

ÚJ TECHNOLÓGIÁKKAL, ÚJ TARTALMAKKAL A JÖVŐ DIGITÁLIS TRANSZFORMÁCIÓJA FELÉ

32. Networkshop: országos konferencia

2023. április 12–14.
Pannon Egyetem, Veszprém

ÚJ TECHNOLÓGIÁKKAL, ÚJ TARTALMAKKAL A JÖVŐ DIGITÁLIS TRANSZFORMÁCIÓJA FELÉ

32. Networkshop: országos konferencia

2023. április 12–14.
Pannon Egyetem, Veszprém

Szerkesztette: Tick József, Kokas Károly, Holl András

HUNGARNET Egyesület
Budapest, 2023





Szerkesztette: Tick József, Kokas Károly, Holl András

Tipográfia és tördelés: Vas Viktória

Networkshop
2023. április 12–14. Pannon Egyetem, Veszprém konferencia előadásainak közleményei

ISBN 978-615-82243-1-4
DOI: 10.31915/NWS.2023

Kiadja a HUNGARNET Egyesület
az MTA Könyvtár és Információs Központ közreműködésével
Budapest
2023

Borítókép: freepik.com

TARTALOMJEGYZÉK

Előszó 5
Király Sándor, Balla Tamás: Flipped classroom az sqlsuli.hu-ban 7
Wirágh András: Abaújszántótól Zsombolyáig.
Megjegyzések egy új sajtóadatbázishoz 14
Albert Ágota Katalin: Az EGT-tagállamok adatvédelmi felügyeleti hatóságainak
szankcionálási gyakorlata az oktatási szektorban a GDPR alkalmazása óta 19
Simon András: Digitális dokumentumok gyűjteménykezelési gyakorlatának támogatása
a digitális tartalmak számossága, mérete és féleségeik vizsgálatával 24
Bódog András: Az Annif gépi tárgyszavazó rendszer magyarországi adaptációjának
feltételei és lehetőségei 31
Dezső Krisztina: A Pécsi Egyetemtörténeti Gyűjtemény online adatbázisai
és digitális gyűjteményei 36
Ungváry Rudolf, Király Péter: Nemzeti könyvtárak és az OSZK MARC21 állományainak
összehasonlító elemzése néhány adatmező alapján 42
Szemes-Révész Enikő Evelin: Kapocs a tudáshoz – A könyvtár szerepe a civilek
és a tudomány kapcsolatában 50
Tóth Zoltán: Az RO-Crate alapú kutatási objektum csomagolás keretrendszere
az ELKH ARP platformban 54
Király Roland, Király Sándor, Palotai Martin Marcell:
Neurális hálózatok oktatási alkalmazását támogató keretrendszer Virtual (VR) és
Augmented Reality (AR) eszközökkel 60
T. Nagy László: Mesterséges intelligencia, multimédia, tanulástámogatás 69
Horváth Péter: Egy automatikusan generált rímszótár fejlesztése
és a magyar kanonikus költészet rímszavainak néhány jellemzője 77
Héjja Balázs, Tóth-Jávorka Brigitta, Tóth Máté: Digitális tartalomfejlesztés
közkönyvtári környezetben 85
Koczka Ferenc: Szemelvények egy felsőoktatási rendszer informatikai
védelmének tapasztalataiból 91
Bolya Mátyás: A digitális gyűjtésrekonstrukció lehetőségei: az Ethiofolk projekt 99
Dobás Kata, Sidó Zsuzsa, Szabó-Reznek Eszter: A Kolozsvári Állami Magyar Színház
jelmezterveinek digitalizációja és felvitele az ITIdata adatbázisba 108
Köpösdi Zsuzsa: H5P-ben létrehozható interaktív és adaptív tananyagok 116
Fülöp Tiffany, Molnár Tamás, Hoczopán Szabolcs: Komplex kutatástámogató
szolgáltatási portfólió az SZTE Klebelsberg Könyvtárban 122
Vass Johanna: Az Open Science könyvtári vonatkozásai 129
Antal Péter, Czeglédi László: A digitális oktatás módszertana a gyakorlatban 135
Máray Tamás: A szuperszámítástechnika mint európai stratégiai ágazat 143
Frankó Máté, Zeller Rozália: Szoftveres Cutter-keresés
az SZTE Klebelsberg Könyvtárban 151
Zsiborács Judit, Dési Ádám Dániel, Nagy Attila Árpád, Urbán Katalin:
Tudománymetriai műhely könyvtári környezetben 157

CC BY

Palkó Gábor, Szekrényes István, Bobák Barbara: A Digitális Örökség Nemzeti Laboratórium webszolgáltatásai automatikus kézírás-felismertetéshez 164
Szűcs Kata Ágnes: Adatvizualizációs lehetőségek a bölcsészettudományban 170
Leitgéb Mária: A BME Építészettörténeti és Műemléki Tanszék repozitóriuma 178
Mihály Eszter, Micsik András: Szerkesztői környezet TEI-alapú szövegkiadásokhoz 186
Dobás Kata, Fellegi Zsófia, Palkó Gábor: A kis gömböc meséje – az ITIdata irodalomtudományos adatbázis fejlesztése 2022–2023-ban 192
Alföldi István, Szemigán Dorottya Henrietta, Palkó Gábor, Fellegi Zsófia: Kutatói e-mail hagyaték archiválása és feldolgozása 199

ELŐSZÓ

Tisztelt Olvasó!

Mekkora a visszhangja a Networkshop konferenciakötetekben megjelent cikkeknek? Szükség van-e olyan konferenciakiadványra, amelyik nem rendelkezik Q értékkel?

A magyar tudományos szaknyelv életben tartása nem képzelhető el magyar szakcikkek, könyvek kiadása nélkül. Minden szakmának – így a felsőoktatási/közgyűjteményi informatikának is – szüksége van publikációs lehetőségre, eredményei írásos közlésére. Szüksége van még akkor is, ha a Networkshop résztvevői és előadói többnyire nem érdekeltek bibliometriai mutatóik növelésében.

A hatodik szerkesztett, lektorált Networkshop konferenciakötet 29 tanulmányt tartalmaz – visszatérő és új szerzők választották előadásuk írott változatának közlésére ezt a kiadványt. Az öt korábbi kötetben összesen 120 cikk jelent meg – az évi cikkszám némileg változik, de a szerzők érdeklődése fennmarad.

Mi a helyzet az olvasókkal? A REAL-ból az eddig megjelent öt kötetet, valamint a külön-külön is elérhető cikkeket eddig több, mint 14 ezerszer töltötték le. (Az anyagok letölthetőek az MTA KIK OCS rendszeréből és a szegedi Contenta-ból is, tehát a letöltések összege ennél minden bizonnyal több.) Az MTMT-ben az öt Networkshop kötet cikkeire rögzített hivatkozások száma 24 – ami lehetne nagyobb is, ha a szerzők felkutatnák és az adatbázisba felvinnék ezeket.

Statisztikák ide vagy oda – reméljük az olvasó örömmel forgatja (görgeti?) ezt a kötetet (és ha talál publikációiban felhasználható információkat, idézi is...)!

Budapest, 2023. december 6.

A szerkesztők



 DOI: 10.31915/NWS.2023.1

Flipped classroom az sqlsuli.hu-ban

Flipped classroom in sqlsuli.hu

Király Sándor, Balla Tamás
Eszterházy Károly Katolikus Egyetem, Informatikai Kar
kiraly.sandor@uni-eszterhazy.hu
balla.tamas@uni-eszterhazy.hu

Absztrakt

Az új Magyar Nemzeti Alaptanterv (NAT2020) a korábbi NAT2012-től eltérően kötelezővé teszi az SQL nyelv ismeretét a diákok számára az emelt szintű érettségin, amely kihívások elé állítja mind a tanulókat, mind a tanáraikat. Elkészült és már használható egy új oktatási portál, az sqlsuli.hu, amely azzal a céllal készült, hogy segítse a tanulókat az SQL (Strukturált Lekérdező Nyelv) elsajátításában. Ennek az online oktatási platformnak a keretrendszere olyan értékelő rendszerrel rendelkezik, amely segíti a tanulókat az SQL parancsok kipróbálásában, tesztelésében, akár mindenfajta tanári beavatkozás nélkül is. A létrehozott mintaadatbázisok a Harry Potter és a Csillagok háborúja filmekhez kapcsolódnak, mely várhatóan növeli a tanulók elkötelezettségét és motivációját a nyelv elsajátításában. A portál felhasználható a Flipped Classroom pedagógia alkalmazására az SQL oktatásában. A tananyag szövege gazdag képi illusztrációkat, idézeteket, érdekes információkat tartalmaz mind a Varázslóvilágról, mind a Galaxisról, így kiválóan felhasználható az otthoni tananyagok, például videók elkészítésére is. A portál értékelőrendszere, csevegő csatornája, és az, hogy a tanár is láthatja az egyes diákok próbálkozásait, megoldásait teszi alkalmassá a portált a tanórai gyakorlásra, csoportmunkára.

Kulcsszavak: Learning Management System, online platform, Flipped Classroom, SQL.

Abstract

The new Hungarian National Base Curriculum (NAT2020), unlike the previous NAT2012, makes knowledge of the SQL (Structure Query Language) language mandatory for students in the advanced final examination, which poses new challenges for both students and their teachers. An educational portal, sqlsuli.hu, has been developed and launched in order to help students to learn SQL. The framework of this online learning platform has an extensive grader tool that helps students test their SQL commands without intervention from a teacher, thus providing a flexible learning experience. The created sample databases are based on Harry Potter and Star Wars data, which is expected to increase students' engagement and motivation for learning. The portal can be used for applying the Flipped classroom pedagogy in the SQL education. The text of the course material contains rich pictorial illustrations, quotes, and interesting information about both the wizarding world and the Galaxy, so it can be used excellently for preparing home course materials and videos. The evaluation system and the chat channel of the portal, and the fact that the teacher can also see their students' attempts and solutions, can be used for class practice and group work. The course is available in Hungarian language.

Keywords: Learning Management System, online platform, Flipped Classroom, SQL.

CC BY

1. Bevezetés

A Flipped Classroom a hagyományos oktatás egyfajta megfordításának tekinthető, ennek megfelelően „megfordított tanterem” vagy „tükrözött osztályterem” lehet a magyar nyelvű megfelelője. Olyan tanulásszervezési megoldásnak tekinthető, melynek során a diákok otthon tekinthetik meg a tanár által elkészített tananyagot, beleértve az online forrásokat, a tanteremben pedig a hagyományos oktatásban egyébként otthonra szánt feladatok kerülnek megoldásra. Ennek megfelelően a tanteremben az interaktív tevékenység, a kollaboratív munka áll a középpontban [3],[20]. A megvalósításhoz a tanárnak a tananyagokat otthoni megtekintésre elérhetővé kell tennie. A tananyagok nem feltétlenül csak videófolyamok lehetnek, a módszer kezdeti alkalmazásakor (80-as, 90-es évek) képekkel illusztrált szöveges állományok kerültek megosztásra. Az osztályteremben történik a kiscsoportos feldolgozás, közösen, az oktató által támogatva, egyéni tanulási utak követésével és egyéni igények figyelembevételével [5],[13],[19]. A tükrözött osztályterem esetében nagyon fontos, hogy a diákok az otthoni felkészülés során az előzetes ismeretszerzés szakaszában teljesítsék a feladatokat, azaz elolvassák a közzé tett anyagot, megnézzék a tanár által közzétett videókat. Ez minden munkának a kiinduló alapja.

A 2020-as NAT alapján az informatika oktatás egyik nagy változása az, hogy emelt szinten az adatbázis-kezelési feladatokat csak az SQL nyelv ismeretével lehet megoldani, a korábban használt Access QBE (*Query-By-Example*) rács nem használható. Egy programozási nyelv elsajátítása nehezebb lehet, mint az SQL utasításoké, ugyanakkor a 18 éven aluli diákok számára az SQL megtanulása is időigényes, és nem is olyan egyszerű, mint gondolnánk. Ez részben az SQL természetéből adódóan, illetve abból a tényből fakad, hogy alapvetően különbözik a középiskolások tanulmányai során elsajátított többi készségtől [18]. A nyelv elsajátítását az sem segíti, összehasonítva egy programozási nyelvvel, hogy az SQL utasítások eredménye jobb esetben egy tábla adatokkal, rosszabb esetben egyetlen sornyi adat vagy egy szám, ami egyáltalán nem motiváló. Az elsajátítás megkönnyítésére Al-Shuaily és Renaud az SQL minták alkalmazását javasolta [2], Mitrovic egy tudásalapú tanítási rendszert fejlesztett az SQL számára [15], Quer és társai pedig egy olyan szoftvereszközt implementáltak, a LearnSQL-t (*Learning Environment for Automatic Rating of Notions of SQL*), amely lehetővé teszi az automatikus és hatékony e-learninget és a relációs adatbázis-készségek értékelését [17]. Garner és Mariani egy olyan grafikus felhasználói felületet valósított meg, amelynek középpontjában egy lekérdezés szöveges fordítása áll, és amely megkönnyíti az SQL megértését a tanulók számára [8].

Az sqlsuli.hu portálon a diákok megtanulhatják, hogyan kell alkalmazni a különböző SQL utasításokat mint például a *Select*, *Update*, *Delete* stb. A tananyag a *Select* utasításra helyezi a hangsúlyt, mivel a diákoknak elsősorban ezt kell tudniuk az érettségin. 24 fejezeten keresztül a tanulók megtanulhatják az utasítás használatát az alapvető használaton túl, a *Join* műveletek és a beágyazott *Select* utasítások segítségével. A DDL (*Data Definition Language*) és a DML (*Data Manipulation Language*) utasításaival is megismerkedhetnek a portál felhasználói. A DML utasításokat (*Update*, *Delete From*, *Insert*, *Insert* és *Select* együtt segítségével) négy fejezetben mutatjuk be gyakorlatokkal. Egy fejezetben a DDL utasítások (*Create table* és *Alter table*) is bemutatásra kerülnek egy gyakorlati feladattal.

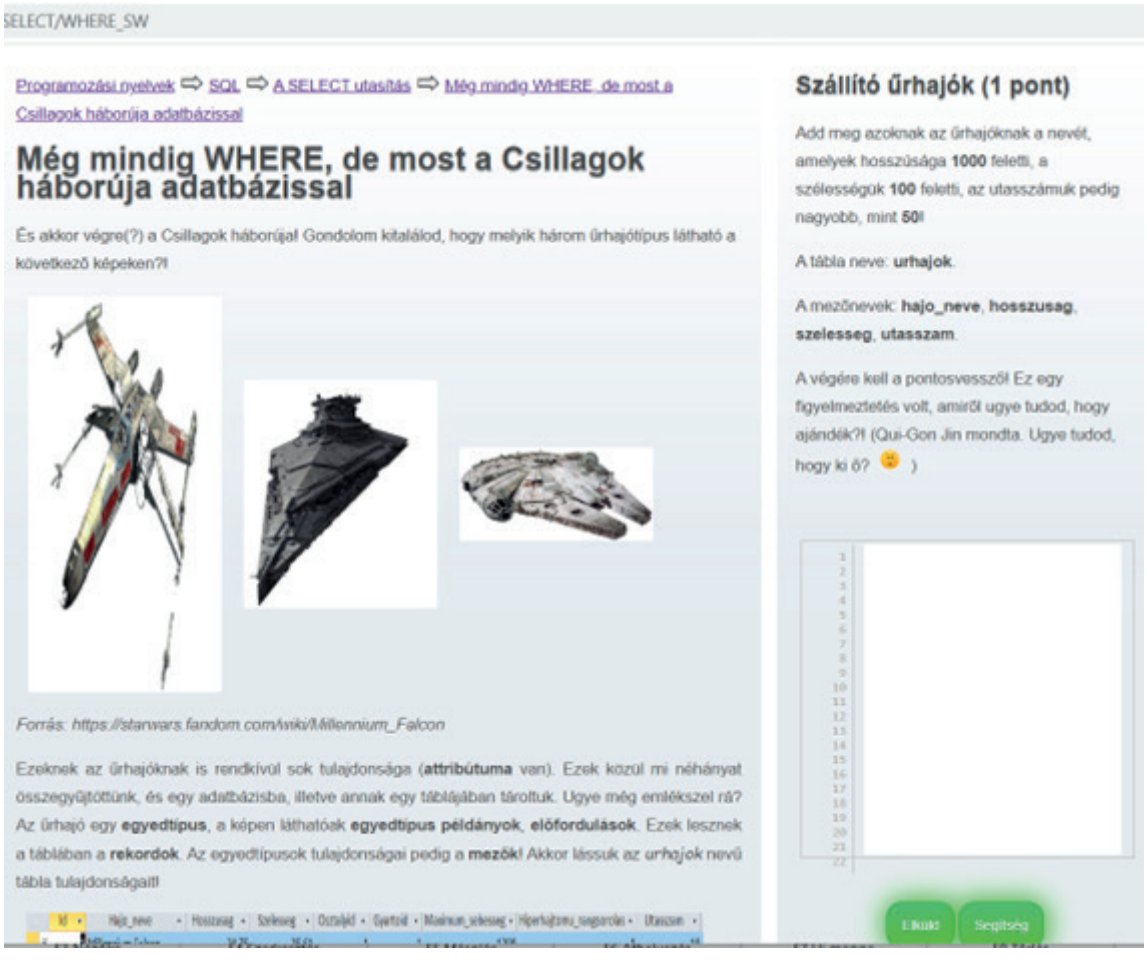
Bár az adatmodellezés nem tartozik az informatika érettségi követelményei közé, a portálon megmutatjuk a diákoknak, hogyan készültek a Harry Potter és a Star Wars adatbázisok. A tananyag a törlési és módosítási anomáliákat, valamint a normálformákat is tartalmazza.

2. Tanulás és gyakorlás az SQL suliban

A főoldalon három figyelemfelkeltő videó található, amelyek vonzóbbá teszik webhelyünket, és ezek közül az egyik mindig véletlenszerűen jelenik meg. A portálra lehet tanárként és diákként is regisztrálni. Az előbbi esetben a rendszer generál egy kulcsot a tanár számára, amelyet megadhat a diákjainak, akik a regisztrációjuk során vagy megadják ezt a kulcsot vagy nem. Előbbi esetben a tanáruk látni fogja, hol tartanak a tananyag feldolgozásában, valamint a sikeres és a sikertelen megoldásaikat is.

Bejelentkezés után a diákok el tudják olvasni az SQL nyelv rövid leírását, majd ki kell választaniuk, hogy adatbázistervezést vagy SQL-t szeretnének tanulni. Az elsajátítandó anyag a képernyő bal oldalán található. Az aktuális témához kapcsolódó feladatok a jobb oldalon találhatók (1. ábra).

A tanulóknak a jobb oldali szövegdobozba kell beírniuk a helyes SQL utasítást. Ha ez megtörtént, a *Küldés (Elküld)* gombra kattintva küldhetik el a rendszernek a megoldásukat. Ha a megoldás helyes, az aktuális témához tartozó következő feladat jelenik meg, de ha ez az utolsó feladat, akkor a következő alfejezet jelenik meg. Ha a megoldás helytelen, figyelmeztető üzenet jelenik meg. A *Help (Segítség)* gombra kattintva az oldal megjeleníti a megoldás egy részét, így segítve a hiba javítását. Ha a tanuló esetleg nem tudja a megoldást, áttérhet egy másik alfejezetre, és később visszatérhet ide, hogy újra megpróbálja megadni a helyes megoldást. A tanulók próbálkozásait a rendszer tárolja, a tanárok ezeket megtekinthetik.

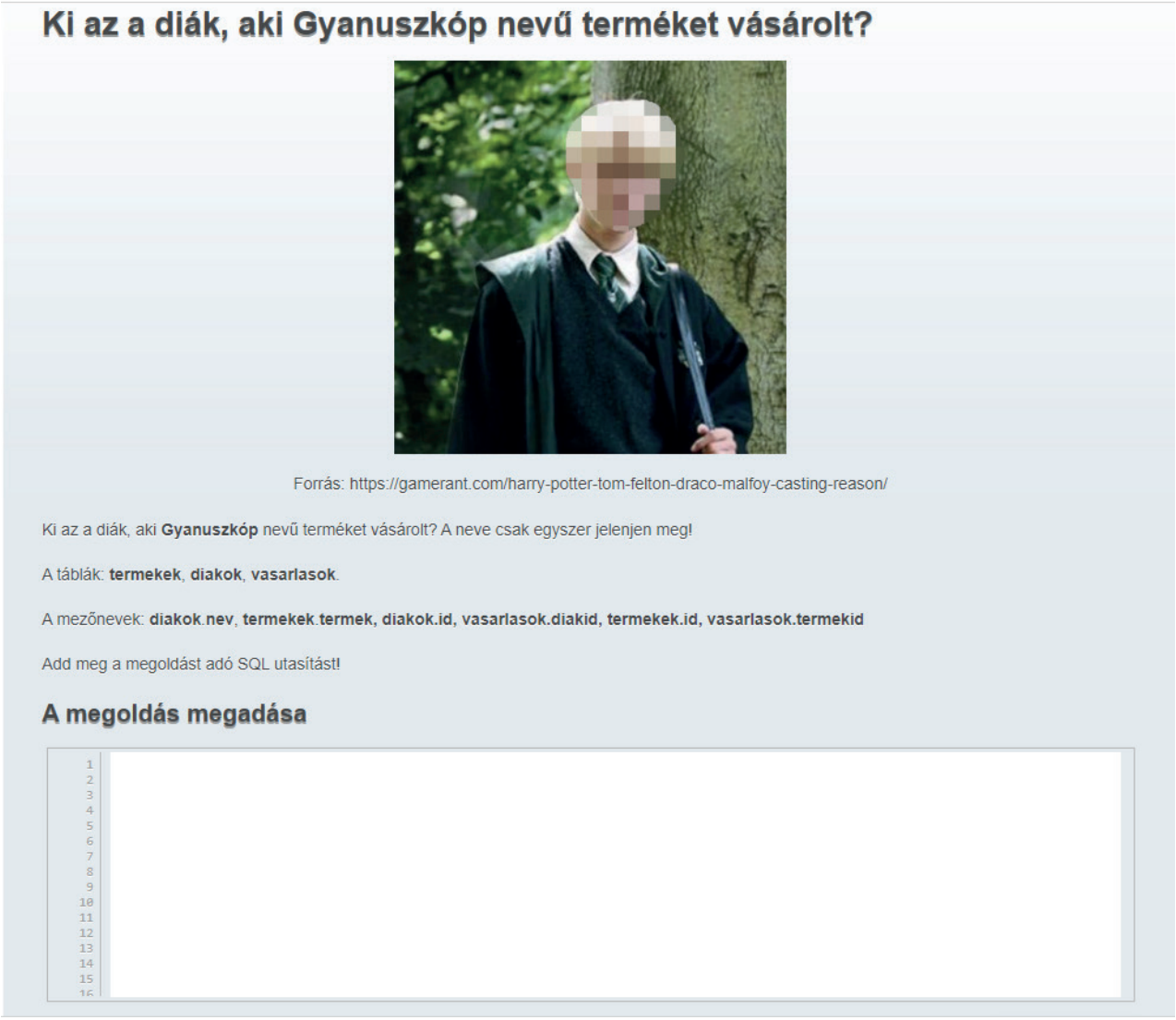


1. ábra: A tananyag szövege a bal oldalon, a jobb oldalon a feladat.

CC BY

Az egymást követő leckék (alfejezetek) egy fejezetet alkotnak. Háromféle tananyagot különböztetünk meg (amelyek egy alfejezetet jelentenek): **tananyag feladatokat**, önálló **feladat** és önálló tananyag. A tananyagok (alfejezetek) szövegblokkokat, a tanulási folyamatot segítő audio-vizuális elemeket, illetve játékokat is tartalmazhatnak. A leckék többsége feladatokat tartalmaz, ilyenkor a tananyag mellett a jobb oldalon egy feladat is megjelenik, amely segíti a tanulókat a lecke elsajátításában (lásd 1. ábra). A legtöbb lecke kettőnél több gyakorlatot tartalmaz. Ezekben a leckékben a felhasznált táblázatok szerkezetét és tartalmát is a szövegben mutatjuk be.

Miután a tanulók feldolgozták a huszonnégy leckét, további sok érdekes információt szerezhetnek Roxfortról és a Galaxisról a következő két fejezetben (jelenleg 23 alfejezetben), amelyek önálló feladat típusú tananyagok. Például a helyes megoldások feltöltése után megjelenik annak a diáknak a neve, aki Gyanuszkópot vásárolt (lásd 2. ábra). Ezekben a feladatokban az előző leckékben használt táblákat használjuk, és a feladatok megoldásához minden információ (a táblák nevei, a mezők pontos neve, mezőnevek a táblák közötti kapcsolatok beállításához) elérhető a feladat szövegében.



2. ábra. Egy önálló feladat tananyag nélkül.

Az önálló feladatok megoldása után érdemes a diákoknak feldolgozni a DDL, DML utasításokat bemutató tananyagokat, majd az adatbázistervezés fejezeteit.

A portálba integrálásra került a DISQUS rendszer, így a felhasználók különféle témákban kommunikálhatnak egymással, valamint megoszthatják egymással a kódolás során szerzett tapasztalataikat, vagy együttműködhetnek egymással [1],[14]. A fórum elolvasásával a tanárok is segíthetnek a diákoknak.

3. A tananyag

Az online környezetben a tanulók teljesítményét számos tényező befolyásolhatja, amelyek a tanulók egyéni jellemzőiből és szokásaiból fakadnak. Ide tartozik a tanulók azon képessége, hogy fenntartsák belső tanulási motivációjukat és figyelmüket [4],[6],[7],[11]. Ezen tényezők befolyásolása online környezetben is lehetséges, hiszen lehetőségünk van az elkötelezettség növelésére, ami a tanuló kognitív folyamataként, a tanulási folyamatban való aktív és érzelmi részvételeként definiálható [4],[16],[21]. Az elköteleződés mindhárom tényezője (kognitív, viselkedési és érzelmi) növelhető, ha a tananyag szövege nemcsak szakmailag korrekt, hanem jó stílusú és személyes hangvételű is [10]. A szöveg elgondolkodtató hatását elősegíthetik jó metaforák, érdekes példák, olykor meghökkentő és provokatív felvetések, kérdések, ellentmondások, meglepő és sok esetben humoros fordulatok szerepeltetése [11],[12]. Az sqlsuli.hu-n a direktív e-learninget részesítik előnyben, mivel az oldal szövege teljesen új a résztvevők számára.

Tananyagunk szövege képzeletbeli párbeszédeket tartalmaz a tananyag készítői és a hallgató között, beleértve a provokatív kérdéseket és megjegyzéseket, valamint az érdekes példákat. A módosított Star Wars és Harry Potter idézetek, mint például „A Sithek mindig ketten vannak, mint most a feladatok.” vagy „A kíváncsiság nem bűn, de legyünk óvatosak a kíváncsiságunkkal, viszont nem ezen a portálon”. Minden leckébe a filmekhez kapcsolódó képeket, valamint hangulatjelek sokaságát helyeztünk el. A videók az *INNER JOIN* jobb megértése érdekében készültek. A tananyagban, ahogyan azt már említettük, a *Select* utasításon kívül a DML és a DDL utasítások is ismertetésre kerülnek.

4. Az SQL suli felhasználása a Flipped Classroom-ban

A portál tananyagának szövege a képekkel együtt könnyen és gyorsan kimásolható a formázások megtartásával, csak a képeket kell átméretezni. Az így kapott dokumentumot ki lehet egészíteni további online forrásokkal, képekkel. Ebben segítenek az Érdekes információk a varázslók világából, valamint az Érdekes információk a Galaxisról fejezetekben található képek, és azok forrásai. A tananyagban található, a filmekből átvett idézetek megkeresését is kérő tananyag a motivációt is növelheti. Például a „A Sith-ekből mindig kettő van”, vajon melyik részben és mikor hangzik el? „A kíváncsiság nem bűn, de óvatosan kell bánnunk a kíváncsiságunkkal.” melyik Harry Potter filmben hangzik el és mikor?

Kihasználható, hogy a portálon **az új ismeretek átadására** a Harry Potter adatbázist használjuk, a megértés mélységét ellenőrző feladatok szintén ezt az adatbázist használják. A Csillagok háborúja adatbázist ugyanakkor az **elsajátított új ismeretek elmélyítésére** használjuk a portálon. Ennek megfelelően az otthoni anyagban mindkét adatbázis felhasználható, az órai megoldásra szintén, így 6-8 feladat megoldására kerülhet sor az órán a portál felhasználásával. Különösen előnyös, hogy az otthonra kiadott tananyag nagy része a feladatok megoldása alatt is átható, hiszen rendelkezésre áll a portálon. Az órai munkára felhasználhatók a korábban már említett, önálló feladatok is. A fenti módon, az órán a tanulók alkalmazhatják az új tudást, a megoldások során az órán kérdéseket tehetnek fel, így jobban megértik az anyagot.

CC BY

A kollaboratív munkát segíti, hogy a tanár láthatja, hogy a diákok mely feladatokat oldják meg éppen, melyek sikerültek már, melyek még nem. Ez utóbbi esetben láthatja a hibás megoldást is. Ez csak akkor valósulhat meg, ha a diákok a regisztrálás során megadták a tanári kódot, amelyet a rendszer akkor generált, amikor a tanár regisztrált. Azaz a tanár az órán tud differenciálni, eldöntheti, hogy az egyes tanulókkal mennyi időt tölt, így egy nagyobb létszámú csoport is jobban kezelhető [9].

A diákok együttműködését segíti a beépített Fórum, ahol a felhasználók különféle témákban kommunikálhatnak egymással, és megoszthatják egymással a kódolás során szerzett tapasztalataikat, vagy együttműködhetnek egymással [1],[14]. A fórum elolvasásával a tanárok segíthetnek a diákoknak. Így több lesz a csoportban az interakció (tanár-diák, diák-diák). A módszer további előnye, hogy a diákok saját tanulási folyamatukat irányítják.

5. Összegzés

A tanulmány célja egy új, interaktív platform, az sqlsuli.hu bemutatása, valamint ennek a felhasználása a Flipped Classrom módszer keretén belül. Ez a webhely olyan SQL tananyagot kínál, melynek elsajátítása után a diákok meg tudják oldani az emelt szintű informatikai érettségi adatbázis-kezelési feladatait. A tananyag és a feladatok Harry Potter és Star Wars adatbázisokra épülnek. A kifejlesztett LMS és a Fórum támogatja a tanulók előrehaladásának nyomon követését az órákon és elősegíti az anyag fejlesztését.

A portált már bárki elérheti regisztráció után, de jelenleg a szerzők diákjai használják. Az anyag jobb megértése érdekében oktatási játékok fejlesztését is tervezzük, amelyeket beillesztünk a fejezetek közé.

Irodalom

[1] Arefin, Ahmed Shamsul (2015): "*Pedagogy of Computer Programming: An Interactive and Collaborative Learning Approach.*", Macquarie University Postgraduate Certificate of Higher Education(2015). EDCN 871 Final Project

[2] Al-Shuaily, H., Renaud, K. (2010): "*SQL Patterns: A New Approach For Teaching SQL.*", 8th International Workshop. Teaching, LEarning and Assessment Of Databases. TLAD.

[3] Alvarez, B. (2011): "*Flipping the classroom: Homework in class, lessons at home.*", Archived 2011-12-22 at the Wayback Machine. *Education Digest: Essential Readings Condensed For Quick Review*, **77** (8): 18–21.

[4]. Balla, T., Király, S. (2020): "*A discussion of developing a programming education portal.*", Central-European Journal of New Technologies in Research, Education and Practice (2020): Volume 2, Number 2. DOI: https://doi.org/10.36427/CEJNTREP.2.2.833

[5] Bergmann, J., & Sams, A. (2012): „*Flip your classroom: reach every student in every class every day.*", Washington, DC: International Society for Technology in Education.

[6] Faragó, B.: "*Tanulói aktivitás, aktív tanulás és tevékenység online környezetben.*", In: Papp-Danka, Adrienn; Lévai, Dóra (szerk.) Interaktív oktatásinformatika, 2015.

[7] Clark, R.C., Mayer, R.E. (2011): "*E-learning and the science of instruction.*", Pfeiffer, San Francisco, (2011)

[8] Garner, P., and Mariani, J.A. (2015): "*Learning SQL in steps.*", In: Journal on Systemics, Cybernetics and Informatics, Vol. 13, No. 4, 2015, p. 19–24.

[9] Hartyányi, M és társai (2018): "Fordított osztályterem a gyakorlatban.", Letölthető: https://www.flip-it.hu/sites/default/files/Public/partner_files/forditott_osztalyterem_a_gyakorlatban_hu.pdf

[10] Héjja-Nagy, K. (2015): "*Tanulási stratégiák és a tanulói aktivitást befolyásoló egyéni feltételek online környezetben.*", In: Papp-Danka, Adrienn; Lévai, Dóra (szerk.) Interaktív oktatásinformatika p. 33–49 (2015).

[11] Király, S. (2016): "*Tanulás támogatása digitális környezetben*.", OKTATÁS-INFORMATIKA 2016: 1 pp. 29–40., 12 p. (2016)
[12] Király, S. (2016): "*How to Implement an E-learning Curriculum to Streamline Teaching Digital Image Processing*.", ACTA DIDACTICA NAPOCENSIA 9: 2 pp. 13–22., 10 p. (2016)
[13] Lakmal, A., Phillip, D. (2015): „*Motivation and cognitive load in the flipped classroom: definition, rationale and a call for research*", Higher Education Research & Development. **34** (1): 1–14. DOI: 10.1080/07294360.2014.934336
[14] McDowell, C., Werner, L., Bullock, H., & Fernald, J (2002): "*The effects of pair-programming on performance in an introductory programming course*.", SIGCSE '02 Proceedings of the 33rd SIGCSE technical symposium on Computer science education (2002): Pages 38–42.
[15] Mitrovic, A (2022): "*A Knowledge-Based Teaching System for SQL*.", https://www.researchgate.net/publication/2425011_A_Knowledge-Based_Teaching_System_for_SQL. Accessed on 05 May 2022.
[16] Pellas, N. (2014): "*The influence of computer self-efficacy, metacognitive self-regulation and self-esteem on student engagement in online learning programs*.", Evidence from the virtual world of Second Life Computers in Human Behaviour, 35, 157–170, (2014).
[17] Quer, C., et al. (2017): "*E-Assessment of Relational Database Skills by Means of Learn SQL*.", 9th International Conference on Education and New Learning Technologies, 2017. https://doi.org/10.21125/edulearn.2017.0779
[18] Renaud, K., Biljon, J. (2004): "*Teaching SQL - Which Pedagogical Horse for This Course?*", Lecture Notes in Computer Science, 2004. DOI: 10.1007/978-3-540-27811-5_22
[19] Schullery, N. M., Reck, R. F., Schullery, S.E. (2011): "*Toward Solving the High Endrollment, Low Engagement Dilemma: A Case Study in Introductory Business*.", International Journal of Business, Humanities and Technology, vol. 1(2), 1–9.
[20] Smith, B.L., MacGregor, J. T. (1992): "*What is collaborative learning*?", In M. Maher, A.M. Goodsell & V. Tinto (Eds.), Collaborative learning: A sourcebook for higher education. National Center on Postsecondar Teaching, Learning and Assesment.
[21]. Wolf, M. (2007): "*Learning to Think in a Digital World*.", In: Bauerlein, M. (ed.): The digital divide: arguments for and against Facebook, Google, texting, and the ages of social network. Jeremy P. Tarcher/Penguin, New York. 34–37., (2007)

CC BY
III.02 DOI: 10.31915/NWS.2023.2

Abaújszántótól Zsombolyáig. Megjegyzések egy új sajtóadatbázishoz

Wirágh András
ELKH Bölcsészettudományi Kutatóközpont, Irodalomtudományi Intézet
viragh.andras@abtk.hu

Absztrakt

A magyar sajtótörténet aranykorát jelentő századforduló alapvető forrásai a mai napig hiányosak, illetve hiányoznak. Az egyik legégetőbb problémának az 1868 és 1910 közötti időszak hiányzó annotált sajtóbibliográfiája nevezhető, hiszen ennek hiányában a kutató vagy az érdeklődő csak töredékes források segítségével állíthatja össze például egy-egy régió vagy település időszaki kiadványainak listáját. Egy több lépcsőben 2018-tól készülő adatbázis részben orvosolhatja ezt a problémát. A jelenleg 2278 rekordból álló, az ITIdata részét képező adatbázisban a Budapesten 1888 és 1918 között, vidéken pedig 1896 és 1929 között megjelenő időszaki kiadványok alapadatai (cím, megjelenési hely, címváltozás, beolvadás stb.) találhatóak meg. Az adatbázis összeállítását több jelentős forrás segítette, de sok időbe telt az ezekben található következetlen adatok pontosítása, illetve egyes specifikus események összehangolása az adatbázissal. Ideális körülmények között az adatbázis még óhatatlanul töredékes állapotában is hasznos és többirányú kutatást is lehetővé tevő segédfelületté fejleszthető.
Kulcsszavak: magyar irodalom, századforduló, sajtóhálózat, adatbázis, ITIdata

Abstract

The basic sources of the fin de siècle, the golden age of Hungarian press history, are still missing or incomplete. One of the most pressing issues is the lack of an annotated press bibliography for the period between 1868 and 1910, due to which researchers or those interested can only compile a list of periodical publications of a region or settlement with the help of fragmentary sources. A database, to be developed in several stages from 2018, could partially remedy this problem. The database, which currently consists of 2278 records and is part of ITIdata, contains the basic data (title, place of publication, change of title, merger etc.) of periodicals published in Budapest between 1888 and 1918 and in rural areas between 1896 and 1929. The database was compiled with the help of a number of important sources, however, it was necessary to clarify inconsistencies and to synchronize specific events with the database. Under ideal circumstances, the database, even in its fragmented state, could be developed into a useful tool for multi-directional research.
Keywords: Hungarian literature, fin de siècle, press network, database, ITIdata

I. Körülmények

A mindenkori sajtótörténeti kutatások bázisát a hivatalos (akadémiai) *történeti szintézis* és az egyes korok sajtótermékeinek alapadatait megadó komplex (minden régiót és minden laptípust felölelő) *annotált sajtóbibliográfia* jelenti. Míg utóbbi online elérhetőség esetén még a gyors keresést, sőt, az adatbázissá formálást is lehetővé teszi, addig előbbi kulcsfontosságú szerepet vállal abban, hogy megismerjük az annotált sajtóbibliográfia felgyarapodását elősegítő protokollokat, illetve az ebben szereplő szempontok (közreműködők személye és státusa, periodicitás, megszűnés, újraindulás, címváltozás, szünetelés stb.) – további vizsgálódásokhoz elengedhetetlen – kontextusát.

Bár a kapcsolódó segédanyagok mennyiségi és minőségi értelemben sem elhanyagolhatók, a századforduló magyar sajtójával kapcsolatos *alapforrások* rendkívül hézagosak, illetve jószerivel hiányoznak. A történeti szintézis 1892-ig jutott a főnarratíva tárgyalásában, míg teljes annotált sajtóbibliográfiával az 1867-es kiegyezésig, az 1911 és 1920 közötti, illetve az 1921-től kezdődő időszakra vonatkozóan rendelkezünk.[1] A modern magyar irodalom kezdeti időszakával, az 1893 és 1910 közti évekkel kapcsolatban a kutatók és érdeklődők nem fordulhatnak megbízható alapforrásokhoz, de az időszak töredékes feldolgozottsága miatt egymásnak ellentmondó adatokkal is gyakorta szembesülhetnek.

A 2007-ben Szegedy-Maszák Mihály vezetésével elinduló Kosztolányi-kutatás volt az első olyan grandiózus vizsgálat, amely – ha ezek eddig nem is voltak nyilvánvaló tények – felszínre hozta századfordulós szépirodalom és sajtó szoros kapcsolatát, illetve az ehhez fűződő adathiányt. A kutatásban résztvevő adatgyűjtők rengeteg, addig csak korabeli lapokban lappangó szöveget gyűjtöttek fel, de feladatukat folyamatosan nehezítette az, hogy nem tudtak egy, a korszak lapterm̀ését megbízhatóan listázó alapforráshoz fordulni. Igaz, a komoly kutatómunka folyamán részben orvosolták az egyik fent említett hiányt, lévén az eddig hat kötetből álló forrásjegyzék[2] a századforduló majdani annotált sajtóbibliográfiájának egyik legfontosabb forrásává lépett elő.

2018-ban lehetőséget kaptam egy, a Kosztolányiénál jóval csekélyebb mennyiségű és minőségű, de ismeretlenebb korpusz, Cholnoky László szövegeinek vizsgálatára. Örömmel konstatáltam, hogy a több tízmillió digitalizált és archivált hírlapoldalnak köszönhetően szabad szemmel, otthonról férhettem hozzá a szerző szövegeihez, de sokszor csak időigényes adatbányászat árán deríthettem fel, hogy az egyes régiókban vagy településeken melyik lapoknak kellene még utánanéznem. Éppen ezért a Cholnoky-szöveglista összeállításával párhuzamosan nekiálltam egy – erősen szűrt – laplista elkészítésének, amelyet végül adatbázisba szerkesztettem. A jelen tanulmányban bemutatott adatbázis első változata a Petőfi Irodalmi Múzeum könyvtári adatbázisában (KOHA) volt hozzáférhető, ennek bővített és átalakított verziója jelenleg a Bölcsészettudományi Kutatóközpont Irodalomtudományi Intézetében fejlesztett ITIdata részét képezi.

