyezni, hogy a különböző metrikák eltérő módon érzékenyek a szöveghosszra. Az utóbbi szempontból az MSTTR-, HDD- és MTLD-kalkulációk a legstabilabbak.[21]

[18] ELKH Nyelvtudományi Kutatóközpont, „ELKH Nyelvtudományi Kutatóközpont által anyakönyvi bejegyzésre alkalmasnak minősített utónevek jegyzéke," hozzáférés: 2021.12.15, http://www.nytud.mta.hu/oszt/nyelvmuvelo/utonevek/index.html.

[19] Ha az adatbázisunkban vannak hiányos szerzői nevek (pl. csak a vezetéknév adott), de a felhasználó tudja a szerző nemét, akkor ily módon is pontosíthatjuk a szerzői nevek felismerését. Például ha csak a Shakespeare név van megadva, akkor ezt az algoritmus ismeretlen nemű kategóriába sorolja. Ezt mi korrigálhatjuk, ha Shakespeare-t hozzáadjuk a férfi nevek listájához.

[20] A lemmatizáláshoz a spaCy nyelvmodelljeit és a *lemmagen* Python-csomagot használjuk. spaCy, hozzáférés: 2021.12.15, https://spacy.io/usage/models; Matjaž Juršic, et al., „Lemmagen: Multilingual Lemmatisation with Induced Ripple-down Rules," *Journal of Universal Computer Science* 16. 9 sz. (2010): 1190–1214; *Lemmagen*, hozzáférés: 2021.12.15, https://pypi.org/project/Lemmagen/.

[21] George Udny Yule, *The Statistical Study of Literary Vocabulary* (Cambridge: Cambridge University Press, 1944); Edward H. Simpson, „Measurement of Diversity," *Nature* 163 (1949): 688, https://doi.org/10.1038/163688a0; Gustav Herdan, „A New Derivation and Interpretation of Yule's 'Characteristic' *K*," *Zeitschrift für angewandte Mathematik und Physik* 6, 4. sz. (1955): 332–334, https://doi.org/10.1007/BF01587632; Heinz Dieter Maas, „Über den Zusammenhang zwischen Wortschatzumfang und Länge eines Textes," *Zeitschrift für Literaturwissenschaft und Linguistik* 2, 8 sz. (1972): 73–96; Fiona J. Tweedie and R. Harald Baayen, „How Variable May a Constant Be? Measures of Lexical Richness in Perspective," *Computers and the Humanities* 32, 5. sz. (1998): 323–352, https://doi.org/10.1023/A:1001749303137; Philip M. McCarthy and Scott Jarvis, „vocd: A Theoretical and Empirical Evaluation," *Language Testing* 24, 4. sz. (2007): 459–488, https://doi.org/10.1177/0265532207080767; Michael A. Covington and Joe D. McFall, „Cutting the Gordian Knot: the Moving-Average Type-Token Ratio (MATTR)," *Journal of Quantitative Linguistics* 17, 2. sz. (2010): 94–100, https://doi.org/10.1080/09296171003643098; Philip M. McCarthy and Scott. Jarvis, „MTLD, vocd-D, and HD-D: A Validation Study of Sophisticated Approaches to Lexical Diversity Assessment," *Behaviour Research Methods* 42, 2. sz. (2010): 381–392, https:

 DOI 10.31400/dh-hun.2021.4.3530

Minden elemző esetében vannak javasolt alapbeállítások, amelyeket a felhasználó tetszőlegesen módosíthat. Az egyes szövegelemzők ki és bekapcsolhatóak. Ha nincs szükségünk valamelyik elemzésre, akkor ily módon gyorsíthatjuk az előfeldolgozást. Az összes előfeldolgozás során megadott paramétert és adatot exportálhatjuk egy JSON-fájlba, amelyeket importálhatunk is a későbbi elemzések, kísérletezések során.

Az előfeldolgozási paraméterek megadása után feltölthetjük az adatbázisainkat többféle formátumban. Az AVOBMAT automatikusan importálja a *Zotero*-gyűjteményeket CSV- és RDF-formátumban (teljes szöveggel), valamint a számos könyvtárban használt EPrintses-adatbázisokat (EP3 XML a teljes szövegek URL-jeivel). A *Zotero* 20 különböző típusú bibliográfiai formátumot (pl. MARC, BibTex) tud importálni, amelyeket a felhasználók gyűjteménybe rendezhetnek. Ezeket a gyűjteményeket manuálisan vagy automatikusan tisztíthatják, bővíthetik új metaadatokat és szövegeket tartalmazó tételekkel.[22] A 87 bibliográfiai metaadatmezőt (pl. szerző, kiadó, publikáció cím) tartalmazó *Zotero* alapú CSV-struktúrát számos egyéb metaadatmezővel kiegészítettük (pl. könyvkereskedő, publikációk gyakorisága). A különböző bibliográfiai metaadatstandardok közötti különbségeket egyeztettük. Például az EP3 XML „Publication" mezője megegyezik a Zotero „Publication Title" mezőjével, így mindkettő egy közös „Elasticsearch" mezőbe kerül „publicationTitle" néven.[23]

Az adatbázisok feltölthetők egy egyszerű CSV-fájl segítségével is. A teljes szövegek hozzáadása többféleképpen történhet: (i) a szövegek rögzíthetők egy külön erre a célra létesített CSV-s mezőben; a szövegekre mutató (ii) relatív vagy (iii) internetes útvonalat is megadhatjuk egy másik mezőben. A második opció esetében a teljes szövegeket tartalmazó mappát és a metaadatokat tartalmazó CSV-fájlt tömörítve kell feltöltenünk. Az AVOBMAT minden olyan szövegformátumot tud importálni, amelyet az *Apache Tika*[24] program képes kezelni, mivel ez alakítja át egyszerű szöveggé a bemeneti fájlokat.

4. A keresés és kiválasztás

Az *Elasticsearch* motort használó alkalmazásban a kutatók kereshetnek a gazdagított bibliográfiai adatokban és az előfeldolgozott szövegekben fazettás, összetett és parancssori keresések segítségével. A fazettás keresésnél minden metaadatmező esetében megjelenik az értékkel nem rendelkező tételek (pl. a szerző nincs mindenhol

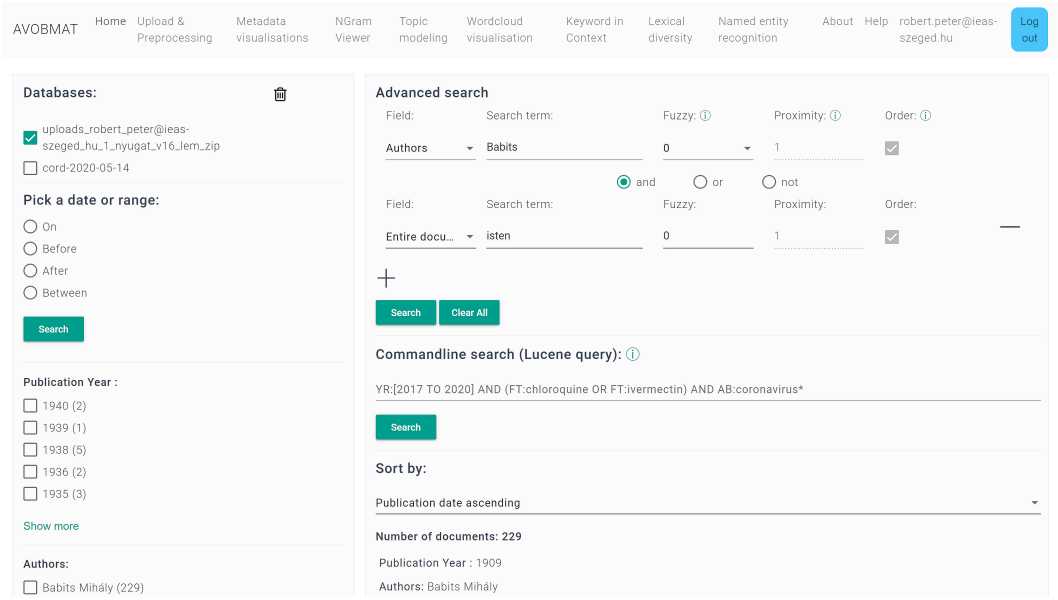
://doi.org/10.3758/BRM.42.2.381; Joan Torruella and Ramon Capsada, „Lexical Statistics and Tipological Structures: A Measure of Lexical Richness," *Procedia: Social and Behavioral Sciences* 95 (2013): 447–454, https://doi.org/10.1016/j.sbspro.2013.10.668; Kristopher Kyle, *Lexical diversity*, hozzáférés: 2021.12.15, https://github.com/kristopherkyle/lexical_diversity. Az AVOBMAT a fenti metrikákhoz tartozó értékek mellett jelzi az egyes szövegekhez tartozó szavak számát (token) valamint a különböző szóalakok számát (típus) is táblázatos formában.