II. Megvalósítás

Feltehetően a mai szemmel felfoghatatlan századfordulós lapbőség az oka annak, hogy az időszakról sem egy részletes és „konszenzusos" összefoglalással, sem egy, a több ezernyi lapot listázó sajtóbibliográfiával nem rendelkezünk. Az adatbőség okozta problémákkal magam is szembesültem, így az első adatbázisba csak az 1896 és 1929 között vidéken megjelenő magyar nyelvű, és alapvetően politikai jellegű napilapokat és hetilapokat vettem fel. Míg az időkorlátot elsősorban Cholnoky László pályafutása jelölte ki (1900-tól már bizonyítottan publikált és 1929-ben hunyt el), a laptípusokat a lehetséges és rendszeres szépirodalmi tartalom megléte okán válogattam ki: ezekben a fajta periodikákban szinte állandóan szerepelt szépirodalom a tárcarovatban, vagy a lapszámok hátsó oldalain Csarnok elnevezésű rovatban (nem is beszélve a húsvéti, pünkösdi és karácsonyi lapszámok szórakoztató mellékleteiről).

---

1 *A magyar sajtó története, II.2. 1867–1892*, fszerk. Szabolcsi Miklós, Akadémiai Kiadó, Budapest, 1985., V. Busa Margit, *Magyar sajtóbibliográfia 1850–1867. A Magyarországon megjelent magyar és idegen nyelven megjelent valamint a külföldi hungarika hírlapok és folyóiratok bibliográfiája*, OSZK, Budapest, 1996., Kemény György, *Magyarország időszaki sajtója 1911-től 1920-ig*, Magyar Nemzeti Múzeum, Budapest, 1942., *A magyarországi hírlapok és folyóiratok bibliográfiája, 1921–1944*, összeáll. Ferenczyné Wendelin Lídia, OSZK, Budapest, 2010.

2 *Kosztolányi Dezső napilapokban és folyóiratokban megjelent írásainak jegyzéke 1–6.*, szerk. Arany Zsuzsanna, majd Dobás Kata, Ráció Kiadó, Budapest, 2008–2018.

CC BY

A gyűjtés alapját a *Vasárnapi Újság*ban, majd a *Magyar Könyvszemlé*ben kezdetben Pákh Albert, majd Szinnyei József és munkatársai által évről-évre publikált laplisták szolgáltatták. A listák a magyarországi hírlapirodalom összes típusát felölelték, és bár egy-két esetben hibák csúsztak az összegző statisztikákba (egyes lapok vándoroltak, illetve duplikálódtak a különböző kategóriák között, illetve egyes években beszámolták az országban megjelenő idegen nyelvű lapokat az össztermésbe, máskor nem), sőt 1907 és 1909, majd 1911 és 1913 között csak a megszűnő és újonnan induló lapok adatait rögzítették, a következetlenségek jelentős részét segédanyagok segítségével sikerült tisztáznom. Bár csak a tárolt anyagra vonatkozó információkat tartalmazzák, az Országos Széchényi Könyvtár online katalógusa és a Kolozsvári Egyetemi Könyvtár digitalizált cédulakatalógusa voltak a fő támpontjaim. Alapvető forrásként tekintettem még Lakatos Éva grandiózus sorozatára,[3] illetve Kemény György csonka sajtóbibliográfiájára, míg az 1921 és 1929 közötti anyag tekintetében Ferenczyné Wendelin Lídia minden tekintetben kiváló forrása állt rendelkezésemre (ez utóbbi két forráshoz ld. 1. lábjegyzet).

Az adatbázisba felvett tételek tartalmazták a periodika címét, címváltozatait, megjelenési helyét, periodicitását és jellegét, a feloldhatatlan következetlenségeket pedig megjegyzésként jelöltem. Ez a lista egészült ki 2022 során az 1888 és 1918 között Budapesten megjelent lapok listájával. (Az időkorlátokat ebben az esetben új kutatásom témája – századfordulós írói karriersémák – indokolta, a listába ezúttal felvettem az irodalmi folyóiratokat és szorosan kapcsolódó képes magazinokat – *Az Érdekes Újság*, *Tolnai Világlapja* stb. – is.) A két részből összeálló adatbázis így jelenleg 2278 rekordot tartalmaz.

Az adatbázis összeállításában és a felmerülő kérdések megvitatásában munkatársam, Dobás Kata volt segítségemre. Az adatok előkészítését excel-táblázatokban végeztük el, a tételek migrálásához QuickStatements-et használtunk. Ezen a szinten abban a kérdésben kellett döntenünk, hogy milyen besorolást kapjanak a különböző típusú periodikák (politikai napilapok és hetilapok, vegyes tartalmú képes lapok, irodalmi folyóiratok stb.). Bízva abban, hogy a sajtóadatbázis egyszer majd annotált sajtóbibliográfiává bővülhet, amelynek annotációiban az érdeklődő részletekbe menően értesülhet a megjelenés különböző adatairól és körülményeiről, úgy döntöttünk, hogy első körben a hírlap/folyóirat felosztást preferáljuk. Eszerint tehát hírlap lett az alapvetően nem irodalmi jellegű, naponta, hetente többször vagy hetente megjelenő lap, míg folyóirat lett a heti (vagy ritkább) megjelenésű, de célzottan irodalmi, művészeti, kulturális jellegű orgánum.

Az adattisztítás során azokkal a lapokkal kapcsolatban merültek fel problémák, amelyek megjelenésüket szüneteltették (1), címváltozáson estek át (2), más lapokba olvadtak be, vagy más lapokat olvasztottak be (3), lévén nem volt egyértelmű, hogy ezeket a fontos, az adatbázisban való szabadszavas keresés szempontjából is kulcsfontosságú megjegyzéseket az adatlap mely pontján szerepeltessük.

Az 1-es esetben a megjegyzést az annotációba szúrtuk be, így, ha a lap egy szűrt (adott évre vagy időtartamra való) keresésénél meg is jelenik a találatok között, megjegyzés jelöli a szünetelés tényét. A 2-es esetnél két alesetet (*a* és *b*) kellett megkülönböztetnünk. Az *a* alesetben, mint például a *Nyugat* és a *Magyar Csillag* esetében az utódlap nem folytatta az előzménylap évfolyamszámozását. Hasonló történt két budapesti napilap, a *Hazánk* és *Az Ország*, illetve *Az Újság* és az *Újság* esetében. (A korábbi esetben a *Hazánk*at körbevevő üzleti kör lehetőségei módosultak, ezért döntöttek a *tabula rasa* mellett 1905 végén, míg *Az Újság* a két világháború közötti időszak egyik legnagyobb belpolitikai botrányában, a frankhamisítási botrányban vált kulcsszereplővé, ezért a rendszer betiltotta és csak kissé

3 Lakatos Éva, *Magyar irodalmi folyóiratok*, Petőfi Irodalmi Múzeum, Budapest, 1972–2000. (A 15 kötetes sorozat 4604 magyar folyóirat, köztük ide sorolt vidéki politikai hetilapok adatait tartalmazza.)

átformált címen engedte újból megjelenni. A másik vétkesnek kikiáltott periodikát, a *Világ*ot végérvényesen betiltották.) Minden esetben nyilvánvaló, hogy tulajdonképpen egyetlen lap két variánsával állunk szemben, amelyeket a hasonló szellemiség, a szerkesztőség összetétele, bizonyos esetekben a megjelenési forma is bizonyít. Mivel azonban egy lap életkorához az évfolyamszámozás szolgál alapvető zsinórmértékként, ezekben az esetekben két-két különböző rekordot rögzítettünk úgy, hogy a rekordok leírásaiban minden címváltozat szerepel.

A *b* aleset a következő: számos esetben a címváltozással folytatódott az eredeti évfolyamszámozás. Ezeknél a lapoknál a rekord az első címváltozatot tartalmazza, annotációban, illetve a 'Más néven' résznél jelölve a későbbi címváltozatokat. Noha ez az eljárás furán hat például a legrégebbi ma is megjelenő budapesti napilap, a Népszava esetében, amely *Munkás Heti Krónika*ként indult (így a *Népszava* csak címváltozatként szerepel az adatbázisban), adatromlás vagy adatvesztés azért nem történhet, mert a célzott keresésben feltűnik a lap széles körben ismert címvariánsa is. Nézzünk egy példát a harmadik esetre!

Végül, pár szót a 3-as esetről: 1904-ben *Ellenzéki Hírlap* (EH) címmel indult lap – az ekkor még önálló településnek számító – Újpesten, 1907-ben pedig elindult a Rákospalotán, Gödöllőn és Újpesten is megjelenő *Szabad Hon* (SZH). 1912-ben az EH beolvasztotta az SZH-t, de ugyanekkor elindult a rákospalotai *Palota-Újpest* (PÚ) című lap is, amely a kezdetektől az EH mutációja (más címen, de azonos tartalommal megjelenő „variánsa") volt, viszont 1913-ban és 1914-ben már önálló szöveggel és új évfolyamszámozással jelent meg. Ha a sajtóbibliográfia keretei korlátozottak lennének, kis ügyeskedéssel akár egy rekorddá is lehetne transzformálni a három lap történetét, de az online adatbázis esetében szerencsére szó sincs ilyenről. Az ökölszabály ezúttal is az évfolyamszámozás: a beolvadt és a beolvasztó lapnál is megjegyzés utal az eseményre, de ezen okból kifolyólag a mutációként induló, de aztán külön életre kelő PÚ is külön rekordként szerepelhet.

Az adatbázisban szabadszavas keresés segítségével címre, címváltozatra, illetve a leírásban található adatokra lehet rákeresni, míg a bonyolultabb keresésekre jelenleg a wikibase-alapú SPARQL lekérdezéseket használhatjuk. (Itt nyílhat lehetőség például az 1900 és 1918 között Budapesten megjelenő lapok kilistázására, illetve a találatok térképre és idővonalra helyezésére.)[4]

III. További tervek

Az első és legfontosabb feladat az, hogy az adatbázis a közeljövőben önálló belépési ponttal rendelkezzen az ITIdatán belül. Ez tenné lehetővé azt, hogy a felhasználók a periodikák adatainak keresése közben ne szembesüljenek az adatbázis más jellegű rekordjai által keltett „zajjal", azaz a keresés eredményei, a találatok egy már eleve megszűrt korpuszból érkezzenek.

Miközben a századfordulós sajtóadatbázis ideális távlati célja az, hogy tartalmazza az 1868 és 1920 között Magyarországon megjelenő összes időszaki kiadvány adatait, illetve ennek az időszaknak az annotált sajtóbibliográfiájaként is funkcionálhasson, ez a kimenet ma – dacára az összeállítást megkönnyítő forrásoknak és technikáknak – csak több körben látszik teljesíthetőnek. Ebből kifolyólag a visszamenőleges adatbővítés helyett a jelenleg az adatbázisban szereplő rekordok annotációnak, ezen belül is a laphoz kötődő személyekre (főszerkesztők, felelős szerkesztők, munkatársak) koncentráló rövid leírások elkészítését látom kivitelezhetőnek. Mindezzel, illetve a személyek névtérbe kapcsolásával lehetővé válna közelebbről megtekinteni a századfordulós sajtó hálózatos rendszerét, az erre való ránagyítással pedig egyszerre válna lehetővé a korban kiemelkedő jelentőségű, de mára

4 Az adatbázis elérhetősége: https://itidata.abtk.hu/wiki/Main_Page

CC BY

elfeledett hálózati csomópontok, lapok és lapalapítók, esetleg települések és régiók kiemelése, ezzel együtt a századfordulós írók karrierépítési stratégiáinak újraértékelése.

Felhasznált irodalom:

A magyar sajtó története, II.2. 1867–1892, fszerk. Szabolcsi Miklós, Akadémiai Kiadó, Budapest, 1985.

A magyarországi hírlapok és folyóiratok bibliográfiája, 1921–1944, összeáll. Ferenczyné Wendelin Lídia, OSZK, Budapest, 2010.

Kosztolányi Dezső napilapokban és folyóiratokban megjelent írásainak jegyzéke 1–6., szerk. Arany Zsuzsanna, majd Dobás Kata, Ráció Kiadó, Budapest, 2008–2018.

Kemény György, *Magyarország időszaki sajtója 1911-től 1920-ig*, Magyar Nemzeti Múzeum, Budapest, 1942.

Lakatos Éva, *Magyar irodalmi folyóiratok*, Petőfi Irodalmi Múzeum, Budapest, 1972–2000.

V. Busa Margit, *Magyar sajtóbibliográfia 1850–1867. A Magyarországon megjelent magyar és idegen nyelven megjelent valamint a külföldi hungarika hírlapok és folyóiratok bibliográfiája*, OSZK, Budapest, 1996.

 DOI: 10.31915/NWS.2023.3

Az EGT-tagállamok adatvédelmi felügyeleti hatóságainak szankcionálási gyakorlata az oktatási szektorban a GDPR alkalmazása óta

Albert Ágota Katalin
4-in-1 Szolgáltató Kft
dralbertagota@gdprszakszeruen.hu

Absztrakt

Jelen tanulmány célja annak bemutatása, milyen fejleményeket hozott a köznevelési és felsőoktatási szektorban az Európai Gazdasági Térség tagállamai területén az általános adatvédelmi rendelet (GDPR) alkalmazása, különös tekintettel a bírságolási gyakorlatra.
Kulcsszavak: GDPR, adatvédelem, felügyeleti hatóságok, közigazgatási bírság, adatvédelmi incidens, érintetti jogok

Abstract

The purpose of this paper is to present the developments in the public education and higher education sectors in the Member States of the European Economic Area in the application of the General Data Protection Regulation (GDPR), with a particular focus on the fining practices.
Keywords: GDPR, data protection, supervisory authorities, administrative fines, data breach, data subjects' rights

A GDPR és a közigazgatási bírság

Az általános adatvédelmi rendelet [GDPR[1]] egyik kulcsfontosságú eleme a felügyeleti hatóságok erősebb jogérvényesítési hatásköre, valamint jelentős mértékű bírságokat határoz meg, illetve rendelkezik a bírságok tagállamok közötti harmonizációjáról is. A bírság mértékéről a GDPR 83. cikke rendelkezik, azonban a tagállamoknak az adott tagállami székhelyű közhatalmi vagy egyéb, közfeladatot ellátó szervvel szemben kiszabható közigazgatási bírság esetében van lehetőségük ezen bírságmértéktől eltérni, így például hazánkban költségvetési szervek esetében legfeljebb 20 millió forint lehet a bírság összege [Infotv.61.§ (4) bekezdés].

Az egységes jogalkalmazás érdekében iránymutatás[2] segíti a hatóságokat, a bírságok nagyságrendjének egységesítése érdekében jelenleg készül az Európai Adatvédelmi Testület iránymutatása[3]. Ez az egységesítés sem fogja megszüntetni a költségvetési szervek „kedvezményét", így például hasonló jogsértésért egy magánegyetem nagyságrendekkel nagyobb büntetést kaphat hazánkban, mint egy állami egyetem.

1 az Európai Parlament és a Tanács (EU) 2016/679 rendelete (2016. április 27.) a természetes személyeknek a személyes adatok kezelése tekintetében történő védelméről és az ilyen adatok szabad áramlásáról, valamint a 95/46/EK irányelv hatályon kívül helyezéséről (általános adatvédelmi rendelet, EGT-vonatkozású szöveg)

2 A 29. cikk szerinti adatvédelmi munkacsoport: Iránymutatás a 2016/679 rendelet szerinti közigazgatási bírság alkalmazásáról és megállapításáról, WP 253

3 Guidelines 04/2022 on the calculation of administrative fines under the GDPR (Adopted - version for public consultation

CC BY

A szankcionálási gyakorlat alakulása

A vizsgált jogsértések egy része bármely adatkezelőnél előfordulhatott volna, míg mások az intézmények alaptevékenységével függtek szorosan össze. Ezen jogsértések általában súlyosabb megítélésűek, mivel az ügyek többsége kiszolgáltatott személyeket érint (gyermekeket, munkavállalókat), illetve gyakran egy eset számos érintett jogaira és szabadságaira jelent kockázatot. Gyakran eredményezett adatvédelmi incidenst a modern technológia használata, például az alkalmazások nem megfelelő jogosultságmenedzsmentje, tesztelési hiányosságok vagy egyéb biztonsági hibák. Ezen ügyek közös jellemzője, hogy többnyire megelőzhetőek lettek volna, ha az intézmények megfelelő technikai és szervezési intézkedéseket vezetnek be, illetve a meglévő intézkedéseknek érvényt szereznek. Egy dán egyetemen egy tesztelés nélküli szoftverfrissítés miatt a HR-rendszerben megváltoztak a jogosultsággal kapcsolatos beállítások, ezért 400 személyes fájlhoz férhetett hozzá két hétig az egyetem több mint hétezer alkalmazottja. Az egyetem nem naplózta a hozzáféréseket, így utólag nem lehetett ellenőrizni, volt-e jogosulatlan hozzáférés az információkhoz. *(*szankció: megrovás; *Datatilsynet; 2021-442-13989*[4]*)*

Norvégiában az adatvédelmi hatóság több, mint 158 ezer eurós bírságot szabott ki egy önkormányzatra. 2018 májusában egy 12 éves diák jelezte, online talált egy olyan mappát több mint 35 ezer személy felhasználónevével és jelszavával, az iskola vezetősége azonban nem tett lépéseket. Augusztusban a diák az iskola igazgatójaként bejelentkezett az oktatásmenedzselő rendszerbe és üzenetet küldött több személynek azért, mert a korábbi jelzését az iskola nem vette komolyan. Amikor az iskola tudomására jutott az üzenetküldés, értesítették a rendőrséget, akik kiderítették, a diák egyszerűen kitalálta az igazgató jelszavát. Az adatvédelmi hatóság a sajtóból értesült az incidensről, majd a vizsgálata során megállapította, az iskola nem engedélyezte a kétfaktoros hitelesítést annak ellenére, hogy a felügyeleti hatóság 2013-14-ben kampányolt e témában az oktatási ágazatban. Az ügy nagy médiafigyelmet kapott azért, mert az iskola feljelentette a diákot, amelyet később visszavont, amikor a rendőrség már befejezte a nyomozást. *(Datatilsynet; 18/02140*[5]*)*

A belga adatvédelmi hatóság szerint a célhoz kötöttség elvét sérti, ha az iskolától a szülők olyan hírlevelet kapnak, amelyben az összes szülő e-mail címe látható („CC"). Az ilyen adatkezelés nem felel meg a szülők észszerű elvárásainak, mivel a kapcsolati adataikat az iskolával kapcsolattartás céljára adták meg, nem pedig azért, hogy azt megosszák az iskola összes többi szülőjével. (szankció: megrovás; *APD/GBA*; *DOS-2020-00608*[6]*)*

Az olasz felügyeleti hatóság gyermekek személyes adatainak iskolai homlokzatra kifüggesztése miatt szabott ki 2 ezer eurós bírságot *(Garante per la protezione dei dati personali*; *9445324*[7]*)*, a spanyol adatvédelmi hatóság pedig azt sérelmezte, hogy egy iskola száz felvett diák személyes adatait tette elérhetővé a nyilvánosság számára a homlokzatán és a honlapján is, így a kiválasztási eljáráshoz nem kapcsolódó harmadik felekkel közölte ezeket a személyes adatokat. (*AEPD*; *PS/00024/2019* [8]*).*

4 https://www.datatilsynet.dk/afgoerelser/afgoerelser/2022/maj/alvorlig-kritik-af-syddansk-universitets-utilstraekkelige-testning-af-softwareopdatering

5 https://www.datatilsynet.no/contentassets/67033efe6b8a48d7aa679be2c8fd436d/18-02140-13-vedtak-om-overtredelsesgebyr---melding-om-avvik-hos-bergen-kommune-253778_15_1.pdf

6 https://gegevensbeschermingsautoriteit.be/publications/beslissing-ten-gronde-nr.-03-2021.pdf

7 https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9445324

8 https://www.aepd.es/es/documento/ps-00024-2019.pdf

A dán adatvédelmi hatóság egy rossz helyre küldött levél miatt adott megrovást, mivel az e-mailben lévő információ generálta pletyka jelentős kockázatot jelentett az érintett diák jogaira és szabadságaira, valamint az incidenst nem jelentették a hatóságnak *(Datatilsynet; 2021-32-2067*[9]*)*, az olasz adatvédelmi hatóság pedig 4 ezer eurós bírságot szabott ki egy középiskolára, mert a hivatalos honlapján közzétette a tanárok teljes névsorát, e-mail címüket, adóazonosító számukat és egészségi állapotukra vonatkozó adatokat, külön jelzést téve a fogyatékkal élő, illetve mozgássérült tanárok neve mellé *(Garante per la protezione dei dati personali; 9283014*[10]*)*

Egy magánóvoda a beíratott gyermekek szüleit tájékoztatta az egyik pedagógus terhességéről, emiatt az intézményt ezer eurós bírsággal sújtotta az olasz adatvédelmi hatóság, figyelembe véve az oktatási ágazat sajátos helyzetét a világjárvány idején. *(Garante per la protezione dei dati personali; 9776444*[11]*).*

Az adatvédelmi incidensek mindennapos formája a személyes adatok „elvesztése" is, a horvát adatvédelmi hatóság egy mesterdiploma elvesztése esetében állapította meg a GDPR rendelkezéseinek megsértését. *(AZOP; Decision of 31 May 2022*[12]*)*

Egyre elterjedtebb **a biometrikus adatokat kezelő** rendszerek használata, melyeket az intézmények általában a jelenlét ellenőrzésére használják. Egy katalán állami középiskolában **a** rendszer – a diákok szüleinek hozzájárulásával – az első évfolyam arcvektorait gyűjtötte össze (csak ennél az évfolyamnál használták), kiegészítve az ikrek ujjlenyomatának adataival, mivel az arcuk azonos, de az ujjlenyomatuk különböző volt. A hatóság megállapította – többek között –, hogy lett volna a privát szférába kevésbé behatoló módja az ellenőrzésnek (lásd hagyományos jelenlét ellenőrző módszerek), illetve az iskola nem tájékoztatta megfelelően a szülőket. (figyelmeztetés; *APDCAT; PS 49/2019*[13]*)*

Svédországban 20 ezer eurós bírságot szabott ki az adatvédelmi hatóság egy iskolára, mert az arcfelismerő technológiát használt a diákok jelenlétének regisztrálásához. A hatóság szerint a diákok nem adhatnak érvényes hozzájárulást a biometrikus adataik kezeléshez, mivel nem volt szabad lehetőségük arra, hogy megtagadják vagy visszavonják a hozzájárulásukat, illetve a technológia a tanulók integritásába való beavatkozás aránytalan a jelenlét regisztrálásának feladatához képest, amely kevésbé privát szférába hatoló módon is elvégezhető. *(KamR Stockholm; Case No. 5888-20*[14]*)*

Az **elektronikus megfigyelőrendszerek** alkalmazása is gyakran eredményez bírságot. Görögországban egy magániskola igazgatója figyelte meg az idegen nyelvet oktató alkalmazott Zoom online kurzusait anélkül, hogy erről megfelelően tájékoztatta volna őt, illetve annak ellenére tette ezt, hogy az alkalmazott emiatt tiltakozott. A munkáltató nem tudta bizonyítani, hogy az igazgató ezen tevékenysége megfelelő és szükséges eszköz az

9 https://www.datatilsynet.dk/afgoerelser/afgoerelser/2021/okt/klage-over-sikkerhedsbrud-hos-falkonergaardens-gymnasium-og-hf

10 https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9283014

11 https://www.gpdp.it/web/guest/home/docweb/-/docweb-display/docweb/9776444

12 https://azop.hr/wp-content/uploads/2022/09/RJESENJE-gubitak-osobnih-podataka.pdf

13 https://apdcat.gencat.cat/web/.content/Resolucio/Resolucions_Cercador/Resolucions/Documents/ca_ps_2019_049.pdf

14 https://www.imy.se/globalassets/dokument/beslut/beslut-ansiktsigenkanning-for-narvarokontroll-av-elever-dnr-di-2019-2221.pdf

CC BY

intézmény jogos érdekének gyakorlásához, nem vette figyelembe a munkavállaló adatkezelés elleni tiltakozását, valamint megsértette a tájékoztatási kötelezettségét. (szankció: 2 ezer Euró; *HDPA; 12/2022*[15])

A GDPR 15. cikke alapján **az érintett jogosult arra, hogy az adatkezelőtől visszajelzést kapjon arra vonatkozóan, hogy személyes adatainak kezelése folyamatban van-e,** és ha ilyen adatkezelés folyamatban van, jogosult arra, hogy a személyes adatokhoz és az ezen cikkben felsorolt információkhoz hozzáférést kapjon, illetve és ennek során ellenőrizze az adatkezelés körülményeit. Azonban ezen **hozzáférési jog gyakorlása** sem mindig zökkenőmentes. A gelsenkircheni közigazgatási bíróság úgy döntött, hogy a vizsgázók jogosultak a saját vizsgájuk ingyenes másolatára, mivel a szakma gyakorlásához szükséges alapképesítés megszerzése érdekében valamely uniós tagállamban a szakma gyakorlására jogosító szakmai vizsgák elvégzése és az ebből eredő személyes adatok kezelése kizárólag az alapvető szabadságok és a belső piac szempontjából való absztrakt jelentőségük miatt valószínűsíthetően az uniós jog hatálya alá tartozó tevékenységnek minősül, ezért a vizsgázó jogosult arra, hogy a jogi államvizsgájával kapcsolatban ingyenes másolatot kapjon papíron vagy szabványos elektronikus formátumban. (*VG Gelsenkirchen; 20 K 6392/18*[16])

Az érintett azonban vissza is élhet a hozzáférési jogával. Spanyolországban egy érintett egy olyan egyetemet panaszolt be a hatóságnál, amely esetében többek között volt alkalmazott, fegyelmi eljárás alá vont alkalmazott, egyetemi hallgató, mesterszakos hallgató, kurzusasszisztens, közigazgatási eljárásokban érdekelt fél, peres fél stb. A hozzáférési kérelmét az egyetem részlegesen teljesítette és kérte adja meg, milyen további információkra van szüksége. Az érintett válaszul megismételte eredeti kérelmét, mire az egyetem joggal való visszaélésnek minősítette az esetet. A hatóság megállapította, az érintett rosszhiszeműen, visszaélésszerűen gyakorolta jogait. *(AEPD; E/00739/2021*[17])

Számtalan büntetést eredményezett a **COVID-19 világjárvány** köszönhetően annak, hogy a vészhelyzetben az adatvédelmi jogszabályok ugyanúgy alkalmazandók voltak, mint a vészhelyzet előtt és után.

Az olasz adatvédelmi hatóság 200 ezer eurós bírságot szabott ki egy milánói egyetemre, mert a járvány idején a hallgatók online vizsgájának ellenőrzésére nem megfelelő felügyeleti rendszereket használt („Respondus" szoftver), és ha a hallgatók nem járultak hozzá ezen rendszer használatához, nem tehettek online vizsgát. A szoftver videofelvételt rögzített a vizsgáról, valamint véletlenszerű időközönként pillanatképeket készített webkamerával, illetve felvételeket a hallgató képernyőjéről, azonosítva és egy zászlóval – további vizsgálat céljából – megjelölve azokat a pillanatokat, amikor szokatlan és/vagy gyanús viselkedést észlelt. A hatóság szerint az ilyen szoftverek használata – tekintettel a járványra – elfogadható, de biztosítani kell az adatvédelmi elvek betartását, különös tekintettel a különleges adatkategóriák kezelésével, a profilalkotással és a nemzetközi adattovábbítással kapcsolatos lehetséges kockázatokra. Az egyetem viszont – többek között – nem tájékoztatta a vizsgázókat az Egyesült Államokba történő adattovábbításról (ahol a szoftvert biztosító vállalat székhelye található), és nem adott magyarázatot a szoftver profilalkotásának logikájáról sem. A hozzájárulás nem volt hivatkozható jogalap, tekintettel a hallgatók és az egyetem közötti hatalmi egyensúlyhiányra, illetve az adatkezelés nem felelt meg az adattakarékosság és a korlátozott tárolhatóság elvének sem. Ezen kívül jogellenesen továbbították a hallgatók

15 https://www.dpa.gr/sites/default/files/2022-03/12_2022anonym_0.pdf
16 https://www.justiz.nrw.de/nrwe/ovgs/vg_gelsenkirchen/j2020/20_K_6392_18_Urteil_20200427.html
17 https://www.aepd.es/es/documento/e-00739-2021.pdf

személyes, köztük a biometrikus adatait az Egyesült Államokba *(Garante per la protezione dei dati personal; 9703988[18])*. Az izlandi hatóság Zoom-vizsga miatt részesített megrovásban egyetemet *(Persónuvernd; 2020112830[19])*, azonban voltak olyan felsőoktatási intézmények is, amelyek az illetékes hatóságok szerint megfelelő intézkedéseket hoztak az online vizsgák adatvédelmi megfelelősége érdekében *(Rb. Amsterdam, C/13/684665/KG ZA 20-481[20], Gerechtshof Amsterdam, 200.280.852/01[21], Datatilsynet, 2020-432-0034[22])*.

Összegzés

A köznevelési és a felsőoktatási intézmények ugyanúgy felelősek az adatkezeléseikért, mint bármely más adatkezelő, EGT-tagállami szinten pedig nem a felelősség megállapításában, hanem a büntetés mértékében van különbség. Ezen felelősség mértékét a világjárvány sem csökkentette.

Az elmúlt öt évben az egyik leggyakoribb probléma a megfelelő biztonsági intézkedések hiánya miatt történt adatvédelmi incidensek, melyek gyakran számos érintett jogaira és szabadságaira jelentettek kockázatot, azonban számos más okra visszavezethető esetben is szankcionáltak a hatóságok. A jogsértések között számos olyan eset van, amelyek az érintetti jogok figyelembe nem vételére, az alapelvek sérelmére, illetve a nem megfelelő adatkezelési konstrukciókra vezethetők vissza, akár rendszer szintű hiba következményeként (például megfelelő intézkedések hiánya, rossz gyakorlat, adatvédelmi kultúra hiányosságai), akár emberi hiba miatt. Hazánkban is születtek elmarasztaló határozatok, az esetek számának csökkentésére pedig csak az adatvédelmi tudatosság növelése, az adatkezelési folyamatok megfelelő szabályozása, az új technológiák értő alkalmazása, az érintetti sajátosságok figyelembevétele, a biztonság szem előtt tartása és a jó gyakorlatok kialakítása lehet a megoldás.

18 https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9703988

19 https://www.personuvernd.is/urlausnir/rafraen-yfirseta-profa-haskolans-i-reykjavik-ekki-i-samraemi-vid-log

20 https://uitspraken.rechtspraak.nl/#!/details?id=ECLI:NL:RBAMS:2020:2917

21 https://uitspraken.rechtspraak.nl/#!/details?id=ECLI:NL:GHAMS:2021:1560

22 https://www.datatilsynet.dk/afgoerelser/afgoerelser/2021/jan/universitets-brug-af-tilsynsprogram-ved-online-eksamen

CC BY

DOI: 10.31915/NWS.2023.4

Digitális dokumentumok gyűjteménykezelési gyakorlatának támogatása a digitális tartalmak számossága, mérete és féleségeik vizsgálatával

Where to write the inventory number? Practice of handling born digital documents in recent collection management systems.

Simon András
Monguz Információtechnológiai Kft Szeged
andras.simon@qulto.eu

Absztrakt

Folyóirat cikkek, tanulmányok, levéltári iratok, magánlevelek, fényképek és a különféle kisnyomtatványok sokasága keletkezik ma már digitálisan. Ezek a dokumentumok különösen számottevő részét teszik ki az egyes közösségek saját kulturális örökségének, melyet a társadalom részéről jelenleg nagy figyelmet kapó terület, a helytörténet kezel. A digitális objektumok látszólag kevés gondoskodást igényelnek, főként a könnyű tárolhatóság, és az olcsó, veszteségmentes sokszorosítás lehetősége miatt, ugyanakkor a közgyűjteményi állományok legsebezhetőbb, legveszélyeztetettebb részét alkotják. Tartós, intakt, hiteles és dokumentált megőrzésük nagy felelősséggel járó munka, amely nagy terhet ró a emlékezetintézményekre. A gyűjteménykezelés korszerű szabályozásához, és a megfelelő gyakorlati megoldások kialakításához elengedhetetlen megfelelő információk szerzése ezen dokumentumok mennyiségéről, elterjedtségéről, sokféleségéről. A tanulmány szerzője a Monguz Kft munkatársaként négy éve, a doktori disszertációjához folytatott kutatásai keretében vizsgálja a cég múzeumi és könyvtári ügyfeleinek adatbázisait, egyebek mellett az őrzött digitális tartalmak számossága mérete és féleségeik tekintetében. A kutatás keretében 2020 őszén lefolytatott egyik vizsgálatot végzi el most el újra, megpróbálva megállapítani a közgyűjteményi tartalomkezelés legfontosabb trendjeit. A tanulmányban ezen vizsgálat eredményéről készül beszámolni.
Kulcsszavak: *közgyűjteményi informatika*, *tartalomkezelés*, *digitálisan keletkezett dokumentum*, *tartalomelemzés*, *fájlformátum*

Abstract

Most of the journal articles, studies, archival documents, private letters, photographs and various small print documents are being produced digitally in the recent days. These documents make up a particularly significant part of the cultural heritage of each community, and local history, managing these types of objects, is one of the areas currently receiving considerable attention from society. Digital objects seem to require little care, mainly because of their easy storage and the possibility of cheap, lossless reproduction, but they are the most vulnerable and endangered objects of public collections. Storing them for a long time unharmed, authentic and documenting the interventions is a major responsibility and a big task for heritage institutions. To ensure modern collection management and to develop appropriate practices, it is essential to obtain adequate information on the quantity and diversity of the digital objects and also information is necessary, how widely these documents are used. The author of the study, as the employee of Monguz Ltd. in 2020 examined the databases of the company's museum and library clients, including the size and diversity of the digital contents of the databases. This examination was the part of his doctoral dissertation researches. The study is now being carried out again after almost

three years, trying to establish key trends in content management in public collections. The presentation will report on the results of this investigation.

A közgyűjteményekben tárolt digitális objektumok kezelése és tárolása növekvő jelentőségű feladatot jelent az intézmények számára. A digitális tartalmak száma egyre nagyobb, és méretük is növekszik. Ennek oka, hogy egyrészt egyre szélesebb a digitálisan keletkezett dokumentumok köre, másrészt melléjük sorakoznak a megőrzési célból digitalizált analóg dokumentumok digitalizált példányai. Különösen nagy szerep hárul a közgyűjteményi munkafolyamatok informatikai igényeinek kiszolgálásakor az intézményben használt integrált rendszerekre, melyek tervezésekor és fejlesztésekor az alábbi problémák megoldására kell felkészülni:

- A digitális és analóg dokumentumokat együttesen kell kezelni, egyelőre ugyanis - főként a múzeumokban - még nincsenek önálló gyűjtemények kialakítva a digitálisan keletkezett, vagy csak digitálisan őrzött dokumentumok számára,
- A tartalomszolgáltatás fejlesztése, összhangban a korszerű felhasználói (fogyasztói) igényekkel, a szerzői jogi szabályozásokkal, a személyiségi jogi elvárásokkal és az emlékezet intézmények saját hosszútávú érdekeivel,
- A mennyiségi növekedés kezelése. Ezek az objektumok óriási mennyiségben keletkeznek, és az analóg tartalmak retrospektív digitalizálása is egyre nagyobb ütemben halad. Az átlagos fájlméret bizonyos mértékben még ma is növekszik, a válogatás feladata tehát egyre kevésbé megkerülhető,
- A formai és tartalmi feltárás egyre költségesebb, a mennyiségi növekedés és az egyre nagyobb kvalifikációt igénylő emberi munkaerő miatt.
- A tartós megőrzés, (műszaki használhatóság, szolgáltathatóság eredetiség és hitelesség biztosítása) költségei is a kereslet folyamatos bővülése miatt állandóan az inflációt meghaladó ütemben növekszenek.

A korszerű integrált gyűjteménykezelő rendszerekben tehát többféle tartalom kezelésére kell felkészülni. Az őrzött dokumentumok egyrészt lehetnek:

- Analóg,
- szolgáltatási célból digitalizált,
- megőrzési célból digitalizált, illetve
- digitálisan keletkezett dokumentumok.

Másrészt az őrzés szempontjából el kell különíteni a

- Sorozatgyártással tömegesen keletkezett,
- sorozatgyártással kis mennyiségben keletkezett, illetve az
- egyedi dokumentumokat.

Az anyag jellege szerint megkülönböztethető, de könyvtári, múzeumi és levéltári közgyűjteményi ágakban egyaránt kezelt dokumentumféleségek az alábbiak lehetnek:

- Könyv,
- periodika,
- különlenyomat,
- térkép,
- audiovizuális felvétel,
- levéltári irat,
- kézirat,

CC BY

- kisnyomtatvány (pl. gyászjelentés, rendezvény meghívó),
- képeslap,
- fotó,
- történeti dokumentum (bármilyen történeti értékű szöveges anyag),
- történeti tárgy,
- vizuális dokumentum.

A történeti tárgy kivételével ezek mindegyike keletkezhet ma már digitálisan is.

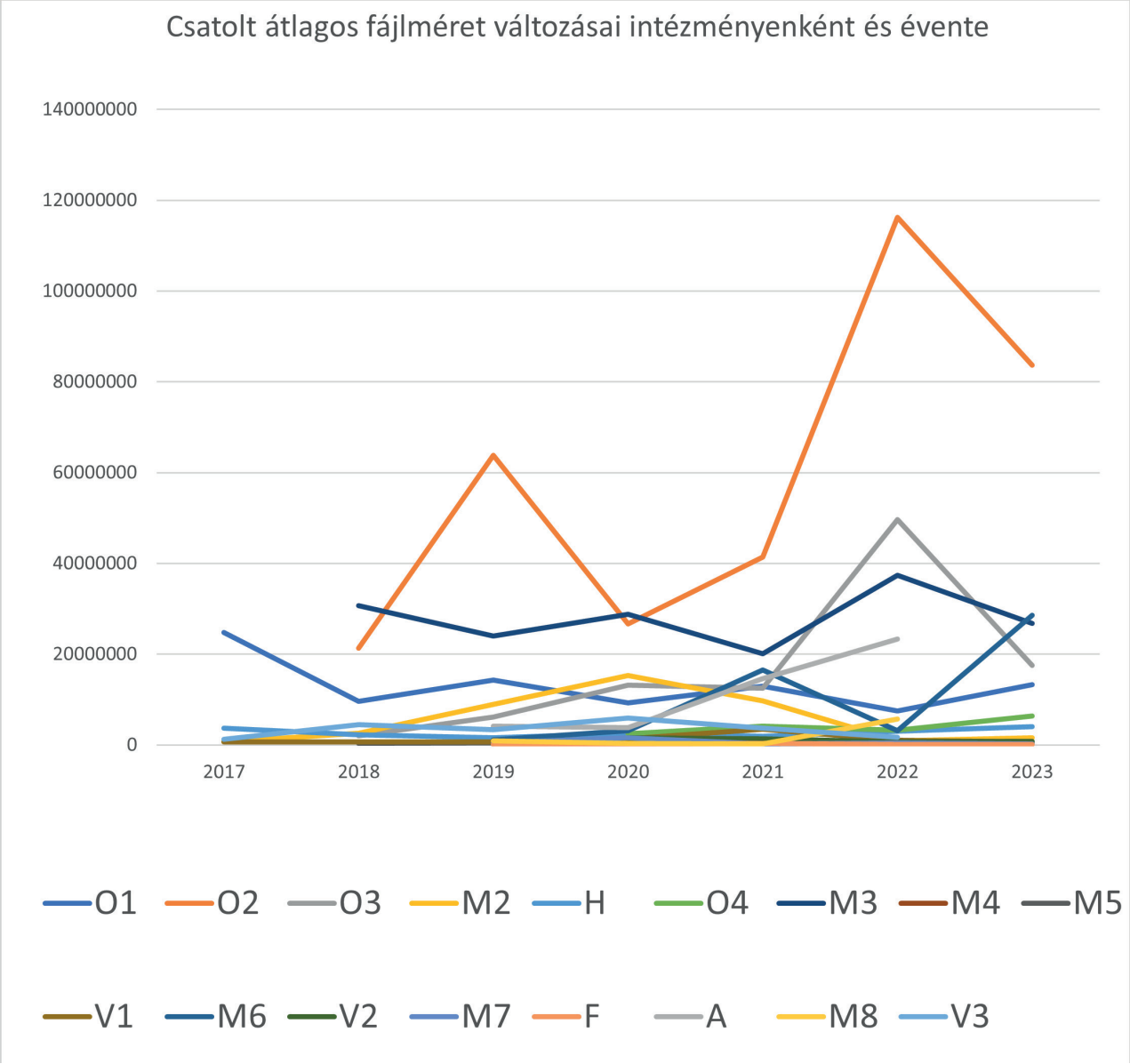
Tekintettel a közgyűjteményekben kezelt digitális dokumentumok hasonlóságára, a közgyűjteményekben hasonló megoldási stratégiák vannak kialakulóban, és emiatt a tartalomkezelést az egyes gyűjteményi ágak esetében párhuzamosan kell vizsgálni. A hasonló elvárásoknak megfelelő hasonló megoldások jelentőségét növeli a közösségek kulturális örökségének mind hangsúlyosabb jelenléte a - mindinkább digitálisan létező - helytörténeti gyűjteményekben. Ezek a tartalmak, melyek korábban nehezen voltak elérhetők, most könnyen közzé tehetők, és a kényelmes és bárhonnan bármikor elérhető felületek révén egyre növekvő közönség számára váltak hozzáférhetővé. A digitalizációnak részint okaként, részint következményeként a kulturális örökség fogalomköréhez tartozó, a helyi közösség saját története szempontjából fontos dokumentumok jelentősége egyre nagyobb.

A közgyűjtemények közös történelmi küldetése annak megakadályozása, hogy az emberiség egy évtizede jelentős részben már digitálisan keletkező kulturális örökségét ne nyelje el a „digital black hole", a digitális fekete lyuk. Ezek az állományok az analóg tartalmaknál sokkal veszélyeztetettebbek, egy részük megfelelő gondosság hiányában örökre elveszhet. Hallatlanul nagy a sérülés, az adatvesztés kockázata. A gyűjtésre, illetve az őrzésre kialakult megoldások, hosszú ideje bevált gyakorlatok a dokumentumok újszerű volta miatt még nem állnak rendelkezésre. Ráadásul a digitális tartalmak hosszútávú, dokumentált és hiteles formában való megőrzése, illetve tovább hasznosíthatóságának hosszú időtartamra való biztosítása igen költséges.[1]

Ezen tartalmak katalógusokban való jelenlétének vizsgálatakor természetesen nem tekinthetünk el attól, mind szakmai, mind fenntartói oldalról nagy az érdeklődés aziránt, hogy mekkora a kihasználtságuk. Fontos rámutatni arra, hogy a digitális tartalmak interneten keresztül történő használata nehezebben mérhető, mint a hagyományos könyvtári kölcsönzés. Ennek nyomon követése a fizikai kölcsönzés esetében alapvető lehetőségként adódik az integrált könyvtári rendszer működéséből. A könyvtári ügyviteli munkát támogató integrált könyvtári rendszer ugyanis minden kölcsönzési eseményt, annak valamennyi adatát tranzakciós rekordokban őrzi, mint könyvtár és olvasó közötti gazdasági eseményt naplózó elektronikus bizonylatot. Bármilyen webes keresőfelületet érintő megkeresés ezzel szemben természetesen csak a rengeteg műszaki információt tartalmazó naplófájljába kerül beírásra. Mivel ez a fájl óriási, emiatt csak rövid ideig van őrizve. A tartós őrzésre érdemes hasznos statisztikai adatokat ebből a fájlból kell kibányászni, addig amíg a fájl egyáltalán megvan. Mindez inkább a szabadpolcos állomány helyben való használatának méréséhez hasonlítható, amikor a könyvtárosok az olvasótermi asztalokon hagyott könyveket a polcra visszahelyezve, erről az eseményről statisztikát vezetnek. A mérés alapja a fizikai kölcsönzési eseménnyel ellentétben tehát nem egy alapfunkcióként tartósan őrzött tranzakciós rekord, hanem a rendszer naplófájljaiba „mellékesen" rögzített adat.

1 Weber, H. Chrobak, Lennart: Legal Implications of Digital Heritagization. Implications juridiques de la patrimonialisation numérique. = RESET [Online], 6 | 2017, Patrimoine et patrimonialisation numériques. http://journals.openedition.org/reset/826 ; DOI : 10.4000/reset.826

Mindez alátámasztja azt, hogy nagyon fontos minél pontosabb ismeretekkel rendelkeznünk a digitális tartalmakról féleség, méret és kor tekintetében, lehetőleg felvázolva az elmúlt évek trendjeit is. Ezen szándék vezetett az elmúlt évek során végzett kutatásaimban, melyeket munkaadóm a Monguz Információtechnológiai Kft ügyfeleinek adatbázisaiban végeztem. Az adatbázisokban megjelenő csatolt fájlok számosságára, féleségére és méretére vonatkozó kutatásaimat 2020 őszén végeztem el, ezeket ismételtem meg 2023 tavaszán ugyanazokra az adatbázisokra ugyanazokat a lekérdezéseket elindítva. A kutatás körülményeiről és módszertanáról hely hiányában itt nincs módom beszámolni, de korábbi publikációkban, ezekről részletesen írok.[2]



1. ábra Az átlagos fájlméret változásai

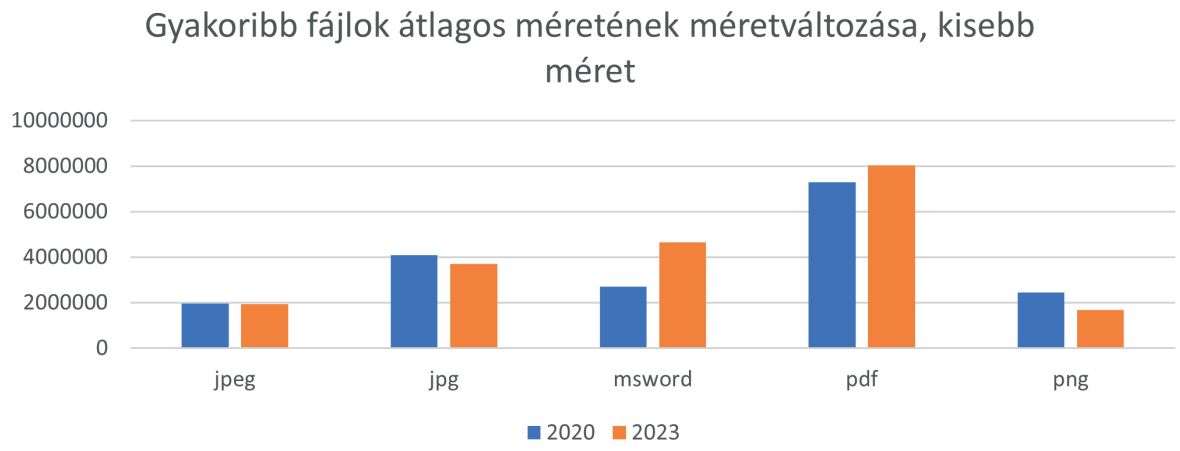
Az első ábrán a digitális katalógus rekordokhoz csatolt fájlok átlagos méretének változásait szemléltetem. A vízszintes tengelyen a fájl a katalógus rekordhoz való csatolásának (nem a létrejöttének) az éve látható. A függőleges tengelyen a fájl mérete olvasható byte-ban kifejezve. Különféle színekkel az egyes vizsgált intézményeket jelöltem. A rövidítések

2 Simon András: Integrált Könyvtári Rendszerben tárolt tranzakciós rekordok felhasználása a könyvtárhasználat statisztikai elemzésére= TMT 66. évf. 2019. 12. sz.

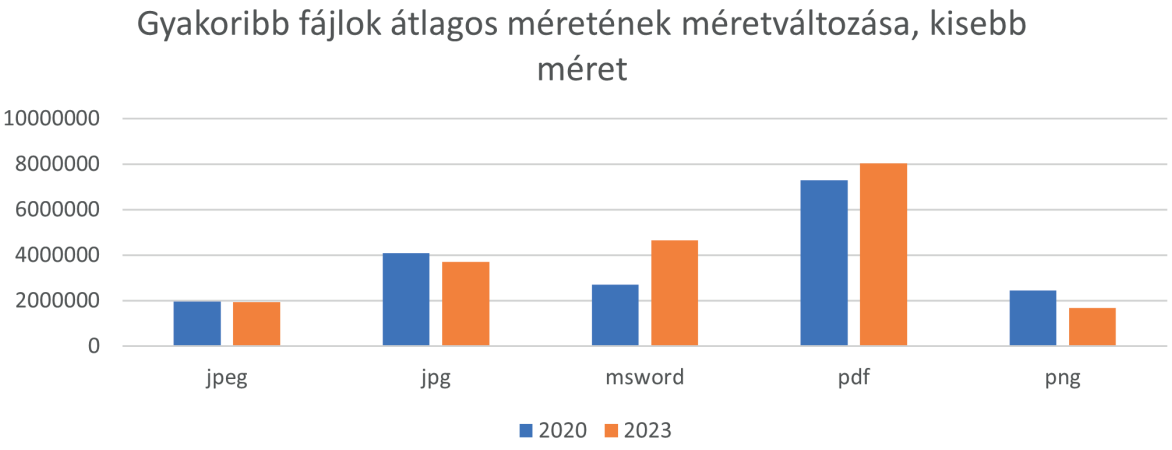
CC BY

feloldása: „*O*” (országos intézmény 4 db), „*M*” (Megyei hatókörű városi intézmény 7 db), „*H*” (Helytörténeti intézmény), „*V*” Városi intézmény 3 db), „*F*” (Felsőoktatási intézmény 1 db), „*A*” (archívum 1db). Az ábra jól mutatja, hogy átlagos növekedés még enyhe mértékben sem feltétlenül állapítható meg három év távlatából. A két kiugró értéket mutató országos intézmény egyébként nagy mennyiségben gyűjt és kezel, audiovizuális anyagokat.

Az 1. számú ábrán látható diagramon felsorolt intézmények adatai láthatók a további ábrákon is.

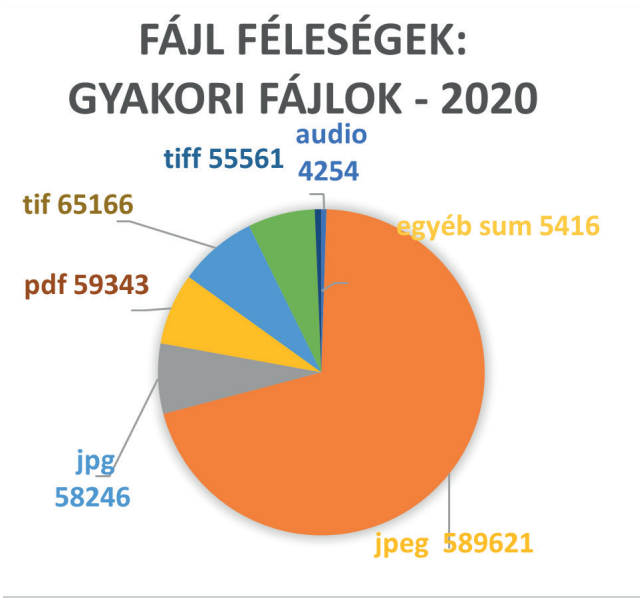


2. ábra Fájlméretek méretváltozásai, (kisebb méretű fájlok)



3. ábra Fájlméretek méretváltozásai, (nagyobb méretű fájlok)

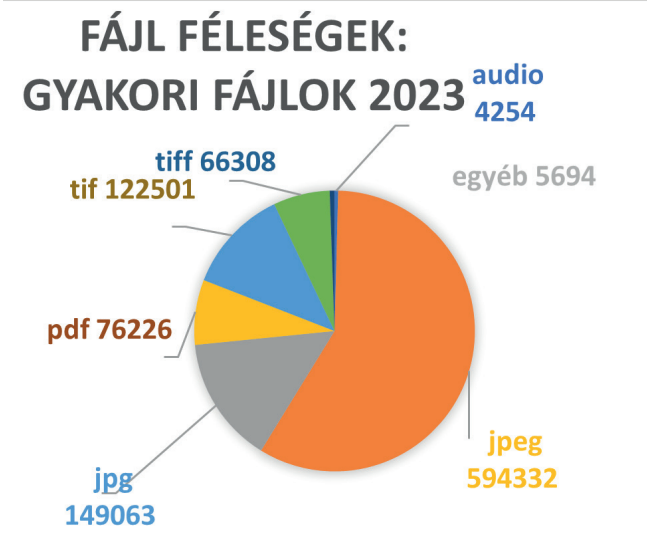
A 2. és 3. ábrán a leggyakoribb fájltípusok átlagos méretváltozását mutatom be. Az áttekinthetőség kedvéért bizonyos fokú egységesítést végeztem a fájlnevekben, összevontam kezelem pl. a tif és tiff formátumot. A nagyobb és kisebb méretű fájlokat a megfelelő szemléltetés céljából különválasztottam. Az 1. számú ábrához hasonlóan itt is szignifikánsan kimutatható átlagos méretnövekedésről nem beszélhetünk.



4. ábra Fájl féleségek 2020 – gyakrabban előforduló fájlok

A 4., 5. 6. és a 7. ábrákon látható kördiagramokon a fájl féleségeket, a különféle fájltípusok előfordulásait vizsgálom. Külön választottam a gyakrabban és a ritkábban előforduló fájlokat, hogy a diagramon a féleségeket, a két adott évben egymáshoz viszonyított méretbeli arányukat jobban szemléltethessem. A gyakoribb előfordulású fájlokat bemutató diagram „egyéb" kategóriája van kibontva a ritkább fájlokat bemutató diagramokon. Itt a fájlnév kiterjesztéseket a diagram készítésekor nem módosítottam, úgy láthatók itt, ahogyan a fájl csatolásakor az adatbázisba a program regisztrált ezeket.

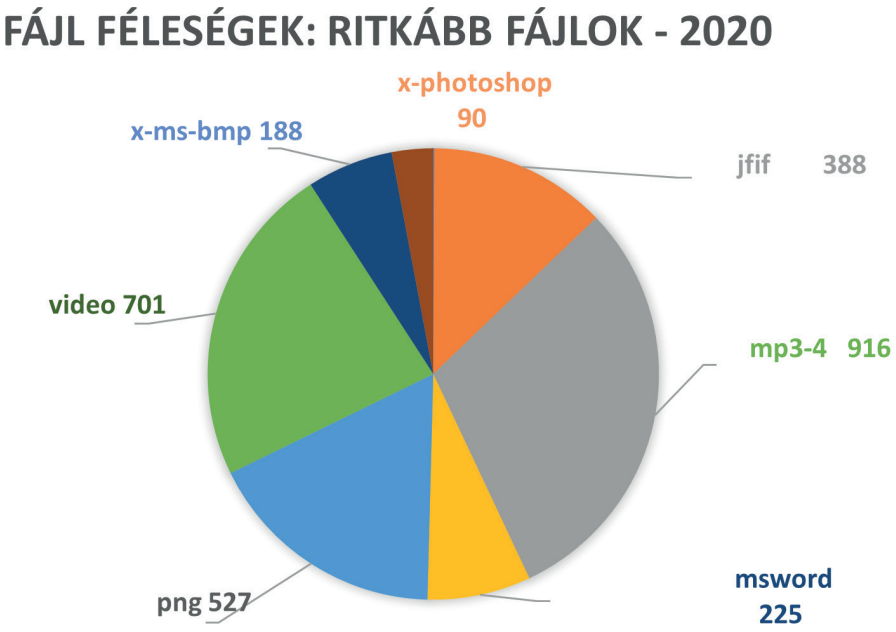
A 4. és 5. ábrán jól látható, hogy a jpeg és jpg formátumok együtt háromnegyed körüli arányt képviselnek.



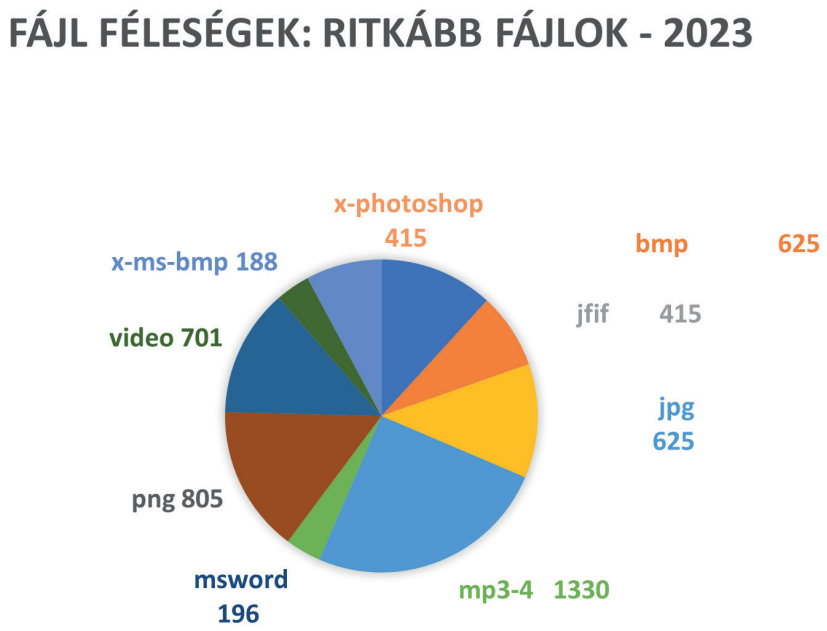
5. ábra Fájl féleségek 2023 – gyakrabban előforduló fájlok

Darabszámra a pdf-et leszámítva gyakoribb előfordulásúnak csak a képfájlok tekinthetők. A csatolt fájlok túlnyomó része kép, alapvetően múzeumi számítógépes leltárkönyvi tételekhez csatolt digitális műtárgyfotó. A fájlnév kiterjesztések mellett a fájlok a vizsgálat ideje alatt az egyes intézményekben megállapított darabszámának összege látható. Az összes intézmény összesített fájl darabszámára általánosságban kijelenthető, hogy három év alatt is jelentős a növekedés. A négy diagramon jól látható, hogy jelentős elmozdulás három év alatt nem volt észlelhető, a ritkábban előforduló fájlok között tapasztalható a 2023-as év esetében a három évvel korábbi helyzethez képest kicsit nagyobb sokféleség, de ez csak a ritkábban előforduló fájlok viszonylagosan nagyobb számával magyarázható. Így pl. a bmp formátum a 2023-ra 625 előfordulással már felkerült a diagramra. Egészen új fájltípusok az elmúlt három évben még kisebb számban sem jelentek meg. Egyes fájlok esetében számszerű csökkenés is észlelhető. Ha intézményekre lebontva vizsgálnánk a fájlok összesített darabszámát, azt látnánk, hogy még inkább tetten érhető az egyes intézményeknél bizonyos fájltípusokban a darabszám csökkenése, ez a csatolt tartalmak rendszeres, minőségi szempontokat követő cseréjével magyarázható.

CC BY



6. ábra Fájl féleségek 2020 - ritkábban előforduló fájlok



7. ábra Fájl féleségek 2023 - ritkábban előforduló fájlok

Három év nem nagy idő, ez alatt komoly trendváltozást láthatóan nem tapasztalhatunk. Az adatok zöme múzeumi adatbázisokból származik, könyvtári katalógusokban lévő bibliográfiai rekordokhoz kevés tartalom lett egyelőre még csak csatolva. A felsőoktatási és szakkönyvtárak ugyanis sok esetben önálló a könyvtári katalógustól független intézményi repozitóriumot vezetnek. Az intézmények egy részében ma is tetten érhető a kampányszerű digitalizálás jelensége, a csatolt fájlok darabszáma, az egyes években lényegesen különböző mértékben növekszik. Mindezek miatt a kutatást mindenképpen érdemes lesz megismételni néhány évvel később, amennyire lehet figyelmet fordítva a párhuzamosan épített repozitóriumokra is.

 DOI: 10.31915/NWS.2023.5

Az Annif gépi tárgyszavazó rendszer magyarországi adaptációjának feltételei és lehetőségei

Requirements and possibilities for the adaptation in Hungary of the Annif automated subject indexing tool

Bódog András
Országos Széchényi Könyvtár, Digitális Bölcsészeti Központ,
Közgyűjteményi Szabványosítási Iroda
bodog.andras@oszk.hu

Abstract

A recurring annual international event for library information science professionals, the *Semantic Web in Libraries (SWIB)* conference is a showcase for the latest innovations in linked data and semantic web applications. A regular feature of this conference is the tutorial on the *Annif automated subject indexing tool*, to which the National Széchényi Library's Office for GLAM Standards has had the opportunity to delegate a member.

Developed by the National Library of Finland, the open source Annif has quickly become one of the flagship for artificial intelligence solutions in libraries. It is already used on routine basis in Finland, more and more libraries are adapting the system, primary in Europe. The secret of its success is both the open source nature of the Annif and the flexibility to integrate with existing components, such as various vocabularies and natural language processing (NLP) algorithms and other AI developments.

The Annif therefore is a very useful tool that can significantly facilitate and even revolutionize content discovery, but the key to successful adaptation is the conscious preparation for this task by the library system. It is not only to provide the technical requirements and standardized indexing and cataloging, but also a paradigm shift is needed to enable the current Hungarian library system to adapt such an innovative solution, which brings a new approach to librarians everyday professional life in Hungary. In my paper, summarizing the experiences of the SWIB workshop, I explore these issues and summarize the basic features of the Annif.

Kulcsszavak: tezaurusz, gépi tárgyszavazás, helyi viszonyokra alkalmazás, Köztaurusz, nemzeti könyvtár
Keywords: thesauri, automated subject indexing, local adaptation, national library

Bevezető

Napjainkban reneszánszukat élik a különböző mesterséges intelligencián alapuló megoldások. Könyvtári közegben a számos felhasználási terület egyike a gépi tárgyszavazás. Könnyen megérthető, hogy miért: egyrészt az elektronikus dokumentumok olyan mértékben gyarapodnak, hogy az azokat feldolgozó intézmények egyszerűen már nem bírnak vele emberi erőforrás tekintetében lépést tartani. Érv még a konzisztensebb tartalmi feltárás biztosítása, továbbá a gépi tárgyszavazással és osztályozással erőforrás és idő takarítható

CC BY

meg, különösen abban a kontextusban, hogy a túlterheltség és a szakemberhiány miatt előfordulhat, hogy már eleve nem jut kapacitás tartalmi feltárásra. A Finn Nemzeti Könyvtár mindezeket szem előtt tartva fejlesztette ki az *Annif* gépi tárgyszavazó rendszert.[1]

Az Annif ismertetése

Az Annif egy nyílt forráskódú, Apache 2.0 licenccel közzétett,[2] moduláris felépítésű, szótárfüggetlen gépi tárgyszavazó és osztályozó rendszer közgyűjtemények számára. A rendszer többféle természetes nyelvű szövegfeldolgozó (NLP) és gépi tanuló modult kombinál. Többnyelvű (finn, svéd, angol), emellett a tárgyszavazáshoz többféle szótárformátum alkalmazható, az egyszerű szöveges TSV-től,[3] a SKOS-ig[4].[5] A fejlesztés során a nyíltforráskód-alapú fejlesztést és közzétételt, a modulszerű felépítést, a webes felhasználói felületet és a REST API révén a más rendszerekbe való integrálhatóságot tartották szem előtt.[6] Az Annifot elsőként a Jyväskylä Egyetem *JYX* repozitóriumában használták félautomata tárgyszavazásra. A szakdolgozataikat feltöltő hallgatóknak a dolgozat szövege alapján ajánl a rendszer tárgyszavakat, amelyeket a hallgatói kiválasztást követően egy könyvtáros validál.[7] A *Finto AI* nevű webalkalmazás 2020-tól üzemel. Bemásolt, vagy URL-alapján weboldalról kinyert finn, svéd vagy angol nyelvű szöveghez ajánl tárgyszavakat. Azóta egyre több intézmény adaptálta az Annif-ot, többek között a finn közmédia (Yle), a svéd (Kungliga bibliotek) és a német nemzeti könyvtár (Deutsche Nationalbibliothek), illetve a ZBW – Lebniz Közgazdasági Információs Központ (Leibniz-Informationszentrum Wirtschaft).[8]

Az Annif moduljai

A *konfigurációs modul* szolgál az Annif által ellátott feladat vagy feladatok (projektek) paraméterezésére, mivel minden egyes projekt a többi modul egyedi beállítását igényli.[9] Az Annif alapvető követelménye egy a tárgyszavak bázisát adó szótár használata, amely egy egyszerű, URI-kkal kiegészített[10] tárgyszójegyzéktől kezdve egy komplex tezauruszig bezárólag bármi lehet. Finnországban szótárként a finn általános ontológiát[11] (YSO) alkalmazzák, amely az 1980-as évektől épített finn általános tezaurusz (YSA) az ISO 25964 tezauruszszabvány előírásait követő, SKOS-formátumba átalakított, többnyelvű

---

1 Suominen, Osma [et al.] *Annif and Finto AI: Developing and Implementing Automated Subject Indexing.* = JLIS.It, vol. 13. (2022) no. 1, p. 266.

2 GitHub. *NatlibFi / Annif* https://github.com/NatLibFi/Annif/; PyPI. annif 0.61.0. https://pypi.org/project/annif/; Quay.io. *natlibfi / annif* Docker-képfájl https://quay.io/repository/natlibfi/annif (2023.05.30)

3 TSV, *Tab Separated Values* (tabulátorral elválasztott értékek). A Kongresszusi Könyvtár sztenderd szöveges formátuma. https://www.loc.gov/preservation/digital/formats/fdd/fdd000533.shtml (2023.05.30.)

4 *Tudásszervezési rendszerek reprezentációjára alkalmazott W3C sztenderd.* W3C weboldala. *SKOS Simple Knowledge Organization System – Home Page* https://www.w3.org/2004/02/skos/ (2023.05.30.)

5 *Annif weboldal* https://annif.org/ (2023.05.30.)

6 Suominen [et al.] i.m. p. 266–267.

7 Suominen [et al.] i.m. p. 275.

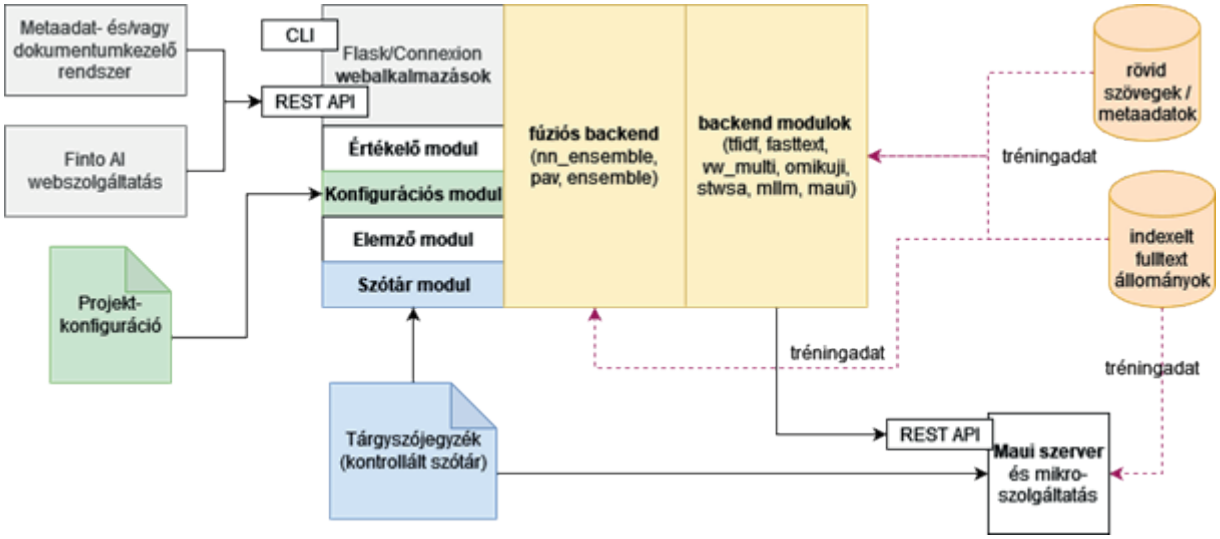
8 *Annif* weboldal

9 Suominen, Osma. *Annif: DIY automated subject indexing using multiple algorithms.* = LIBER Quarterly, vol. 29. (2019) no. 1., p. 6–7.

10 A legegyszerűbb TSV-formátumban is követelmény, hogy az egyes deszkriptorok rendelkezzenek egységes forrásazonosítókkal (URI). Lásd: Github. NatLibFi / Annif. *Subject vocabulary formats*

11 *Az ontológiát az ISO 25964-2:2013 tezauruszszabvány mint a fogalmi rendszerek explicit formális leírását definiálja. A finn megoldás esetében gyakorlatilag SKOS-formátumban, RDF-tripletekkel leírt tezauruszról beszélünk, amely magában foglal minden szabványos tezauruszrelációt.* A fogalommeghatározást lásd = ISO 25964-2:2013 p. 10.

kapcsoltadat-alapú továbbfejlesztése. A finn szemantikus webkörnyezet infrastruktúráját megalapozó FinnONTO kutatási projektre építve jött létre a több hasonló kialakítású szótárat is szolgáltató *Finto* nevű központi szótárszolgáltatás. Az YSO emellett, általános felső ontológiaként, a gerincét képezi a *KOKO* nevű összetett ontológiának,[12] amely utóbbi 13 további szaktezaurusszal alkot egységes hierarchiát. A Finto-szótárak megfelelő, de felhasználóbarát megjelenítését és használatát a szintén saját fejlesztésű és nyílt forráskódú *Skosmos*[13] szoftver teszi lehetővé.[14]



1. ábra: Annif-modulok
(Suominen et al.: Annif and Finto AI. Developing and Implementing Automated Subject Indexing ábrája alapján, JLIS vol. 13. no. 1. p. 267.)

A tárgyszavazni kívánt dokumentumok szövegét előzetesen fel kell készíteni a gépi tárgyszavazó algoritmusok futtatására. Erre szolgál a szövegfeldolgozási módszereket alkalmazó *elemző modul*, amely a szöveget szótövekre bontja (stemmelés) vagy szótári formára hozza (lemmatizálás).[15] A tárgyszavazó algoritmusokat *backendként* implementálták. Kétféle megközelítésen alapulnak: lexikai megközelítésen, amely a szövegszavakat párosítja a szótári alakkal, illetve asszociatív megközelítésen, amely statisztikai vagy gépi tanulási módszerekkel, hatalmas méretű[16] manuálisan tárgyszavazott tanítókorpuszt hasznosítva azonosítja a szöveg által érintett tárgyköröket és az azokkal párosított tárgyszavakat. Még hatékonyabb eredményt ad a két módszer kombinálása a fúziós megközelítés révén. Az *értékelő modul* az algoritmusok eredményét értékeli, többek között az F1-érték, vagyis a pontosság (precise) és a teljesség (recall) arányának számításával. Az értékeléshez és a tanításhoz a korpuszt az ún. gold standard adja, egy gondosan válogatott, megfelelően tárgyszavazott, teljes szövegű, rövid szöveges vagy csak metaadatokat tartalmazó dokumentumhalmaz. E több százezer

12 Lappalainen, Mikko [et al.]. Reuse of library thesaurus data as ontologies for the public sector. IFLA2014, Lyon, (2014), p. 1–8.

13 Skosmos weboldal https://skosmos.org/; GitHub. NatLibFi / Skosmos https://github.com/NatLibFi/Skosmos (2023.06.01.)

14 Finto – Finnish Thesaurus and Ontology Service. *KOKO Ontology.*

15 *Az Annifban jelenleg alkalmazott elemző algoritmusok részletet összehasonlítását lásd:* Suominen, Osma – Koskenniemi, Ikka. Annif Analyzer Shootout: Comparing text lemmatization methods for automated subject indexing. = Code4Lib Journal, vol. 54. (2022)

16 *A SWIB 2022 konferencia Annif-workshopján Osma Suominen a szótár lexikai egységeinek százszorosát tartotta optimálisnak a korpusz tokenszám szerinti méretét illetően* = SWIB 2022 online konferencia. Introduction to the Annif automated indexing tool. 2022.11.30. https://swib.org/swib22/programme.html#day3 (2023.06.02.)

CC BY

dokumentumot bontják tovább *tréning-*, *validációs* és *tesztadatok* halmazaira. Az Annif használata rendszerekbe integrált módon *alkalmazásprogramozási interfészen*, vagy különálló *Annif-szolgáltatásokban* (pl. a Finto AI) valósul meg. Az elektronikus dokumentumokból kinyert szöveget elemzi a rendszer, majd az algoritmusok révén a konfigurált szótárból vett tárgyszavakat ajánlja a feldolgozó könyvtárosnak, tehát a kontroll végig a szakember kezében van.[17]

Annif Magyarországon? Feltételek és lehetőségek

A hazai adaptáció alapvető feltétele, hogy legyen egy megfelelő kontrollált szótár a tárgyszavak számára, illetve álljanak rendelkezésre magyar nyelvű természetes nyelvfeldolgozó algoritmusok. A szótárak tekintetében olyat kell választani, amelynek használata kellően elterjedt ahhoz, hogy elérhető legyen egy jelentős mértékű „gold standard" állomány, jóllehet arra is akad precedens, hogy egyes szakkönyvtárak a nulláról kezdve adaptálnak egy szótárat, és vezetnek be erre alapozva gépi tárgyszavazó rendszert. Kiváló példa erre a Társadalomtudományi Kutatóközpont Könyvtár és Dokumentációs Központja, akik a SZTAKI fejlesztőinek támogatásával, szintén a *Skosmos*t alkalmazó európai társadalomtudományi tezauruszt (CESSDA ELSST)[18] fordították magyarra, és használták fel eddig feldolgozatlan gyűjteményük tartalmi feltárására.[19] A nemzeti könyvtár vonatkozásában az OSZK által is használt *Köztaurusz*[20] általános tezaurusz jöhet számításba. Ez a szótár funkcióját tekintve párhuzamba is állítható a finn YSO-val, ám alapvető különbség, hogy Magyarországon sajnos koránt sincs olyan egységes tárgyszavazási gyakorlat, mint Finnországban. A Könyvtári Intézet 2016-os felmérése szerint kiugróan magas a kontrollált szótárakat nem alkalmazó szabad tárgyszavazás mértéke. A Köztaurusz használata a nemzeti könyvtáron kívül a megyei könyvtárak többségére és néhány nagyobb könyvtárra korlátozódik, azonban még mindig ez tekinthető a leggyakoribb, több intézmény által is használt közös tárgyszórendszernek.[21] Szabványos tezauruszként[22] a Köztaurusz strukturális szempontból alkalmas a gépi tárgyszavazásra, az egyes deszkriptorok azonban nem rendelkeznek URI-kkal.[23] A Relexből a tezauruszállomány SKOS-formátumban is exportálható, e formátum hátránya ugyanakkor, hogy problémás az ÉS/VAGY logikai operátorok megfelelő kezelése. E probléma kiküszöbölésére indítunk egy kutatási projektet, amelynek célja, hogy az ISO 25964 szabvánnyal összhangban megvizsgálja, hogyan lehet a hagyományos formátumú tezaurusz teljes szemantikus kapcsolatrendszerét leképezni kapcsoltadat-formátumban. A mintegy 84 ezer lexikai egységet tartalmazó Köztaurusz durván 8 és fél millió tokenes tanítókorpuszt igényel. További kutatás tárgya a gyakorlatban is megvizsgálni a Köztaurusszal már tárgyszavazott korpuszként is használható állományokat. E területen távlati tervünk a

17 Suominen [et al.] i.m. p. 267–274.

18 Consortium of European Social Sciences Data Archives. *ELSST Thesaurus (Version 3 – 2022)*

19 Egyed-Gergely Júlia [et al.]. Szociológia, kutatási adatok, mesterséges intelligencia: lehetőségek és tapasztalatok. = Valós térben – Az online térért. Networkshop 31: országos konferencia. 2022. április 20–22. Debreceni Egyetem. p. 163. DOI: 10.31915/NWS.2022.20

20 Naprakész változata a Relex webes felületén érhető el: http://mokka.hu/relex/guest.html

21 Bognár Noémi, Tóth Máté. *Tartalmi feltáró eszközök használata a magyarországi könyvtárakban*. = Könyv, könyvtár, könyvtáros. 25. évf. (2016) 6. sz., p. 22–24.

22 A Közgyűjteményi Szabványosítási Iroda e sorok írása idején végzett a Köztaurusz szabványossági vizsgálatával a hatályos nemzetközi tezauruszszabvány (ISO 25964) vonatkozásában. A fogalmak egyes egyedeire vonatkozó előfordulás-reláció (instance relationship) a korábbi magyar szabványban nem került szabványosításra, ezt leszámítva a Köztaurusz jelenleg is szabványos tezaurusznak számít.

23 2010–2011 környékén az OSZK Nektár katalógusába készült a Köztaurusz akkori változatáról URI-névtér (https://nektar.oszk.hu/auth/valami a csúcsdeszkriptor), a Relex SKOS-exportja is ezt alkalmazza a deszkriptorok azonosítására, azonban ez nem naprakész állomány.

Magyar Elektronikus Könyvtár (MEK) felhasználása tanítókorpuszként, majd ez alapján a gépi tárgyszavazás bevezetése a MEK-ben. Az Annif-ban jelenleg elérhető szövegfeldolgozó algoritmusok közül a következők bírnak valamilyen magyar nyelvű támogatással: az alapértelmezett stemmelő NLTK Snowball, a spaCy,[24] Simplemma, Stanza, UDPipe[25]. Mivel az előfeldolgozott szövegek tanítása és értékelése hardverigényes feladat, ezért erre célirányos környezetet kell kiépíteni, vagy más szolgáltatótól igénybe venni.

Ami a Közgyűjteményi Szabványosítási Irodát érinti, az elsődleges feladata, hogy a kurrens kihívásoknak megfelelően szabványos formában fejlessze a Köztauruszt. Célunk, hogy, a tezauruszformátumban rejlő szemantikus lehetőségek teljes biztosítása mellett, a Köztaurusz kapcsolatadat-formátumú tezauruszként, tárgyi authoritykontrollra is alkalmas szótárállományként szolgálja tovább a magyar könyvtári rendszert.

Felhasznált források

Annif – Tool for automated subject indexing and classification https://annif.org/ (2023.05.30.)

Bognár Noémi, Tóth Máté. *Tartalmi feltáró eszközök használata a magyarországi könyvtárakban*. = Könyv, könyvtár, könyvtáros. 25. évf. (2016) 6. sz., p. 18-30.

Consortium of European Social Sciences Data Archives. *ELSST Thesaurus (Version 3 – 2022)* https://thesauri.cessda.eu/elsst-3/en/?clang=hu (2022.06.02.)

Egyed-Gergely Júlia [et al.]. *Szociológia, kutatási adatok, mesterséges intelligencia: lehetőségek és tapasztalatok*. = Valós térben – Az online térért. Networkshop 31: országos konferencia. 2022. április 20–22. Debreceni Egyetem. p. 161-169. DOI: 10.31915/NWS.2022.20

Finto – Finnish Thesaurus and Ontology Service. https://finto.fi/en/ (2023.06.01.)

Finto AI (angol nyelvű felület) https://ai.finto.fi/?locale=en (2023.05.30.)

GitHub. *NatlibFi / Annif* https://github.com/NatLibFi/Annif/ (2023.05.30.)

ISO 25964-1:2011 *Information and Documentation. Thesauri and interoperability with other vocabularies. Part 1: Thesauri for information retrieval*. Geneva, ISO, 2011. 152 p.

ISO 25964-2:2013 *Information and Documentation. Thesauri and interoperability with other vocabularies. Part 2: Interoperability with other vocabularies*. Geneva, ISO, 2013. 99 p.

Lappalainen, Mikko [et al.]. *Reuse of library thesaurus data as ontologies for the public sector*. IFLA2014, Lyon, (2014), p. 1-3. URL: https://library.ifla.org/id/eprint/819/1/086-lappalainen-en.pdf

Suominen, Osma – Koskenniemi, Ilkka. *Annif Analyzer Shootout: Comparing text lemmatization methods for automated subject indexing*. = Code4Lib Journal, vol. 54 (2022) URL: https://journal.code4lib.org/articles/16719

Suominen, Osma [et al.] Annif and Finto AI: Developing and Implementing Automated Subject Indexing. = JLIS.It, vol. 13. (2022) no. 1, p. 265–282. DOI: https://doi.org/10.4403/jlis.it-12740

Suominen, Osma. *Annif: DIY automated subject indexing using multiple algorithms*. = LIBER Quarterly, vol. 29. (2019) no. 1., p. 1–25. DOI: https://doi.org/10.18352/lq.10285

SWIB 2022 online konferencia. *Introduction to the Annif automated indexing tool*. 2022.11.30. https://swib.org/swib22/programme.html#day3 (2023.06.02.)

24 Lásd a HuSpaCy projektet a Githubon: https://github.com/huspacy/huspacy (2023..06.05.)

25 European Language Grid: UDPipe Hungarian: Morphosyntactic Analysis of Raw Text https://live.european-language-grid.eu/catalogue/tool-service/438/overview/ (2023.06.05.)

CC BY

DOI: 10.31915/NWS.2023.6

A Pécsi Egyetemtörténeti Gyűjtemény online adatbázisai és digitális gyűjteményei

Online databases and digital collections of the Pécs University Historical Collection

Dezső Krisztina
Pécsi Tudományegyetem Egyetemi Könyvtár és Tudásközpont (Pécs)
dezso.krisztina@lib.pte.hu

Absztrakt

Pécsi Egyetemtörténeti Gyűjteményben őrzött műtárgyak, történeti dokumentumok, fotók digitális formában is megőrzésre kerülnek. A digitalizáláson és tudományos feldolgozáson túl az online látogatók, kutatók számára több platformon is elérhetővé tettük gyűjteményeinket (múzeumi adatbázis, egyetemtörténeti topográfia, virtuális kiállítások). A különböző platformok egymásra is épülnek, gyakran egy-egy műtárgy mindhárom felületen megjelenik, akár különböző aspektusokban, történetekben elhelyezve. A tanulmányban ezen digitális gyűjteményeink létrejöttét, a bennünk feldolgozott gyűjteménytesteket és a világhálón elérhető digitális adatbázisainkat, gyűjteményeinket mutatom be.
Kulcsszavak: digitális gyűjtemények, múzeumi gyűjtemények, egyetemtörténet. múzeumi digitalizálás

Abstract

Artifacts, historical documents and photographs retained in the Pécs University Historical Collection are also preserved in digital form. In addition to scientific processing, we have made our collections available to online visitors and researchers on several platforms (museum database, topography of university history, virtual exhibitions). The different platforms are also built on one another, and an artefact often appears on all three platforms. In this paper I will present the creation of these digital collections, the types of artefacts we process as well as the digital databases and collections we make available on the net.
Keywords: digital collections, museum collections, university history, museum digitisation

Az első pécsi gyűjtemények és kiállítások létrejötte[1]

1967-ben, a középkori pécsi egyetem 600 éves évfordulójának ünnepségei kapcsán merült fel az igény, hogy a pécsi egyetemekhez kapcsolódó dokumentum-, és tárgyanyagot össze kellene gyűjteni. 1975-ben hozta létre a Pécsi Tudományegyetem az Egyetemi Archívumot, mely az Egyetemi Könyvtáron belül került elhelyezésre. Az Egyetemi Archívum célja az egyetem működésére vonatkozó dokumentum, fotó és tárgyi anyag összegyűjtése volt. Az archívum gyarapodása az 1990-es évek közepétől megállt, majd 2010-től a Pécsi Egyetemtörténeti Gyűjteménybe került az összegyűjtött anyag.