[22] A metaadatok minőségellenőrzésével kapcsolatban lásd Király Péter publikációit és programkódjait. Például Péter Király und Rudolf Ungváry, „Bemerkungen zu der Qualitätsbewertung von MARC-21-Datensätzen," in Michael Franke-Maier, Anna Kasprzik, Andreas Ledl und Hans Schürmann, Hg., *Qualität in der Inhaltserschließung*, 177–228 (Berlin: De Gruyter Saur, 2021), https://doi.org/10.1515/9783110691597−011 és https://github.com/pkiraly.

[23] *Elasticsearch*, hozzáférés: 2021.12.15, https://www.elastic.co/.

[24] *Apache Tika*, hozzáférés: 2021.12.15, https://tika.apache.org/.

megadva) száma is egy külön metaadatmezőben *(missing_value)*, ami segíti a felhasználót az adatok és eredmények kritikus értelmezésében. Az AVOBMAT támogatja a paraméterezhető közelítő *(fuzzy)*, szószomszédsági *(proximity)* kereséseket, valamint a Lucene-szintaxist használó parancssori lekérdezéseket is. Így lehet használni például helyettesítő karaktereket (pl. ? vagy *) és reguláris kifejezéseket (melyeket *slash* közé kell tenni) is a kereséseknél.[25] Az utóbbi segítségével például egy szó összes ragozott alakjára is rákereshetünk. A paraméterezhető közelítő keresés különösen hasznos OCR-ezett dokumentumok esetén vagy régi szövegek elemzése során a nem standardizált helyesírás miatt. A szótávolság-keresés figyelembe veszi a megadott szótávolságban lévő szavak sorrendjét. Az összetett keresőben kombinálhatjuk a szótávolság- és a közelítő keresési funkciókat a Boole-operátorokkal (AND, OR, NOT) összekapcsolt keresésekkel. A metaadatmezők esetében csak azok jelennek meg a grafikus felületen, amelyeket egy adott adatbázis tartalmaz. A feltöltött adatbázisokat a keresés során egyszerű kijelöléssel lehet egyesíteni. A különböző keresési funkciók segítségével leszűkített korpuszon végezhetők el a metaadat- és szövegelemzések.



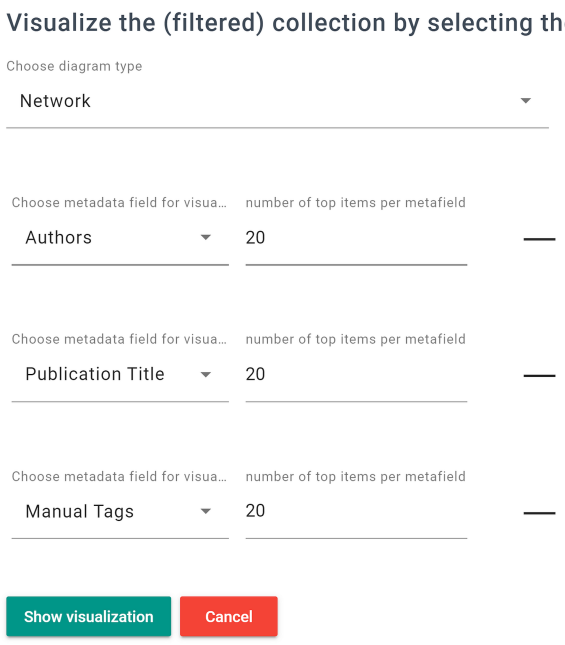
1. ábra. Az AVOBMAT grafikus felülete

[25] A keresési szintaxissal kapcsolatos dokumentáció az alábbi linken elérhető: `https://www.elastic.co/guide/en/elasticsearch/reference/7.17/query-dsl-query-string-query.html#query-string-syntax`.

 DOI 10.31400/dh-hun.2021.4.3530

5. Metaadat-elemzés és vizualizáció

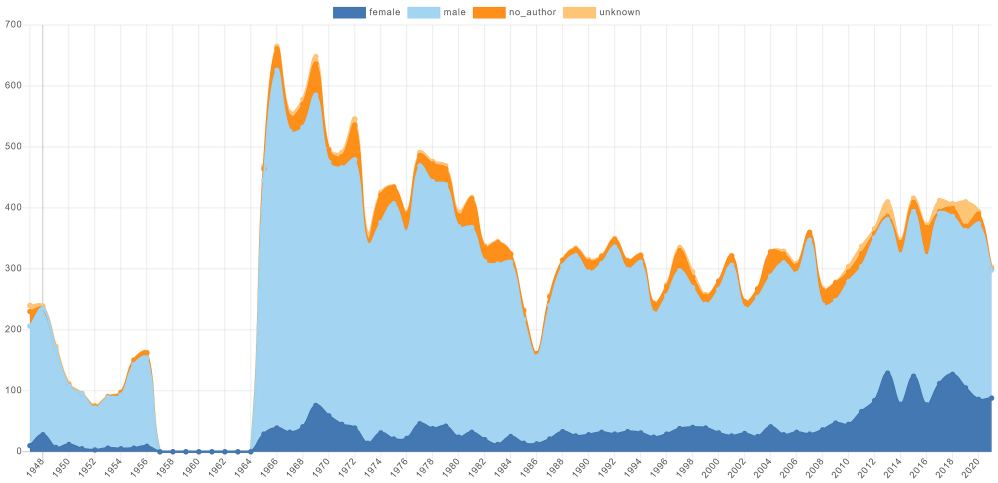
A felhasználók elemezhetik és vizualizálhatják a bibliográfiai adatokat (i) kronologikusan, vonal- és területdiagramokon, normalizált és aggregált formátumban;[26] (ii) interaktív hálózati elemzést készíthetnek legfeljebb három metaadatmező segítségével; (iii) a megadott paraméterek alapján tetszés szerinti kör-, sáv- és oszlopdiagramokat készíthetnek a bibliográfiai adatok felhasználásával. Az elemzők esetében a kutató határozza meg, mely metaadatmezőket szeretne elemezni és az adott mezőhöz tartozó adatsoron belül az első hány leggyakrabban előforduló tételt szeretne megjeleníttetni. Az adatpontok azért vannak külön ábrázolva, mert ezekre kattintva megjelennek a hozzájuk tartozó értékek. Az ábrákon található színmagyarázatok egyben szűrőként is funkcionálnak: minden vizualizáció esetében az egyes megjelenített metaadatok interaktív módon ki-be kapcsolhatóak, és ugyanez vonatkozik a hiányzó értékek *(missing_values)* és egyéb értékek *(other_values)* mezőkre is. Az utóbb azokra a „kimaradt" értékekre utal, melyeket a felhasználó dob el a paraméterezés során, amikor kiválasztja, hány leggyakrabban előforduló tételt szeretne ábrázolni egy metaadatmezőn belül. A diagramok egyes pontjaira kattintva a program megjeleníti az adott ponthoz tartozó értékeket (pl. név, szám, százalék).