1 A gyűjtemények történetéről lásd: Dezső Krisztina – F. Dárdai, Ágnes: A Pécsi Egyetemtörténeti Gyűjtemény. *Per Aspera Ad Astra*, 2016. ***3***(1), 36–52. hozzáférés 2023. 06. 16. https://doi.org/10.15170/PAAA.2016.03.01.03

1992-ben nyitotta meg kapuit a Pécsi Orvostudományi Egyetemen az első állandó kiállítás, mely a korábbi nagy gyűjtőmunka eredményeként mutatta be az Erzsébet Tudományegyetem Orvostudományi Karának, majd az 1950-ben megalakult Pécsi Orvostudományi Egyetemnek a történetét. A kiállítás 2022-ben az épület átépítése kapcsán megszűnt, anyaga a Pécsi Egyetemtörténeti Gyűjteménybe került át.

2000-ben új korszak kezdődött a pécsi felsőoktatási intézmények történetében. A Janus Pannonius Tudományegyetem, a Pécsi Orvostudományi Egyetem, valamint a szekszárdi Illyés Gyula Pedagógiai Főiskola egyesülésével megalakult az integrált Pécsi Tudományegyetem és hamar felmerült az igény, hogy ne csak az orvosi fakultás, hanem a többi kar történetének bemutatása is lehetővé váljon. A Vasváry-házban karonként ismerhették meg a látogatók a jogelőd, majd az integrált egyetem karainak történetét. A kiállítás 2005-ig működött a Vasváry-házban, anyaga szintén az egyetemtörténeti gyűjteménybe tagozódott be.[2]

Pécsi Egyetemtörténeti Gyűjtemény

2005-ben az egyetemi könyvtár szervezeti keretei közé került az egyetemtörténeti múzeumi gyűjtemény. Ekkor indultak meg egy új állandó kiállítás készítésének munkálatai is. 2010-ben ünnepélyes keretek között nyílt meg az új tárlat. Ugyanebben az évben közérdekű muzeális gyűjteményi rangot is kapott az intézmény.

Az állandó kiállítás mellett évente több időszaki kiállítást is készítenek a gyűjtemény munkatársai. De nemcsak a gyűjtemény épületében, hanem a Pécsi Tudományegyetem karain is több kihelyezett kiállítást készítettek a muzeológusok az adott kar történetéhez kapcsolódva. Az eltelt időszakban mintegy 30 időszaki kiállítás keretében láthatták az érdeklődők a gyűjtemény tárlatait.

A Pécsi Egyetemtörténeti Gyűjtemény gyűjtőköre a Pécsi Tudományegyetem és jogelőd intézményeinek tárgyi, dokumentum és könyvanyaga. A gyűjteményben az alábbi egyetemek, főiskolák általános és kari anyaga található meg:

- Magyar Királyi Erzsébet Tudományegyetem (1912–1950),
- Pécsi Tudományegyetem (1950–1982),
- Pécsi Orvostudományi Egyetem (1950–2000),
- Pécsi Pedagógia Főiskola, Pécsi Tanárképző Főiskola (1948–1982),
- Janus Pannonius Tudományegyetem (1982–2000),
- Pollack Mihály Műszaki Főiskola (1970–1986),
- Illyés Gyula Tanítóképző Főiskola (1977–2000),
- Pécsi Tudományegyetem (2000-től folyamatos).

Dokumentumgyűjtemény
A Pécsi Tudományegyetem és jogelődjeinek intézményeihez, karaihoz, az oktatókhoz, diákokhoz köthető hivatalos és személyes dokumentumanyag található a gyűjteményben. Fontosabb dokumentumtípusok: hivatalos dokumentumok, oklevelek, tanulmányi iratok, személyekhez köthető dokumentumanyag. Egy mintegy 3000 db-os képeslapgyűjtemény is gazdagítja a gyűjteményt.

2 Az orvoskari és a Vasváry-házban található kiállítások ismertetője: Benke József: *Pécsi Tudományegyetem Egyetemtörténeti Múzeuma*. Pécs, 2004.

CC BY

Fotógyűjtemény
A dokumentumtár mellett a leggazdagabb anyaggal a fotógyűjtemény rendelkezik. A fotóanyag tematikáját tekintve több egységre bontható: az egyetem hivatalos eseményein készült felvételek, arcképgyűjtemény, épületfotók gyűjteménye. A hagyományos papírképek mellett jelentős a digitális fényképek száma is.

Történeti tárgyak gyűjteménye
E sokszínű gyűjteményben az oktatáshoz használt eszközöktől a szobrokig nagyon sokféle tárgy található. Az egyetem hivatalos jelvényei mellett az egyetemi ajándéktárgyak, az egyetemi hallgatók sport- és művészeti eredményeinek elismeréseként kapott kupák, serlegek, bélyegzőgyűjtemény, képzőművészeti gyűjtemény, professzorok személyes tárgyai kerültek itt elhelyezésre.

Műszaki és technikatörténeti gyűjtemény
Ebben a gyűjteményben nagyrészt az orvostudományi karhoz köthető, a gyógyításban, illetve az oktatásban, kutatásban használt orvosi eszközök, mérőműszerek, mikroszkópok, kézi eszközök találhatóak. De a Műszaki és Informatikai Kar, illetve a Természettudományi Kar által használt mérőeszközök, a hallgatók által készített munkák is bekerültek a gyűjteménybe.

Numizmatikai gyűjtemény
A Pécsi Tudományegyetem és jogelőd intézményei által készíttetett jubileumi vagy rendezvényekhez kapcsolódó plakettek, emlékérmek, kitűzők, jelvények képezik a gyűjtemény egyik felét. A másik nagy csoporthoz az egyetem oktatóinak, professzorainak személyes anyagában található kitüntetések, érmek, plakettek tartoznak.

Digitális gyűjtemények

Gyűjteményi adatbázis
A muzeális anyag digitalizálása 2014-ben kezdődött meg a fotó- és képeslapgyűjtemény feldolgozásával. Az anyagok a Corvina katalóguson belül kialakított különgyűjteményekben voltak elérhetőek. A feldolgozásnál nehézséget jelentett, hogy a fényképeket és a képeslapokat az alapvetően könyvek feldolgozására kifejlesztett mezőkben kellett leírni MARC-formátumban. A leírások tartalmazták az alapadatokat és tárgyszavak alapján kereshetőek voltak az egyes tételek, de a szabványos múzeumi leírást ezekben a rekordokban nem lehetett megvalósítani. Az e-KéPEK[3] és a Reuter-képeslapadatbázis 2019-ig gyarapodott, mintegy 2416 fényképet és 687 képeslapot tartalmazott.

2019-ben a két már meglévő digitális gyűjtemény átemelésre került a Qulto/Huntéka adatbázisba.[4] Ez az adatbázis lehetőséget ad rá, hogy az egyes műtárgyakat a típusának megfelelő űrlapokban írják le a muzeológusok, alkalmas leíró kartonok generálására, a mozgatás nyilvántartására is. 2021-ben elérhető lett az adatbázis nyilvános felülete. Az átemelt képeslap- és fotógyűjtemény mellett dokumentumok, tárgyi, numizmatikai, valamint műszaki és technikatörténeti anyag is feldolgozásra került.

3 Az adatbázisról részletes ismertető: Dezső Krisztina – Schmelczer-Pohánka, Éva: A Pécsi Egyetemtörténeti Gyűjtemény képadatbázisa (eKéPEK). *Per Aspera Ad Astra*, 2014. ***1***(1), 171–176. hozzáférés: 2023. 06. 16. https://doi.org/10.15170/PAAA.2014.01.01.11

4 Az adatbázis elérhetősége: https://egyetemtortenet.lib.pte.hu/online-collection/-/results/init, hozzáférés: 2023. 06. 16.

A múzeumi KDS fotózáshoz kapcsolódva mintegy 200 műtárgyról készült jó minőségű digitális fénykép. A műtárgyak fotóval, részletes adatokkal, leírásokkal ellátott rekordjai nemcsak a saját adatbázisunkban, hanem a Museumap oldalán is elérhetőek az érdeklődők számára.

A gyűjtemény Qulto/Huntéka adatbázisában jelenleg több gyűjteményrész adatai találhatóak meg:

- Adattári fényképgyűjtemény: 64 db
- Eredeti, archív fényképek gyűjteménye: 2535 db
- Képeslapgyűjtemény: 700 db
- Műszaki és technikatörténeti gyűjtemény: 40 db
- Numizmatikai gyűjtemény: 68 db
- Történeti dokumentumok gyűjtemény: 807 db
- Történeti tárgyak gyűjteménye: 87 db

A digitalizált műtárgyanyag felhasználási köre igen széleskörű lehet. A hagyományos és virtuális kiállítások mellett kiadványokban, filmekben, múzeumpedagógiai programok során is használják a digitális képeket.

Virtuális kiállítások
Az első virtuális kiállítások a TGYO Blogon[5] 2019-ben készültek el. Az Öttorony című, a pécsi irodalmi életet bemutató[6], illetve a Somogyi Ferenc munkássága[7] előtt tisztelgő kiállítások virtuális változatai kerültek fel a blogra. A bejegyzések vitrinfotókkal, valamint a kiállított könyvek, dokumentumok és tárgyak fényképeivel, diasorozataival kiegészítve kerültek fel az egyes témakörökhöz. Végül ezek a kisebb témákat felölelő írások egy nagyobb közös felületen összefűzve alkották a virtuális kiállításokat.

2022-ben lehetőség nyílott rá, hogy a Storytelling program segítségével valódi virtuális kiállítások formájában lehessen bemutatni az intézmény tárlatait. A 2022 és 2023-as évben négy kiállítást lehetett így a nagyobb nyilvánosság számára is elérhetővé tenni.[8] A virtuális kiállítások másik nagy előnye, hogy a kiállítás zárása után is megtekinthető marad a bemutatott anyag. Sőt a virtuális változatban általában olyan érdekesebb dokumentumok, tárgyak, fényképek is bemutatásra kerülnek, amelyek kiegészítik a kiállítótérben látható anyagot, teljesebbé teszik a képet. Lehetőség nyílik rá, hogy egy fotósorozat valamennyi képét bemutassák a kurátorok, egy-egy érdekes kötetbe belenézzenek, beleolvassanak a látogatók, vagy egy-egy műtárgy apróbb részletét felnagyítva is láthassák. A virtuális kiállítások kiegészítéseként online katalógusok is készülnek a tárlatokhoz, melyeket szintén a TGYO Blogon érhetnek el a látogatók.

---

5 A TGYO Blog elérhetősége: https://tgyoblog.lib.pte.hu/, hozzáférés: 2023. 06. 16.

6 Öttorony virtuális kiállítás főoldala: https://tgyoblog.lib.pte.hu/ottorony-a-pecsi-irodalmi-muveltseg-a-kezdetektol-a-huszadik-szazadig/, hozzáférés: 2023. 06. 16.

7 Somogyi Ferenc kiállítás főoldala: https://tgyoblog.lib.pte.hu/somogyi-ferenc-emlekkiallitas-virtualis-kiallitas/, hozzáférés: 2023. 06. 16.

8 Virtuális kiállítások: *34. Eucharisztikus Világkongresszus* virtuális kiállítása. (https://tgyoblog.lib.pte.hu/eucharisztikus-virtualis-kiallitas/); *Pécs – Debrecen – Szeged. Klebelsberg Kuno és a magyar felsőoktatás* virtuális kiállítása (https://tgyoblog.lib.pte.hu/pecs-debrecen-szeged-klebelsberg/); *Erzsébet királyné, egyetemünk egykori névadója* virtuális kiállítás (https://tgyoblog.lib.pte.hu/erzsebet-kiralyne/); *Füvésztudomány a Mecsek lábánál* virtuális kiállítás (https://tgyoblog.lib.pte.hu/fuvesztudomany/), hozzáférés 2023. 06. 16.

CC BY

A Pécsi Egyetemtörténeti Gyűjtemény kiállításaihoz múzeumpedagógiai órák is készülnek. Ezen foglalkozások iránt a járványidőszak alatt is nagy volt az érdeklődés, ezért a múzeum munkatársai azokat a foglalkozásokat, amely alkalmasak voltak rá, elkészítették online verzióban is. Ezekhez részletes leírásokat, feladatlapokat, kisfilmeket készítettek a múzeumpedagógusok. A pandémia alatt a *Legyél te is könyvkötőmester!*, a *Pécsi aprónyomtatványok az 1848-49-es forradalomról*, illetve a *Szövegekből épülő város* című foglalkozások váltak elérhetővé. A 2022-es évben nyitott Klebelsberg Kunóról szóló kiállításhoz kapcsolódó jelenléti múzeumpedagógiai óra virtuális változata is nyilvános. Ez utóbbi foglalkozáshoz a LearningApps felületen készültek az online kitölthető, megoldható feladatok, melyeknek megoldásához a kiállítás virtuális változatát tudják használni az iskolai csoportok.[9]

Egyetemtörténeti topográfia

Az Erzsébet Tudományegyetem Pécsre költözésének centenáriuma alkalmából készült az *Egyetem a városban. A pécsi Erzsébet Tudományegyetem topográfiája a két világháború között* című egyetemtörténeti topográfiai bemutató[10], illetve az ehhez kapcsolódó séták, amelynek során virtuálisan, illetve a helyszínen az applikáció segítségével bejárhatóak az egyetem két világháború között működő karainak oktatási épületei, a klinikák, a tanári lakóházak, az egyetemhez köthető emléktáblák, szobrok. Az 1923-ban Pécsre költöző egyetem a szűkös anyagi körülmények miatt nem tudott új épületeket emelni, ezért a város, az állam és a püspökség meglévő épületeket ajánlott fel az egyetem számára. Ezek jellemzően a 19. század végén, a 20. század elején épültek, és a város különböző pontjain helyezkedtek el.

Az adatbázisba több mint száz épület, emlékhely, szobor került be. A rekordokban az adott látványosság megnevezése, címe és GPS-koordinátái mellett a térképen is látható elhelyezkedése. Az egyes épületek, látványosságok történetét a rekordokban részletes leírások ismertetik, kiemelt figyelmet fordítva a látnivalók egyetemhez való kapcsolatára. A rekordokban archív- és újonnan készített fotók is segítik a megismerést. Az adatbázisban már több mint száz pécsi egyetemhez kötődő helyszínről lehet információkat kapni.

A látványosságokat túrákká lehet egybefűzni. Az egyetemi centenáriumhoz kapcsolódva négy túra érhető el nyilvánosan az Egyetemtörténeti topográfia felületén. Három az Erzsébet Tudományegyetem két világháború között működő karainak történetét követi nyomon a pécsi helyszíneken: *A jogi kari útja a két világháború között*, *Az eltűnt bölcsészkar nyomában*, illetve az *Erzsébet Tudományegyetem klinikái*. A negyedik tematikus túrán Klebelsberg Kuno pécsi látogatásának útvonalát lehet végigjárni. A túrák mellé játékokat is lehet készíteni, az állomásokon az egyes látványosságokhoz kapcsolódó feladatot oldanak meg a résztvevők. A 2023-as év folyamán egy középiskola számára készült 15 helyszínt bejáró tematikus túra és játék.

A tematikus túrák jelenleg az Erzsébet Tudományegyetemhez köthető helyszíneket mutatják be, a tervek szerint az 1950-es évek utáni egyetemi helyszínekkel és túrákkal fog még bővülni az adatbázis.

9 Az online múzeumpedagógiai órák elérhetősége: https://tgyoblog.lib.pte.hu/category/muzeumpedagogia/, hozzáférés 2023. 06. 16.

10 Az egyetemi topográfia elérhetősége: https://varosfoglalo.pecs.monguz.hu/, hozzáférés 2023. 06. 16.

Összegzés

A Pécsi Egyetemtörténeti Gyűjtemény létrejötte óta törekszik arra, hogy a Pécsi Tudományegyetem és jogelőd intézményeinek emlékanyagát minél szélesebb nyilvánosság számára tegye elérhetővé. Az utóbbi két-három évben ezt a lehetőséget az online gyűjtemények megteremtésével igyekeznek bővíteni a muzeológus munkatársak. Az online múzeumi adatbázis, a virtuális kiállítások, az online múzeumi foglalkozások és az egyetemtörténeti topográfiai adatbázis nemcsak az egyetemi polgároknak és a pécsi látogatóknak, hanem az ország, illetve a világ minden pontjáról érkező érdeklődők számára bármikor elérhetővé teszik a pécsi egyetemtörténet emlékeit.

Felhasznált irodalom

Benke József: *Pécsi Tudományegyetem Egyetemtörténeti Múzeuma*. Pécs, PTE, 2004.

Dezső Krisztina – F. Dárdai, Ágnes: A Pécsi Egyetemtörténeti Gyűjtemény. *Per Aspera Ad Astra*, 2016. 3(1), 36–52. hozzáférés 2023. 06. 16. https://doi.org/10.15170/PAAA.2016.03.01.03

Dezső Krisztina – Schmelczer-Pohánka, Éva: A Pécsi Egyetemtörténeti Gyűjtemény képadatbázisa (eKéPEK). *Per Aspera Ad Astra*, 2014. 1(1), 171–176. hozzáférés: 2023. 06. 16. https://doi.org/10.15170/PAAA.2014.01.01.11

CC BY

DOI: 10.31915/NWS.2023.7

Nemzeti könyvtárak és az OSZK MARC21 állományainak összehasonlító elemzése néhány adatmező alapján

Ungváry Rudolf
Országos Széchényi Könyvtár
ungvaryr@gmail.com

Király Péter
Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG), Göttingen
peter.kiraly@gwdg.de

Bevezető

Korábbi tanulmányokban összehasonlítottuk néhány nemzeti és tudományos könyvtár katalogizálási gyakorlatát. 2020-ban ennek a feltárásnak a *tartalomra vonatkozó* adatelemeit (043, 045, 052, 072, 080, 082, 084, 085, 505, 520, 6XX) és a velük összefüggő adatelemeket (007, 008, 034, 041) elemeztük. 2022-ben a rekordfej (06, 07 és 17), továbbá a kódolt fizikai jellemzők (007) és a meghatározott jellemzők és információs adatok (008) *tartalomra vonatkozóan fontosabbnak tekintett* adatelemeit, végül pedig a 648 és 65X almezőit vizsgáltuk összehasonlítva a kiválasztott pozíciókon szereplő értékeket.

Jelen összehasonlítás fókuszában az Országos Széchényi Könyvtár katalógusa áll. Miben hasonló és miben különbözik katalogizálási és feldolgozási gyakorlata más nemzeti könyvtárétól?

A következő MARC21 mezőket[1] és pozícióikat vetettük össze (1. táblázat):

Mező	*pozíció/almező*	*elnevezés*
Rekordfej	06	a dokumentum jellege (=a rekord típusa)
	07	bibliográfiai szint
	17	a leírás jellege
007	00	kódolt fizikai jellemzők, dokumentumkategória
	01	a dokumentum speciális megjelölése: szöveg
	01	a dokumentum speciális megjelölése: térkép
008	18–21	könyvek, illusztráltság
	25–27	könyvek, tartalmi jellemzők
	33	könyvek, műfaj
600		személynév tárgyszó
648	A	kronologikus tárgyszó
650	A	szaktárgyszó
651	A	földrajzi tárgyszó
653		szabadon választott tárgyszó
655	A	formai tárgyszó (dokumentumtípus)

1. táblázat. Az összehasonlításba bevont MARC21 mezők és pozíciók.

1 Az OSZK-ban a HUNMARC-ot használják, de az összehasonlítása bevont mezők esetében ezek azonosak a MARC21 mezőivel.

A rekordfej és a vezérlő mezők adatait a jelentőségükhöz képest a végfelhasználók csak nagyon nehézkesen tudják használni, ezért le is mondanak a használatukról. Munkánk célja, hogy néhány elérhető nemzeti könyvtár vonatkozó állománya esetén a felszínre kerüljenek eme adatmezők kitöltésének jellegzetességei, és az esetleg belőlük fakadó problémák, anomáliák, hogy mindezzel empirikus adatok álljanak rendelkezésre mind a katalógusok, mind a kezelőrendszereik, mind pedig a MARC21 fejlesztéshez.

A vizsgálatba bevont nemzeti könyvtárak: Kongresszusi Könyvtár (LC), Német Nemzeti Könyvtár (DNB), Osztrák Nemzeti Könyvtár (ÖNB), Finn Nemzeti Könyvtár, nemzeti bibliográfia (NFI), Svéd Nemzeti Könyvtár és közös katalógus (KBS), Lengyel Nemzeti Könyvtár (BNPL), Belga Nemzeti Könyvtár (KBR), Brit Nemzeti Könyvtár (BL), Izraeli Nemzeti Könyvtár (NIZ), Holland Nemzeti Könyvtár (KB), Cseh Nemzeti Könyvtár (CSH), Országos Széchényi Könyvtár (OSZK).

1. Rekordfej és a vezérlő mezők

1.1 A rekordfejen belüli pozíciók

A *rekordfej pozíciói tekintetében a feldolgozottság 100 %-os*, a rekordfej minden pozíciójának kötelező ugyanis értéket adni. Most csak a rekordtípus (rekordfej 07) és a bibliográfiai szint, ill. a leírás jellege (szintje) összevetésének eredményeit ismertetjük.

A bibliográfiai szint egyes típusai elsősorban a nyelvi anyagra jellemzők. A kétdimenziós nem kivetíthető és a zenei hangzó anyagok között találni összefoglaló, alárendelt (kötet, részegység) és sok monografikus szintű feldolgozást. A kétdimenziós anyagok (pl. fényképek, grafikák) esetén számos könyvtárban (OSZK, LC, ÖNB, BNPL KBR, BL, KB, CSH) túlnyomórészt fényképekről van szó. A zenei hangzó anyagok többnyire nem komolyzenei dalok, és DNB-re meg a BL-re jellemzők, kisebb arányban az OSZK-ra is.

A második összehasonlításban feltűnő, hogy általában milyen kicsi a számítógép által kezelt állomány (Rf/06=m^2). Az OSZK a 2000 tételes elektronikus típusú állományával a hasonló nagyságú könyvtárakkal (ÖNB, NFI) van egy sorban, holott ezen források elterjedtsége már rendkívül nagy. Ennek oka bizonyára a MARC21 szabályozása. Ha ugyanis van egy jelentős szempont, amely miatt a forrás egy másik rekordtípusba is tartozik, akkor az *m* kód helyett a másik szempont típusával kell osztályozni, például ha az állomány nyelvi anyag, grafika, kartográfiai anyag, hangzó anyag, mozgókép. Vagyis rekordtípusként nem lehet több szempont szerint osztályozni, ami az egy dimenziós osztályozási rendszerek alapvető hátránya.

Figyelemre méltó az OSZK-ban a két dimenziós grafikák (Rf/06=k, nagyrészt képes levelezőlapok) viszonylag gazdag leírási szintű feldolgozása: teljes autopsziával és anélkül (Rf/17=#, ill. l), egyszerűsített autopsziával és anélkül (Rf/17=5, ill. 2). Az ismeretlen minősítés is gyakori, de az OSZK-ban egyáltalán nincs ilyen; mivel a szoftver automatikusan ad egy alapértelmezett minősítést – mindez visszaköszön más, itt nem tárgyalt adatelemek esetén is. A KBR és BL a rekordok egy részét a leírás szintjét illetően ismeretlennek (Rf/17=u) nyilvánították – ami a könnyebbik megoldás (más adatmezőnél is sok példa van erre). A BL-ben sok a rövidített leírási szint is (Rf/17=3). A KBS-ben a rekordok többségének leírási szintje minimális (Rf/17=u) vagy rövidített (Rf/17=3).

2 Az OSZK-ban erre az *l* kódot használják.

CC BY

A leírás szintje szempontjából, nem meglepően, szinte mindenhol jelentős a nyelvi anyag feldolgozottsága, az LC és DNB esetében kitűnik még a zenei hangzó anyag is. Kevésbé feltűnően, de az OSZK-t, NIZ-t és CSH-t is ugyanez jellemzi. Az angol nyelvű (LC, BL) és inkább brit orientációjú könyvtárakban (NFI, KBS, NIZ) észrevehető inkább, hogy a többi rekordtípusban is megjelenik a leírás szintjének minősítése. Az OSZK itt a középmezőny alján foglal helyet.

A legkiegyensúlyozottabb helyzet az LC-ben van: feltűnő a zenei anyag és a fényképek minimális jelenléte. Az OSZK ezzel szemben az éllovasok között van, aminek oka a gondos aprónyomtatványtári képeslap- és a zeneműtári feldolgozás. A zenei anyagok és a grafikák vonatkozásában az ÖNB van hasonló pozícióban – az OSZK a középmezőny alján foglal helyet, holott maga a zenei anyag feldolgozása a Zeneműtárban alapos. A legegyszerűbb igényeket kielégítő slágerek, illetve rajzos, többnyire történelmi eseményeket ábrázoló grafikák minősítése a „nincs” (rekordfej/17=z). Egyebütt ezen dokumentumoknak nincs nyomuk, holott legalábbis a fényképeknek idővel többnyire történelmi értéke lesz.

Az LC, BL és KBS minden jel szerint változatosan, a különféle leírási szinteket figyelembe véve osztályoz, a DNB, ÖNB, BNPL, OSZK esetében feltehetően az alapértelmezések az uralkodók. Valójában nem sok jelentőséget tulajdonítanak a rekordtípus pozícióinak, ami még inkább jellemző a 007/008 adatmezőkre is.

1.2 A rekordfej-pozíciók és a 007/008 adatmezők értékei

Mivel a 00X adatmezők kitöltése nem kötelező, alacsony a számuk. A nemzeti könyvtárak között ennek nyomán keletkezett egyenetlenségek példaként a rekordfej 06-os pozíciójának és a dokumentumkategóriának (007/00) az összehasonlítását tárgyaljuk.

Feltűnő, hogy az LC, DNB, ÖNB, NFI, BNPL KBR, BL és NZI esetében viszonylag gazdag a 007 mező dokumentumkategóriáinak kitöltöttsége a rekordtípus Rf/06 pozíciójához viszonyítva. Több könyvtárban ráadásul jelentős az elektronikus dokumentumok nyelvi anyagként való feldolgozása is. Az OSZK, és a kisebb nemzeti könyvtárak katalógusából ez teljesen hiányzik. Könyvek esetében a meggyökeresedett hagyomány az ok. Csak a nyomtatott és kéziratos kartográfiai anyagok közel teljesek, ami nemzetközi viszonylatban is intenzív térképtári munkára utal. Jelentős még az OSZK-ban (és a BNPL-ben) a grafikák (Rf 06=k, „két dimenziós”) és a nem zenei hangfelvételek egymáshoz viszonyított feldolgozottsága. Az első az aprónyomtatványtárban a képeslapok feldolgozottságát dicséri, a hangfelvételeké pedig a Zeneműtárat. A BNPL-ben a két dimenziós rekordtípusok (rekordfej 06=k) feldolgozottsága a mozgókép dokumentumkategóriájában (007/00=k) a leggazdagabb.

Néhány kisebb nemzeti könyvtárhoz hasonlóan az OSZK sem minősíti a nyelvi anyagot (Rf/06) szöveges dokumentumkategóriaként (007/00=t).

Érdekes, hogy a két dimenziós nem kivetíthető ábrázolásokat (lényegében a fényképeket, Rf/06=k) minden jel szerint csak az LC, BL és KBS osztályozza változatosan, a különféle leírási szinteket figyelembe véve. A DNB, ÖNB, BNPL és OSZK esetében feltehetően az alapértelmezések az uralkodók. Lehet, hogy nem sok jelentőséget tulajdonítanak a rekordfej értékeinek.

A nem kötelezőség miatt a *feldolgozottság mértéke olyan kicsi, hogy valójában nem lehet érvényes megállapításokat tenni. Ez még inkább érvényes, ha az összehasonlításban mindkét oldalon nem*

kötelező adatmezők szerepelnek. Valószínű, hogy a szabályozások eleve azért teszik lehetővé, hogy kihagyják a 00X-es mezőket mert nem tulajdonítanak fontosságot nekik.

1.3 Szöveges dokumentumok tartalmi jellemzői (007/01 t vs. 008/24–27)

Az 00X-es mezőkben olyan alacsony az értékkel rendelkező adatelemek száma, hogy statisztikailag értékelhetetlenek. Példánk a szöveges dokumentum speciális megjelölésének (007/01) a könyvek tartalmi jellemzőinek (008/24–27) összehasonlítása. Noha a szöveges dokumentumok tartalmi jellemzők alapján való keresése a legjellemzőbb végfelhasználói igény, alig van a könyvtárakban feldolgozás, a DNB csupa kódolatlan (008/24–27= |), az ÖNB csupa ismeretlen minősítést ad meg.

Csak a DNB, ÖNB és CSH esetén nagy a minősítések száma, de többségük alapértelmezésnek látszik. Az OSZK a szöveges dokumentumok tartalmi jellemzőit nem minősíti. Más értékek esetében alig néhány százalék az előfordulás. Itt is lehetséges, hogy a tartalmi jellemzők szerinti dokumentumosztályozást inkább a 655 formai tárgyszavak mezőjében, a kutatási jelentéseket az 513 $a almezőben oldják meg. A 00X mezők kereshetősége egyáltalán nem megoldott, a könyvtárak minden jel szerint nem akarnak ezekkel vesződni, különösen, hogy számos adatmező redundáns a 00X mezők pozícióinak tartalmával (2.táblázat).

A rekordfejben és a 00X vezérlőmezőkben	*A szöveges megjegyzésmezőkben*
Rekordfej (rekordtípus 06 c=nyomtatott zenemű; d=kéziratos zenemű)	254 $a A kotta típusának megevezése
Rekordfej (rekordtípus 06 m=számítógép által kezelt állomány; 007/00 dokumentumkategória c=elektronikus dokumentum	256 $a és $b Elektronikus dokumentumok jellemzői
007/01 a dokumentum speciális megjelölése: térkép 007/01 a dokumentum speciális megjelölése: szöveg 008/18–21 könyvek: illusztráltság	300 Terjedelem, fizikai jellemzők $a A dokumentum fizikai hordozójának a fajtája $b Egyéb fizikai jellemzők, illusztrációk 588 Megjegyzés a terjedelemről/fizikai jellemzőkről
008/24–27 tartalmi jellemzők: könyvek	513 Megjegyzés a kutatási jelentés típusáról és idejéről $a A jelentés típusa.
Rekordfej (rekordtípus 06 m=számítógép által kezelt állomány); 007/00 dokumentumkategória c=elektronikus dokumentum	516 Megjegyzés az elektronikus dokumentum típusáról vagy adatairól

2. táblázat. A rekordfej és a 00X vezérlő mezők dokumentumtípusok osztályozására való pozíciói és a velük összefüggő szöveges megjegyzésmezők

CC BY

2. Tárgyszómezők (tárgyi melléktételek)



1. A 6XX mezők fontosabb almezői összevont buborékdiagramban, a jobb áttekintés érdekében.

Az 1. ábrán jól látható a 650-es szaktárgyszók dominanciája, de kiemelkedő a földrajzi név, a szabadon választott és különösen a formai tárgyszó gyakorisága is. Az OSZK a középmezőny alsó részében foglal helyet, ugyanakkor almezők kitöltésének dolgában feltűnően jó, noha lennének egyszerűen megoldható javítandók, mint a $2 forrásadatok, a $0 és a $1 azonosítók megadása, továbbá a formai tárgyszavak gyakoribb használata.

A tárgyszómezők a tartalom, forma (dokumentumtípus), idő, földrajzi elhelyezkedés szerinti legfontosabb mezőcsoportot alkotják. Feltűnő, hogy a LC, BNPL és NIZ mennyivel több személynévvel tárgyszavazza a tartalmat, ráadásul az utóbbi kettő nem is tartozik a legnagyobb nemzeti könyvtárak közé. Az OSZK a középmezőnyben van.

A kronologikus kiegészítő gyakoribb használata (BNPL NIZ, LC, KB) a személynév tárgyszó gondos szerkesztésére utal. Ahol ez az érték nagyobb, ott a $c foglalkozás kiegészítő értéke is nagyobb. Feltűnő, hogy a név és a mű kapcsolatát leíró, szerepjelölő $e használata milyen elenyésző, számos könyvtárban egyáltalán nem használják (a HUNMARC nem ismeri).

A $j az OSZK-ban az utónevet tartalmazza, feltűnően nagy értékét ez magyarázza; a MARC21-ben az egyéni nevet minősítő adat és jóformán nem használják. A $q teljes név az LC, BL és NIZ könyvtárakban használatos, ahol felveszik önállóan a rövidített nevet.
Fontos jelenség, hogy újabban bekerült a MARC21-be a besorolási adatok $0 azonosítója, melyet a legtöbb könyvtárban az új rekordokban mindig megadnak, egyedül a KB, CSH és az OSZK nem alkalmazza, ami előbb-utóbb komoly lemaradással jár. Ez vonatkozik a $1 fizikai tárgyak azonosítójára is.

Egyes könyvtárakban nagy gondot fordítanak arra, hogy minden személynév tárgyszóhoz megadják a forrást (DNB, BNPL, NIZ, CSH). Az OSZK-ban nem, mivel a HUNMARC-ból hiányzik. Érdekes, hogy az LC is ritkán használja ezt az almezőt, ami csökkenti az állomány elemezhetőségét, mivel nem lehet kapcsolatot találni az alkalmazott tárgyszó rendszerek, tezauruszok és a katalogizált rekordok között.

Az ÖNB és KBR személynév-tárgyszavak (továbbá a szak- és földrajzi tárgyszavak) használata minimális, jelentős viszont a szabadon választott tárgyszavak és a formai tárgyszavaké. A testületi- és rendezvénynevek mindenhol csak a rekordok kisebb százalékához kapcsolódnak.

A BNPL, CSH, KB és NFI átlagon felül használ kronologikus tárgyszavakat. A könyvtárak többségében a forrásadatot is gondosan megadják. Egyedül az OSZK és az LC nem minősíti külön a kronologikus tárgyszavakat, előbbi azért, mert az AMICUS-ban nem definiálták.

Feltűnő, hogy a nemzeti könyvtárak többségéhez képest, ahol a rekordok több, mint harmada-fele kap 650-es szaktárgyszót, az OSZK-ban csak a rekordok negyede, amivel a középmezőnyben van, holott itt hagyományosan színvonalas a tartalomfeltáró munka. Ennek oka, hogy csak 2004-től kezdve adnak meg tárgyszavakat, de ha retrospektív feldolgozásra kerül sor, akkor majd pótolják.

Még több könyvtárra érvényes, hogy igyekszenek megadni a szaktárgyszó $2 forrását és az új rekordok esetén a $0 besorolási rekordazonosítót. A szaktárgyszavakhoz az OSZK-ban mindig megadják a forrást (besorolási rekordazonosító almező nincs a HUNMARC-ban).

Földrajzi nevek dolgában a szaktárgyszavakhoz hasonló a helyzet: vannak könyvtárak, melyekben aránylag gyakoriak, és a forrást is pontosan megadják. Érdekes, hogy a szaktárgyszavakkal és dokumentumtípussal szemben a földrajzi nevekhez az OSZK nem adja meg a forrást, pedig ezek is a Köztauruszból származnak.

A DNB, ÖNB, KBS, KBR és BL nagy arányban használ szabadon választott tárgyszavakat, ami némileg magyarázza, hogy ezekben a többi tárgyszótípus használata relatíve gyengébb. Még feltűnőbb azonban, hogy mennyire ritkán fordul elől, hogy megadják a $2 forrást.

A CSH és BNPL feltűnően sok esetben megadja a dokumentumtípust. Más könyvtárakban is feltűnik a gyakoriság, de csak a speciálisabb típusokat esetén. Az OSZK is jól áll, ráadásul rögzíti, hogy ezek forrása az OSZK által kezelt dokumentumtípus-tezaurusz.

Összegzés

Az OSZK helye

Az OSZK a tárgyalt adatmezők feldolgozásának dolgában a középmezőnyben, annak az alsó részében foglal helyet. Ez – a Nemzeti Könyvtár nemzetközi viszonylatban gyenge anyagi lehetőségeinek fényében – inkább meglepő. Összefügghet azzal, hogy munkatársai révén egyelőre jelentős hozzáértési szellemi tőke halmozódott fel. A könyvtár ezt éli fel.

A 00X mezők kitöltését alapértelmezéseket tartalmazó „sablonok” alapján végzik, amit ideje lenne fölülvizsgálni. Az alapértelmezéseket a könyvtári szoftver tartalmazza. Noha a többi nemzeti könyvtár alapértelmezései se közismertek (csak az adatbázisaik elemzése alapján állapíthatók meg), úttörő munka lenne az OSZK alapértelmezéseit az indoklásukkal együtt nyilvánosságra hozni.

CC BY

Kiemelkedő az aprónyomtatvány-, zeneműi- és térképtári feldolgozói munka. Erre, valamint a könyvek monografikus és az időszaki kiadványok feldolgozására egyaránt jellemző, hogy közös tárgyszórendszert, a Köztauruszt használják. Az összes magyarországi könyvtár által használható, 150.000 szemantikai szócikkekben elrendezett lexikai egységet tartalmazó Köztaurusz az ezredfordulón az összes online elérhető magyar tezaurusz és tárgyszójegyzék egyesítésével és egységesítésével készült. Követő tartalmi karbantartása egyelőre biztosítva van, szoftverének (Relex) a sorsa azonban nincs még megnyugtatóan rendezve.

A MARC21 a szemantikus web és a mesterséges intelligencia jövőbeli alkalmazhatóságát előkészítendő, bevezetett két új általános metaadat-azonosítót. A $0 metaadat-azonosító, $1 valós fizikai tárgy azonosítója. A nemzeti könyvtárak jelentős része már elkezdte használni ezen azonosítókat, ideje, hogy erre az OSZK-ban is sor kerüljön.

Az elemzés fontosabb konkrét tapasztalatai

A rekordfej és a vezérlő mezők (00X) pozícióin a dokumentumok típusainak egyetlenszer alkalmazható, egy karakteres, mesterséges nyelvű osztályozási jelzeteit rögzítik, amik pozícióként meghatározott szempont szerinti tipológiát képviselnek. Ez a jelzetrendszer csak nagyon általános, durva besorolást tesz lehetővé.

Az összehasonlítások alapján az egyes osztályok meghatározásai elégtelennek tűnnek. Például nem látszik teljesen világosan, hogy mi értendő alárendelt részegységnek kéziratok vagy plakátok esetén.

A pozíciók kitöltését feltehetően helyi szabályzatok vagy szóbeli gyakorlat határozza meg, ami a nyilvánosságának hiányában nagyban korlátozza az elemzésben felvetődő problémák átfogó megtárgyalását, konszenzusos megoldását, nehezíti a nemzetközi gyakorlat egységesülését. A MARC Tanácsadó Bizottság honlapján[3] jól követhetők a tételfejjel és a 00X adatmezőkkel kapcsolatos változtatások. Átfogó átdolgozásra irányuló igénynek, a 00X tartalmi típusrendszerének kritikájára vonatkozó törekvéseknek azonban nincs nyoma.

A felvetődő problémák és megoldásuk nincs ellentmondásban azzal, hogy lehetséges egy, a jelenleginél jobban kidolgozott, átfogó tipológiai osztályozás a vezérlő mezőkben, nem utolsósorban annak érdekében is, hogy a katalógusok keresőrendszereit az ezekben a mezőkben való felhasználóbarát keresésre felkészítsék.

Sok esetben a tételfej dokumentumtípusával, illetve a 007 mező dokumentumtípusaival párhuzamosan léteznek szöveges mezők is (mint a 254, 256, 300, 513, 516, 588), melyekben ugyancsak osztályozhatók dokumentumtípusok, a 655-ös mezőben kötött tárgyszavakkal is. Jelenleg nincs szabályozva ezeknek a mezőknek a viszonya, t.i. hogy melyiket milyen értelemben célszerű használni a feldolgozás egyértelműsége érdekében.

A vezérlőmezők pozícióin kis számban, de sok katalógusban fordulnak elő –feltehetően nem kellően ellenőrzött importból származó – definiálatlan értékek.

A 00X kódolt fizikai és meghatározott jellemzők a kereshetőség változatosságának lehetőségei. Általuk magától értető, átfogó tartalmi tájékozódás lehetséges. Főképp a keresés elején játszhatnának hasznos szerepet: a végfelhasználó számára kiderülhetne, milyen sokféle irányban indulhatna el dokumentum- és tartalomtípusok szerint. Egyszerű

3 https://www.loc.gov/marc/mac/index.html

megoldásokkal lehetne közelebb hozni a felhasználóhoz ezeket a lehetőségeket, melyek a Kongresszusi Könyvtárat kivéve, ma még a nagy nemzeti könyvtárakban sincsenek igazán kihasználva. Az LC keresőoldalán többek között dokumentumtípusok szerint lehet szűkíteni a keresést.

Végül: a katalógusok, repozitóriumok stb. metaadatsémáiba be kell építeni az adott intézményben kialakult/kidolgozott egyedi használati szokásrendjének tárolására és kezelésére alkalmas részt, úgy, hogy azon keresztül a katalogizálási rendszerbe beépített alapértelmezéseket a könyvtáros felhasználó tudja módosítani.

Irodalom

HUNMARC. A bibliográfiai rekordok adatcsere formátuma. KSZ 4/1. 2002. március.