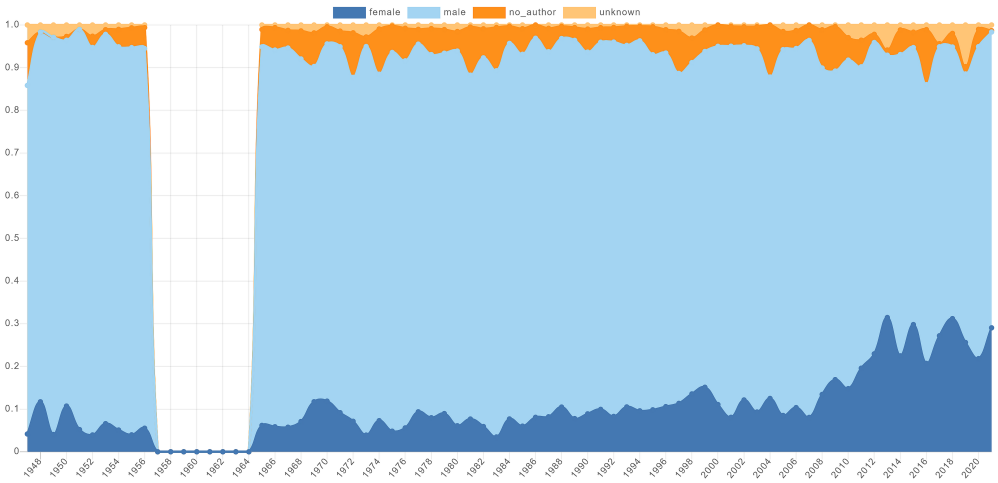
2. ábra. A metaadat-vizualizációs beállítási panel

[26] Az idősoros ábrázolás esetében ez például azt jelenti, hogy az aggregált módban az adott évhez tartozó nyers adatok száma jelenik meg a grafikonon, normalizált ábrázolás esetén pedig a relatív gyakoriságot láthatjuk százalékokban kifejezve. Ha például egy napilap cikktípusait (hír, hirdetés stb.) jelenítjük meg, akkor az aggregált ábrázolás során az adott évhez tartozó összes hirdetés száma jelenik meg a függvényen, míg a normalizált verzióban azt láthatjuk, hogy az adott évben megjelent összes cikk hány százaléka volt hirdetés.

Nézzünk néhány konkrét példát a különböző típusú metaadat-vizualizációkra:



3. ábra. Női, férfi, szerző nélküli és azonosíthatatlan nemű szerzők a *Tiszatáj* folyóiratban (aggregált) 1948 és 2021 között



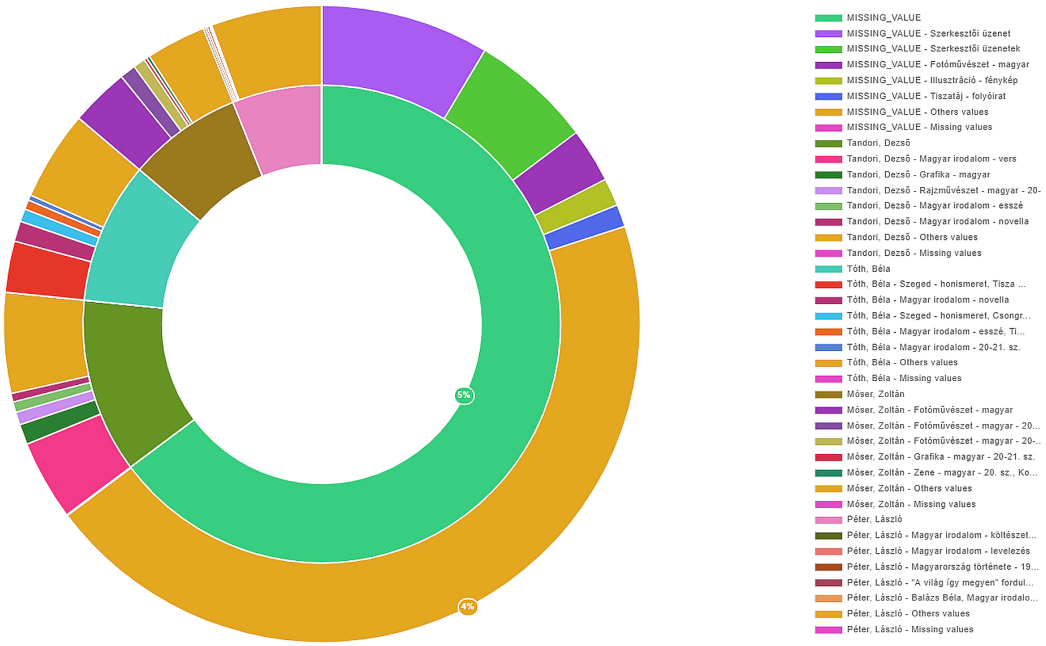
4. ábra. Női, férfi, szerző nélküli és azonosíthatatlan nemű szerzők a *Tiszatáj* folyóiratban (normalizált) 1948 és 2021 között

A *Tiszatáj*ban publikáló női és férfi szerzők eloszlása mellett a fenti ábrákon az is látszik, hogy 1957 és 1964 között nincsenek értékek a diagramokon. Ennek az az oka, hogy ebben az időszakban formátumot váltott a lap, a korábbi, oldaltól oldalig terjedő forma helyett ebben a néhány évben hasábos formában jelent meg, így a szegedi könyvtárosok nem darabolták szét, nem bontották cikkekre ezeket a számokat.

Az alábbi többszintű gyűrűdiagrammon a *Tiszatáj*ban publikáló 5 leggyakoribb szerző és a műveikhez analóg módon rendelt kulcsszavak eloszlását láthatjuk. A felsorolásban megfigyelhető, hogy az egyes kulcsszavak nem minden esetben vannak

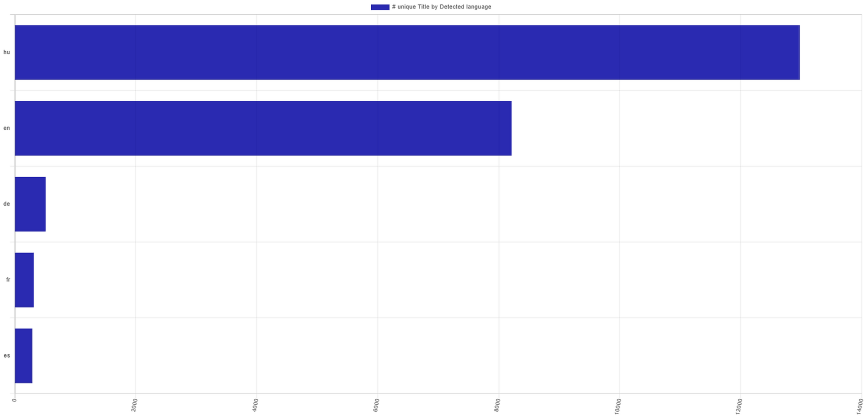
 DOI 10.31400/dh-hun.2021.4.3530

egységesítve (pl. a „Szerkesztői üzenet” és a „Szerkesztői üzenetek” külön kategóriákat alkotnak). Az ilyen pontatlanságok beazonosítását és kijavítását követően, a metaadatok tekintetében megtisztított, normalizált adatbázist újra feltölthetjük az AVOBMAT-ba.



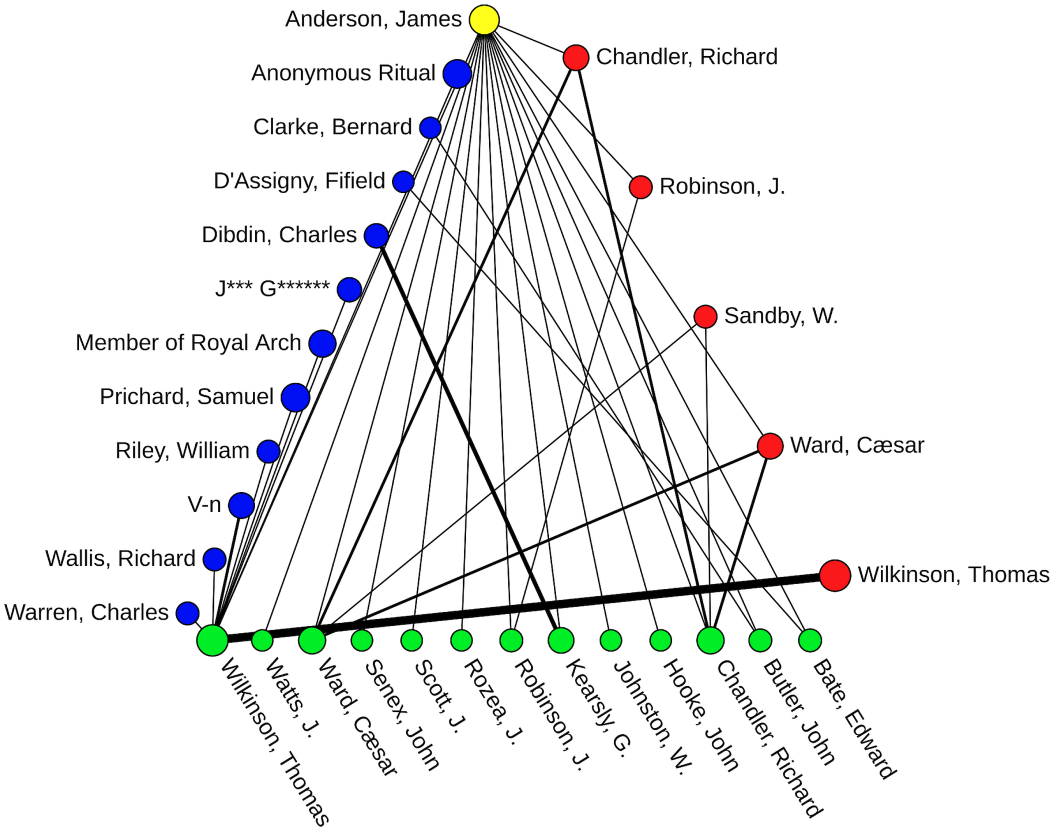
5. ábra. A *Tiszatáj*ban publikáló 5 leggyakoribb szerző és a műveikhez rendelt kulcsszavak (egyéb értékek kikapcsolva)

Tandori Dezső írásait sötétzölddel láthatjuk a belső gyűrűben. A gyűrűsáv méretéből látható, hogy ő publikálta a legtöbb írást a folyóiratban. Az adott gyűrűrészre kattintva a pontos százalék is megjelenik. A Tandori-gyűrűsáv külső gyűrűben található folytatásában a szerző által írt művek kulcsszavazott kategóriái találhatóak: pl. a sötétebb rózsaszín jelöli a „magyar irodalom – vers” kategóriát, ahogy az a jobb oldali színskála színeihez rendelt jelöléseknél is megfigyelhető.



6. ábra. Az SZTE *Egyetemi Kiadványok* repozitóriumában az 5 leggyakoribb nyelven 2000 után írt cikkek száma (automatikus nyelvfelismerés)

A metaadatokhoz tartozó tételek hálózati kapcsolatát is megjeleníthetjük, s ezekhez tartozó alhálózatokat is vizualizálhatunk. A hálózati csúcsoknál található kör mutatja a csúcshoz tartozó kapcsolatok számát, amely a kör méretével arányos. A csúcsokat összekötő vonalak (élek) vastagsága a kapcsolatok számát jelzi, melyek száma megjelenik, ha egy adott élre kattintunk.



7. **ábra**. A 18. századi brit és ír szabadkőműves könyvek szerző–nyomdász–könyvkereskedő alhálózata, James Anderson szerző kapcsolati hálója

A bibliográfiai adatok elemzése amellett, hogy feltárja a különböző metaadatmezőkben tárolt rekordok közötti, korábban ismeretlen összefüggéseket, és rávilágít az eleddig figyelmen kívül hagyott trendekre, fényt deríthet az adatbázisok (például bibliográfiai adatokat érintő) hiányosságaira, korlátaira, beleértve az adatbázis-készítésnél alkalmazott előítéleteket (például a gyűjteménybe kerülő szövegek kiválogatásával vagy osztályozásával kapcsolatban).[27] A legtöbb adatszolgáltató (könyvtárak és profitorientált cégek egyaránt) vagy nem ismerte fel, vagy ha igen, nem szívesen hozza nyilvánosságra ezeket az információkat. Ilyen típusú problémák előzetes számítógépes feltérképezése segíti a kutatókat abban, hogy megalapozott szakmai döntéseket hozzanak projektjeikről, és kritikusan elemezzék a digitális gyűjtemények tartalmát. Továbbá az adatgazdáknak is lehetőséget nyújt a bibliográfiai adatok minőségének javításához.

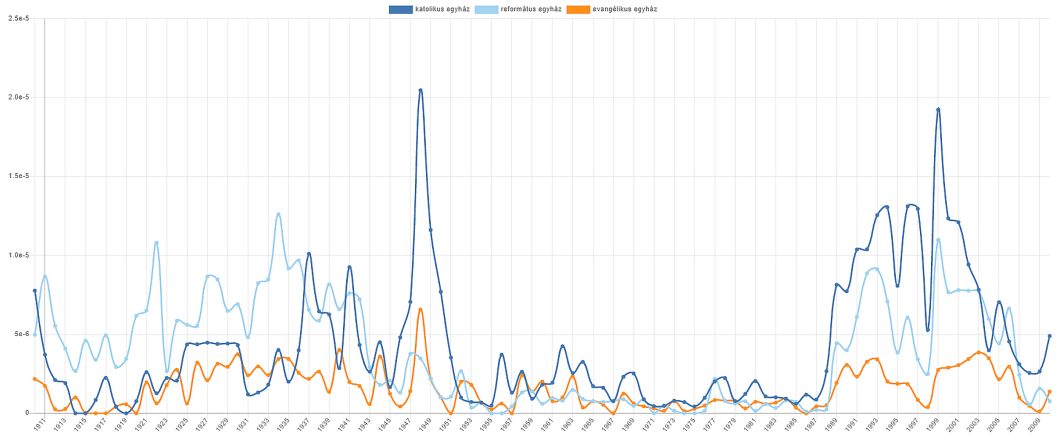
[27] Katherine Bode, „Why You Can't Model Away Bias?" *Modern Language Quarterly* 81, 1 sz. (2020): 95–124, https://doi.org/10.1215/00267929–7933102.

 DOI 10.31400/dh-hun.2021.4.3530

6. Szövegelemzés és vizualizáció

6.1. N-gram elemzés

A szövegek diakronikus elemzését az AVOBMAT n-gram elemzője támogatja. Idősoron megjeleníti a felhasználó által megadott – teljes szövegben található – n-gramok (itt egymás után következő *n* darab szó) éves eloszlását aggregált és normalizált módon. A legfeljebb öt szó hosszúságú n-gramokat az előfeldolgozási szakaszban azonosítja a program. A normalizált nézet esetében a százalékos gyakoriságot úgy kapjuk meg, hogy az adott évben fellelhető, felhasználó által keresett n-gramok számát elosztjuk az ugyanahhoz az évhez tartozó szövegekben található szavak számával.



8. ábra. A *katolikus egyház, református egyház* és *evangélikus egyház* bigramok normalizált eloszlása a *Délmagyarország* napilapban, 1911 és 2009 között[28]

6.2. Témamodellezés

A témamodellezés segítségével rejtett és absztrakt témákat, szemantikai információkat fedezhetünk fel szövegekben. Az algoritmus statisztikai módszereket használ a szövegekbe ágyazott témák feltárására, valamint e témák kapcsolatainak és időbeli változásainak feltárására.[29] Az AVOBMAT rendelkezik egy böngészőbe épített Latent Dirichlet Allocation (LDA) funkcióval, amely a *jsLDA*-könyvtárra[30] támaszkodik a témamodellek kiszámításánál és grafikus ábrázolásánál. Az LDA a felhasználó által megadott számú látens témát azonosít, ahol minden dokumentum e témák keverékének tekinthető. A módszer az együtt előforduló szavakat csoportosítja témákba, a dokumentumokhoz pedig valószínűségekkel hozzárendeli az egyes témákat. A témaelemzés mellett a modellezés eredményeit is különböző módon tudja megjeleníteni az