Király Péter, Jakob Voß, et al. (2017-) QA catalogue. v0.7.0 (2023). https://doi.org/10.5281/zenodo.8159388

Király, Péter: Validating 126 million MARC records. = DATeCH2019 3rd Int. Conf. on Digital Access to Textual Cultural Heritage. ACM, 2019. pp. 161-168. https://doi.org/10.1145/3322905.3322929

MARC21 Format for Bibliographic data. Library of Congress. Update No. 30 (May 2020) https://www.loc.gov/marc/bibliographic/

Az OSZK tezaurusz és a KÖZTAURUSZ. = Könyvtári Figyelő, 2001. 01. – p. 11–40. http://ki.oszk.hu/kf/kfarchiv/2001/1/ungvary.html

Rob Styles, Dany Ayers, Nadeem Shabir: Semantic MARC, MARC21 and the Semantic Web. = WWW2008 Workshop on Linked Data on the Web. CEUR, 2008. https://ceur-ws.org/Vol-369/paper02.pdf

Tezauruszkezelő programok és a RELEX. = TMT 2001. január. – p. 3–16. http://tmt.omikk.bme.hu/show_news.html?id=1620&issue_id=26

Ungváry Rudolf: Besorolási, szabványosított, normatív vagy „autorizált". = TMT, 2019. 06. 24., 66.évf., 6. sz., p. 328–342. https://tmt.omikk.bme.hu/tmt/article/view/12309/14064

Ungváry Rudolf: MARC21 tartalmi adatmezők használata jelentősebb nagykönyvtárakban. Egy elemzés néhány tanulsága. = Networkshop (2020): 33-53. https://doi.org/10.31915/NWS.2020.4

Ungváry Rudolf: Ismeretszervező-könyvtári rendszerek tartalmi feltárásának összehasonlító vizsgálata MARC21 környezetben. = TMT, 2020. (67. évf.) 11. sz. pp. 655-680. https://tmt.omikk.bme.hu/tmt/article/view/12776/14514

Ungváry Rudolf, Király, Péter. Bemerkungen zu der Qualitätsbewertung von MARC-21-Datensätzen. = M. Franke-Maier, A. Kasprzik, A. Ledl, H. Schürmann (eds.) Qualität in der Inhaltserschließung. De Gruyter Saur. 2021. pp. 177-227. https://doi.org/10.1515/9783110691597-011

Ungváry Rudolf, Király, Péter. A MARC21 tételfejének és kódolt tartalmi jellemzőinek feldolgozási minősége néhány nemzeti könyvtárban. Egy elemzés tanulságai. = TMT, 2022. augusztus 8. https://doi.org/10.3311/tmt.13174

CC BY

DOI: 10.31915/NWS.2023.8

Kapocs a tudáshoz – A könyvtár szerepe a civilek és a tudomány kapcsolatában

Link to knowledge – The role of the library in the relationship between citizens and science

Szemes-Révész Enikő Evelin
Pécsi Tudományegyetem Egyetemi Könyvtár és Tudásközpont
revesz.eniko@lib.pte.hu

Absztrakt

A nagy mennyiségű adat- és információáramlás, valamint a társadalmi igények átalakulása következtében a könyvtárak számára elengedhetetlen az új kapcsolódási lehetőségek feltérképezése, melynek segítségével továbbra is nélkülözhetetlen szereplői maradnak a civil és tudományos közeg számára. A nyílt tudomány ernyője alá tartozó közösségi tudomány lehetőséget ad a könyvtár szerepkörének bővítésére, valamint társadalmi kohéziójának javítására. Az alábbiakban a közösségi tudomány fogalmi keretei és a benne rejlő lehetőségek kerülnek bemutatásra a PTE Egyetemi Könyvtár és Tudásközpont munkacsoportjának elmúlt két évben végzett tevékenysége, valamint a releváns nemzetközi szakirodalom áttekintésének segítségével.
Kulcsszavak: közösségi tudomány, nyílt tudomány, társadalom, társadalmi szerepvállalás

Abstract

With the influx of data and information and the changing needs of society, it is essential for libraries to explore new ways of connecting with each other, so that they can remain indispensable players in the civil and scientific community. Citizen science, under the umbrella of open science, is an opportunity to expand the role of the library and improve its social cohesion. In the following, the conceptual framework of citizen science and its potential are presented with the help of the activities of the working group of the University of Pécs Library and Knowledge Centre over the last two years and a review of the relevant international literature.
Keywords: citizen science, open science, society, citizen engagement

Bevezetés

Napjainkban a digitalizáció folyamatos fejlődése, valamint a mesterséges intelligencia megjelenése és használatának egyre szélesebb körben történő elterjedése következtében átalakul a társadalom információigénye, mely egyben egy folyamatosan változó tényező is. Egyrészt az internetnek köszönhetően lehetőség van az azonnali, naprakész tájékozódásra, másrészt pontosan ezen könnyen elérhető és gyorsan változó impulzusok befolyásolják a társadalmi igények folyamatos változását és alakulását. Ebben a helyzetben a könyvtárak számára is új szerepkörök megtalálására kell törekedni, mellyel fenn tudják tartani a társadalom ellátásában betöltött helyüket. Minderre jó lehetőséget nyújt a közösségi tudomány, melynek segítségével a könyvtárak lehetnek az összekötő kapocs a tudomány, azaz a hiteles információk forrása, és a társadalom, azaz az információt igénylők között.

A közösségi tudomány alapjai

A közösségi tudomány az angol „citizen science” megfelelője, melyet pontos fordítás esetén polgári, vagy állampolgári tudományként emlegethetnénk, ez azonban nem tükrözné a fogalom tényleges jelentését, mely szerint a kutatásokban nem csak a tudományos oldal képviselői, hanem a téma iránt érdeklődő, amatőr állampolgárok is részt vesznek elsősorban önkéntes alapon. Éppen ezért a közösségi tudomány kifejezés fedi le leginkább az angol fogalmi kereteket *[3]*. A közösségi tudomány a nyílt tudomány (open science) ernyője alá tartozik, mely egyre fontosabb szerepet tölt be az európai, így a magyarországi egyetemek és könyvtárak életében egyaránt. A közösségi tudomány fogalmának ismerete azonban Magyarországon jelenleg még gyerekcipőben jár, de látva az európai példákat, a nyílt tudományhoz kapcsolódó célkitűzések és projektek sikeréhez és megvalósításához elengedhetetlen a közösségi tudomány alkalmazása. Mivel a közösségi tudomány egyszerre segíti a kutatási eredményekhez és azok publikációihoz való nyílt hozzáférést, valamint a civilek aktív részvételét a kutatásokban, így eszközként, ugyanakkor célként is tekinthetünk rá a nyílt tudomány szempontjából *[1]*.

Az első közösségi tudományos projektek megjelenését már az 1900-as évektől számolják, azonban ekkor még más elnevezést használtak. A „citizen science” kifejezés első feljegyzett használata 1989-ben jelent meg *[6]*. Szintén meghatározó mérföldkő az Európai Civil Tudomány Egyesületének[1] 2014-ben történő létrejötte, melynek célja a tudomány demokratizálása, valamint a közösségi tudományos projektek számának növelése Európában, ezzel a lakosság minél szélesebb körben történő bevonása a tudományos kutatásokba. Az egyesület tíz pontban határozta meg a közösségi tudományos projektek jellemzőit, mely pontokkal jól körülhatárolható a közösségi tudomány fogalma is. Mindezt röviden bemutatva elmondható, hogy a közösségi tudományos projektnek „közösségi” oldalát a korábban már említett civil lakosság adja, akik az adott projektben való részvételükkel hozzájárulnak a tudományos kutatás megvalósulásához, az eredmények létrejöttéhez. Előnyös számukra a projektben való részvétel, hiszen beleláthatnak egy-egy tudományos munkába a kezdetektől egészen a projekt lezárásáig, valamint bővíthetik tudásukat akár gyakorlati szinten is. A közösségi tudomány lényege, hogy cselekvő módon vonja be a lakosságot a projekt minden egyes szakaszába, továbbá, hogy kikérje véleményét az egyes részek megvalósítása során ezzel is biztosítva annak érthetőségét, átláthatóságát. A projekt „tudományos” oldalán az adott tudományterület kutatói állnak, akik számára elengedhetetlen a civilek részvétele a kutatási projektben. Számukra előnyös a kutatás ilyen módon történő megvalósítása, hiszen ezáltal nyitni tudnak a társadalom felé a tudományos témakörökben, ezzel elősegítve a hiteles tájékoztatást és tájékozódást, továbbá megismerhetik a civil oldal nézőpontját a kutatott problémával kapcsolatban, ezzel akár új perspektívát azonosítva a kutatási probléma megközelítéséhez. Mindezek alapján tehát elmondhatjuk, hogy a közös munkával mindkét fél nyer. Ezek a projektek tudományos eredménnyel bírnak, a projektek adatai és metaadatai pedig nyilvánosan elérhetőek. Fontos továbbá, hogy az eredmények bármilyen típusú közlése során mindig meg kell említeni a projekt civil résztvevőit, hiszen nélkülük az nem valósulhatott volna meg *[2,6]*. A közösségi tudományos projektekre számos példát láthatunk már, de elsősorban a természettudományok területén találkozhatunk ilyesfajta kutatásokkal, ahol főleg applikációk segítségével kérnek adatokat a lakosságtól például szúnyogokra, kullancsokra, vagy épp vízfolyásokra vonatkozóan.

1 European Citizen Science Association (ECSA)

CC BY

A könyvtárak szerepe a közösségi tudományban

Az amerikai lakosság körében 2008-ban és 2018-ban is végeztek felmérést, melyben a könyvtárak megítélését vizsgálták. Az eredmények[2] alapján elmondható, hogy a válaszolók az olyan hagyományosnak tekinthető tulajdonságok, mint a csendes, tanulásra alkalmas terek, megfelelő internetkapcsolat mellett egyre inkább közösségi helyként tekintenek a könyvtárakra, melyek ezzel olyan találkozási pontokká válnak a közösség tagjai számára, amelyek számos egyedi programlehetőséget nyújtanak *[5]*. A könyvtárak részéről pedig éppen ez a közösségi hely funkció az, ami kitörési lehetőséget ad a digitalizáció és modernizáció okozta korlátokból, és segít megtartani a könyvtárak társadalomban betöltött funkcióit. A közösségi tudomány pedig hozzájárul ennek a funkciónak a kiteljesedéséhez azáltal, hogy kihasználhatóvá teszi a könyvtár versenyelőnyt jelentő tulajdonságait. Egyrészt a könyvtár kiváló találkozási pont a tudományos és a civil oldal képviselői számára, hiszen mindkét fél által egyformán elérhető, szabadon látogatható, és ez a fajta szabadság és elérhetőség adja a könyvtár előnyét, hiszen semleges térként funkcionál, mivel sem a tudományos oldal képviselői, sem pedig a civilek nem éreznek frusztrációt, ha a könyvtári terekben kell találkozniuk, mivel mindkettőjük számára ismerős, otthonos a terep. Mindemellett a könyvtárak képesek kiszolgálni mindkét fél igényeit, ezáltal kellő tájékoztatást nyújtva a tudományos és hétköznapi témákban egyaránt. Tehát a könyvtár egyfajta kapocs szerepet tud betölteni a kutatások, innovációk és a társadalom különböző tagjai között. Ez a fajta kapocs vagy híd szerep pedig a közösségi tudományos projektekben elengedhetetlen, mely szerep betöltéséhez szükséges kompetenciákkal és tudással rendelkeznek a könyvtárak. Egyrészt adott a könyvtárosok szakmai tudása és kommunikációs készsége, mellyel támogatni tudják egy-egy projekt során a kutatókat, továbbá adott a könyvtár kapcsolati hálója a civil lakosság felé, hiszen a könyvtárba betérő olvasók könnyen elérhetőek és tájékoztathatóak az adott projektről. Mindemellett a könyvtár rendelkezik azokkal a fizikai terekkel, melyek kihasználhatóak a projekt során megvalósuló előadások, konferenciák, tájékoztatók alkalmával, ezzel is erősítve a könyvtár harmadik hely funkcióját. Mindemellett pedig lehetőséget nyújt a kutatókkal történő szorosabb együttműködésre a könyvtárak adatgyűjtési és tárolási tevékenysége és tapasztalata, mellyel szintén úgy tudják támogatni az adott kutatási projektet, hogy az nem jár többletberuházással *[1,3,4]*.

A PTE Egyetemi Könyvtár és Tudásközpont munkacsoportjának tevékenysége

A cikk ezen része néhány gyakorlati elemmel kívánja szemléltetni, hogy a könyvtár hogyan tudja alkalmazni a rendelkezésére álló kompetenciákat és erőforrásokat a közösségi tudományos tevékenység kialakítására és fejlesztésére.

A PTE Egyetemi Könyvtár és Tudásközpont közösségi tudománnyal foglalkozó munkacsoportja 2021-ben jött létre azzal a céllal, hogy feltérképezze a nyílt tudományhoz kapcsolódó közösségi tudományban rejlő potenciális lehetőségeket. A nemzetközi példákat áttekintve látható volt, hogy a Könyvtár jó pozícióban van abból a szempontból, hogy egyetemi könyvtárként adott számára a tudományos oldal képviselőivel a kapcsolat, mindeközben pedig, mivel a társadalom minden tagja által szabadon látogathatóak egységei, így a civil lakossággal való kapcsolatteremtésre is megvannak a feltételei. A közösségi tudományos tevékenység megvalósítására számos lehetőség nyílik. A PTE Egyetemi Könyvtár és Tudásközpont célul tűzte ki, hogy felkutassa elsőként az Egyetem karain és intézeteiben működő olyan projekteket, melyek azonosítottan közösségi tudományos projektek, illetve

2 OCLC and American Library Association (2018): From Awareness to Funding: Voter Perceptions and Support of Public Libraries in 2018. *Dublin*, OH: OCLC. https://doi.org/10.25333/C3M92X.

azokat is, melyek tagjai ugyan még nem ismerik a közösségi tudomány fogalmát, így nem azonosították ilyen szinten a kutatótevékenységüket, azonban megfelelnek ezen fogalmi kereteknek. Ezt követően pedig már lehetőség van arra, hogy a városban és a régióban is felkeresse az ilyen típusú projekteket. A Könyvtár a meglévő erőforrásainak hatékonyabb kihasználásával törekszik a projektek támogatására. Így elsősorban marketingtevékenységet nyújt a projektek számára, mely keretei között a honlapján létrehozott egy közösségi tudományos aloldalt[3], ahol megtalálhatóak a projektek részletei, továbbá közösségi média felületein és blogoldalain rendszeresen népszerűsíti partnerei tevékenységét, eseményeit és célkitűzéseit. Ez a fajta támogató tevékenység kiegészül közös programok szervezésével, illetve a partnerek programjai számára helyszín biztosításával a könyvtári terekben, ezzel is megteremtve a lehetőséget, hogy minél több könyvtárhasználóhoz eljusson a közösségi tudományos projektekben való részvételi lehetőség.

Konklúzió

Összességében elmondható, hogy a digitalizáció és a modernizáció térnyerése következtében a könyvtárak számára kitörési lehetőséget nyújt a nyílt tudományhoz tartozó közösségi tudomány, melynek segítségével a könyvtárak tölthetik be a kapocs, híd szerepét a tudományos oldal és a civil társadalom képviselői között ezzel is támogatva harmadik missziós tevékenységét és megtartva pozícióját a társadalmi igények kiszolgálása területén. A közösségi tudomány és az ahhoz kapcsolódó projektek lehetőséget adnak a könyvtárak számára, hogy meglévő erőforrásaik és kompetenciáik felhasználásával építsenek ki partnerkapcsolatokat a közösségi tudományos projektek résztvevőivel, ezáltal bővítve szolgáltatásaikat és szerepkörüket, valamint növelve a könyvtárhasználók elégedettségét, akik részéről egyre inkább megjelenik az igény, hogy a könyvtár találkozási hely funkciót töltsön be, tehát a hagyományos szolgáltatások mellett lehetőséget biztosítson a közösségépítésre is. A lehetőségek adottak, csupán minden könyvtárnak meg kell találnia azt a pontot, ahol be tud csatlakozni a közösségi tudományos tevékenységbe.

Felhasznált irodalom

[1] Cigarini A. et. al. (2021): Public libraries embrace citizen science: Strengths and challenges. *Library and Information Science Research*, 43. évf. 2. sz. 101090. https://doi.org/10.1016/j.lisr.2021.101090 (2023.06.14.)

[2] European Citizen Science Association (2015): 10 Principles of Citizen Science. https://www.ecsa.ngo/documents/ (2023.06.13.)

[3] Gaálné Kalydy Dóra (2020): Civilek a kutatásban- közösségi tudomány a könyvtárban. *Könyvtári figyelő*, 30. (66.) évf. 1. sz. pp.54-57. https://epa.oszk.hu/00100/00143/00359/pdf/ (2023.06.13.)

[4] Ignat T.- Cavalier D.- Nickerson C. (2019): Citizen science und bibliotheken: Walzer tanzen auf dem weg zur zusammenarbeit. *VOEB-Mitteilungen*, 72. évf. 2. sz. pp. 328-336. https://doi.org/10.31263/voebm.v72i2.3047 (2023.06.14.)

[5] OCLC and American Library Association (2018*)*: From Awareness to Funding: Voter Perceptions and Support of Public Libraries in 2018. Dublin, OH: OCLC. https://doi.org/10.25333/C3M92X (2023.06.14.)

[6] Vohland, K. et al. (2021): Editorial: The Science of Citizen Science Evolves. In: Vohland, K., et al. The Science of Citizen Science. Springer, Cham. https://doi.org/10.1007/978-3-030-58278-4_1 (2023.06.13.)

3 https://www.lib.pte.hu/hu/service/citizen_science-273

CC BY
III.07 DOI: 10.31915/NWS.2023.9

Az RO-Crate alapú kutatási objektum csomagolás keretrendszere az ELKH ARP platformban

The framework of research object packaging based on RO-Crate in the ELKH ARP platform

Tóth Zoltán
SZTAKI – DSD
toth.zoltan@sztaki.hu

Absztrakt

A FAIR irányelveknek való megfelelés több kutatási területen megkerülhetetlen tényezővé kezd válni, és ezzel a magyar kutatók is egyre gyakrabban szembesülnek publikációs tevékenységük kapcsán. Ezek az irányelvek alapvetően azt célozzák meg, hogy a kutatásokat alátámasztó adatok megtalálhatóak és feldolgozhatóak lehessenek számítástechnikai eszközökkel, akár emberi beavatkozás nélkül is. Előremutató lépés a kutatási adatok hagyományos adatrepozitóriumban történő tárolása és metaadatolása, ez azonban nem feltétlenül elegendő az irányelvek követéséhez, ugyanis az ott elvárt metaadatok jellemzően a repozitóriumba feltöltött adatcsomag egységekre vonatkoznak, a finomabb, akár fájl-szintű értelmezésnek nincsen ezekben a rendszerekben szabványosított, bevett módja. A FAIR Digitális Objektumok megoldást jelenthetnek erre a problémára. Ennek egyik lehetséges megvalósítása az RO-Crate kutatási objektum csomagolás, melyet az ELKH ARP (ELKH Adat Repozitórium Platform) projekt bevezet az Adat Repozitórium Platformba. Ismertetésre kerül az RO-Crate formátum, valamint az, hogy ezzel a kutatók mi módon találkoznak majd a repozitóriumban végzett tevékenységeik során.
Kulcsszavak: FAIR irányelvek, adatrepozitórium, FAIR Digitális Objektum, RO-Crate, metaadat

Abstract

Compliance with FAIR principles is becoming an unavoidable factor in multiple research areas, that Hungarian researchers also face in their publishing activities. These principles primarily aim to ensure that the data supporting research can be located and processed using computational tools, even without human intervention. Storing and providing metadata for research data in a traditional data repository is a progressive step. However, this alone may not be sufficient to adhere to the guidelines, as the expected metadata in these systems typically pertain to the units of data packages uploaded to the repository, and there is no standardized, established way to interpret finer details, such as file-level information. FAIR Digital Objects can provide a solution to this problem. One possible implementation is the research object packaging based on RO-Crate, which the ELKH ARP (ELKH Data Repository Platform) project introduces into the Data Repository Platform. The RO-Crate format will be presented, as well as how researchers will encounter it in their activities within the repository.
Keywords: FAIR principles, data repository, FAIR Digital Object, RO-Crate, metadata

Bevezetés

A FAIR irányelvek [1], elsősorban az OpenScience kapcsán, egyre inkább előtérbe kerülnek. A betűszó a Findable, Accessible, Interoperable és Reusable angol szavakból áll

össze, és kicsit egyszerűsítve azt írják le, hogy a kutatási adatoknak ahhoz, hogy a kutatás tisztasága és ellenőrizhetősége biztosítva legyen, megtalálhatónak, egyszerű eszközökkel hozzáférhetőnek és feldolgozhatónak kell lenniük, valamint a kutatási eredményekhez vezető út ellenőrizhető módon megismételhető kell legyen. Belátható, hogy az OpenScience kezdeményezés örömmel tűzte zászlajára ezeket az elveket, hiszen ezek leegyszerűsítik a már publikálásra került adatok feldolgozását minden hozzáértő személy számára (még akkor is, ha a FAIR irányelvek követése nem garantálja a kutatási adatok közkinccsé tételét [3]). A FAIR irányelvek kapcsán egy fontos kitételt meg kell még említeni: az irányelveknek való megfelelést nem csak emberi feldolgozás esetén kell teljesíteni, hanem automatikus eszközök számára is biztosítani kell a hozzáférhetőséget és esetenként az automatikus feldolgozhatóságot is. Ez utóbbi kitétel megvalósulása viszont olyan szemantikus címkézését feltételezi a kutatási adatoknak, ami a jelenleg használt adatrepozitóriumokban nem, vagy csak nagyon sok kompromisszum árán valósulhat meg. Az ELKH ARP [13] repozitóriumában bevezettük az RO-Crate formátum olyan támogatását, amivel az adatrepozitóriumok ezen hiányossága áthidalható.

FAIR digitális objektumok

Az Európai Unió egy akciótervben 2018-ban bevezette a FAIR Digitális Objektumok fogalmát [2]. A FAIR Digitális Objektumok (FAIR Digital Object - FDO) olyan digitális objektumok, amik adott környezetben megvalósítják a FAIR irányelveket: „Az adatok, szoftver és más erőforrások reprezentációja.”... „Társítva vannak hozzá perzisztens azonosítók, metaadatok és kontextuális dokumentáció, ami lehetővé teszi a felderíthetőséget, idézést és újrahasznosítást.”

Ez a definíció, ismételten csak kicsit egyszerűsítve azt jelenti, hogy az FDO-nak az adott digitális környezetben meg kell tudnia mondania magáról mind emberek, mind pedig automatikus feldolgozó eszközök számára, hogy mi is valójában. Az újrahasznosítás és reprodukálhatóság kritériuma miatt olyan szintű metaadatolást kell tudni biztosítani, amivel formálisan meghatározhatóak a kutatási adatok feldolgozásához szükséges lépések, valamint megjelölhető mind a forrásadatok, mind pedig a feldolgozás eredménye is. Már ez a kritérium is olyan terhet ró az egyszerű generikus adatrepozitóriumokra, ami nehezen teljesíthető, ott ugyanis jellemzően a repozitálás és formális metaadatolás szintje az adatcsomag, amiben az egyes fájlokról nehéz megállapításokat tenni.

RO-Crate

Az RO-Crate (Research Object Crate) [4] csomagolástechnika az FDO egy lehetséges megvalósítása, amit az ELKH ARP projekt során kiválasztottunk a FAIR irányelvek támogatására. Kifejlesztése során a Kutatási Objektumok (Research Object - RO [11]) alapelveket ötvözték a DataCrate [5] csomagolással.
Az RO alapelvek előtérbe helyezik az azonosíthatóságot, az aggregációt és az annotációt:

- Azonosíthatóság
  Minden egyes objektumnak valamilyen egyedi azonosítója van.
- Aggregáció
  A kutatások eredménye nem csak maga a publikáció, hanem hozzá kell érteni a kutatás teljes folyamatában mindent, a forrásadatoktól a köztes lépéseken át a végső konklúzióig.
- Annotáció
  Mindenről, ami része a Kutatási Objektumnak, szemantikusan meg kell tudni mondani, hogy mi. Ez az eredeti Kutatási Objektum koncepció esetén Schema.org szemantikus annotáció társítását jelentette az egyes objektumokhoz.

A DataCrate egy csomagolástechnológia, ami az adatfájlok egységbe foglalását, tömörített tárolását (Bagit technológia [6], illetve JSON-LD [7] formátumú metaadatleírás), és szintén Schema.org annotációját jelentette.

CC BY

A RO-Crate tehát az RO és DataCrate egyfajta evolúciója. Ez az eredeti RO elvekhez képest praktikus egyszerűsítéseket tartalmaz (elegendő csak a kutatás célirányos leírásához szükséges fájlokat/objektumokat megfelelően annotálni), valamint nem ragaszkodik kifejezetten a Schema.org annotációkhoz, bármilyen publikus séma szerinti annotációt megenged a szemantikus leírásokhoz. Technikailag egy tömörített hierarchikus fájl-halmazt jelent, melyet egy JSON-LD leírás lát el akár fájl szintű metaadatokkal. Koncepcionálisan a következő elemekből áll:

- Adat entitások
  - Könyvtárszerkezet (kezdve egy lokális root elemmel);
  - Fájlok ebben a könyvtárszerkezetben;
  - Távoli URI-kkal beazonosítható objektumok.
- Kontextuális entitások
  - Olyan entitások, amik a digitális világon kívül is léteznek (pl. emberek, helyek);
  - Elsősorban metaadat formájában létező leírások (pl. geokoordináták).
- JSON-LD leírás
  - Összekapcsolja az adat és kontextuális entitásokat valamilyen publikus séma szerint tipizálva azokat.

Az RO-Crate felépítése a következő [8]:

```
<RO-Crate gyökér könyvtár>/
|  ro-crate-metadata.json    # RO-Crate Metadata Fájl – kötelező elem
|  ro-crate-preview.html     # RO-Crate Website honlap – javasolt elem
|  ro-crate-preview_files/   # Javasolt elem(ek)
|   | [other RO-Crate honlap fájlok]
|  [fájlok és könyvtárak]  # 0 vagy több
```

Tipikus felhasználási esetben adott kutatás adatai egy hierarchikus fájlrendszerbe kerülnek, ehhez kapcsolódnak a teljes adathalmazt leíró metaadatok (RO-Crate gyökér szintű metaadatleírás), illetve az egyes könyvtárakat, fájlokat leíró metaadatok, valamint olyan erőforrás-leírások, amik URI-kon keresztül beazonosíthatóak és nem kerülnek közvetlenül bele a kutatási adatok közé.

RO-Crate JSON-LD leírás

A kutatási adatok valamilyen hierarchiába rendezése/rendeződése gondos adatmenedzsment, tervezés útján magától is kialakulhat, és nem különösebben különbözik attól, ahogy egy generikus adatrepozitóriumba történő feltöltés során az adatok formáját és hierarchiáját elképzelhetjük. Az annotáció formája viszont már jelentős hozzáadott értéket képvisel ehhez az alapkoncepcióhoz képest. Ez az annotáció az RO-Crate gyökér könyvtárában található ro-crate-metadata.json fájlban valósul meg. Ez egy egyszerű flat JSON-LD leírás (azaz az adat és kontextuális entitások vektorszerűen vannak felsorolva benne szerializálva, egyedi azonosítókkal ellátva, és a hierarchikus struktúrát ezeknek az objektumoknak a leírása, és egymásra történő hivatkozása adja). Ettől az egyszerűsített leírástól az RO-Crate technológia fejlesztői a konkrét implementációk megkönnyítését remélik. A fejlesztők továbbá azt a megkötést is alkalmazták, hogy bár bármilyen ontológia szerint annotálhatóak a leírásban található entitások, de az annotáció URI-ja mellett azok címkéjének a megadása is kötelező. Ez azt eredményezi, hogy külön navigáció nélkül is értelmezhető az RO-Crate-ek tartalma emberi olvasók számára is.

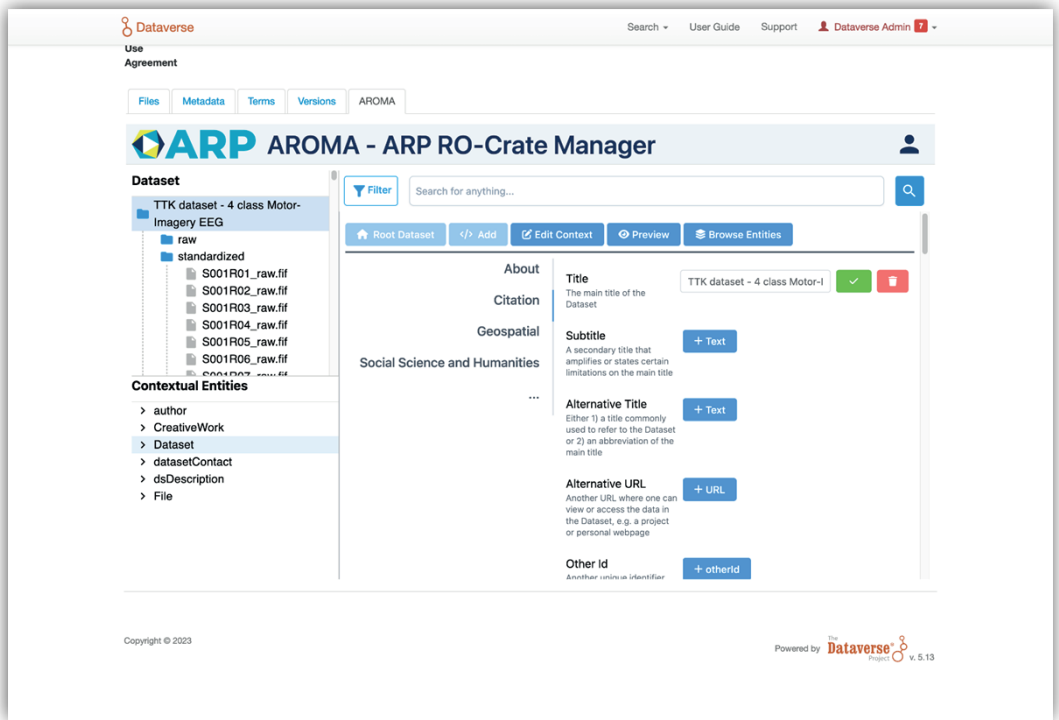
RO-Crate csomagok támogatása az ELKH ARP repozitóriumában

A generikus repozitóriumok lehetőséget adnak, hogy abban a felhasználók adatcsomagokat hozzanak létre, az adatcsomagokban fájlokat helyezzenek el (akár valamilyen hierarchiába rendezve azokat), és az adatcsomag egészére metaadatleírást adjanak. Ebben az értelemben az RO-Crate adatcsomagok, mint fájlok is elhelyezhetőek bennük. Az RO-Crate adatcsomagok gyökér szintű metaadatleírása vonatkozik a teljes adatcsomagra, azaz ez pontosan megfeleltethető lehetne a repozitóriumokba kerülő adatcsomagok metaadatainak. A jelenleg használt adatrepozitóriumokba RO-Crate csomagokat feltöltve a fájl szintű metaadatok nem kerülnek értelmezésre (az RO-Crate adatcsomag egyetlen fájlnak fog látszani), az RO-Crate gyökér szintű metaadat leírása pedig teljesen különválik az adatcsomag metaadatleírásától. Az ELKH ARP repozitóriuma ezekre a hiányosságokra, valamint a fájl szintű metaadatok kereshetőségére ad megoldást.

A mi értelmezésünkben egy RO-Crate adatcsomag megfelel egy repozitóriumi adatcsomagnak, azaz egy-egy kutatás aggregált, publikálásra szánt eredményének. Az ELKH ARP repozitóriumának alapja egy Dataverse adatrepozitórium [9], és ebben a szokásos repozitóriumi funkciók mellett lehetőséget biztosítunk RO-Crate adatcsomagok importjára, exportjára, valamint helyben történő szerkesztésére is.

Import során létrejön egy olyan repozitóriumi adatcsomag, aminek az adatcsomag-szintű metaadatai kitöltésre kerülnek az RO-Crate gyökér szintű metaadataival. Ennek feltétele, hogy a metaadatolásra használt séma az általunk üzemeltetett, és egyébiránt a felhasználóink által is bővíthető sémaregiszterből (CEDAR [10]) származzon. Az importáláskor kibontásra kerül az RO-Crate, és létrejön a fájlok szokásos könyvtárhierarchiája a repozitóriumban. Ezzel együtt egy külön szerkeszthető objektumként megtekinthető az RO-Crate metaadatleírás, melyben elvégezhető a fájl-szintű metaadatolás.

Akár RO-Crate-ként, akár egyszerű repozitóriumi műveletekkel lett létrehozva egy-egy adatcsomag, annak az RO-Crate jellegű exportja szintén megoldott a rendszerből. A leírás megengedő ebből a szempontból, azaz ha nem történt meg a fájl szintű metaadatolás, az egyébként kötelező RO-Crate gyökér-szintű metaadatolás akkor is létrejön az adatcsomag metaadataiból, és ez kerül bele az RO-Crate JSON-LD leírásába.



1. ábra: A Dataverse installációba integrált ARP RO-Crate Manager

CC BY

A repozitórimban lehetőséget biztosítunk az RO-Crate metaadatok bármilyen szintű szerkesztésére az általunk fejlesztett AROMA (ARP RO-Crate Manager) komponens segítségével (1. ábra). Ez egy faszerű nézetet biztosít az adott adatcsomaghoz, melyben az egyes elemekhez megadhatóak a megfelelő metaadatok. Amellett, hogy így fájl szintű metaadatok megadása is lehetséges, az is megoldott, hogy a RO-Crate gyökér szintű metaadatainak módosítása közvetlenül módosítja a Dataverse adatcsomag szintű metaadatleírásokat is, illetve ez fordítva is megtörténik, azaz a Dataverse adatcsomagleírás változtatása közvetlenül módosítja a kapcsolódó RO-Crate metaadatokat.

A metaadatokban történő keresés a Dataverse lehetőségeinek felel meg a repozitórium felületén belül, azaz közvetlenül kereshetőek a teljes adatcsomagra vonatkozó metaadatok. Ezen felül viszont a platform részét képezi egy keresőfelület (közös kereső), ami tudásgráffá konvertálja az RO-Crate metaadatleírásokat, és amin keresztül kereshetővé válnak a fájl szintű metaadatok is.

Felhasználói esetleírás

Az ELKH ARP repozitóriumának előzménye a SZTAKI DSD (Számítástechnikai és Automatizálási Kutatóintézet - Elosztott Rendszerek Osztály) által üzemeltetett Dataverse alapú CONCORDA (Concentrated Cooperation on Research Data) adatrepozitórium [12]. Ebben felmerült az a felhasználói igény, hogy egy kutatás során létrejött nagy mennyiségű képi adatot kereshető formában el lehessen látni geokoordinátákkal. A geokoordináták társítására lehetőséget biztosít a Dataverse is, de csak adatcsomag szinten. Az alkalmazott “rossz gyakorlat” az volt, hogy felhasználóink egy fájlos adatcsomagokat hoztak létre, melyekhez adatcsomag szinten társították a kívánt adatokat. Ettől százas nagyságrendben jöttek létre olyan adatcsomagok, amik hivatkozása publikáció esetén meglehetősen nehézkes.

Ez a “rossz gyakorlat” teljes egészében kiváltható az új rendszerben az RO-Crate ábrázolással. Itt már lehetséges a publikálásnak megfelelő fájl-aggregáció, majd a fájlok egyenként történő metaadatolása, azaz a képállomány ellátása a megfelelő metaadatokkal. A kereshetőséget a geokoordinákra a közös kereső felülete biztosítja.

Konklúzió

A FAIR irányelvek követése a repozitóriumokkal szemben új elvárásokat támaszt. Az ezeknek való megfelelés új eszközkészletet igényel, és egy ilyen eszköz az FDO-k implementálása, az RO-Crate csomagolás bevezetése. Ezt, valamint a keresést segítő infrastruktúrát vezeti be az ELKH ARP repozitóriuma a kutatási adatok kezelésére Magyarországon, amivel világszinten is előremutató szolgáltatás jön létre. Az ELKH ARP platform jelenleg fejlesztési és tesztelési fázisban van, a fejlesztés várható befejezése 2023. december vége.

Bibliográfia

[1] Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

[2] European Commission, Directorate-General for Research and Innovation, Turning FAIR into reality : final report and action plan from the European Commission expert group on FAIR data, Publications Office, 2018, https://data.europa.eu/doi/10.2777/1524

[3] Putu Hadi Purnama Jati, Yi Lin, Sara Nodehi, Dwy Bagus Cahyono, Mirjam van Reisen; FAIR Versus Open Data: A Comparison of Objectives and Principles. Data Intelligence 2022; 4 (4): 867–881. doi: https://doi.org/10.1162/dint_a_00176

[4] Stian Soiland-Reyes, Peter Sefton, Mercè Crosas, Leyla Jael Castro, Frederik Coppens, José M. Fernández, Daniel Garijo, Björn Grüning, Marco La Rosa, Simone Leo, Eoghan

Ó Carragáin, Marc Portier, Ana Trisovic, RO-Crate Community, Paul Groth, Carole Goble (2022): Packaging research artefacts with RO-Crate. Data Science 5(2) https://doi.org/10.3233/DS-210053
[5] The DataCrate specification for packaging research data, https://github.com/UTS-eResearch/datacrate (letöltve: 2023.06.18.)
[6] Kunze, J., Littman, J., Madden, E., Scancella, J., and C. Adams, The BagIt File Packaging Format (V1.0), RFC 8493, DOI 10.17487/RFC8493, October 2018
[7] JSON for Linking Data, https://json-ld.org/ (letöltve: 2023.06.18.)
[8] RO-Crate Structure, https://www.researchobject.org/ro-crate/1.1/structure.html (letöltve: 2023.06.18.)
[9] Harvard Dataverse Repository, https://dataverse.harvard.edu/ (letöltve: 2023.06.18.)
[10] CEDAR - Center for Expanded Data Annotation and Retrieval, https://more.metadatacenter.org/ (letöltve: 2023.06.18.)
[11] Bechhofer, S., De Roure, D., Gamble, M. et al. Research Objects: Towards Exchange and Reuse of Digital Knowledge. Nat Prec (2010). https://doi.org/10.1038/npre.2010.4626.1
[12] CONCORDA - Concentrated Cooperation on Research Data, https://concorda.hu/ (letöltve: 2023.06.18.)
[13] ELKH Adatrepozitórium Platform, https://science-research-data.hu/ (letöltve: 2023.06.18.)

CC BY

DOI: 10.31915/NWS.2023.10

Neurális hálózatok oktatási alkalmazását támogató keretrendszer Virtual (VR) és Augmented Reality (AR) eszközökkel

Framework for creating, working and teaching Artificial Neural Networks using augmented reality (AR) and virtual reality (VR) tools

Király Roland, Király Sándor, Palotai Martin Marcell
Eszterházy Károly Katolikus Egyetem, Informatikai Kar
kiraly.roland@uni-eszterhazy.hu, kiraly.sandor@uni-eszterhazy.hu,
palotaimartinm@gmail.com

Absztrakt

A mesterséges intelligencia alapú rendszerek nagyjából 2010-es évek második felétől kezdődően, az informatikai fejlesztések elválaszthatatlan részévé váltak. Az ipari szereplők mindegyike valamilyen MI (Mesterséges Intelligencia) alapú fejlesztésbe kezdett, vagy olyan usecase-eket írt elő a munkatársainak, amelyek kapcsolódnak a területhez. Ebben a környezetben az oktatás szereplőinek is fel kell készülniük a gazdasági folyamatok támogatására és számolniuk kell az MI és a hozzá kapcsolódó munkaterületek szakembereinek a képzésével. A hatékony oktatási tevékenység megvalósításához tehát hamarosan szükség lesz olyan módszerek és szoftverek kidolgozására, amelyek felgyorsítják a neurális hálózati modellek programozását kipróbálását és tesztelését. Mindezért egy egyszerűen paraméterezhető, széles körben használható és oktatási célokra is alkalmazható eszközt készítettünk, amelyet szeretnénk továbbfejleszteni egy a nagyközönség, valamint oktatási intézmények számára is használható szoftvercsomaggá. A célközönségünk olyan közép és felsőoktatási intézmények, amelyek a neurális hálózatok használatának tanítását tűzték ki célul. A munkánk során elkészített keretrendszer amellett, hogy a neurális hálózatok konstruálására és tanítására is tartalmaz modulokat, segítségével a rendszerben létrehozott neurális hálózatokat vizualizálni lehet. A vizuális megjelenítés – és természetesen az interakciók – Augmented Reality (AR) és Virtual Reality (VR) környezetben valósulnak meg, vagyis a virtuális teret használjuk a hálózatok megjelenítésére és kezelésére.
Kulcsszavak: Neurális hálózatok oktatása, Virtual Reality, Augmented Reality

Abstract

Deep learning is a very popular topic in the computer sciences courses though it is often challenging for beginners to take their first step due to the complexity of understanding and applying Artificial Neural Networks (ANN). We present SNet, a framework for creating and training neural networks for solving different problems of real life and also for research. The visual presentation - and of course the interactions – of ANNs created in our framework can be visualised takes place in an Augmented Reality (AR) and Virtual Reality (VR) environment thus we use virtual space to display and manage networks. The system graphically monitors the neural network and as an edge-labelled directed graph can also display it. Assessing the impact of the AR/VR experience via a formative test and survey revealed that students overwhelmingly reported positively on the engaging nature and interactivity of AR/VR.