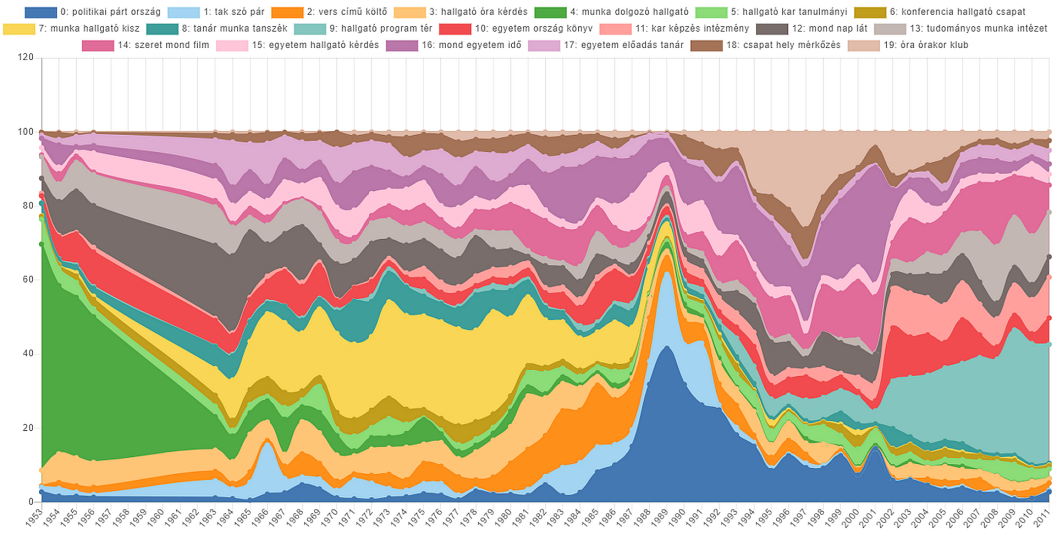
[28] 1920 és 1925 között csak részben vagy egyáltalán nem jelent meg a *Délmagyarország*, ekkor *Szeged* néven volt elérhető napilap. 1956. november 20. és 1957. április 30. között pedig a *Szegedi Néplap* váltotta fel a *Délmagyarország*ot. Az n-gram elemzés ezen újságok cikkeit is tartalmazza.

[29] David M. Blei, Andrew Y. Ng and Michael I. Jordan, „Latent Dirichlet Allocation," *Journal of Machine Learning Research* 3 (2003): 993–1022.

[30] *jsLDA*, hozzáférés: 2021.12.15, `https://mimno.infosci.cornell.edu/jsLDA/`.

AVOBMAT. Megmutatja az egyes témákhoz kapcsolódó legrelevánsabb szavakat és dokumentumokat, megjeleníti e témák eloszlását idősoron, vizualizálja a különböző témák közötti korrelációkat, és különböző formátumokban exportálja az eredményeket. A bibliográfiai adatok felhasználása lehetővé teszi, hogy diakronikus témamodellezéseket végezzünk, amelyek általánosabb szemantikai mintákat tárnak fel a nyelvhasználatban, mint amilyeneket a gyakorta nagy méretű digitális gyűjtemények szoros olvasása nyújtana.

Az eredeti *jsLDA*-implementáció paraméterként a témák számát és az iterációkat igényli. Ezt három új paraméterrel bővítettük. A felhasználó beállíthatja az elemezni kívánt korpuszban a szavak minimális előfordulási számát. Ha ez a minimum nagy, az algoritmus gyorsabb lesz a szerver és a böngésző közötti csökkentett adatközlés miatt, de hátránya, hogy elveszíthetjük a dokumentumokra vonatkozó információk egy részét. A leggyakrabban előforduló (stop)szavakat interaktív módon távolíthatjuk el a „Vocabulary” ikonra kattintva. Az ilyen szűrés után mindig újra kell futtatni az elemzőt. A felhasználók beállíthatják az alfa és béta LDA-hiperparamétereket is: az alfa a dokumentum–téma sűrűséget, a béta pedig a téma–szó sűrűséget jelöli.[31] A *jsLDA* programot még kiegészítettük egyrészt azzal, hogy a témák időbeli eloszlását aggregált és normalizált módokon is ábrázolhatjuk, másrészt az egyes témákhoz kapcsolható dokumentumok alapvető bibliográfiai adatait is megjeleníthetjük a témákra vonatkozó dokumentumokhoz tartozó valószínűségi értékek mellett.



9. ábra. A *Szegedi Egyetem* folyóirat egy témamodellezése, 1953–2011 (témák száma: 20, alfa = 0,1; béta = 0,01)

A fenti témamodellezés esetén így is értelmezhetjük az alábbi témákhoz tartozó szavakat: [0] *politikai, párt, ország, kérdés, tart, lát, helyzet* – pártpolitikai hírek; [2] *vers,*

[31] Szimmetrikus Dirichlet-eloszlást feltételezve, az alacsony alfa érték nagyobb súlyt helyez arra, hogy minden dokumentum csak néhány domináns témából álljon, míg a magas érték sokkal több viszonylag domináns témát ad vissza. Hasonlóképpen, egy alacsony béta érték nagyobb súlyt helyez arra, hogy az egyes témák csak néhány domináns szóból álljanak. Ha magasabb a béta érték, a témák nagy számú – korpuszban található – szóból állnak.

 DOI 10.31400/dh-hun.2021.4.3530

című, költő, kötet, szerző – a folyóiratban megjelent versek, költeményes kötetek bemutatása; [5] *hallgató, kar, tanulmányi, ösztöndíj, félév, szociális, támogatás, szak* – hallgatói támogatásokkal, ösztöndíjakkal kapcsolatos hírek; [7] *munka, hallgató, KISZ, kollégium, bizottság, főiskola, tag, éves, feladat, tevékenység* – KISZ-es eseményekhez, tagsághoz köthető témák; [18] *csapat, hely, mérkőzés, pont, bajnokság, második, játékos, együttes, verseny* – egyetemi sportbajnokságokkal kapcsolatos híradás.

6.3. Szóstatisztikai elemzések

A szófelhők hatékony eszközök lehetnek egy korpuszban valamilyen szempontból prominens szavak kiemelésére. Háromféle szóstatisztikai elemzőt integráltunk az AVOBMAT alkalmazásba. A legegyszerűbb vizualizáció a szógyakoriság alapján készíti el a szófelhőt és mutatja az egyes szavakhoz tartozó gyakorisági adatokat. Minden egyes vizualizáció esetében megadhatjuk, hogy hány darab szó jelenjen meg a szófelhőben. A második elemző *(Significant text)* azt mutatja, hogy milyen, az átlagostól jelentősen eltérő gyakoriságú szavak különböztetik meg egy digitális gyűjtemény általunk szűréssel kiválasztott részhalmazát a korpuszban található összes szövegtől.[32] A harmadik elemző *(TagSpheres)* lehetővé teszi a felhasználók számára, hogy egy szó kontextusát vizsgálják.[33] A különböző szófelhők mellett a szóstatisztikai adatokat oszlopdiagrammokban is láthatjuk, és az itt szereplő adatsorokat exportálhatjuk.

A *Significant text* elemző egy lekérdezés által definiált alkorpuszra leginkább jellemző (jelentősen eltérő gyakoriságú) szavakat azonosítja. Például ha a felhasználó az AVOBMAT keresési lehetőségeit használva kiválaszt egy szerzőt a korpuszból, akkor ez az eszköz megmutatja azokat a szavakat, amelyek e szerző műveihez legszignifikánsabban kapcsolódnak (jelentősen eltérő a gyakoriságuk) a teljes korpuszban található szövegekhez képest. Az előbbi részhalmazt előtérhalmaznak *(foreground set),* a dokumentumok teljes halmazát pedig háttérhalmaznak *(background set)* nevezzük.[34] Az *Elasticsearch* ezen halmazok statisztikai összehasonlításával rangsorolja az egyes szavakat. A következő képlet mutatja az úgynevezett JLH-érték kiszámítását, amelyet a szavak rangsorolásához alkalmazunk:

$$JLH = \left(p_{előtérhalmaz} - p_{háttérhalmaz}\right) \frac{p_{előtérhalmaz}}{p_{háttérhalmaz}}$$

ahol a $p_{előtérhalmaz}$ a relatív gyakorisága az előtérhalmazban található kifejezésnek, míg a $p_{háttérhalmaz}$ a relatív gyakorisága ugyanennek a kifejezésnek a háttérhalmazban. Az

[32] A *significant text* elemző dokumentációját lásd, hozzáférés: 2021.12.15, `https://www.elastic.co/guide/en/elasticsearch/reference/8.0/search-aggregations-bucket-significanttext-aggregation.html`.