Bevezetés

A neurális hálózatok matematikai modellje, grafikus megjelenése és a tudományos kutatásokban, ideértve az orvosi és biológiai alapú felhasználást, az élelmiszerekkel kapcsolatos tudományokat, valamint a vírusok terjedésével kapcsolatos kutatásokat is, egyre nagyobb százalékban jelenik meg a neurális hálózatok használatának igénye. Ez a tudományterület tehát erősen inter- és multidiszciplináris skilleket. Sajnos a tudományos munkát végző szakemberek a fentebb említett tudományterületeken az informatikai tudás hiányában, és a neurális hálózatokat bemutató vizuális toolok használata nélkül csak elméleti szinten rendelkeznek ilyen ismeretekkel. A másik probléma az utánpótlás képzése. Ahhoz, hogy a fiatal kutatók tapasztalattal is rendelkezzenek a neurális hálózatokról és azok működéséről, szükségük van már alap, vagy középszintű oktatásuk során is arra, hogy a neurális hálózatok működését megértsék. A tudományos háttér ismeretével az oktatási tevékenység során nem minden esetben számolhatunk, így a vizuális tanulási képességeket sem tudjuk elméleti ismeretek elsajátítására felhasználni a megfelelő eszközök és módszerek hiányában. Ahhoz tehát, hogy a fenti célokat megvalósítsuk, valós időben és térben elérhetővé kell tenni a „virtuális világ azon részét, ahol a neurális hálók futnak". Ehhez az eszközünk az, hogy a neurális hálózatokat a virtuális térben megjelenítve és kezelő felületet biztosítva a használatukhoz szó szerint "kézzel foghatóvá" tesszük. Több évtizedes felsőoktatási és kutatási tapasztalataink azt mutatják, hogy a neurális hálózatok és a hozzájuk szorosan kapcsolódó deeplearning (A. G. Abulrub et al., 2011; Wadee Alhalabi. 2016) modellek elsajátítása még a viszonylag magasan képzett szakemberek számára is okozhat nehézségeket. Az oktatásban a műszaki területeken is jelentkező probléma, hogy a mélytanulási algoritmusok, valamint a neurális hálózatok oktatására nincsenek kidolgozva hatékony módszerek, vagy ha vannak is, azok nem terjedtek el széles körben.

A tudásmegosztás legtöbb esetben a matematikai modellek ismertetésén alapul. A neurális hálózatokat felrajzolják a táblára, vagy statikus prezentációkat elemeznek és az ábrákat kiegészítve magyarázatokkal, megpróbálják azok működését bemutatni.

Műszaki területeken a legtöbb esetben előkerülnek a programkódok. A Python (VanRossum, Guido, and Fred L. Drake, 2010) programozási nyelven implementált algoritmusok hatékonyan segítenek a modellek megértésében és futtatásában, de csak olyanok számára, akik képesek programokat olvasni és értelmezni, valamint rendelkeznek némi rutinnal a programírásban.
A Python nyelvhez készült TensorFlow (M. Abadi, A. Agarwal, 2016) és Keras (Chollet, Francois, 2018) nevű csomagok alkalmasak neurális hálózatok készítésére, és azok betanítására. Tartalmaznak tanító halmazokat, amelyek előre elkészített bemenő adatok használatát teszik lehetővé bárki számára, így nagyon hatékony módszereket adnak a programozók és az oktatók számára a neurális hálózatok tanítására és megismerésére.

Sajnos azonban az említett eszközöket nem lehet sikeresen alkalmazni az orvosi, művészeti, vagy természettudományi szakterületeken általános oktatási eszközként.

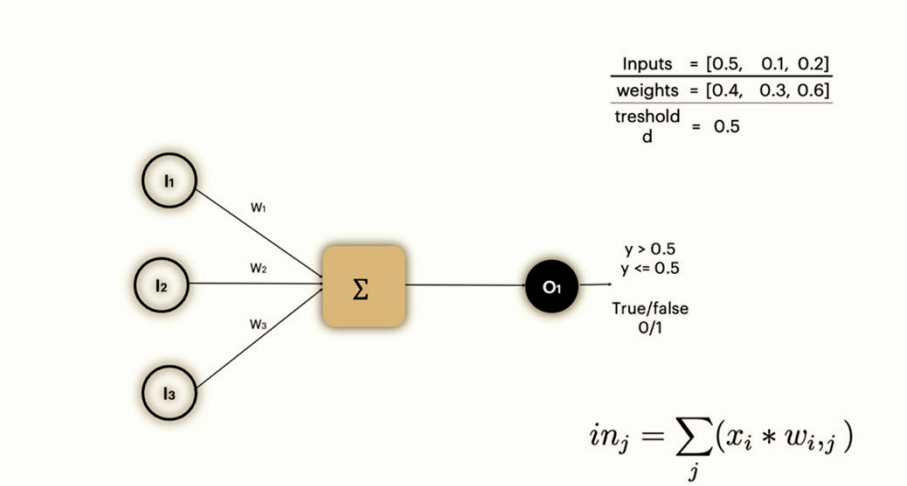
Neurális hálózatok szemléltetése az oktatásban

A neurális hálózatok matematikai modellje, grafikus megjelenése és implementációja teljesen eltérnek egymástól. Nem azonos térben és dimenzióban léteznek. Ahhoz, hogy ezt a problémát megérthessük, vegyük az alábbi példát: Készítsük el egy egyszerű perceptron modelljét, amely néhány bemenetből áll, és tartalmaz egy kimenetet.

CC BY

A perceptron feladata, hogy a bemenetére érkező értékek alapján – amely értékek lehetnek pl. egy kép egyetlen pixelét leíró színkódok, vagy egy pont térbeli helyzetét leíró adatok – eldöntse, hogy azok összesített és súlyokkal transzformált értéke meghaladja-e a 0.5 értéket (ezt az értéket treshold-nak nevezzük). Amennyiben igen, az eredmény igaz, tehát pl. az adott pixel színe sötétnek detektálható, ellenkező esetben világosnak.

A modell az 1. ábrán látható. Tartalmaz három bemeneti neuront, amelyek mindegyike egy élen keresztül kapcsolódik egy belső számítást végző matematikai függvényhez, amely a döntést fogja meghozni. A függvény lehet egyszerű szigmiod függvény (Olah, Chris, 2014), vagy RElu függvény (Olah, Chris, 2014; Schmidt-Hieber, Johannes, 2020), a lényeg, hogy a bemenetek alapján ki tudja számolni az eredményt, ami egy nulla és egy közé eső érték.



1. ábra Perceptron formális modellje

A három bemenet sorrendben (0.1, 0.5, 0.2), az élekhez tartozó súlyok (0.4, 0.2, 0.6). Láthatjuk, hogy mindkét halmaz, vagy tömb esetén három értéket kapuk, amelyeket sorrendben össze kell szorozni egymással, majd a kapott szorzatokat összeadni. Ez lesz a kimeneti neuron bemenete, amiről el kell döntenie, hogy az kisebb-egyenlő, mint 0.5, vagy nagyobb. A formális modell az 1-es ábrán látható. Az elsőt jelöljük a logikai TRUE, a másodikat a logikai FALSE értékkel. A bemeneti értékek minden esetben adottak, a súlyok pedig arra valók, hogy segítségükkel finom hangolhassuk a modellt.

Amennyiben a bemeneti értékekkel reprezentált pixelről tudjuk, hogy az sötét, viszont a neurális háló által adott eredmény y = TRUE, akkor a súlyokat úgy kell átalakítanunk, hogy az eredmény megfeleljen az általunk elvárt értéknek ^Y = FALSE.

Tehát a hálózat tanításához annyit kell tennünk, hogy a súlyokat átalakítjuk, és addig próbálgatjuk az eredményt, amíg a kapott érték meg nem felel az elvárt értéknek, vagy annyira meg nem közelíti, ami a számunkra már elfogadható.

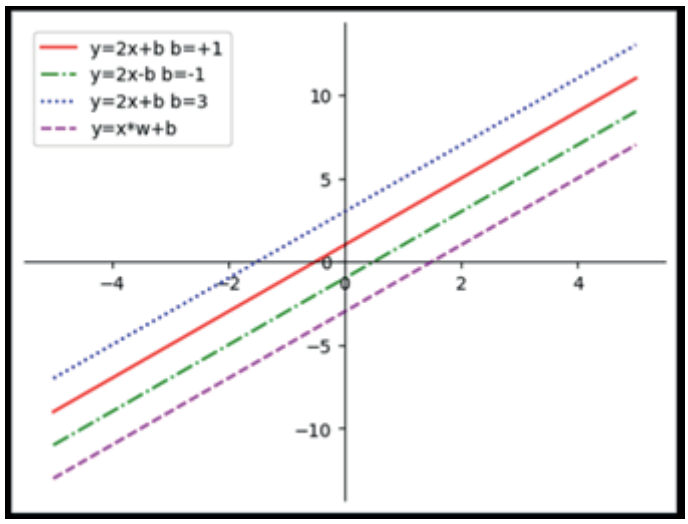
Ezt a módszert hívjuk tanításnak vagy tanulásnak, mert segítségével úgy paraméterezzük a modellt, hogy az minden esetben (a tanító halmaz bármelyik elemére) a számunkra elvárt eredményhez közeli kimenetet adja, vagyis megtanítjuk arra, hogy felismerje a bemeneti adatokat.

Ugyanezzel a módszerrel eldönthető egy képről, hogy az kutyát, vagy macskát ábrázol, de azt is meg tudjuk állapítani a betanított neurális hálózatok segítségével, hogy a kép mely pontjain találhatók arcok, vagy milyen tárgyak láthatók a képen. Nyelvi elemzést is tudunk végezni az ilyen modellek segítségével, és szövegeket is tudunk generálni bemeneti adatok alapján. Látható, hogy a modell maga nem túl bonyolult, de a megértése hosszas magyarázatot és ábrák felrajzolását, valamint matematikai képletek és algoritmusok elemzését és megértését követeli meg.

Térjünk vissza az eredeti problémára, vagyis a modell megértésére. A rajz alapján könnyedén elképzelhetjük annak felépítését. Természetesen nagyobb és bonyolultabb hálózatok esetében ez nem ilyen egyszerű feladat, de nem is ez a lényeg. Vegyük észre, hogy a bemeneti értékek egy tömbben, vagy bonyolultabb esetben egy mátrixban elhelyezhetők. Az élekhez tartozó súlyok is egy tömb, vagy mátrix elemei. A modellben az eredmények kiszámítása így visszavezethető mátrixok és tömbök szorzására. A tanítási algoritmus is kimerül abba, hogy a súlymátrixok értékeit addig változtatjuk, amíg jó eredményeket nem kapunk a bemenő adatok és az elvárt értékek alapján.

Az egyes részeredmények kiszámításához tehát elég egy olyan függvény, ami képes mártixokat összeszorozni. A neurális hálózatok alkalmazási területein nem mátrixnak, hanem tensor-nak (Developers, 2022) neveztük az adatokat tároló adatszerkezeteket. A tensor és a mátrix nem azonos. A tensor a mátrix, vagy tömb specializációja. A modell végeredményének a kiszámításához pedig egy olyan függvényre van szükségünk, amely egy számról eldönti, hogy az kisebb-egyenlő, vagy nagyobb, mint a treshold értéke.

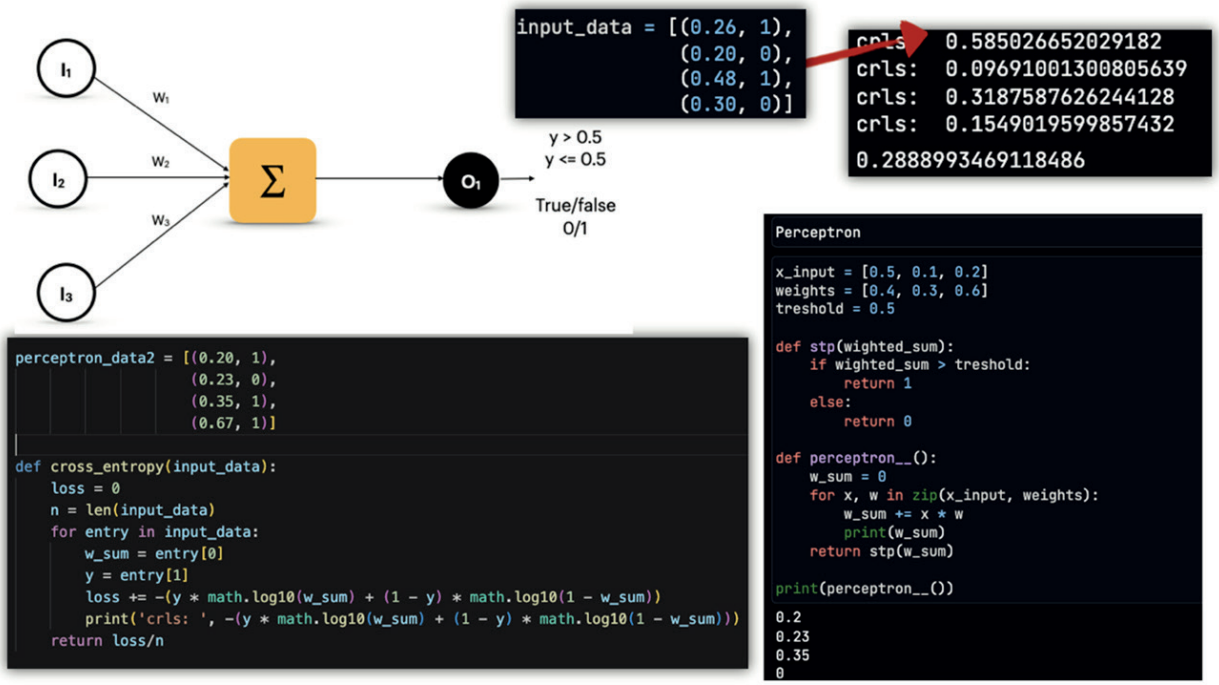
A feladat leegyszerűsítve és matematikai modellé konvertálva a következő: ***y = Input * weights + b***, ahol a *b*, vagyis a bias (Geman at al., 1992) egy olyan érték, amely segítségével az eredményt a megfelelő tartományba tudjuk transzformálni. A legtöbb esetben ez egy konstans érték, ami a jobb eredmény elérése érdekében kerül a képletbe. Ez a matematikai konstrukció szinte teljesen azonos az egyenes egyenletével, így ahhoz hasonló módon ábrázolni tudjuk. Az eredmény a 2-es ábrán látható.



2. ábra Az eredmény kiszámítása függvénnyel ábrázolva

CC BY

A modell végeredménye egy másik nagyon egyszerű függvény segítségével eldönthető, amely függvénynek az implementációját is megadtuk, ahogyan az a 3-as ábrán látható programban megfigyelhető. A forrásszöveg a teljes modell programját tartalmazza, és látható rajta a veszteség kiszámítására használható függvény implementációja. Ez a programkód azért került az ábrára, mert megmutatja a tanítási algoritmusban szereplő matematikai módszer implementációját.



3. ábra A hálózat struktúrája, a matematikai modell és a forráskód eltérő megjelenítése

Miért fontos ez számunkra? Ha összefoglaljuk a fentieket és levonjuk a következtetéseket, arra a megállapításra juthatunk, hogy a neurális háló modellje – a 3-as ábra alapján –, a matematikai megoldások, amelyeket a grafikus modell alapján készítettünk, valamint az implementáció olyan mértékben eltér egymástól, hogy azok mindegyike külön magyarázatot igényel. Legalábbis akkor, ha a modell működését szeretnénk megérteni.

Ha a matematikai alapokkal kezdjük az oktatást, akkor elveszik a hálózat eredeti struktúrája, és hamar azon kaphatjuk magunkat, hogy mátrixok szorzását próbáljuk meg elmagyarázni a hallgatóság számára.

Amennyiben a programírás útját választjuk, akkor ugyanez a helyzet, és sajnos a matematikai modell implementációjára helyeződik át a hangsúly. Ez szintén nem szerencsés. Viszont, ha a grafikus modellel kezdünk, akkor az algoritmus és a matematikai modell magyarázata lesz gyakorlatilag meseszerű elbeszélés, ami szintén nem adja vissza egzakt módon a modell működését és nem feltétlenül vezet el a megfelelő szintű megértéshez.

A három módszer együttes alkalmazása és a vizuális megjelenítés, pl. animációk már sokat segíthetnek, de sajnos ezek az animációk nem működés közben mutatják meg a neurális hálózatokat. Szemléltetni tudjuk segítségükkel a hálózat struktúráját, el tudjuk magyarázni a működést, fel tudjuk írni a matematikai modell elemeit, és implementálni is tudunk. Ez mind igaz, de egyik esetben sem a valódi hálózatot mutatjuk meg működés közben, csak annak

egy egyszerű szimulációját, vagyis egy kicsit szofisztikáltabb módon, de ugyanazt tesszük, mint a táblarajzokkal az előző esetben.

A példánkban szereplő modell viszonylag egyszerű, és nem tartalmaz olyan bonyolultabb módszereket, mint a megerősítéses, vagy a megerősítés nélküli tanulás, esetleg a back propagation (Hecht-Nielsen et all., 2009) algoritmusok, vagy a konvolúciós függvények (Gu, Jiuxiang, et al., 2018) alkalmazása, amelyek tovább mélyítenék a problémát.

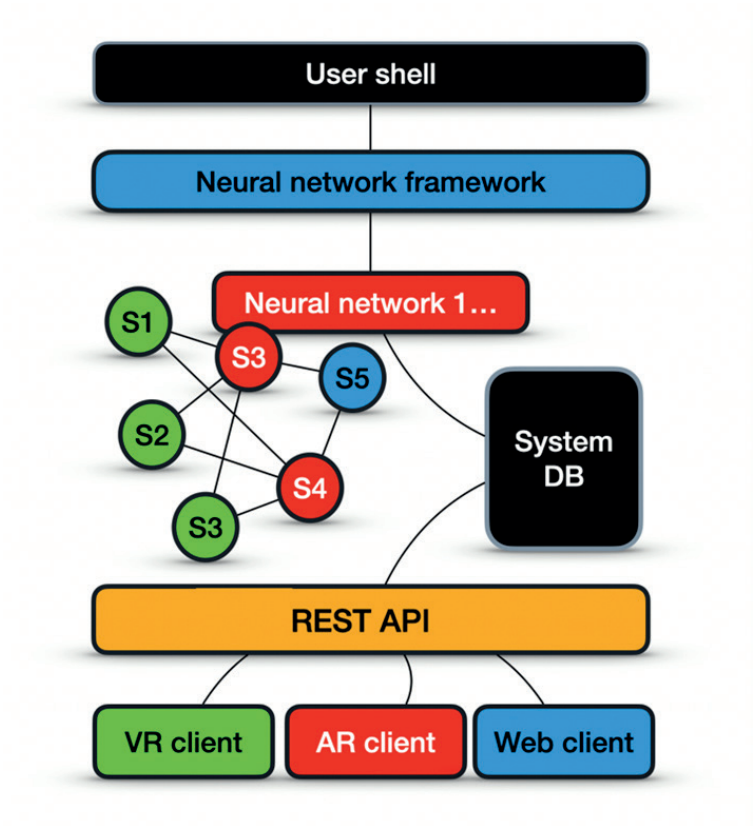
Amennyiben nem műszaki területen, vagy az oktatás alacsonyabb szintjein próbáljuk meg bemutatni ugyanezt, a matematikai modell és az implementáció használata szinte egyáltalán nem lehetséges. Egész egyszerűen a hallgatóság nem tud programozni és nincs olyan szintű matematikai tudása, ami a fenti módszer használatához szükséges volna.

Véleményünk szerint, és ahogy azt a bemutatott példa is bizonyítja, a probléma nehezen áthidalható a hagyományos, vagy a viszonylag modern oktatási módszerek és eszközök használata mellett. Mindezért szükségesnek tartottuk egy olyan eszköz kifejlesztését, amely széles körben és az oktatási rendszer szinte minden szintjén segít abban, hogy a neurális hálózatok és a mélytanulási algoritmusok hatékony oktatására lehetőség nyíljon. A probléma megoldására kidolgoztunk egy modellt, valamint kifejlesztettünk egy olyan keretrendszert (4-es ábra), amely képes működés közben megmutatni a neurális hálózatokat és a virtuális térben elérhetővé teszi a felhasználók számára azok manipulálását és megismerését. A keretrendszer segítségével elérhetjük, hogy az alapvető programozói tudással nem rendelkező, vagy azt viszonylag alacsony szinten művelő hallgatók számára is lehetőséget biztosítsunk a neurális hálózatok megismerésére és professzionális szinten való alkalmazhatóságára.

A vizuális megjelenítés - és természetesen az interakciók - Augmented Reality (AR) és Virtual Reality (VR) környezetben valósulnak meg, vagyis a virtuális teret használjuk a hálózatok megjelenítésére és kezelésére. AR és VR eszközök használatával a keretrendszerben létrehozott neurális hálózatokat virtuális környezetben meg tudjuk mutatni és ugyanezen felületen elérhetővé válik azok irányítása, programozása és kezelése is. Ez a lehetőség amellett, hogy nagyon látványos eredményt produkál a külső szemlélő számára, oktatástechnológiai szempontból is nagy jelentőséggel bír.

A keretrendszer oktatási célú felhasználásával elérhetjük, hogy az alapvető programozói tudással nem rendelkező, vagy azt viszonylag alacsony szinten művelő hallgatók számára is lehetőséget biztosítsunk a neurális hálózatok megismerésére és professzionális szinten való alkalmazhatóságára. A cikk szempontjából azonban legfontosabb tulajdonsága, hogy a rendszer grafikusan monitorozza a neurális hálózatot és élcímkézett irányított gráfként is képes azt megjeleníteni.

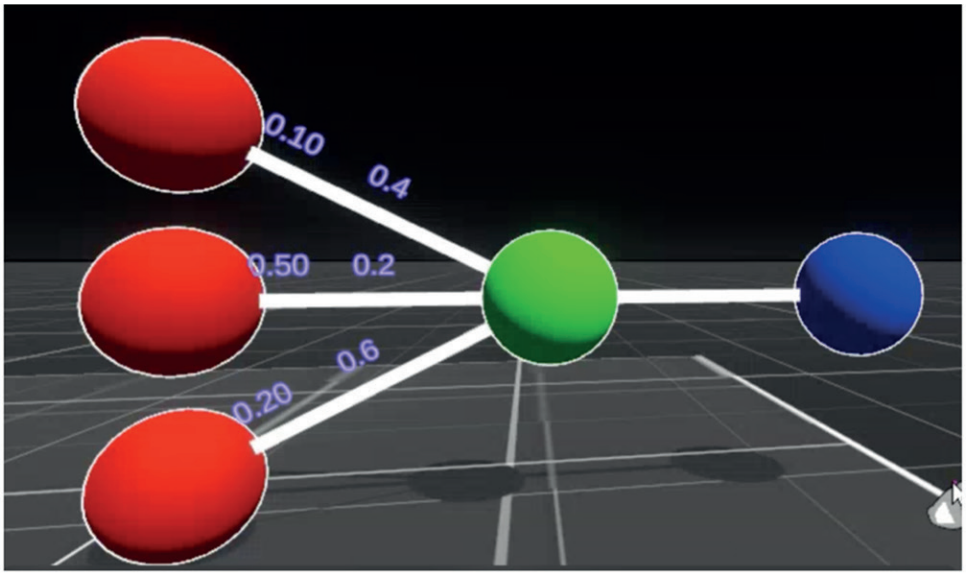
CC BY



4. ábra Az általunk készített keretrendszer felépítése

A modell (4-es ábra) szemléltetéséhez készítettünk mobil, valamint Unity alapú, VR szemüvegen keresztül használható interfészt a neurális hálózat kezeléséhez.

A neurális háló modelljének a kidolgozása során azt tartottuk szem előtt, hogy ne matematikai modellek alapján számoljunk eredményeket, hanem a neurális hálózat valódi struktúráját modellezzük, és az implementációban is neuronokra osztott rendszerben és valós időben zajló számítási folyamatokban gondolkoztunk.



5. ábra VR alapú megjelenítés

A hálózat VR alapú megjelenése a 5-ös ábrán látható, valamint az alábbi videón is megtekinthető: https://youtu.be/UlyVR76NgI8

A VR modul Unity Engine alatt fut, ami képes megjeleníteni a neurális hálózatokat a virtuális környezetben. Ez az alrendszer lehetőséget biztosít tanuló hálózatok generálására is. A vizuálisan megjelenő hálózatokat automatikusan színezi, template-ezi, de lehetőséget biztosít egyéni template-ek bevezetésére. A virtuális valóságban az egyes node-okat meg lehet fogni, megérinteni, forgatni és kezelni lehet az egyes funkcióikat. A VR-ban kétféleképpen lehet a környezettel interakcióba lépni. Az egyik módszer a virtuális kezünk, a másik pedig egy lézer pointer eszköz amivel egyszerűen megfoghatunk távoli objektumokat, majd közelebb húzhatjuk azokat magunkhoz. Ez segíti a nagyobb hálózatok kezelést is. A rendszerhez készült egy generátor program is, mellyel random hálózatokat tudunk generálni. A hálózat generátorhoz készült egy megjelenítő modul, így valós időben is láthatjuk a szemléltetési célból készült hálózatokat. Az ilyen hálózatok esetén a pozíció és élgenerálás randomizált, viszont a generált node-ok számát, éleit, dimenzióit és elhelyezkedését be lehet állítani a modul szintén VR alapú felületén. A generált hálózat ebben az esetben is SQL fájlba exportálható, így azt a keretrendszer más részei is azonnal használni tudják a fájlok feltöltését követően.

Összefoglalás

Ahogy azt a bevezetőben említettük, az általunk kifejlesztett és a fentiekben bemutatott rendszert jelen cikk szerzői is használják és a közeljövőben használni kívánják oktatási célokra. Az oktatásban a célterület a neurális hálózatok megismertetése, valamint alapvető használata. A Reality Interface lehetőséget ad arra, hogy az oktatási feladatokat korosztályokhoz, vagy elért oktatási eredményekhez tudjuk kötni. Segítségével az oktatási céljainknak megfelelő hálózatokat és tartalmakat tudunk létrehozni és publikálni azokat a Laravel interface-en keresztül a különböző e-learning-rendszerek számára, így:

- lehetőségünk van példahálózatokat generálni a vizuális megjelenítéshez.
- Az általunk használt oktatási környezethez képesek vagyunk testreszabni a hálózat megjelenését, figyelembe véve a korosztály életkori sajátosságait.
- A rendszer Laravel kommunikációs interface modulja segítségével, endpointok használatával publikálhatjuk az elkészített tananyagot bármilyen REST API használatára felkészített e-learning rendszer, vagy webes felület számára is.

A következő fejlesztési tervben már szerepel az adatok online és realtime-jellegű frissítése, ami lehetővé teszi a folyamatos hálózat megfigyelést a kommunikációs interfészen keresztül. Ebben a verzióban, amelynek a fejlesztői változata már létezik, a REST API felől folyamatos szinkron frissítések érkeznek a kliensek felé, így azok folyamatosan konzisztensen tartják a hálózatot annak vizuális másolatával. Viszont ez a megoldás nagy erőforrás igénnyel jár, és a hálózatban szereplő neuronok és kapcsolataikat reprezentáló élek pozícióját folyamatosan újra kell kalkulálni. Az AR és VR környezetben újra kell rajzolni a hálózat megváltozott részeit, ami a kliensekre nagy terhet ró. Ezeket a számításokat, valamint a hálózat futtatását pontosan ezért szeretnénk a HPC szuperszámítógépre portolni. Viszont ez a lépés a hálózat futtató környezetének teljes újraimplementálását vonja magával, ami komoly tervezést igényel. Megtettük ez irányba a megfelelő lépéseket, a KIFÜ (Kormányzati Informatikai Fejlesztési Ügynökség) munkatársaival felvettük a kapcsolatot és hozzáférést szereztünk az elérhető legnagyobb teljesítményű szuperszámítógéphez, a Komondorhoz. Jelenleg az implementáció programozási nyelvének kiválasztásán és a modell, valamint az architektúra átdolgozásán munkálkodunk.

CC BY

Irodalom

Chollet, Francois. Deep learning with Python. Simon and Schuster, 2021.Olah, Chris. „Neural networks, manifolds, and topology, 2014." URL http://colah. github. io/posts/2014-03-NN-Manifolds-Topology (2018).

Graph Stream graph vizualizer, Last download: 2022.03.12, https://graphstream-project.org/doc/Tutorials/Getting- Started/

Laravel system documentation, Rest API, http ://laravel.org

Abadi, Martín, et al. „Tensorflow: Large-scale machine learning on heterogeneous distributed systems." arXiv preprint arXiv:1603.04467 (2016).Wadee Alhalabi. 2016. Virtual reality systems enhance students' achievements in engineering education. Behaviour & Information Technology 35, 11 (July 2016), 919–925. https://doi.org/10.1080/0144929X.2016.1212931.

VanRossum, Guido, and Fred L. Drake. The python language reference. Amsterdam, Netherlands: Python Software Foundation, 2010. VanRossum, Guido, and Fred L. Drake. The python language reference. Amsterdam, Netherlands: Python Software Foundation, 2010.

Schmidt-Hieber, Johannes. „Nonparametric regression using deep neural networks with ReLU activation function." (2020): 1875-1897.

Geman, Stuart, Elie Bienenstock, and René Doursat. „Neural networks and the bias/variance dilemma." Neural computation 4.1 (1992): 1-58.

Gu, Jiuxiang, et al. „Recent advances in convolutional neural networks." Pattern recognition 77 (2018): 354-377.

Developers, TensorFlow. „TensorFlow." Zenodo (2022).

Hecht-Nielsen, Robert. „Theory of the backpropagation neural network." Neural networks for perception. Academic Press, 1992. 65-93.

 DOI: 10.31915/NWS.2023.11

Mesterséges intelligencia, multimédia, tanulástámogatás

Artificial intelligence, multimedia, learning support

T. Nagy László
Debreceni Református Hittudományi Egyetem (Debrecen)
t.nagy.laszlo@drhe.hu

Absztrakt
A mesterséges intelligencia (MI), mint tényező egyre fontosabb és napról-napra jelentősebb szerepet tölt be a mindennapjainkban. A digitális eszközök és általában a „digitális kultúra" szerepe az oktatásban is ugrásszerű fejlődésen ment és megy keresztül. Elég csak az utóbbi évek pandémia által generált hatását említeni az online oktatás fejlődésére. Természetesen az oktatásban felhasználható számtalan lehetőség mellett, a tanulás támogatásában is hasonló mértékű pozitív változások tapasztalhatóak.

A MI több megnyilvánulása multimédiás képességekkel is rendelkezik, ez a legtöbb esetben médiakonverziót is jelent. Jelen tanulmányban megvizsgálom és összegzem a leggyakoribb médiakonverzió irányokat és lehetőségeket, amelyek a MI alkalmazásai által napjainkban elérhetőek, továbbá néhány példán keresztül megemlítem a tanulás támogatásában könnyen hasznosítható megoldásokat is.

Kulcsszavak: mesterséges intelligencia, tanulástámogató eszközök, digitális oktatás, online tanulás, multimédia

Abstract
Artificial intelligence is becoming more and more important in our everyday lives. The role of digital tools and ‚digital culture' in education in general has been evolving rapidly. One need only think of the impact of the pandemic in recent years on the development of online education. Of course, along with other, almost countless opportunities in education, there have been similar positive changes in the way learning is supported.

Many applications of AI also have multimedia capabilities, in most cases involving media conversion. In this paper, I will consider and summarize the most common media conversion directions and possibilities that are available through AI applications today, and I will mention examples for those that can be easily used to support learning.
Keywords: artificial intelligence, learning support tools, digital education, online learning, multimedia

Bevezetés

A mesterséges intelligencia vívmányai az elmúlt évtizedekben, de különösen az elmúlt 5-10 évben szinte észrevétlenül szivárogtak be a mindennapjainkba. A legtöbb ember sok esetben úgy kezdett mesterséges intelligencián (MI) alapuló terméket vagy szolgáltatást használni, hogy igazából nem is sejtette (vagy nem is gondolkozott el rajta) hogy a háttérben az eredményeket részben vagy egészben mesterséges értelem alkotja. A ChatGPT megjelenésével ez hirtelen megváltozott.

CC BY

A témával hosszabb ideje foglalkozók láthatták a fokozatos és apró eredményeken az egyre fejlettebbé váló MI-át. A ChatGPT 2022 novemberi szabad és ingyenes elérhetővé tétele, majd az azt követő erős médiavisszhang, olyan rétegek látókörébe is eljuttatta a mesterséges intelligenciát és annak vívmányait, akik eddig kevés tudomást szereztek róla vagy egyáltalán nem érdeklődtek iránta. (ChatGPT 2023)

Jelen kutatásban azt a célt tűztem ki, hogy megvizsgálom az interneten elérhető (főleg ingyenes) mesterséges intelligenciát használó eszközök milyen multimédiás és médiakonverziós képességekkel rendelkeznek, és ezek a lehetőségek hogyan (és mi módon) tudják támogatni a tanulási és/vagy oktatási folyamatokat.

Alapelvek

A mesterséges intelligencia jelenleg hozzáférhető megnyilvánulásai az elméleti háttér szempontjából szinte minden esetben fekete dobozként kezelendők, hiszen a bemeneti kérdés és a kimeneten megjelenő válasz között nem ismerjük a pontos működést. Nem tudjuk azt sem, hogy milyen alap információ(adat)bázisból áll össze, azaz mire támaszkodik az output, ezért a kutatás során – empirikus módszerrel – az MI-nak adott célzott feladatok és kérdések heurisztikus kiértékelésével keresem a válaszokat. A cél elsősorban nem a működés pontos feltárása, hanem annak megállapítása, hogy az adott entitásnak milyen multimédiás és médiakonverziós képességei vannak, valamint alkalmas lehet-e tanulástámogató eszköznek? A kutatás első fázisában fontosnak éreztem pontosítani mi is a mesterséges intelligencia, vagyis mit tekintünk MI-nak? A ChatGPT erős médiavisszhangja miatt sokan a kérdésre „válaszolós” nyelvi modellt (ChatGPT) tekintik „a mesterséges intelligenciának”, mint az első olyan konkrét találkozást a témával, ahol tudatosan használnak egy olyan rendszert, amiről tudják, hogy MI alapú. Azonban, hogy mitől tekintünk egy megoldást MI-nak, azt fontos definiálni.

Az MI fogalmának alkotása egy alapvető szempontrendszer összeállítását jelentette esetemben, hiszen egy fogalommeghatározás sokszor szubjektív és a témamegközelítés irányától (pl. diszciplína) is függhet.

- általánosságban elmondhatjuk: egy entitást akkor tekintünk MI-nak, ha képes az emberi értelem, vagy gondolkodás utánzására, azaz olyan válaszokat ad a kimenetén, amit akár ember is adhatott volna. (ld. például a Turing teszt módszertanát) (Turing teszt 2023, Turing-teszt (1))
- Tehát számunkra fontos az, hogy a rendszer képes legyen a működését, azaz az eredményét „célszerűen” az adott feladatnak megfelelően (és természetesen megismételhető módon) megváltoztatni.
- A mesterséges intelligencia (mindamellett, hogy humán kognitív képességekkel rendelkezik) legyen képes megoldani számára ismeretlen feladatokat is. Tehát az MI saját, helyzetnek, feladatnak megfelelő (túlnyomórészt helyes, vagy legalábbis ellentmondásmentes) válaszokat produkáljon.
- A saját válaszok elvárása azt az igényt is magával hozza, hogy az MI legyen képes a saját tapasztalatai alapján javítani a hatékonyságán, azaz egyre pontosabb válaszokat adjon, vagyis legyen képes tanulni. Az MI gépi tanulása olyan tanulóalgoritmusok alkalmazását jelenti tehát, amelyek képesek szabályosságok és összefüggések felismerésére tanítópéldák halmaza alapján.

A bemeneti adathalmazban lévő összefüggés(ek) megtalálását leginkább úgy reprezentáljuk, hogy a rendszer egy bemenetre generál egy kimenetet, a tanulási folyamatot pedig az jelzi, hogy egyre több helyes (elvárt) kimenet keletkezik. Fontos kiemelni, hogy a **lényeg** nem a konkrét tanulópéldák megtanulása (ld. túltanulás), hanem **a helyes általánosítás** a tanulás során nem látott (nem ismert) példákra is! (Tóth 2020)

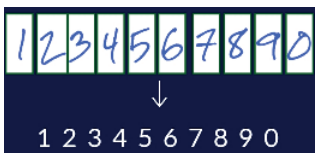
A gépi tanulás a mesterséges neurális hálózatok egy alapvető és erős tulajdonsága, azaz a környezetükből tanulni tudnak, ami azt jelenti, hogy képesek tanulással javítani a működésüket. A többrétegű neurális hálózatok (deep learning) alkalmazása a mai MI-ák fontos alapköve. Lényeges, hogy viszonylag (nyers) kevés bemeneti adatból is tudnak dolgozni. (Tóth 2020)
Tehát az MI használatának olyan esetekben van igazán értelme, amikor nem tudunk (vagy nem is akarunk) az összes lehetséges várható be és kimenetre felkészülni (pl. azok nagy számossága miatt), vagy olyan feladatot akarunk rábízni az algoritmusra, amelyek hagyományos programozói megoldása is igen nagy emberi/számítási erőfeszítéseket igényelne. Megállapítható, hogy a mesterséges intelligencia lehetséges és alkalmazott felhasználási területeit az adatelemzés/adatfeldolgozás műveletein kívül, a kimeneti oldalon alapvetően két konverziós, működési formára tudjuk bontani, ezek a:

- transzformáció
- tartalom generálás

A mély neurális hálók fontos képessége a hatékony mintafelismerési vagy mondhatjuk úgy is, összefüggés-felismerési tulajdonsága, ezzel bizonyos területeken (pl. a kép és a beszéd/szöveg-feldolgozás) sokkal jobb eredményeket érnek el, mint a korábbi gépi tanulási megoldások. Ebből a mintafelismerési tulajdonságból adódóan pedig képesek az adattranszformációra és tartalom generálásra (adatszintetizációra). Meg kell azonban jegyezni, hogy igazán élesen nem minden esetben választható szét e két fogalom, inkább azt mondhatjuk, valamelyik jobban jellemző. Például tetszőleges képet a bemenetre adva egy kutyáról, egy kondicionált mély neurális háló képes azt a szöveges kimenetet adni: „kutya”. Ez esetben a képpontokból összeálló adathalmazt, mint képi információt („értelmezi”) transzformálja a nyelv „kutya” szavára, amelyet karaktersor formájában prezentál kimeneti eredményként. Ez esetben inkább adattranszformációnak értelmezzük a műveletet (a kutyát ábrázoló képből (pixelhalmazból), a kutya – betűkből álló – szóképe lett), ha viszont fordítva: azt kérem szöveges formában a MI-tól: „Rajzolj kutyát!” inkább tartalomgenerálásnak definiáljuk a műveletet. Azaz, ha csökken a jellemzőtér transzformációnak, ha pedig nő generálásnak értékeljük a kimenetet. Azt is meg kell jegyezni, hogy konverzióknál általában médiaváltás (pl. képből szöveg, jpg>txt) is bekövetkezik, de nem minden esetben!

Transzformációk

Milyen fajta konverziókra képes tehát az MI? Milyen irányokba képzelhető el a konverzió? A jelenlegi alkalmazásokat megvizsgálva leginkább az alábbi transzformációs irányokkal találkozhatunk, melyeket az alábbiakban rendszerezek:

képből > karakter (karakterfelismerés) 

Ez esetben gondoljunk az optikai karakterfelismerésre (OCR) ahol a releváns (szöveges) információ változatos grafikus (képi) formában jelenik meg a bemeneten, például a képen egy

CC BY

szó vagy szöveg részeként, a kimeneten pedig a betűk felismerésének megfelelő karaktert várjuk. (OCR 2023) A karakterfelismerés funkció több dokumentumkészítő alkalmazásban és a Google Lens azonnali fordítás opciójában is megtalálható. E médiakonverzió igen széleskörűen integrálható tanulástámogató eszköz, a lefotózott szövegek – karakteres transzformáció után – akár szerkeszthetők, másolhatók lesznek, vagy másik szöveges alkalmazás bemenetére küldhetők.

képből > szöveg (objektumfelismerés)



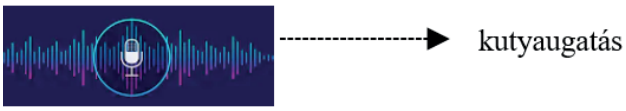
Ez esetben a képi pixelhalmazból szöveg (karakterhalmaz) lesz a képi tartalomnak megfelelően. (Amit akár a kép metaadataként is fel tudunk használni.) Az objektumfelismerés az önvezető járművek egyik igen fontos alapképessége. Ezen túl az eszközök, tárgyak, egyéb dolgok neveinek beazonosításában is segíthet akár rögtön idegennyelvű válaszokkal. Széleskörűen alkalmazható tanulástámogató eszköz (pl. tárgykép-szókép-jelentés összekötés, szótanulás stb.) a Google Lens-ben szintén integrált megoldás. (Google Lens 2023)

beszédből > írott szöveg (beszédfelismerés)



Az MI alkalmazásának egy fontos területe ez a konverzió. Ami technikailag azért nehéz feladat, azaz nem igazán standardizálható, mert a hangzó beszéd – egy nyelvének – beszélői többféle akcentussal, kiejtéssel, tempóval, hangerővel, beszédhibákkal, háttérzajokkal stb. terhelten beszélnek. Ez esetben a beszéd karakteres (szöveges) verziójának előállítását várjuk az MI-tól, tetszőleges audió bemenet alapján. Klasszikusan diktálás funkciók is kiválthatóak vele, (ld. pl. jegyzet, vázlat, feljegyzéskészítés mint tanulástámogató dokumentumok) vagy meglévő hangfájl szöveges leirata, vagy videófájl feliratozása is elkészíthető vele. Az általános célokon túl hallássérülteknek és nyelvtanulóknak is hasznos funkció.

hangból > szöveg, (hangfelismerés)

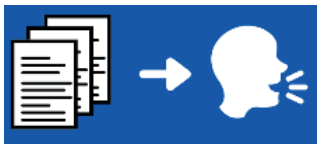


Ez esetben a bemeneti hang(ok)nak (nem beszéd) várjuk a szöveges leírását: pl. lehetséges szöveges kimenetek a bemenet alapján: *fűnyíróhang*, *madárfütty*, *hajókürt* stb. Hallássérülteknek szöveges leírás (felirat) készíthető pl. videófájlra, vagy általános vezérlő parancsoknak is felhasználható a kimenet.

zenéből > szöveg (zenefelismerés)



Lejátszott zeneszám esetében pl. a kimenet metaadat; szerző/előadó, cím: *Abba – Dancing Queen*. Számos zenefelismerő megoldás működik nagy népszerűséggel, akár komolyzenei adatbázissal is. Más esetben az énekhang szöveges leírását kaphatjuk eredményként, hasonlóan ahogy a beszédből > szöveg bekezdésnél már említettem.

szövegből > beszéd (gépi felolvasás) 

Az írott szöveges információt a MI felolvassa. Betűk, számok vagy szavak kiejtése egyszerű, – akár analóg – hozzárendeléssel is megoldható (ld. pl. vasúti hangosbemondó, ahol pl. a vonat indulási idő számokat gép olvasta fel). A szöveg „felolvasása" beszédszintetizációval is történhet, ahol minden betűnek külön képezi a kiejtését a gép (pl. digitális) hangszintetizátor segítségével.

Az előző felolvasási módszereknél jóval fejlettebb megoldás az MI alkalmazása a témában. Ez esetben a felolvasás úgy történik, hogy élő személy tetszőleges (pl. 10-100 oldalas) felolvasásának hangfájlját felhasználva (mint tanítási adatbázis), a mély háló bármilyen (bemenetre adott) szöveget „fel tud olvasni" a tanító adatbázisban szereplő ember természetes hangján, amelyet az valójában soha nem olvasott fel (vagy legalábbis biztosan nincs benne az MI tanulási adatbázisában). Ez esetben is a deep learning hálózat fejlett mintafelismerési és predikciós képességét használjuk fel a kívánt kimenet elérése érdekében.

Ebben a konverziós esetben is inkább transzformációról beszélhetünk, hiszen egy írott szöveg felolvasását várjuk eredményként. Azonban – ha jobban belegondolunk – a konverziónál a hangok szintetizációja zajlik valójában. Amiért mégis inkább transzformáció ez a konverzió: a leírt szöveg hanghű „felolvasása" történik, új szavak, hangok, gondolatok hozzáfűzése nélkül. Ez a módszer hasznos tanulástámogató lehetőségeket rejt magában, hiszen akár utazás közben is meghallgathatjuk, (megtanulhatjuk, átismételhetjük) a leírt szöveges tartalmakat, továbbá látássérülteknek is igen hasznos eszköz. Pl. Natural Readers, Voicemaker (Natural Reader 2023), (Voicemaker 2023)

Tartalomgenerálás

A legfontosabb transzformációs irányok áttekintése után az alábbiakban a főbb tartalomgeneráló irányokat ismertetem.

kérdésre > válasz (szöveg > szöveg)

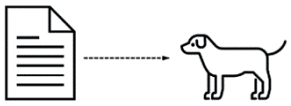
Ez esetben nincs médiaváltás, azaz egy szövegesen megfogalmazott kérdésre szöveges választ várunk (ld. pl. ChatGPT 3.5 nyelvi modellt). Viszont a kérdésre emberi nyelven „értelmes választ" szeretnénk kapni, tehát a MI a kérdés alapján szintetizálja (generálja) a várt szöveges választ. Mint tudjuk, a ma elérhető mesterséges intelligenciák nem „értik" valójában a kérdést, ahogy mi emberek tesszük, hanem a feltett kérdés alapján próbálják megjósolni a lehetséges kimeneti választ (predikció) az adatbázisukban fellelhető releváns információk alapján. A működés jellegéből adódóan előfordulhat, hogy a MI (pl. ChatGPT) olyan választ ad a kérdésre „nagy magabiztossággal", ami nyilvánvalóan nem helyes. Ilyen esetben nem szándékos a félrevezető eredmény, hanem szakszóval élve a MI „hallucinál". Mivel nem tudja értelmezni a feltett feladatot/kérdést, – tehát valójában nem érti mit kérdeztünk – nem tudja elbírálni a válasz jóságát sem. Egy a ChatGPT-hez hasonló nyelvi modell esetében a kimenet ugyanis a kérdésre: az adatbázisban talált információk alapján összeállított lehetséges (valószínű) szósorrend/mondathalmaz csupán.

Mindezek „ellenére" a ChatGPT konkrét összetett válaszokat képes adni a kérdéseinkre. A kérdések akár összetett mondatok is lehetnek, sőt konkrét szövegrészek értelmezését és

CC BY

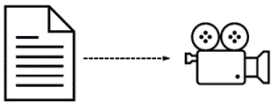
elemzését is kérhetjük sok egyéb más mellett. A tanulási céltól függően, a hagyományos keresőnek (Pl. Google) és a nyelvi modellnek feltett kérdések válaszait szabadon variálhatjuk, ugyanis mindegyik megoldásnak vannak előnyei és hátrányai.

szövegből > kép (képgenerálás), képből > szöveg (képleírás)



Az első esetben szöveges bemenet utasításai alapján alkot a MI, azaz képet generál szövegből. Az utóbbi időben több olyan publikusan elérhető megoldás jelent meg, amely igen valósághű képi világot képes előállítani leírás alapján (MidJourney, DALL-E, Stable Diffusion, Nightcaffe, replicate, stb.), (MidJourney 2023), (DALL-E 2023), (Stable Diffusion 2023), (Nightcaffe 2023), (replicate 2023). Számos módon felhasználható az oktatásban és a tanulásban is. Készíthetünk ábrákat, rajzokat, változatos formában képeket illusztrációként, példaként, szóképként stb. szöveges utasításokkal. A másik irány, amikor a MI a képről szöveges leírást, azaz képleírást készít az objektumfelismeréshez hasonlatos, de annyival több nála, hogy a képi tartalmak viszonyát, kapcsolatát, helyzetét, arányát, jelentését stb. akár szimbolikusan is értelmezi. Mindkét konverziós esetben médiaváltás történik.

szövegből > videó (videógenerálás)



A szöveges bemenet utasításai alapján a MI videó médiát alkot. Ez a forma ma még kevéssé elterjedt, de számos megoldása létezik.

képből > kép, képből > videó

A MI képből egy másik képet vagy videót alkot, az utasításoknak (pl. szöveges) megfelelően. Pl. hozzáad vagy elvesz/átalakít képet, képi objektumokat. A szövegből kép bekezdésben említett, jelenleg elérhető megoldások közül, több is képes bemeneti képek felhasználására, átalakítására a célnak megfelelően.

szövegből > hang



A szöveges bemenet utasításai alapján az MI hangot(kat) vagy akár zenét alkot. Igen érdekes kísérleti területe az MI kutatásának a „zeneszerzés". Több irányban folynak kísérletek: az MI utasítások alapján ír zenét pl. egy adott stílusban, vagy egy adott előadót utánozva alkot.

MI a tanulástámogató lehetőségei, összegzés

Az előzőekben a konverziós irányok figyelembevételével történő elemzés után, összegzésként néhány olyan elérhető komplex megoldást is szeretnék megemlíteni, amelyek többféle adatkonverzióra is képesek és igen sokoldalú tanulástámogató eszközök lehetnek számos tudományterületen.

ChatGPT: A megjelenésével elérhetővé tett komplex predikciós nyelvi modell általános, komplex kérdésekre is képes válaszolni. Olyan kérdésekre, amelyekre a megjelenése előtt

lévő korábbi megoldások válaszai (vagy eredményei) nem igazán voltak – a tanulás vagy oktatástámogatás esetében – relevánsak vagy komolyan vehetők. A ChatGPT legutóbbi verziói azonban olyan eszközt adtak a kezünkbe, amelyeket nem lehet figyelmen kívül hagyni. A nyelvi modell elemzése messze meghaladja e tanulmány kereteit, azonban az kijelenthető, hogy a ChatGPT igen széles körben alkalmazható általános tanulástámogató eszköz, amely képes – a legtöbb tudományterületen – konkrét kérdések megválaszolására, elemzések, dolgozatok, esszék elkészítésére, vagy akár nyelvtanulás támogatására is. A 4-es verziótól a modell „lát", azaz képes értelmezni a bemenetére adott képeket is, ezzel mind szélesebb körben és komplexebb válaszok lehetőségét biztosítva. Például tekinthetjük metaforikusan egy «magántanárnak» vagy «konzulensnek» az MI-át, akivel „meg lehet beszélni" sok mindent, akár idegen nyelven is! (ChatGPT4 2023)

Google Lens: Sokoldalú tanulástámogató eszközként is használhatjuk az alkalmazást, hiszen segítségével átalakíthatunk képeken lévő szövegeket karakteres formába, majd azt rögtön le is fordíthatjuk idegen nyelvre. A fordítást az eredeti szöveg helyére a rendszer beilleszti – mintegy virtuális valóságként – többszörös adatkonverziót végezve akár valós időben, online. Képes a képen lévő objektumok (tárgyak, élőlények, emberek stb. felismerésére), amely funkcióval ismeretlen dolgokról kereshetünk információkat a kívánt nyelven, akár feladatok megoldásához. (Google Lens 2023)

MidJourney: Ha képi információt szeretnénk generálni szövegből a MidJourney igen széles körben társ ebben, felhasználhatjuk bemenetként a meglévő képeinket is. A rendszer több változatot is készít egy témára, amit jól lehet finomhangolni a kívánt végeredmény szempontjából. Kreatív feladatok esetében jó ötletekkel szolgálhat, vagy oktató, szemléltető (tanulástámogató) anyagok elkészítésében is hasznos segítség lehet. (MidJourney 2023)

Replicate: Ha képet szeretnénk átalakítani, rajzolni, képről szöveget alkotni az oldal széleskörű lehetőségeket kínál. Egy képről akár egyszerű fogalmazást is készíttethetünk vele. (Az oldalon találhatóak videó és hangzóanyag készítő MI megoldások is.) (replicate 2023)

Az előzőekben tehát áttekintettem és összegeztem a MI főbb adatkonverziós képességeit, valamint példaként megemlítettem néhány, az egyes konverziókhoz kapcsolódó tanulástámogatásban felhasználható lehetőséget. Végezetül négy konkrét, – jelenleg igen népszerű – MI-át használó alkalmazást emeltem ki, amelyek többféle adatkonverziós képességükkel kreatívan és széleskörűen alkalmazható megoldások lehetnek a tanulástámogatásban vagy bizonyos esetekben az oktatásban is.

Hivatkozások

ChatGPT (2023) Letöltés (2023. 06. 18.) https://openai.com/blog/chatgpt
ChatGPT4 (2023) Letöltés (2023. 06. 18.) https://openai.com/research/gpt-4
DALL-E (2023) Letöltés (2023. 06. 18.) https://openai.com/dall-e-2
Google Lens (2023) Letöltés (2023. 06. 18.) https://lens.google/intl/hu/#translate
MidJourney (2023) Letöltés (2023. 06. 18.) https://www.midjourney.com/home/?callbackUrl=%2Fapp%2F
Natural Reader (2023) Letöltés (2023. 06. 18.) https://www.naturalreaders.com/
Nightcaffe (2023) Letöltés (2023. 06. 18.) https://creator.nightcafe.studio/text-to-image-art
replicate (2023) Letöltés (2023. 06. 18.) https://replicate.com/
Stable Diffusion (2023) Letöltés (2023. 06. 18.) https://stablediffusionweb.com/

CC BY

Tóth László (2020) Mesterséges neuronhálók és alkalmazásaik, Elektronikus tananyag, Szegedi Tudományegyetem, Letöltés (2023. 06. 10.) https://www.inf.u-szeged.hu/~tothl/bevmely/01.%20Gepi%20tanulas,%20neuron,%20neuronhalo.pptx

Turing-teszt (1) Turing teszt ismertető, Letöltés (2023. 06. 18.) http://www.mestersegesintelligencia.hu/doc/Turing%20teszt.php

Turing-teszt (2023) Wikipédia szócikk, Letöltés (2023. 06. 18.) https://hu.wikipedia.org/wiki/Turing-teszt

Voicemaker (2023) Letöltés (2023. 06. 18.) https://voicemaker.in/

 DOI: 10.31915/NWS.2023.12

Egy automatikusan generált rímszótár fejlesztése és a magyar kanonikus költészet rímszavainak néhány jellemzője

Building an automatically generated rhyming dictionary and some characteristics of rhyming words in Hungarian canonical poetry

Horváth Péter
ELTE Digitális Bölcsészet Tanszék, Digitális Örökség Nemzeti Laboratórium
horvath.peter@btk.elte.hu

Absztrakt

A tanulmány a jelenleg 50 kanonikus magyar költő összes versét tartalmazó ELTE Verskorpusz alapján automatikusan generált rímszótárt mutatja be. Ismerteti a verskorpusz automatikus annotálásához használt új rímképletelemző algoritmust, a rímszótárban feltüntetett adatokat, valamint a rímszótár különböző formátumait és online elérhető lekérdezőfelületét. A tanulmány emellett három rövid példával rámutat arra is, hogy a rímszótár hogyan használható a magyar rímelés vizsgálatában.

Kulcsszavak: rímszótár, rímképlet, ELTE Verskorpusz, automatikus annotálás, lekérdező felület, ragrím

Abstract

The paper presents a rhyming dictionary generated automatically on the basis of ELTE Poetry Corpus, which currently contains all the poems of 50 Hungarian canonical poets. It describes the new algorithm annotating the rhyme patterns of the poems in the corpus, the data included in the rhyming dictionary, the different formats of the dictionary, and an online query interface. Trough three examples, the paper also highlights how the rhyming dictionary can be used in the investigation of Hungarian rhyming.

Keywords: rhyming dictionary, ELTE Poetry Corpus, automatic annotation, query interface, inflectional rhyme

1. Bevezető

Míg egy értelmező szótár elkészítése kifejezetten időigényes feladat, amely számos lexikográfus munkáját kívánja meg, addig egy rímszótár létrehozása teljesen automatizálható, amennyiben rendelkezünk egy megfelelő méretű verskorpusszal. Egy ilyen korpuszalapú, automatizáltan elkészítendő magyar rímszótár tervezetét már Mártonfi Attila felvázolta tizenöt évvel ezelőtt [1]. Az ELTE Verskorpusz létrehozásának köszönhetően az automatizáltan létrehozandó magyar rímszótár elképzelése ténylegesen megvalósíthatóvá vált. A rímszótár korpuszát adó ELTE Verskorpusz 50 kanonikus magyar költő összes versét, összesen 13 362 verset tartalmaz a szövegek grammatikai és vershangzáshoz kapcsolódó poétikai annotációival [2]. A korpuszban szereplő legkorábbi szerző Tinódi Sebestyén, a legkésőbbi pedig Radnóti Miklós. A korpusz automatikusan létrehozott annotációi között szerepelnek többek között a versszakok rímképletei a hagyományos, a latin ábécé egymást követő betűit használó jelöléssel (pl. aabbcb), amely a versszövegek mellett a rímszótárt generáló szkript

CC BY

bemenetét adta. A rímszótárt generáló szkript kimenete XML-formátumban tartalmazza a rímpárokat és azok előfordulási helyét, illetve további jellemzőit. Az XML-formátumból további formátumok is létre lettek hozva, megkönnyítve a szótár felhasználását.

2. Az ELTE Verskorpusz rímképletelemző algoritmusa

Az ELTE Verskorpuszban szereplő versek rímképleteinek a felismertetését egy új, az utóbbi egy évben implementált algoritmus végzi el. Míg a régi algoritmus egy szabálykészlet alapján elemezte a versek rímképletét, addig az új algoritmus több – jelenleg nyolc – szabálykészletet használ a rímképlet meghatározásához. A rímképlet felismertetése során az algoritmus célja az, hogy az elemzés kimenete konzisztens legyen, azaz minden versszak ugyanazt a rímképletet kapja meg. Amennyiben az első szabálykészlet alapján történő elemzés nem ad konzisztens eredményt, akkor a program további, egyre lazább szabálykészletek alapján is elemzi a verset. Ha valamelyik szabálykészlet alapján történő elemzés konzisztens eredményt ad, akkor az algoritmus futása leáll, és ezt az eredményt kapja meg a vers annotációként. Az alkalmazott nyolc szabálykészletből az egyiknek kitüntetett a státusza, ugyanis ha a program egyik szabálykészlet alapján sem tudott konzisztens rímképletet rendelni a vershez, akkor a vers rímképleteként ennek a kimenetét adja meg. Ugyanez történik abban az esetben is, ha a versek versszakainak a sorszáma eltérő, vagy csak egy versszakból áll a vers, hiszen ezekben az esetekben eleve nem lehetséges konzisztens rímképlet megadása. Az 1. táblázat mutatja be a nyolc szabálykészletet, abban a sorrendben, ahogyan az algoritmus futása során is alkalmazásra kerülnek.

Szabály-készlet	Utolsó előtti magánhangzó a hosszúságot nem számítva azonos	Utolsó magánhangzó a hosszúságot nem számítva azonos	Utolsó előtti szótag hosszúsága azonos	Szövégi mással-hangzó megléte tekintetében azonos
1	igen	igen	igen	igen
2	nem	igen	igen	igen
3	igen	igen	igen	nem
4	igen	igen	nem	igen
5	nem	igen	igen	nem
6	igen	igen	nem	nem
7	nem	igen	nem	igen
8	nem	igen	nem	nem

1. táblázat. A rímképlet-felismerő algoritmus szabálykészletei

Az 1. táblázatból látható, hogy a legszigorúbb szabálykészlet alapján csak abban az esetben rímel két sor, ha a sorok utolsó és utolsó előtti magánhangzói megegyeznek, ha az utolsó előtti szótagok hosszúsága megegyezik, azaz mind a kettő hosszú vagy rövid szótag, illetve ha mind a két sor végén vagy van mássalhangzó vagy nincs mássalhangzó. Ezzel szemben a nyolcadik, legkevésbé szigorú szabálykészlet – amely valójában csak egy darab szabály alkalmazását jelenti – már abban az esetben is rímelőnek tekint két sort, ha azokban az utolsó magánhangzó a hosszúságot nem számítva megegyezik. Amennyiben egyik szabálykészlet alapján sem tud a program konzisztens elemzést adni, a program a második szabálykészlet alapján meghatározott rímképlettel annotálja a verset. Hasonlóan, amennyiben nem egyenlő a versszakok sorszáma, vagy csak egy versszakból áll a vers, a program csak a második szabálykészlet alapján elemzi le a verset, és az így kapott rímképlettel annotálja azt.

A fent bemutatott, nyolc szabálykészletet használó algoritmus alkalmazásával a cél az volt, hogy a lehető legtöbb konzisztensen, azaz a versszakokat azonos rímképlettel annotált verset kapjuk meg. A 2. táblázat azt mutatja be, hogy a 13 362 versből hány konzisztensen annotált verset kapunk, ha a második szabálykészlet, a második, ötödik és nyolcadik szabálykészlet együttes, illetve mind a nyolc szabálykészlet együttes alkalmazásával végezzük el a rímképletek meghatározását.

Alkalmazott szabálykészlet	Konzisztens rímképlettel elemzett versek száma
2. szabálykészlet	5054
2., 5. és 8. szabálykészlet	5357
Mind a nyolc szabálykészlet	5983

2. táblázat. Az algoritmus által konzisztensen elemzett versek száma

3. A rímszótár adatai és formátumai

A rímszótár a rímpárokat alkotó szóalakok mellett három típusú adatot tartalmaz. Egyrészt tartalmazza a rímpárt alkotó szavak grammatikai és fonológiai jellemzőit: a szótári alakot, a szófajt, a morfoszintaktikai jellemzőket, a szótagszámot, a hangrendet, valamint a fonológiai szerkezet egyszerűsített reprezentációját. Ezek az információk az ELTE Verskorpuszban is szerepelnek annotációkként, a grammatikai jellemzők az e-magyar programmal [3, 4, 5], a fonológiai jellemzők pedig a verskorpuszhoz fejlesztett annotáló programmal lettek felismertetve. Emellett a rímszótár tartalmazza a rímpárokat alkotó szavak pozíciójára vonatkozó jellemzőket: a rímpártagok egymástól való távolságát sorszámban kifejezve (egy rímpár tagjai között maximum négy sor lehet), a rímpártagok versbeli sorrendjét, valamint a rímpártagok közé esetlegesen beékelődő, a rímpártagokkal rímelő sorok számát. Végezetül a szótár úgyszintén tartalmazza a rímpárok előfordulására vonatkozó bibliográfiai információkat, azaz a rímpárt tartalmazó vers szerzőjét, címét, azonosítóját és az ELTE Verskorpusz lekérdezőfelületére mutató URL-jét, valamint a rímpárt tartalmazó versszak azonosítóját.
A rímszótárban szereplő rímpárok ábécésorrendben szerepelnek. A rímpárok ábécérendbe sorolása a következő elvek szerint történt:

1. A rímpárok lemmák alapján vannak ábécérendbe sorolva.
2. Az azonos lemmájú, de eltérő szófajú rímpárok esetében az azonos szófajú lemmákat megvalósító rímpárok egymást követően sorolódnak fel.
3. Az azonos lemmájú és azonos szófajú rímpárok szóalakok szerint vannak ábécérendbe sorolva.

A rímpárokat megvalósító előfordulások felsorolásánál azok az előfordulások szerepelnek előbb, ahol a rímpártagok közötti távolság kisebb.
A rímszótárt létrehozó szkript kimenete egy XML-fájl, amely a rímszótár összes rímpárját és a rímpárokhoz tartozó, fent felsorolt adatokat tartalmazza. Az XML-formátum mellett a rímszótár TSV, SQLite és PDF formátumban is elérhető, az utóbbiban, az emberi olvasás megkönnyítése miatt nem szerepel a rímpárokra vonatkozó összes információ. A rímszótár az összes formátumban letölthető a https://github.com/ELTE-DH/rhyming-dictionary oldalról. A github oldalon szereplő dokumentáció részletesen bemutatja az egyes formátumokat.

CC BY

4. A rímszótár online elérhető lekérdezőfelülete

A rímszótár online lekérdezőfelülete a https://rimszotar.elte-dh.hu címen érhető el. A Python FastApi keretrendszerében programozott lekérdezőeszköz a rímszótár SQLite-formátumú relációs adatbázisában keres, amely a rímszótár XML-verziójából lett generálva.[1] Szóalakok és lemmák alapján is kereshetünk. A találatokat emellett szűrhetjük szerző, szófaj, valamint a rímpártagok pozíciója alapján is. A lekérdezés eredménye TSV-formátumban letölthető, és bármilyen táblázatkezelő programban megnyitható. A lekérdezőfelület Súgójában részletes leírás olvasható a keresőfelület használatáról és a kimenetként kapott adatokról.

5. A rímelés néhány általános mintázata a magyar kanonikus költészetben

A rímszótár adatai alapján különböző, a magyar kanonikus költészet rímelésével kapcsolatos kérdések vizsgálhatók. Az alábbiakban három ilyen kérdés vizsgálatára térek ki röviden. A vizsgálatokban csak azokat a rímpárokat vettem figyelembe, ahol a rímpár két tagja közé nem ékelődik be egy azokkal rímelő sor.

5.1. A rímpártagok hosszúsága

A 3. táblázat azt mutatja be, hogy az ötvenből hány olyan szerző van, akinél a legnagyobb számban azok a rímpárok fordulnak elő, amelyekben a hívó rímszó és a felelő rímszó ugyanolyan szótagszámú, illetve hány olyan, akinél nagyobb szótagszámú hívó rímszóval, illetve nagyobb szótagszámú felelő rímszóval rendelkező rímpárok fordulnak elő a legnagyobb számban. A 4. táblázat nem veszi figyelembe azt az esetet, amikor a hívó és a felelő rímszó ugyanannyi szótagszámú, azaz itt pusztán az szerepel, hogy az ötvenből hány olyan szerző van, akinél nagyobb számban fordulnak elő azok a rímpárok, amelyekben a hívó rímszó a hosszabb, nem pedig a felelő rímszó, illetve akinél nagyobb számban fordulnak elő azok a rímpárok, amelyekben a felelő rímszó a hosszabb, nem pedig a hívó rímszó.

Legnagyobb számú kombináció	hívórím = felelőrím	hívórím > felelőrím	hívórím < felelőrím
Szerzők száma	38	0	12

3. táblázat. Az egyes hosszúságkombinációkat legnagyobb mértékben megvalósító szerzők száma

Nagyobb számú kombináció	hívórím > felelőrím	hívórím < felelőrím	Ugyanannyiszor fordul elő a két eset
Szerzők száma	12	36	2

4. táblázat. A hosszabb hívó rímszóval és a hosszabb felelő rímszóval rendelkező rímpárokat nagyobb mértékben megvalósító szerzők száma

A táblázat adataiból látható, hogy a rímpárok legnagyobb részében a hívó és a felelő rímszó ugyanolyan hosszú. Ha pedig a hívó és a felelő rímszó nem ugyanolyan hosszú, akkor jellemzően a felelő rímszó a hosszabb. Ennek feltételezésem szerint valamilyen pszicholingvisztikai oka lehet, ami miatt az ember egy szó kapcsán könnyebben asszociál olyan rímelő szóra, amely ugyanolyan szótagszámú vagy nagyobb szótagszámú, mint olyanra, amely rövidebb szótagszámú. Ezt a hipotézist további kísérletes vizsgálatokkal lehetne megerősíteni.

---

1 Köszönettel tartozom Indig Balázsnak, aki sokat segített a FastApi és a relációs adatbázisok használatának megértésében, és az elkészült programot a szükséges módosításokkal felrakta a tanszék szerverére. Úgyszintén köszönettel tartozom Nagy Mihálynak, aki számos, a lekérdező készítése során felmerülő programozási kérdésemet megválaszolta.

5.2. A rímszavak hangrendje

Az 5. táblázat azt mutatja be, hogy hány szerzőnél fordulnak elő az adott hangrendbe tartozó rímszavak a legnagyobb mértékben.

Legnagyobb számú hangrend	magas	mély	vegyes
Szerzők száma	49	1	0

5. táblázat. A magas, mély és vegyes hangrendű rímszavakat legnagyobb mértékben használó szerzők száma

A táblázatból látható, hogy egy kivételével az összes szerző esetében a magas hangrendű szavak jelennek meg a legnagyobb mértékben rímhelyzetben. Az egyetlen kivétel Tinódi Sebestyén, akinél a mély hangrendű rímszavak fordulnak elő a legnagyobb számban. Ennek oka minden bizonnyal a vala szó rímhelyzetben való nagymértékű használata [6, 7].

A 6. táblázat az ELTE Regénykorpusz[2] mondatvégi szavainak és az ELTE Verskorpusz rímszavainak a hangrendi megoszlását mutatja be. Összehasonlítási alapul a regénykorpusznak azért a mondatvégi szavait használtam, mivel rímhelyzetben nem akármilyen szó, hanem jellemzően tagmondat végi szavak állnak. A verskorpusz esetében az egyes szerzőknél talált arányok mediánját tüntettem fel.

	magas	**mély**	**vegyes**
Regénykorpusz	44,1%	33,5%	22,1%
Verskorpusz (medián)	48,7%	33,9%	17,0%

6. táblázat. A magas, mély és vegyes hangrendű rímszavak arányainak mediánja összevetve a regénykorpusz mondatvégi szavaival

Látható, hogy a verskorpusz rímszavai esetében nagyobb a magas hangrendű szavak és kisebb a vegyes hangrendű szavak aránya, mint a regénykorpusz mondatvégi szavainál. Ennek az oka az lehet, hogy egy adott hangrendbe tartozó szó általában egy vele egy hangrendbe tartozó szóval rímel, és egy több szót magában foglaló hangrendcsoport esetében könnyebb két egymással rímelő szót találni, mint egy kevesebb szót magában foglaló hangrendcsoport esetében. Vagyis a rímkényszer miatt a regénykorpusznál is meglévő, eleve létező különbségek erősödnek fel.

5.3. Ragrímek

A 7. és 8. táblázat az azonos szófajú hívó és felelő rímszóval rendelkező rímpárok, illetve az azonos szófajú és morfoszintaktikai tulajdonságú hívó és felelő rímszóval rendelkező rímpárok arányainak a mediánját mutatja be. A táblázatokban feltüntettem azt is, hogy a regénykorpusz mondatvégi szavai által alkotott random szópárokból mennyi az azonos szófajú, illetve az azonos szófajú és azonos morfoszintaktikai tulajdonságú tagokkal rendelkező szópárok várható értéke.

	Szófaj	**Szófaj + morfoszintakszis**
Verskorpusz (medián)	0,462	0,285
Regénykorpusz (várható érték)	0,265	0,031

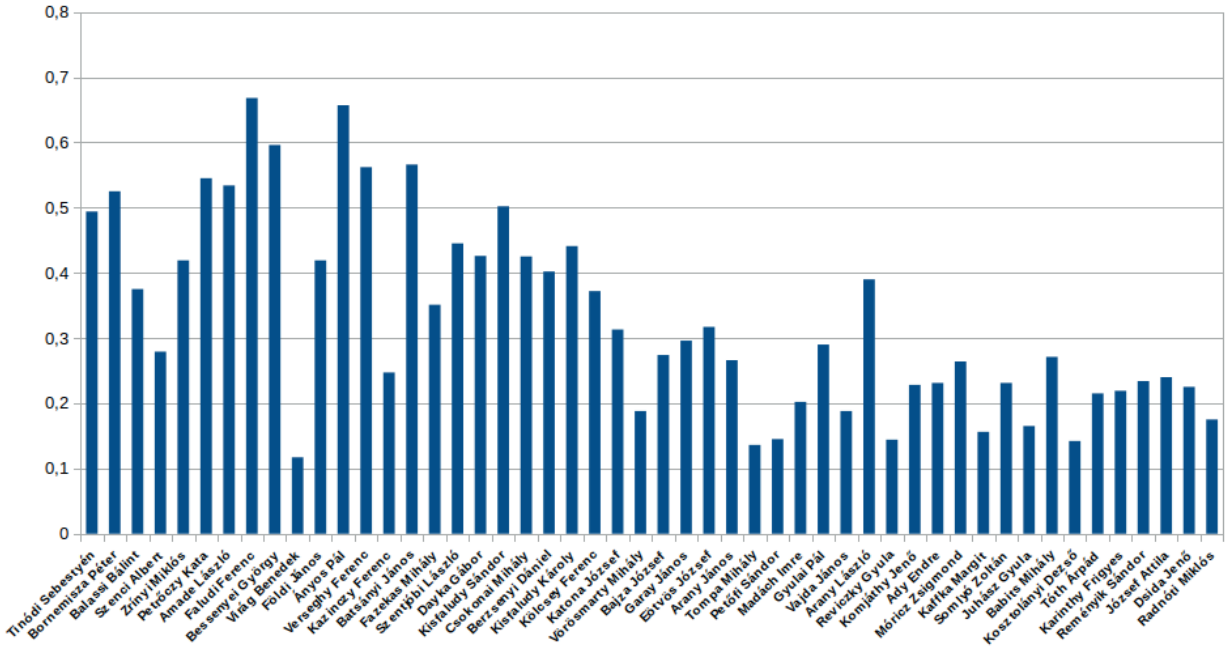
7. táblázat. Az azonos szófajú, valamint azonos szófajú és azonos morfoszintaktikai tulajdonságú rímszavakból álló rímpárok arányainak mediánja összevetve a regénykorpusz random mondatvégi szavaira kapott várható értékkel

2 https://github.com/ELTE-DH/regenykorpusz

CC BY

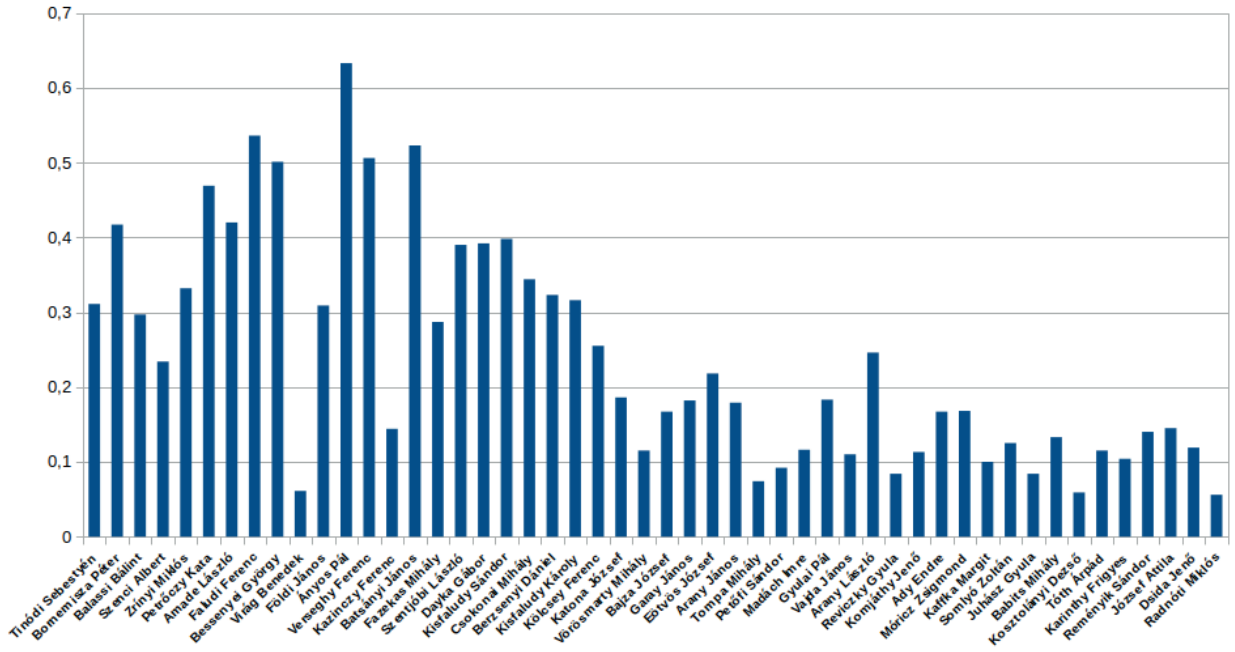
A táblázat számaiból látszik, hogy az azonos szófajú tagokkal rendelkező rímpárok arányainak a mediánja több, mint másfélszerese, az azonos szófajú és morfoszintaktikai tulajdonságú tagokkal rendelkező rímpárok arányainak a mediánja pedig majdnem tízszerese a regénykorpusz random mondatvégi szópárjaira kapott várható értéknek. Az azonos szófajú, illetve azonos szófajú és azonos morfoszintaktikai tulajdonságú tagokkal rendelkező rímpárok nagy aránya a ragrímek használatával magyarázható, hiszen a ragrímek esetében a rímelő tagoknak jellemzően azonos a szófaja, és az azonos toldalékok miatt azonosak a morfoszintaktikai jellemzői (a régi magyar irodalom ragrímeit kvantitatív módon már vizsgálta Seláf és Plecháč [6], valamint Maróthy, Seláf és Plecháč [7]).

Az azonos szófajú és morfoszintaktikai tulajdonságú rímpárok szerzőnként kapott arányait érdemes rávetíteni egy időbeli tengelyre, amiből láthatóvá válik a ragrímek használatának változása a magyar kanonikus költészetben. Az 1. ábra ezt mutatja be oszlopdiagram formájában. Az oszlopdiagramból látható, hogy a 19. században az azonos szófajú és morfoszintaktikai tulajdonságú szavak aránya lecsökken, és a 20. század első felében is alacsony marad.

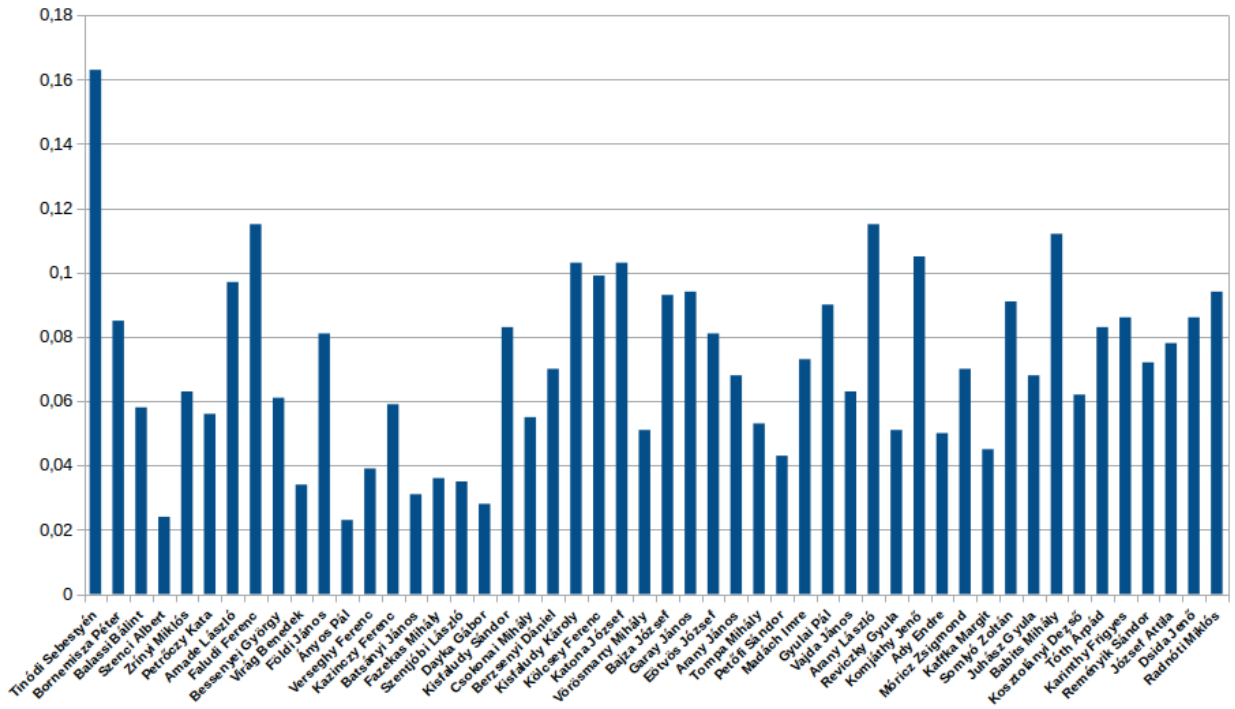


1. ábra. Azonos szófajú és azonos morfoszintaktikai tulajdonságú rímszavakból álló rímpárok arányai

Annak érdekében, hogy biztosak lehessünk abban, hogy az 1. ábrán látható csökkenő tendenciát a ragrímek visszaszorulása okozza, szerzőnként lekérdeztem a főnévi, melléknévi és igei, azonos szófajú és morfoszintaktikai tulajdonságú szavakkal rendelkező rímpárok arányait is oly módon, hogy kizártam az alanyesetű, egyes számú főneveket, az alanyesetű, egyes számú, alapfokú mellékneveket, valamint az E/3, kijelentő módú, jelen idejű, határozatlan igéket, vagyis a három szófaj ragozatlan szóalakjait. Az így nyert gyakorisági adatokat a 2. ábrán szereplő oszlopdiagram mutatja be. Látható, hogy a csökkenő tendencia még erősebb. A 3. ábrán szereplő oszlopdiagram pedig az előző esetben kizárt eseteket, vagyis a főnévi, melléknévi és igei, azonos szófajú és morfoszintaktikai tulajdonságú, de ragozatlan szavak által alkotott rímpárok arányait mutatja be. Ebben az esetben csökkenő tendenciáról egyáltalán nem beszélhetünk, ami megerősíti, hogy a 2. ábrához hasonlóan az 1. ábrán szereplő oszlopdiagram időben csökkenő tendenciáját is a ragrímek visszaszorulása okozza.



2. ábra. A főnévi, melléknévi és igei, azonos szófajú és azonos morfoszintaktikai tulajdonságú, ragozott rímszavakból álló rímpárok arányai



3. ábra. A főnévi, melléknévi és igei, azonos szófajú és azonos morfoszintaktikai tulajdonságú, ragozatlan rímszavakból álló rímpárok arányai

CC BY

Hivatkozott irodalom

[1] Mártonfi Attila: Egy magyar rímszótár terve. In: Bartók István et al. (szerk.) *„Mielz valt mesure que ne fait estultie”*. A hatvanéves Horváth Iván tiszteletére. Budapest: Krónika Nova Kiadó pp. 198–204, 2008.

[2] Horváth Péter, Kundráth Péter, Indig Balázs, Fellegi Zsófia, Szlávich Eszter, Bajzát Tímea Borbála, Sárközi-Lindner Zsófia, Vida Bence, Karabulut Aslihan, Timári Mária, Palkó Gábor: ELTE Verskorpusz – a magyar kanonikus költészet gépileg annotált adatbázisa. In: Berend Gábor, Gosztolya Gábor, Vincze Veronika (szerk.) *XVIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2022)*. Szeged: Szegedi Tudományegyetem TTIK, Informatikai Intézet. pp. 375–388, 2022.

[3] Váradi Tamás, Simon Eszter, Sass Bálint, Mittelholcz Iván, Novák Attila, Indig Balázs, Farkas Richárd, Vincze Veronika: e-magyar – A digital language processing system. In: Nicoletta Calzolari et al. (eds.) *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Paris: European Language Resources Association pp. 1307–1312, 2018.

[4] Indig Balázs, Sass Bálint, Simon Eszter, Mittelholcz Iván, Kundráth Péter, Vadász Noémi: emtsv – Egy formátum mind felett. In: Berend Gábor, Gosztolya Gábor, Vincze Veronika (szerk.) *XV. Magyar Számítógépes Nyelvészeti Konferencia. Szeged: SZTE Informatikai Intézet* pp. 235–247, 2019.

[5] Novák Attila, Rebrus Péter, Ludányi Zsófia: Az emMorph morfológiai elemző annotációs formalizmusa. In: Vincze Veronika (szerk.) *XIII. Magyar Számítógépes Nyelvészeti Konferencia*. Szeged: Szegedi Tudomány-egyetem, Informatikai Intézet pp. 70–78, 2017.

[6] Seláf Levente, Petr Plecháč, Számoljuk meg a valákat! *A históriás ének rímelése. In: Seláf Levente (szerk.) A históriás ének. Poétikai és filológiai kérdések*. Budapest: Gépeskönyv, 2023.

[7] Maróthy Szilvia, Seláf Levente, Petr Plecháč: Rhyme in 16th-Century Hungarian Historical Songs: A Pilot Study. In: Plecháč, Petr – Kolár, Robert – Bories, Anne-Sophie – Říha, Jakub (eds.) *Tackling the Toolkit: Plotting Poetry through Computational Literary Studies*. Prague: Institute of Czech Literature of the Czech Academy of Sciences pp. 43–58, 2021.