[33] Stefan Jänicke and Gerik Scheuermann, „On the Visualization of Hierarchical Relations and Tree Structures with TagSpheres," in José Braz et al., eds., *Computer Vision, Imaging and Computer Graphics Theory and Applications,* 199–219 (Cham Springer International Publishing, 2017), `https://doi.org/10.1007/978-3-319-64870-5_10`.

[34] Ezt az *Elasticsearch*ben használt alapbeállítást az eredmények értelmezésénél figyelembe kell venni. A háttérhalmazt úgy is megadhatjuk az *Elasticsearch* konfigurációjában (a *background_is_superset* paramétert hamisra állítva), hogy ez diszjunkt halmazt képezzen az előtérhalmazzal, így csak azokat a szövegeket tartalmazza, amelyeket nem választott ki a felhasználó. Ezt a választási opciót szeretnénk a grafikus felületre is kivezetni a jövőben.

AVOBMAT ezt az összehasonlító elemzést még három közismertebb metrika *(mutual information, chi square, Google normalized distance)* segítségével is kiszámolja. Az elemzés elvégzése előtt beállíthatunk két paramétert: (i) hány darab szó jelenjen meg a szófelhőben; (ii) hány darab dokumentumot válasszon ki mintaként *(sample)* az előtérhalmazból és a háttérhalmazból.[35]

10. ábra. Szabó Dezső *Nyugat* folyóiratban megjelent 230 írására legjellemzőbb szavak (JLH-metrika, 100 szó). A szófelhő bizonyos szavai jól tükrözik a szerző francia műveltségét.

[35] Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, *An Introduction to Information Retrieval* (Cambridge: Cambridge University Press, 2009), 272–275 *(mutual information)*, 275–277 *(chi square)*, https://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf; Rudi L. Cilibrasi and Paul M. B. Vitányi, „The Google Similarity Distance," *IEEE Transactions on Knowledge and Data Engineering* 19, 3. sz. (2007): 370–383, https://arxiv.org/pdf/cs/0412098v3.pdf, https://doi.org/10.1109/TKDE.2007.48.

 DOI 10.31400/dh-hun.2021.4.3530

Word	Score	%
etc	21451.45	100.00%
szabó	18864.44	87.94%
dezső	14802.03	69.00%
kisott	6006.43	28.00%
parnasse	4671.66	21.78%
irradiál	4290.31	20.00%
kisottá	3432.25	16.00%
kisotto	3432.25	16.00%
donkisottság	3432.25	16.00%
kialakító	3432.24	16.00%
levés	2890.29	13.47%
lisle	2745.79	12.80%
hélas	2745.79	12.80%
sanglots	2745.79	12.80%
soif	2745.79	12.80%
megszenvedett	2745.79	12.80%
kannibál	2681.43	12.50%

11. ábra. Szabó Dezső *Nyugat* folyóiratban megjelent 230 írására legjellemzőbb szavak (JLH-metrika) és az ezekhez tartozó statisztikai adatok

Melyik metrikát válasszuk? A *mutual information* a magas gyakoriságú kifejezéseket részesíti előnyben, még akkor is, ha azok a háttérhalmazban is gyakran előfordulnak. Így ez a stopszavak kiválasztásához is vezethet. A *mutual information* nem valószínű, hogy nagyon ritka kifejezéseket, például helytelen helyesírással írott szavakat emel ki. A *Google normalized distance (gnd)* a magas együttes előfordulási gyakoriságú kifejezéseket részesíti előnyben, és elkerüli a stopszavak kiválasztását; talán jobban alkalmas a szinonimák felismerésére. A *gnd* azonban hajlamos a nagyon ritka kifejezések kiválasztására, amelyek például helyesírási hibákból származnak. A *chi square* és a JLH hozzávetőlegesen a kettő között helyezkedik el.[36]

A hagyományos szófelhők a szavakat egymástól függetlenül kezelik, és elveszítik a szavak közötti kontextuális információt. A szavak szövegkörnyezetének grafikus ábrázolásához a *TagSpheres* programot integráltuk. Ez olyan szófelhőt hoz létre, amely egy megadott keresőszó környezetében együttesen előforduló szavakat mutatja. A különböző szótávolságra található szavakat eltérő színekkel jelöli. A keresőkifejezés mellett a felhasználó megadhatja (i) az együtt előforduló szavak minimális gyakoriságát; (ii) az együtt előforduló szavak maximális szótávolságát a megadott szótól; (iii) a szavak a

[36] „Significant text aggregation," hozzáférés: 2021.12.15, `https://www.elastic.co/guide/en/elasticsearch/reference/8.0/search-aggregations-bucket-significanttext-aggregation.html`.

keresőkifejezéstől csak balra, csak jobbra vagy mindkét környezetben való előfordulását. Ennél az elemzőnél különös jelentősége van annak, hogy az előfeldolgozás során kiszűrtük-e a stopszavakat.

12. ábra. Babits Mihály „Istenképe” a *Nyugat* folyóiratban megjelent művei alapján (3 szótávolság stopszavak nélkül, minimum szógyakoriság: 2). Ilyen típusú elemzést más szerzők esetén is elvégezhetünk és összehasonlíthatjuk az eredményeket.

 DOI 10.31400/dh-hun.2021.4.3530

Word	Word distance	Count	%
isten	0	577	100.00%
tud	1	21	3.64%
ad	1	20	3.47%
ember	1	19	3.29%
ó	1	14	2.43%
isten	1	12	2.08%
szó	1	8	1.39%
hisz	1	8	1.39%
régi	1	7	1.21%
kér	1	7	1.21%
mond	1	7	1.21%
ég	1	7	1.21%
harcol	1	7	1.21%
lát	1	7	1.21%
süket	1	7	1.21%
anya	1	6	1.04%
világ	1	6	1.04%
szeret	1	6	1.04%

13. ábra. Az *Isten* szó környezte Babits Mihály *Nyugat* folyóiratban megjelent műveiben

6.4. Konkordancia

A konkordancia eszköz segíti az elemezni kívánt szövegek szoros vagy lassú olvasását. Megadhatjuk, hogy az adott keresési kifejezés (akár több szó) környezetében hány betűt jelenítsen meg a program, valamint azt is, hogy maximum hány találatot mutasson. A kulcsszavak kontextusát kétféle nézetben jeleníthetjük meg: az egyikben („View occurances line by line") soronként jelennek meg a találatok (így a szövegkontextus kisebb), a másikban („View occurances in context") pedig a szövegdobozban annyi karakter jelenik meg a keresési kifejezés körül, ahányat a felhasználó beállít. Mindkét esetben a találatokat rendezhetjük szerző, megjelenési év és szöveg szerint.

Authors	Title	Publication year	Text
Schöpflin Aladár	A két Vörösmarty	1908	tudatja vele rosszallását s röviden, pregnánsan, rá nézve jellemzően fejti ki a függetlenség értékét az életben, szemben a szolga-szerep adta viszonylagos jóléttel. Egész világfelfogását bizonyos józan egyensúly jellemzi. Vallja a francia forradalom alapelveit, de forradalmi színezet nélkül. A ***magyar nemzet*** függetlenségéért lelkesül, de még a 48 mámora sem ragadja magával, óvatos higgadtsággal áll elébe. Sarkallja a nemzetet haladásra, de ostor nélkül. A jövőbe néz, de ugyanekkor a múltba is bele tud merülni. Szellemi érdeklődésében megfér Horatius és Hugo Viktor; az egyikhez temperamentuma húzza
Junius	Gyulai Pál	1909	amelyhez tartozott egyébként minden gondolkozó agy, amely mérsékletet hirdetett, amely a tisztes kiegyezést többre tartotta volna a kétségbeesés heroikus erőfeszítéseinél, az élet-halál küzdelem kockázatánál. Világos után, teóriában, jogilag legalább nem volt többé Magyarország és nem volt többé ***magyar nemzet.*** Földünkön mindenfelé csak rom és pusztulás; testünk-lelkünk merében vérző seb. Nem csoda, ha nemzetünk a becsülettel megállott heroikus erőfeszítés után zsibbadt, tompa aléltságba esett; ha biztatva magát jövendővel, megadta magát sorsának. A magyarnak egy időre, akkor azt hihettük, hogy örökre
Schöpflin Aladár	A fiatal Gyulai	1909	kell?" S nem ezerszer gúny-e, hogy ez a vitatkozás még ma, ötven év múlva sem vesztette el teljesen aktualitását? Legélesebbé akkor válik a hangja, mikor az ötvenes évek lírikusainak kulturálatlanságát tépdesi. Már akkor ki van benne alakulva az az ideál, amelyet a magyar irodalom elé tűzött: a ***magyar nemzet*** egyéniségét európai színvonalon művészileg kialakító irodalom. A mai napig is ez maradt legfőbb kritériuma minden magyar irodalmi műnek az ő szemében. Általában egész kritikai pályájának mozgató elvei már ekkor megállapodtak benne. Felfogása a későbbi évek nyugodt alkotó munkájában bővült és

14. ábra. Konkordancianézet. A *magyar nemzet* kifejezés a *Nyugat* folyóirat cikkeiben

6.5. Névelem-felismerés

Az AVOBMAT-ba integráltuk a spaCy neurális hálókra épülő nyelvmodelljeit, melyek segítségével szövegekből automatikus módszerekkel kinyerhetünk névelemeket (Named Entity Recogniton: NER), többek között közneveket, tulajdonneveket (pl. személynevek, helyek, szervezetek nevei) és dátumokat. Ez a funkció 16 nyelven működik, a magyar nyelvet is beleszámítva, bár az utóbbinak jelenleg még nincs hivatalos spaCy nyelvmodellje.[37] Az alábbi táblázat mutatja, milyen nyelveken milyen névelemeket azonosít az AVOBMAT. A névelem-felismerés eredményeit többféle módon lehet megjeleníteni. A szemantikus gazdagítás során létrejött névelemek egyes típusai külön metaadatmezőkben tárolódnak és jelennek meg a fazettás és összetett keresőben, valamint a metaadat-vizualizációs beállítási panelben. A szövegben felismert névelemek a teljes szövegben is megtekinthetők: ehhez a találati listában ki kell választanunk egy szöveget és a megjelenési módot a „Named Entity Recognition"-re kell állítani. Ekkor a névelemeket és ezek típusait eltérő színekkel látjuk majd a szövegben. Az AVOBMAT a névelem-felismerés eredményeiről egyszerű statisztikákat is készít. Az „Entities in all documents" funkció az adatbázisunkban vagy annak általunk szűkített részhalmazában mutatja a felismert névelemeket, számukat és azt, hogy hány dokumentumban fordulnak elő. Az „Entities by documents" pedig a névelemek számát mutatja dokumentumonként. A nyelvi modellek frissítése lehetséges. A névelem-felismerés pontossága nyelvenként, ezeken belül elérhető (általában kis, közepes és nagy) modellenként és szövegtípusonként változik.[38]

	Person	Organization	Location	Miscellaneous	Language	Work of art	Geopolitical	National or religious group	Date	Ordinal	Product	Quantity	Time	Money	Infrastructure	Cardinal	Event	Law	Percent	Period	Movement	Phone	Pet name	Title affix
Chinese	X	X	X		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X					
Danish	X	X	X	X																				
Dutch	X	X	X		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X					
English	X	X	X		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X					
French	X	X	X	X																				
German	X	X	X	X																				
Greek	X	X	X				X				X						X							
Italian	X	X	X		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X		X	X	X	X
Japanese	X	X	X	X																				
Lithuanian	X	X	X				X				X		X											
Hungarian	X	X	X	X																				
Norwegian Bokmål	X	X	X				X		X				X											
Polish	X	X	X	X																				
Portuguese	X	X	X		X	X	X	X	X	X	X			X	X	X	X			X				
Romanian	X	X	X	X																				

15. ábra. Névelem-felismerés különböző nyelveken az AVOBMAT-ban

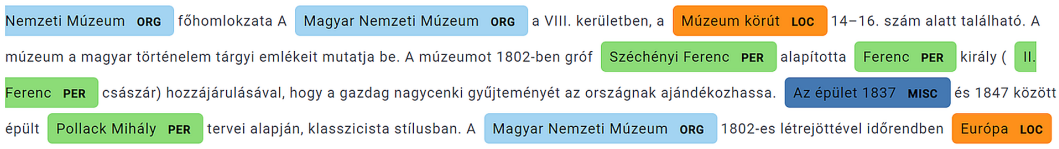
[37] György Orosz, Zsolt Szántó, Péter Berkecz, Gergő Szabó and Richárd Farkas, „HuSpaCy: An Industrial-Strength Hungarian Natural Language Processing Toolkit," in Berend Gábor, Gosztolya Gábor és Vincze Veronika, szerk., *XVIII. Magyar Számítógépes Nyelvészeti Konferencia,* 59–73 (Szeged: JATEPress, 2022), https://rgai.inf.u−szeged.hu/file/427.

[38] „SpaCy Models and Languages," hozzáférés: 2021.12.15, https://spacy.io/usage/models.

 DOI 10.31400/dh-hun.2021.4.3530

Entity	Count ↓	Type	Documents
Szeged	346	LOCATION	63
Juhász Gyula	311	PERSON	52
József Attila	216	PERSON	34
Szegeden	175	LOCATION	52
Illyés	159	PERSON	9
Bálint Sándor	142	PERSON	14

16. ábra. Névelem-statisztika Péter László *Tiszatáj* folyóiratban publikált cikkeire vonatkozóan



17. ábra. A teljes szövegben megtekinthető névelemek a *Wikipédia* „Budapest” szócikkében

7. Jogosultságkezelés és adminisztrátori felület

Az AVOBMAT jelenleg egy egyszerű jogosultságkezelési funkcióval rendelkezik. A regisztráció után az ideiglenes vagy állandó felhasználók feltölthetnek egyéni adatbázisokat. A kutatási projektek keretében létrejött adatbázisokat igény szerint akár publikussá is lehet tenni egy külön adminisztrátori felületen.[39] Az ily módon bárki számára hozzáférhető metaadatok és szövegek innovatív módon, korszerű szövegbányászati eszközökkel elemezhetők egy egyszerű grafikus felületen. Mivel jelenleg a feltölthető adatbázisok méretét nem lehet az adminisztrátori felületen szabályozni és a rendelkezésre álló szerver kapacitása korlátozott, ezért az AVOBMAT még nem tud tömeges felhasználói igényeket kielégíteni. Egyéni egyeztetést követően kutatóknak

[39] Jelenleg egy COVID-19-cel kapcsolatos tudományos cikkeket (63571 darab) tartalmazó adatbázis (cord-2020-05-14) érhető el nyilvánosan. Ezt és az adatbázis korábbi verzióit 44 országban használták a pandémia kezdetén. Fontos megjegyezni, hogy ezen a felületen nem tesztelhető az összes cikkben említett elemző (pl. a névelem-felismerés). Lásd hozzáférés: 2021.12.15, `https://avobmat.hu/covid-19/` és hozzáférés: 2021.12.15, `http://dighum.bibl.u-szeged.hu/avobmat-covid/home`. Lucy Wang, Kyle Lo and Róbert Péter, „COVID-19 Open Research Dataset and AVOBMAT Text Mining Tool,” New York NLP Meet-up, 2020. ápr. 27., hozzáférés: 2021.12.15, `https://www.youtube.com/watch?v=GivUfb8KhZY`; Christopher Nunn, „Research COVID-19 with AVOBMAT,” *OpenMethods: Highlighting Digital Humanities Methods and Tools,* 2020. jún. 8, `https://openmethods.dariah.eu/2020/06/08/research-covid-19-with-avobmat/`.

és kutatócsoportoknak viszont tudunk hozzáférést biztosítani az eszközhöz, az erőforrások függvényében. Az adminisztrátori felületen beállíthatjuk még a metaadatmezők neveit, rövidítéseit és megjelenéseit a különböző elemzőkben és a keresési felületen, valamint új metaadatmezők felvételére is van lehetőség.

8. Összegzés és tervek

E dolgozat a platformfüggetlen AVOBMAT szöveg- és adatbányászati kutatási eszközt mutatta be, amely elsősorban olyan felhasználók számára lett kifejlesztve, akik nem rendelkeznek programozási ismeretekkel. Láttuk, hogy ez a felfedezést segítő alkalmazás számos dinamikus szöveg- és adatbányászati elemzést tesz lehetővé. Az egyszerű, felhasználóbarát grafikus felület interaktív paraméterbeállítást és vezérlést biztosít a normalizálást is támogató előfeldolgozástól az analitikai szakaszokig. A szoros és távoli olvasás kombinálása mellett, az AVOBMAT-eszköztár egyedülálló tulajdonsága, hogy egyesíti a bibliográfiai (meta)adat- és a szövegelemzést számos nyelven. Lehetővé teszi a felhasználók számára, hogy a feltöltött és nyilvánosan elérhető adatbázisokat metaadatok és teljes szövegű kulcsszavas keresések segítségével szűrjék, és a szűrt adatkészleteken elvégezzék az összes metaadat-vizualizációs, hálózati és nyelvtechnológiai elemzést. A felhasználók így könnyen és interaktív módon kísérletezhetnek az elemzések különböző beállításaival a munkafolyamat során. Ezáltal a program segít felismerni a számítógépes szöveg- és adatelemzés episztemológiai kihívásait, korlátait és erősségeit, valamint kritikus módon értelmezni az alkalmazott módszereket és eredményeket. Bár adatvezérelt technológiákat használunk az elemzéseknél, fontos nyomatékosítani, hogy a sokszor hiányos, zajos, esetleges adatokat nem tekinthetjük adottnak vagy objektívnek, az adatok elméletekkel terheltek, a kapcsolatuk a kontextustól függően folyamatosan változik.[40] Ezért olyan funkciókat építettünk be az alkalmazásba, melyek segítik az adatbázisok korlátainak és az adatbázis-építés mögött húzódó előítéletek feltérképezését. A mennyiség nem garantálja a minőséget, az adatok természetéből adódó pontatlanságát számos esetben, bizonyos mértékig kompenzálja az adatok volumene és nagy száma *(big data)*. Az eredmények értelmezésénél nem feltétlenül a pontos számokon van a hangsúly, hanem azokon az esetenként új típusú bizonyítékként szolgáló trendeken, mintákon és modelleken, amelyeket ezek feltárnak. Az AVOBMAT csupán egy eszköz, módszertani apparátus, a digitális forrás- és kódkritika *(digital source criticism, critical code studies)* alkalmazása nélkülözhetetlen az eredmények interpretációja során. Az AVOBMAT export és import funkciói (konfigurációs beállítások, adatsorok, vizualizációk) az eredmények reprodukálhatóságát, valamint az előfeldolgozás és a szövegelemzés átláthatóságát hivatottak megkönnyíteni.

[40] Tobias Blanke and Andrew Prescott, „Dealing With Big Data," in Gabriele Griffin and Matt Hayler, eds., *Research Methods for Reading Digital Data in the Digital Humanities,* 184–205 (Edinburgh: Edinburgh University Press: Edinburgh, 2016), https://doi.org/10.1515/9781474409629-012; Giovanni Schiuma and Daniela Carlucci, eds., *Big Data in the Arts and Humanities* (Boca Raton, FL: Auerbach Publications, 2018); Jennifer Edmond, Nicola Horsley, Jörg Lehmann and Mike Priddy, eds., *The Trouble With Big Data: How Datafication Displaces Cultural Practices* (London: Bloomsbury Academic, 2021), https://doi.org/10.5040/9781350239654.

 DOI 10.31400/dh-hun.2021.4.3530

A további fejlesztési terveket illetően, az ismert apró kódhibák javítását követően a feltöltési lehetőségeket szeretnénk kiegészíteni többek között a TEI XML formátummal, valamint olyan API-k integrálásával, melyek metaadatokat és teljes szöveget tudnak kinyerni különböző, nemzetközi standardokat használó tudományos adatforrásokból. A felhős környezetben működő és skálázható szerverkapacitás függvényében a felhasználóknak lehetőséget biztosítanánk saját, egyénileg tanított spaCys nyelvi modellek feltöltésére. Az AVOBMAT-ba feltöltött adatbázisok szemantikus gazdagítását a szófaji egyértelműsítéssel (POS tagging) szeretnénk bővíteni, melynek a grafikus felületen is kereshető eredményeiről szintén készítenénk statisztikai kimutatásokat. Az egyértelműsített metaadatok, a névelem-felismeréssel és névelem-összekapcsolással (Named Entity Linking: NEL), valamint magyar nyelvű szövegekre finomhangolt *wikifier*rel azonosított és a *Wikidata* szemantikusan összekapcsolt névelemeihez rendelt adatok felhasználásával *Neo4j*-re épülő tudásgráfokat szeretnénk létrehozni.[41] A komoly nyelvtechnológiai kihívásokat okozó névelem-azonosítás és -összekapcsolás hatékonyságának növelését segítik az AVOBMAT többnyelvű automatikus morfológiai elemzői (pl. lemmatizálás). A magyar *DBpedia*[42] és *Nemzeti Névtér*[43] elemeivel is gazdagított tudásgráf-adatbázisok és szemantikus tudáshálók segítségével további, eleddig ismeretlen kapcsolatokat fedezhetnénk fel, többek között szövegek, metaadatok, személyek, helyszínek, események és fogalmak között.[44]

Introducing the AVOBMAT (Analysis and Visualization of Bibliographic Metadata and Texts) Multilingual Research Tool

The objective of this paper is to demonstrate the workflow, different analytical functions and features of the multilingual AVOBMAT (Analysis and Visualization of Bibliographic Metadata and Texts) digital tool. This web application enables researchers to critically analyse bibliographic data and texts at scale with the help of data-driven methods and tools supported by artificial intelligence and natural language processing techniques. The unique features of the AVOBMAT toolkit are that (i) it can preprocess, analyse and (semantically) enrich a huge number of texts and metadata in several languages; (ii) the implemented analytical and visualization tools provide interactive close and distant reading of texts and bibliographic data; (iii) it combines bibliographic data and natural language processing research methods in one integrated, interactive and user-friendly web application. In the preprocessing phase, the user can set nine optional parameters such as lemmatization and stopword filtering. Users can create different configurations for the different analyses and

[41] *Neo4j*, hozzáférés: 2021.12.15, `https://neo4j.com/`.

[42] *DBpedia*, hozzáférés: 2021.12.15, `https://hu.dbpedia.org/`.

[43] *Nemzeti Névtér*, hozzáférés: 2021.12.15, `https://abcd.hu/`.

[44] Lásd pl. Biography Shampoo: A Network of Finnish Biographies on the Semantic Web, hozzáférés: 2021.12.15, `http://biografiasampo.fi/`; David Lindemann and Christiane Klaes, „Zotero to Wikidata Through Wikibase: A Workflow for Publication Metadata LOD-ification Using Free Software," megjelenés alatt.

visualizations. The metadata enrichment includes the automatic identification of the gender of the authors and automatic language detection. Users can search and filter the uploaded and enriched bibliographic data and preprocessed texts in faceted, advanced and command line modes. Having filtered the uploaded databases and selected the metadata field(s), users can (i) analyze and visualize the bibliographic data chronologically in line and area charts in normalized and aggregated formats; (ii) create an interactive network analysis; (iii) make pie, horizontal and vertical bar charts of the bibliographic data. As for the content analysis, the diachronic analysis of texts is supported by the N-gram viewer. Two types of frequency analyses are implemented: the significant text function shows what differentiates a subset of documents from other texts in the corpus, and the TagSpheres enables users to investigate the context of a word. The close reading is also fostered by the Keyword in Context tool. AVOBMAT has an in-browser Latent Dirichlet Allocation function to calculate and visualize topic models. It semantically enriches the texts and metadata by the use of named entity recognition in 16 languages. The export functions of AVOBMAT facilitate the reproducibility of the results and transparency of the preprocessing and text analysis. It helps users realize the epistemological challenges, limitations and strengths of computational text analysis and visual representation of digital texts and datasets.

Keywords:
text and data mining, multilingual digital tool, natural language processing, metadata, semantic enrichment

 DOI 10.31400/dh-hun.2021.4.3530