 DOI: 10.31915/NWS.2023.13

Digitális tartalomfejlesztés közkönyvtári környezetben

Developing digital content in a public library

Héjja Balázs
Hamvas Béla Pest Megyei Könyvtár
informatika@pmk.hu

Tóth-Jávorka Brigitta
Hamvas Béla Pest Megyei Könyvtár
javorka.brigitta@pmk.hu

Tóth Máté
Hamvas Béla Pest Megyei Könyvtár
igazgato@pmk.hu

Absztrakt

A tudományos könyvtárak világában igen elterjedt gyakorlat, hogy az intézmények nemcsak közvetített szolgáltatásként kínálnak máshol előállított tartalmakat, hanem saját maguk is tartalom-előállítóként lépnek fel. Gondoljunk csak a referáló lapokra, az adatbázis-tartalmakra, vagy tutorial videókra, amelyeket az egyetemi és a szakkönyvtárak alapvető funkcióik ellátása érdekében rendszeresen fejlesztenek. A közkönyvtárakban éppen az eltérő funkciók miatt kevésbé jellemző ez az attitűd.

A szentendrei Hamvas Béla Pest Megyei Könyvtárban a COVID-19 járvány idején kényszerűségből indított online közvetítések teremtették meg az olvasói igényt arra, hogy a könyvtár maga is fejlesszen önállóan tartalmakat helytörténeti témákban. Az online előadásokkal komplex, az oktatásban, az ismeretterjesztésben és általában a helytörténeti munkában is jól használható tudásbázis jött létre, amely hosszú távon a jelenleg fejlesztés alatt álló digitális könyvtárunk videótárának alapja lesz.

Az előadás bemutatja a közkönyvtári tartalomfejlesztés lehetőségeit általában, majd a Hamvas Béla Pest Megyei Könyvtár gyakorlatát mint egy esettanulmányt ismerteti az igény felmerülésétől a megvalósításig, kitekintéssel a jövőre, a digitális könyvtárba való integrálás lehetőségeire.

Abstract

Among scientific libraries it is a widespread practice that the institutions are functioning as content creators and not just providers of content which is created by others. While periodical indexes, database content, reference tools and tutorial videos are developed by university and scientific libraries to fulfil their basic functions, the creation of content by public libraries is much less typical.

The emergence of user needs for online streaming services under the pressure of COVID19 pandemic stimulated Béla Hamvas Pest County Library to develop professional content in the field of local history. The online presentations, videos form a continuously growing knowledge

CC BY

base which is well useable in local history work in general and will be a very important part of our digital library which is under development recently in the county library.

The paper presents the possibilities of online content development in a public library context in general and the practice of Béla Hamvas County Library as a case study in particular from the emergence of needs to the realisation, with outlook to the future and to the possibilities of integrating content into a digital library.

A könyvtár mint tartalom-előállító

A Hamvas Béla Pest Megyei Könyvtár a COVID-19 járvány alatt számos egyedi helytörténeti, helyismereti tartalmat fejlesztett és tett közzé az intézmény elektronikus felületein. Ezek a dokumentumok egy idő után egy egyre jelentősebb és egyre értékesebb gyűjteményrészt tettek ki, amelyet mind tudatosabban kezdtünk el menedzselni az elmúlt években. A tanulmány célja, hogy bemutassa a közkönyvtári tartalomfejlesztés lehetőségeit általában, majd a Hamvas Béla Pest Megyei Könyvtár gyakorlatát mint egy esettanulmányt az igény felmerülésétől a megvalósításig. Végül kitekintünk a jövő útjaira, az elkészült tartalmak digitális könyvtárban való megjelenítésének lehetőségeire.

A könyvtárak hagyományosan a publikált tudást hivatottak feldolgozott és rendszerezett formában közzétenni, de a tudományos könyvtárak világában nem ritka, hogy maguk is tartalom-előállítóként funkcionálnak. A könyvnyomtatást megelőzően, amennyiben egy könyvtárnak szüksége volt egy tudományos szövegre, a beszerzés egyetlen útja volt, hogy kézírással lemásoltatták egy újonnan készített dokumentumba. Az Alexandriai Könyvtár tűzte ki célul a világtörténelemben először, hogy a tudományos világ összes szellemi termékét összegyűjti, amelynek érdekében másolatokat készítettek a tartalmakról, új könyveket hoztak létre.[1]

Napjainkban az egyetemi és a szakkönyvtárak a referáló lapokban új tudományos eredményeket nem, de mégis új dokumentumokat hoznak létre, amelyekben reprezentálják az adott tudományterületek aktuális állását. Szintén az egyetemi könyvtárak gyakorlata, amikor az adatbázisok vagy a tájékoztató eszközrendszer használatáról, a publikálásról, a tudományos kommunikáció bármely aspektusáról készítenek tutorial videókat a használóik számára.

Egy közkönyvtárban a fő szempont nem a tudományos tartalmak előállításán vagy közvetítésén van, mégis találhatunk olyan területeket, ahol releváns lehet az egyedi tartalmak előállítása. Ilyen lehet az információs írástudás vagy a kritikus gondolkodás fejlesztése mellett a helytörténeti kutatások eredményeinek bemutatása is. A közkönyvtárak egyre inkább olyan közösségi interakciós térként definiálják magukat, amelyben inspiráló találkozások zajlanak, majd ezek eredményeként új érték jön létre, akár új tartalmak is születhetnek.[2]

A Hamvas Béla Pest Megyei Könyvtár gyakorlata

A szentendrei Hamvas Béla Pest Megyei Könyvtárban a COVID-19 járvány idején kényszerűségből indított online közvetítések teremtették meg az olvasói igényt arra, hogy a könyvtár maga is fejlesszen önállóan tartalmakat, elsősorban helytörténeti témákban.

1 Vallejo, Irene (2022): Papirusz: A könyvek története az ókori világban. Budapest, Magnólia kiadó.

2 Tóth Máté (2022): A könyvtárak társadalmi szerepei empirikus kutatási adatok tükrében. Szentendre, Hamvas Béla Pest Megyei Könyvtár.

A bezárások ideje alatt sem szerettük volna feladni a könyvtári programok megtartását, ezért sok más intézményhez hasonlóan[3] mi magunk is az online közvetítések eszközéhez nyúltunk. Igyekeztünk egy az egyben online térbe helyezni a programjainkat.

Ebben az időszakban keresett meg minket Thiel Katalin filozófia professzor asszony egy Hamvas Béla életéről és munkásságáról szóló online sorozat ötletével. Az előadások célja egy olyan bevezető nyújtása volt, amely a filozófia területén kevésbé jártas olvasóknak is segít értelmezni a könyvtár névadójának műveit. Az előadások Thiel Katalin otthonában kerültek felvételre, majd YouTube premierként kerültek publikálásra, és azóta is elérhetők maradtak *„A nevezetes névtelen”* című lejátszási listában. Azt figyeltük meg, hogy bár élőben csak húsz-harminc, maximum néhány száz fő követte az eseményt, az azt következő napokban több esetben ezer fölé is felkúszott a megtekintők száma.[4] A Hamvas előadás-sorozattal sikerült megszólítanunk azokat a fizikailag távol élő Hamvas Béla rajongókat is, akik személyesen soha nem jöttek volna el a könyvtárunkba.

A Hamvas-sorozat mellett számos ismert személyiséget is vendégül láttunk, akiket vagy a könyvtár olvasótermében zajló beszélgetés közvetítésével, vagy online bejelentkezés útján juttattunk el virtuálisan a járványhelyzet miatt kényszerűségből az otthonukban tartózkodó olvasóinkhoz.

A tartalmak közvetítésén túl megpróbáltuk minél hitelesebben reprodukálni a személyes találkozásokban rejlő lehetőségeket is. A Hamvas-sorozatban a chat funkción érkező kérdéseket már közvetlenül az előadást követően megválaszolta az előadó, a személyes találkozások közvetlenségét pedig az előadást követő élő beszélgetésekkel igyekeztünk helyettesíteni, amelyeknek a Zoom platform adott otthont.

A könyvtárak újranyitását követően úgy döntöttünk, hogy – tekintettel arra, hogy a rendezvényeinket látogatók jellemzően az idősebb korosztályhoz tartoznak – megtartjuk az online közvetítéseket, hogy a vírus által veszélyeztetettebb, ebből kifolyólag emiatt óvatosabb olvasóink se maradjanak programok nélkül. Ekkor párhuzamosan zajlottak az előadások, író-olvasó találkozók a fizikai és az online térben. Az egyik ilyen igen sikeres program volt Török Katalin „Szentendre új megvilágításban” című helytörténeti előadás-sorozata.[5] Ennek az előadássorozatnak azóta a második évadát is megtartottuk, de a korábbi élő közvetítésről egy hetes csúszásban publikálásra váltottunk, hogy a pandémia lecsengésével a motivált résztvevőket a személyes jelenlétre biztassuk.

Szintén ekkor volt a Szentendrei Járás Egészségfejlesztési Irodával együttműködésben a „Mégis kinek az egészsége?” című előadás-sorozat[6], amelyben elsősorban a középkorúak számára nyújtottak a betegségek megelőzésével és az egészséges életmód kialakításával kapcsolatos ismereteket.

---

3 A többi könyvtár tevékenységeiről ld. Bódog András a Könyvtárak a koronavírus-járvány idején című tanulmányát.

4 A nevezetes névtelen. Bevezetés Hamvas Béla életművébe. Thiel Katalin előadássorozata. Lejátszási lista Hamvas Béla Pest Megyei Könyvtár YouTube csatornáján. https://www.youtube.com/playlist?list=PLITGZi7F4UFvLqpzQnaFoyCrqKTTGtYN6

5 Szentendre új megvilágításban. Török Katalin előadássorozata. Lejátszási lista a Hamvas Béla Pest Megyei Könyvtár YouTube csatornáján. https://www.youtube.com/playlist?list=PLITGZi7F4UFvjWklnvj1LRWeCHBIcTZQr

6 Mégis kinek az egészsége? A Szentendrei Járás Egészségfejlesztési Iroda előadássorozata. Lejátszási lista a Hamvas Béla Pest Megyei Könyvtár YouTube csatornáján. https://www.youtube.com/playlist?list=PLITGZi7F4UFvQeOT4QnxZxGe1SbBhmYcA

CC BY

Ezek mellett azonban igyekeztünk valamennyi író-olvasó találkozót, előadást, kiállításmegnyitót online térben is közvetíteni. Miközben mi a pandémia miatt kialakult veszélyhelyzet folyamatos oldódása miatt egyetlen fizikai térben megrendezett programot sem mulasztottunk el virtuálisan is elérhetővé tenni, addig fokozatosan szoktak hozzá az olvasók ahhoz, hogy minden, ami a könyvtárban történik, online is elérhető, élőben is követhető, majd visszanézhető.

Egy-egy kiállításmegnyitó, egy-egy országosan ismert író, költő vagy közéleti személyiség szentendrei fellépése mind olyan alkalmak, amelyek a közösség története szempontjából meghatározó jelentőségűek. A felvételeket egyre tudatosabban kezdtük el készíteni a szentendrei közösség életének dokumentálása érdekében is. Például az óvodás gyerekek rajzaiból készített kiállítás megnyitója nem is elsősorban a jelen érdeklődő közönsége, hanem sokkal inkább a jövőből az intézmény történetére való visszatekintés szempontjából bír jelentőséggel.

2022 tavaszán, a személyes találkozók stimulálása, a személyes jelenlét ösztönzése érdekében bevezettük, hogy minden felvétel csak késleltetve jelenik meg az intézmény YouTube-csatornáján. Összességében minden évben több tucatnyi videó került fel, amelyek nézettsége minden korábbi várakozásunkat felülmúlta és magasan meghaladta a személyes látogatók számát. Egy idő után a könyvtár által előállított nagy mennyiségű tartalom olyan értéket kezdett képviselni, amelynek a menedzselése önmagában is komoly feladatnak tűnt. A programok közvetítése, dokumentálása mellett a könyvtár saját helytörténeti kutatásaiból szintén készült két film, amelyeket a YouTube-csatornánkon tettünk közzé. Az egyik a Szentendrei Betonárugyár száz éves jubileumára készített kutatásunk interjúiból összeállított dokumentumfilm[7], a másik a megyei könyvtár alapításának 70 éves évfordulója alkalmából készített riportfilm volt. Ezeket a tartalmakat a könyvtár munkatársai készítették saját erőforrásból.

A rendelkezésünkre álló technikai feltételeket az igények felmerülésével párhuzamosan folyamatosan igyekeztünk javítani. A szerény anyagi lehetőségeinkhez mérten jelentős beruházást jelentettek a jó minőségű kamera, a térmikrofonok és a reflektorok, amelyekkel, ha nem is professzionális, de az amatőr felvételeknél jóval magasabb színvonalú felvételeket tudtunk készíteni.

Az alábbi technikai eszközöket használjuk a felvételek során:

- Panasonic HC-X 1500 kamera
- Zoom H2N hangrögzítő
- HDMI-USB átalakító (noname)
- Slik kameraállvány
- 4 db BeamZ BT450 ledes PAR reflektor
- Laptop AMD Ryzen processzorral
- OBS stúdió streamer szoftver

A saját fejlesztésű tartalmak közzététele mellett elkezdtük a régi, elavult formátumokban lévő filmek digitalizálását is. A 8 mm-es szalagok digitalizálását Négyesi Gábor, az egykor a megyei könyvtárban működő amatőr filmklub munkatársa önkéntes munkában vállalta, ugyanis nála rendelkezésre álltak a munkához szükséges technikai eszközök. Szintén

7 A Hamvas Béla Pest Megyei Könyvtár SZEBETON projektjéről ld. Tóth Máté. Könyvtári partnerkapcsolatok kialakítása című tanulmányát.

fontos részét képezik a helytörténeti tárunknak azok a VHS kazetták, amelyeket az elmúlt évtizedekben készítettek a könyvtár programjain. Olyan művészek szentendrei látogatásairól szóló felvételeink is vannak, akik azóta már nincsenek az élők sorában (Pl. Cseh Tamás, Faludy György).

Szintén részben az új tartalmak előállítása, részben pedig a közösségekben rejlő szinergiák erősítése hívta életre a Helytörténet és helyi közösség című konferenciát, amelyet a Hamvas Béla Pest Megyei Könyvtár rendezett meg 2023. április 24-én. Ebben a kezdeményezésben fórumot és előadási lehetőséget kívántunk teremteni a Pest megye helytörténetével foglalkozó kutatóknak, hogy bemutassák az eredményeiket a megyei könyvtárban. Valamennyi előadást felvettük és közzétettük a könyvtár közösségimédia-felületein videó formátumban. Szintén kértünk az előadóktól tanulmányt, amelyet olvasószerkesztést követően egy ISBN számmal ellátott elektronikus konferenciakötetben fogunk publikálni. Ezek a tartalmak szintén az intézmény digitális könyvtárát fogják gazdagítani hosszú távon.

A terveink között szerepel, hogy tudatosan készítünk interjúkat, riportokat híres szentendrei lakosokkal, intézményvezetőkkel, a közösség életében meghatározó szerepet játszó személyiségekkel, amelyekből a későbbiek folyamán lehetőség nyílik podcastok vagy rövidebb terjedelmű videóriport összeállítására. Ezeket a tartalmakat is a digitális könyvtárban kívánjuk elhelyezni.

Az egyedileg fejlesztett tartalmak jövője

Az online előadásokkal komplex, az oktatásban, az ismeretterjesztésben és általában a helytörténeti munkában is jól használható tudásbázis jött létre, amely hosszú távon a jelenleg fejlesztés alatt álló digitális könyvtárunk videótárának alapja lesz. A digitális könyvtár fejlesztését 2023 februárjában kezdtük el azzal a céllal, hogy egységes és vonzó megjelenésű felületet kínáljunk azoknak a helytörténeti tartalmainknak, amelyek jelenleg is teljes szöveggel érhetők el a könyvtár katalógusából. Az elektronikus formában elérhető könyvek, aprónyomtatványok, fotók, cikk-kivágatok mellett a videótartalmak egy igen markáns – és reményeink szerint intenzíven használt – részét fogják alkotni a digitális könyvtárnak.

A célunk, hogy ezekre a tartalmakra a jelenleginél sokkal könnyebben rábukkanjon az érdeklődő: ne csak akkor, amikor célzottan erre keres a könyvtár YouTube-csatornáján, hanem akkor is, amikor adott témához keres irodalmat vagy tartalmat a digitális könyvtár felületén. A saját magunk által fejlesztett videódokumentumokat a többi – a felhasználó számára releváns – találat között szeretnénk megjelentetni, ezzel bemutatni az adott tartalom kontextusát képező anyagokat is. A videók mögé kívánjuk tenni az elhangzó szöveg leiratát képező szövegfájlt, amelyben teljes szövegű keresés végezhető.

Összegzés

A digitális tartalmak előállítása kevésbé jellemző tevékenység közkönyvtári környezetben, különösen az országos szakkönyvtárak és az egyetemi könyvtárakhoz képest, pedig a helytörténet vonatkozásában a közkönyvtáraknak is van tudományos funkciója. A tartalmak előállítása a Hamvas Béla Pest Megyei Könyvtár gyakorlatában elsősorban a közösségi programok és a közösségépítés kapcsán kezdődött, majd a későbbiekben önállóan is fontos célként fogalmazódott meg. A könyvtár YouTube-csatornáján tárolt tartalmak a szentendrei helytörténeti kutatás egyre értékesebb forrását jelentik, amelyet tudatosan kell menedzselni.

CC BY

A könyvtár által előállított tartalmak igen elenyésző részét teszik ki a készülő digitális könyvtár állományának. Figyelemre méltó ugyanakkor, hogy ezeknek a tartalmaknak a használata sokszorosa a már publikált dokumentumokénak. Hosszú távon fontos, hogy a tartalmak létrehozása és digitális könyvtárban való közzététele fenntartható módon beépüljön a könyvtár munkafolyamatainak a sorába.

Irodalom

Bódog András (2020): Könyvtárak a koronavírus-járvány idején: Pandémia és infodémia. *Könyvtári figyelő*. 66. évf. 3. sz. pp 419-436. https://epa.oszk.hu/00100/00143/00362/pdf/EPA00143_konyvtari_figyelo_2020_03_419-436.pdf Hozzáférés: 2023. augusztus 13.

Tóth Máté (2022): A könyvtárak társadalmi szerepei empirikus kutatási adatok tükrében. Szentendre, Hamvas Béla Pest Megyei Könyvtár.

Tóth Máté (2023): Könyvtári partnerkapcsolatok kialakítása: Esettanulmány a Hamvas Béla Pest Megyei Könyvtár gyakorlatából. *Tudásmenedzsment*. 24. évf. Ünnepi különszám Varga Katalin 60. születésnapja alkalmából https://journals.lib.pte.hu/index.php/tm/article/view/6320 Hozzáférés: 2023. augusztus 13.

Vallejo, Irene (2022): Papirusz: A könyvek története az ókori világban. Budapest, Magnólia kiadó.

 DOI: 10.31915/NWS.2023.14

Szemelvények egy felsőoktatási rendszer informatikai védelmének tapasztalataiból

Koczka Ferenc
Eszterházy Károly Katolikus Egyetem, Információtechnológiai Tanszék,
Nemzeti Közszolgálati Egyetem, Kiberbiztonsági Tanszék
koczka.ferenc@uni-eszterhazy.hu.

Absztrakt
A felsőoktatásban működő informatikai rendszerek védelmével kapcsolatban meglehetősen kevés tudományos mű áll rendelkezésre. Csak néhány publikusan elérhető nemzetközi forrásban lelhetők fel olyan adatok, melyek részleges képet nyújtanak az oktatási intézményeket érintő informatikai incidensekről. Tekintettel arra, hogy hazai viszonylatban ezek elenyésző mértékben állnak rendelkezésre, a magyar oktatási intézmények kibervédelmi incidenseinek számáról, azok okairól és a támadások motivációiról nincs reális képünk. Adatok hiányában a nemzetközi tapasztalatokra hagyatkozhatunk: az azokból kiolvasható tendenciák várhatóan hazai viszonylatban is érvényesek lehetnek. Cikkemben egy ilyen adatforrás elemzését végzem el.

Kulcsszavak: oktatási intézmények védelme, kibervédelem, informatikai incidensek.

Abstract
There are a limited number of academic resources on the protection of IT systems in higher education. Only a few international public sources provide detailed data, which only give an overview of the number and nature of IT incidents affecting educational institutions. Such data on Hungarian incidents are scarce, so not much is known about cyber security incidents in Hungarian educational institutions and their causes and motivations. In the absence of data, we can rely on international experience, the trends of which may be partly applicable to Hungary.

Keywords: protection of educational institutions, cyber defense, IT incidents.

Bevezetés

Az oktatási intézmények informatikai védelmével kapcsolatos nemzetközi tudományos szakirodalom és adatforrások köre meglehetősen szűkös. Ulven és Wangen 2021-es szakirodalmi áttekintésében [1] 18 tudományos igényű cikket, és 14 egyéb forrást (fehér könyveket, műszaki jelentéseket, szakdolgozatokat, szakmai weboldalakat) kutatott fel. Rahim és szerzőtársai bibliometriai elemzésükben az elmúlt tíz év online forrásból elérhető szakirodalmát vizsgálták. Ezekben 418 dokumentumot azonosítottak, amelyek többségükben nem tudományos igényű cikkek, hanem konferencia előadások voltak, közülük is csak hat volt publikusan is elérhető. A hivatkozott források közt egyetlen magyar sem volt [2], és utalás sem szerepelt a hazai egyetemekre. Bár a hazai és nemzetközi összehasonlításban számos azonosság jelenik meg, melyet a linzi székhelyű Johannes Kepler Universität-en végzett tanulmányutam is megerősített, a hazai felsőoktatás védelmi kérdéseinek vizsgálatakor számos különbség is feltételezhető. Ezek azonosításához fel kell térképezni a felsőoktatás

CC BY

értékeit, a szférát érő informatikai incidenseket, sebezhető pontjaikat és azokat a tényezőket, amelyek következtében a védelmi megoldások szükségszerűen eltérnek más területekétől. Külföldi gyakorlatban sem találtam példát kifejezetten oktatási intézményekre szabott szabályzásra, de egyes országokban elindultak olyan folyamatok, melyek a felsőoktatási intézményeket is érintik. A felsőoktatási intézmények jogszabályi környezetében várhatóan a NIS2 irányelv hoz változást [3]. 2016-os elődjének célja a kiberbiztonság javításával kapcsolatos jogszabályi környezet javítása volt, melyet az informatikai rendszereket ért incidensek számának akkori jelentős növekedése indokolt. A NIS2 számos új követelményt fogalmaz meg, miközben a korábbiak szigorítását javasolja, és jelentősen bővíti az érintett intézmények körét is. Deklarálja a biztonsági intézkedések jóváhagyási és felügyeleti feladatkörét, az egyes szervezeti egységek vezetőinek informatikai biztonsági képzését és az intézményi vezetők személyes felelősségét is, emellett a szervezet bevételével arányos, nagy összegű bírság kiszabásának lehetőségét írja elő.

Kiberfenyegetettségek a felsőoktatásban

Számos egyetem szenvedett már el különböző típusú informatikai incidenseket. A média kibervédelemmel foglalkozó híreiben szinte alig található oktatási intézmény ellen irányuló támadásról szóló híradás, de az egyetemi informatikai üzemeltetők több ilyenről is beszámolnak. Ezek mennyiségi és súlyossági besorolásához, valamint statisztikai módszerekkel történő elemzésükhöz konkrét adatokra van szükség, ugyanakkor ilyenek alig állnak rendelkezésre. Nemzetközi viszonylatban több, elsősorban amerikai adatforrásokra támaszkodhatunk [4] és az ottani tendenciákból vonhatunk le következtetéseket a várható hazai változásokra is. Az Open Security Foundation szerint az összes biztonsági incidens 35%-a a felsőoktatásban történik, ezt személy szerint túlzónak tartom. Giszczak kutatása szerint [5] 2016 első felében 50%-kal nőtt a felsőoktatási adatokkal kapcsolatos jogsértések száma. Munkájában bemutatja, hogy a reputációs veszteség megjelenik a kutatási támogatások és az adományok megszerzésekor, amelynek mértékét kiszivárgott rekordonként körülbelül 300 dolláros kárként határozza meg.

A Verizon 2022-es „Data Breaches in Education” [6] riportjának az oktatási szférát elemző fejezetének főbb pontjai szerint az USA-ban 1.241 incidens történt, ebből 282-t több forrásból is megerősítettek. Eszerint a rendszerekbe történő belépés, alapvető webes alkalmazások támadása és egyéb hibák a jogsértések 80%-át teszik ki. A betörések 25%-át belső szereplők, 75%-ukat külső támadó kezdeményezi, melyek célja 95%-ban anyagi haszonszerzés, és csak 5%-ban valamilyen kémkedési szándék. Az incidensek 63%-a személyes, 41%-a hitelesítő, 23%-a egyéb, 10%-a pedig belső adatok megszerzésére irányul. A jelentés az összegzésében kiemeli: „Az oktatási szolgáltatások kísértetiesen hasonló tendenciát követnek, mint a többi iparág többsége; drámaian megnövekedett a ransomware-támadások száma, mely a jogsértések több mint 30%-a. Ezen túlmenően ennek az iparágnak meg kell védenie magát az ellopott hitelesítő adatokkal és az adathalász támadásokkal szemben, amelyek potenciálisan felfedhetik az alkalmazottak és diákok személyes adatait”.

A hackmageddon.com[1] havi bontásban közöl statisztikákat a szerkesztő által számos különböző forrásból gyűjtött támadásokról és incidensekről. Ez a forrás sem rendelkezik teljes körű adatbázissal, de a vizsgálatom tárgyaként választott időszakban, 2016 és 2022 között nagyszámú, összesen 12.743 kibervédelmi incidenst dokumentált úgy, hogy

1 Hackmageddon. Lásd: www.hackmageddon.com/2021/01/13/2020-cyber-attacks-statistics/

adataiban kiválaszthatók az oktatási intézményeket érintő incidensek és azok részletei is[2]. Ezek elemzése céljából felvettem a kapcsolatot a site üzemeltetőjével, aki kutatási célú hozzáférést biztosított a nyers adataihoz. Sajnos ez nem tartalmazza az oktatási intézmények típusait, így az ez alapján levont következtetések az oktatási szféra egészére érvényesek.

Trendek meghatározhatósága érdekében elsőként az oktatási intézményeket ért incidensek számát évekre bontva gyűjtöttem ki. A *NemOkt* oszlopban az adott évben ismertté vált, nem oktatási intézményekre irányult adatsértések száma szerepel, melyet az adott év oktatási szférát érintő incidensek száma követ (Okt). A két adat százalékos aránya évről évre mutatja az oktatási intézmények az oktatási szférára irányuló támadások részarányát. Az *Éves részarány* a vizsgált évek összes adatsértésének az adott évre eső arányát írja le, mely az adott évben az adott területre jutó adatsértések számának és az összes támadásnak (11.940, illetve 803) százalékos értékben kifejezett hányadosa.

Év	NemOkt	Okt	%	Éves részarány	
2016	1.082	49	**4,5%**	9,1%	6,1%
2017	901	68	**7,5%**	7,5%	8,5%
2018	619	42	**6,8%**	5,2%	5,2%
2019	1.671	135	**8,1%**	14,0%	16,8%
2020	2.169	183	**8,4%**	18,2%	22,8%
2021	2.374	174	**7,3%**	19,9%	21,7%
2022	3.124	152	**4,9%**	26,2%	18,9%
Összesen	11.940	803	**6,7%**	100%	100%

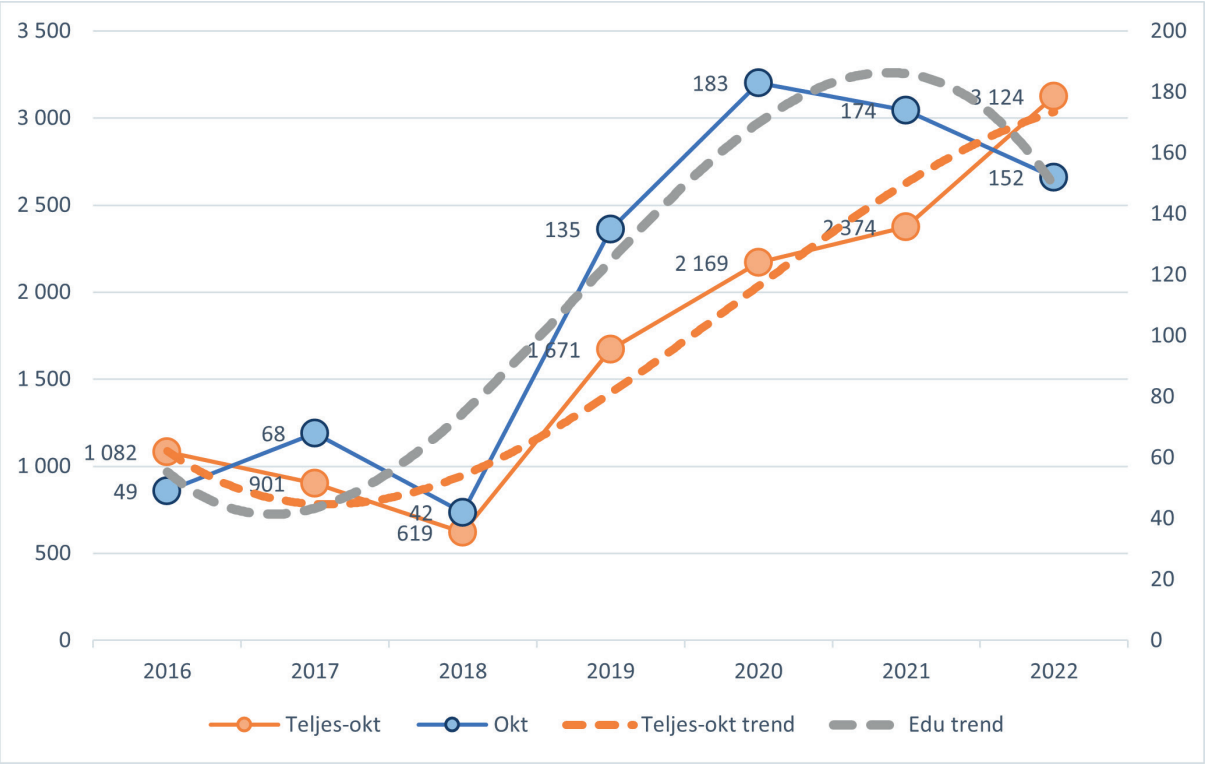
1. táblázat. Az oktatási szektort ért támadások összevetése a támadások teljes számával éves bontásban. Forrás: saját szerkesztés a Hackmageddon adatai alapján.

Az adatok alapján megállapítható, hogy az oktatási intézményeket érő adatsértések aránya a vizsgált időszakban összességében 6,7%, mely az egyes években 4,5-8,4% között változott. Az évről évre emelkedő számú támadások mellett az oktatási szektorra irányulók száma 2021-ben megtorpant, majd csökkenni kezdett. Mindkét adatsort egy diagramban ábrázoltam, melynek pontjait kizárólag a jobb áttekinthetőség érdekében összekötöttem (a két pont közötti értékek alakulásáról az ábra nem ad információt). A diagram az összehasonlíthatóság érdekében az y tengelyen kettős skálázást alkalmaz. Az ábra szaggatott vonalai az adott értéksor harmadfokú regresszióval kifejezett trendjét ábrázolják. Ezek egyértelműen rámutatnak arra, hogy míg a narancsszínnel jelölt, összes támadást leíró trend 2018 óta egyértelmű emelkedő tendenciát követ, az oktatási szektor esetében ez a trend 2021-ben megfordult[3].

2 A forrás harmadik normálformába alakítása az eredeti adatsorok számának növekedését eredményezte. A közölt adat a folyamat végén keletkezett rekordok száma.

3 A harmadfokú regresszió természeténél fogva nem alkalmazható hosszútávú előrejelzésre, a diagramon szereplő trendvonal a szférát ért támadások csökkenő tendenciáját prognosztizálja.

CC BY



1. ábra. A támadások teljes és az oktatási szektorra irányuló adatai és változásainak trendje. Forrás: saját szerkesztés.

Ez a tendencia ellentmond a külföldi szakirodalomban széles körben elemzett, a Covid19 által megkövetelt rapid informatikai változtatások következményeként megvalósított, a távolléti oktatás támogatását szolgáló informatikai fejlesztések biztonságcsökkentő hatásának. Zalat és szerzőtársai tanulmányukban arra a következtetésre jutottak, hogy az online tanulásra való átállás a tanulás támogatásához szükséges informatikai szolgáltatásokban működési zavarokat eredményezett, melyek jellemzően szolgáltatáskiesések vagy szolgáltatás megtagadásos támadások következtében alakultak ki [7]. A 2022-es évben mért csökkenés valószínűsíthető oka pedig az orosz-ukrán háború következményeként a kiberműveletek célpontjainak áthelyezése.

A Hackmageddon adatbázisában a támadások motiváció szerinti besorolását is elvégezték, melyek a Cybercrime (CC), Cyber Espionage (CE), a Cyber Warfare (CW) és a Hacktivism (H) kategóriákba esnek[4]. Ezeket az oktatási intézmények vonatkozásában szintén év szerinti bontásban vizsgáltam annak érdekében, hogy változásuk trendje mellett a támadók motivációinak változásokra is következtetni lehessen. Az adatok elemzése alapján elmondható, hogy az oktatási szférát ért támadásokat a kiberbűnözés, főként az anyagi előnyök megszerzése hajtja. A kiberkémkedésként azonosított esetek kivétel nélkül célzott támadások voltak, melyek többségében a tanulók befolyásolására, vagy kutatóintézeti adatok megszerzésére irányultak. Egy példa erre a 2022.09.01-én rögzített incidens, melynek leírása szerint „Kína feljelenti az Egyesült Államok pekingi nagykövetségét, miután az ország két legjelentősebb kiberhatósága (a kínai Nemzeti Számítógépes Vírus Veszélyhelyzeti Reagáló Központ (CVERC) és a 360 nevű cég) közös jelentésében vádolja a Nemzetbiztonsági Ügynökséget, miszerint érzékeny információkat lopott kínai intézményekből, legfőképp az Északnyugati Műszaki Egyetemről". Az egyetlen Cyber Warfare eset az orosz-ukrán háborúhoz kötődik: a jelentés szövege szerint „a Wordfence kutatói az orosz megszállás

4 Néhány adat besorolása hiányzott, vagy nem volt egyértelmű, ezeket a táblázatban nem tüntettem fel.

kezdete óta hatalmas támadási hullámot regisztráltak ukrán WordPress oldalak ellen, céljuk ezek leállítása és általános morál rombolása". Érdemes megjegyezni, hogy a besorolás nem minden esetben egyértelmű, pl. Vatikán hivatalos honlapjának megtámadását az orosz invázió pápai elítélése után, vagy az NLB hackercsoport által hárommillió orosz iskolás személyes adatainak közzétételét nem a Cyber Warfare-be, hanem a Hacktivizmusba sorolja. Összességében azonban elmondható, hogy ezek szerepe az oktatási szektorban csupán 4%.

	2016	2017	2018	2019	2020	2021	2022	Összesen	%
CC	41	65	40	123	179	173	145	766	95,8%
CE	3	1	1	11	3	1	3	23	2,9%
CW		0	0	0	0	0	1	1	0,1%
H	3	2	0	1	1	0	3	10	1,3%
Összesen	47	68	41	135	183	174	152	800	100,0%
%	5,9%	8,5%	5,1%	16,9%	22,9%	21,8%	19,0%	100,0%	

2. táblázat. Az oktatási szektort ért incidensek motivációinak évek szerinti megoszlása.
Forrás: Hackmageddon adatai alapján saját szerkesztés.

A motivációk elemzését érdemes az oktatási szektoron kívül eső intézményekre is megvizsgálni és azzal összehasonlítani. Bár ott is magas a kiberbűnözés aránya (81,2%) ugyanakkor jelentősen nagyobb számban történnek kiberkémkedés vagy hacktivizmus célú esetek. A kiberhadviselés 4,2%-os értéke pedig arra utal, hogy az ilyen indíttatású támadások ellen ebben a szférában lényegesen hatékonyabb védekezést kell folytatni.

A motivációk ismerete nagyban befolyásolhatja a védekezés módszertanának kidolgozását, a védendő rendszerek azonosítását és a védelmükre szolgáló eszközök kiválasztását is. Ez alapján az oktatási intézményeknek elsősorban azokra a rendszerekre kell koncentrálniuk, melyek a támadóik számára anyagi haszonszerzés lehetőségét kínálják, tehát érzékeny adatok megszerzésére vagy ransomware aktiválásra irányulnak.

Az adatbázis elemzésével az alkalmazott módszerek is azonosíthatók voltak. A támadók által használt eljárásokat 29 támadási technikába sorolják be, viszont ezek több mint felét a vizsgált időszakban csak egyszer alkalmazták.

Az így kapott adatok elemzésével kimutatható, hogy az oktatási intézményekkel szemben leginkább a malware-re alapozott támadási technikákat alkalmazzák, ezek aránya hozzávetőleg 40%. Bár ez a módszer már 2016-ban is megjelent, alkalmazásának növekvő tendenciája valószínűsíti, hogy az hatékony módszert jelent. Ismeretlen marad a támadási technikák közel negyede, és ennek trendje is erősödött az elmúlt években, ráadásul a támadások egyre nagyobb részét ez a típus teszi ki. Az account hijacking során ellopják vagy átirányítják egy személy valamilyen hozzáférését. Annak ellenére, hogy legnagyobb anyagi hasznot a célzott támadások kivitelezésével lehet elérni, azok száma elenyésző, és releváns változás nem is fedezhető fel a vizsgált időszakban. A Covid19 alatt alkalmazott, a távolléti oktatást segítő szoftverek hibáinak kihasználására a támadók az átálláshoz rendelkezésre álló rövid idő okozta zűrzavart igyekeztek kihasználni.

A további technikák aránya az előzetes feltételezéseimet messze alulmúlták. A sérülékenységek általános kihasználását az adatok alig támasztják alá, és kis számban detektáltak a szektorral szemben kezdeményezett túlterheléses támadást, vagy SQL injection-t. A lista utolsó helyén megjelenő jelszófeltörési eljárást pedig csak három esetben regisztrálták.

CC BY

Megjegyzendő, hogy ezek az értékek hirtelen megváltozhatnak, amennyiben a szektorban tömegesen alkalmazott szoftver (esetleg hardver) biztonsága sérül. Magyar viszonylatban ilyen incidens volt az eKréta rendszer elleni támadás, mely során egy megtévesztő levél alkalmazásával, rendszerben jelen levő a többszörös konfigurációs hibák kihasználásával végül magyar tanulók adatai nagy mennyiségben szivárogtak ki. Az eset példa nélküli volt, egy közérdekű adatigénylés tanúsága szerint a Nemzeti Adatvédelmi Hatóság felé 2018. február és 2023. március között jelentett 124 esetből 62 az eKréta rendszer feltörésével volt kapcsolatos, ami az összes jelentett incidens 50%-a.

Technika	2016	2017	2018	2019	2020	2021	2022	Össz.	Arány
Malware	3	18	10	71	101	75	75	353	44,1%
Unknown	20	19	13	18	33	54	53	210	26,3%
Account hijacking	9	24	15	33	22	20	14	137	17,1%
Targeted attack	2	2	2	5	2	0	3	16	2,0%
Zoom bombing	0	0	0	0	9	6	0	15	1,9%
Vulnerability	0	0	1	0	0	12	1	14	1,8%
DDOS	2	1	0	0	7	0	0	10	1,3%
Defacement	2	3	0	1	2	1	1	10	1,3%
SQL Injection	5	0	0	0	1	0	0	6	0,8%
Brute Force	1	0	0	2	0	0	0	3	0,4%

3. táblázat. Az oktatási szektort ért releváns támadási technikák évek szerinti eloszlása.
Forrás: Hackmageddon adatai alapján saját szerkesztés.

Összegzés

A Hackmageddon adatbázisának vizsgálata alapján megállapítható, hogy az elsősorban amerikai, továbbá angol, kanadai, ausztrál, indiai és ír források által szolgáltatott adatok alapján az oktatási intézmények fenyegetettsége 7% körüli mértékre tehető, mely kismértékű ingadozás mellett 2016 óta jelentős mértékben nem változott. A támadók előszeretettel alkalmaznak malware-ekre alapozott támadási módszereket, de lehetőség szerint igyekeznek megszerezni és felhasználni a felhasználók különböző hozzáféréseit. A 2022-ben folyó háború ellenére ezeknek az intézményeknek a kiberhadviselésben nem látszik szerepük. A támadók tevékenysége elsősorban a kibertérre vagy ott elkövetett bűncselekményekre alapozott, így feltehetően az anyagi haszon megszerzésére irányul. Annak ellenére, hogy az elemzésekben bemutatott tendenciák nemzetközi adatokon alapulnak, azok érvényesek lehetnek a hazai intézményekre is, így a bemutatott elemzések és következtetések segíthetik az informatikai rendszerek védelmi pontjainak meghatározását.

Irodalom

[1] G. Wangen és J. B. Ulven: „A Systematic Review of Cybersecurity Risks in Higher Education”, *Future Internet*, 1. kötet 13, 1-40 o., 2021.
[2] N. Rahima, Z. Othmanb és F. Z. Hamidc: „Cyber Security and the Higher Education Literature: A Bibliometric Analysis”, *International Journal of Innovation, Creativity and Change*, 12. kötet 1. szám 2020. 12.
[3] *Az Európai Parlament és a Tanács (EU) 2022/2555 Irányelve*, 2022.
[4] F. Inc.: „Why Cyber Attackers Are Targeting Higher Education, and What Universities Can Do about It. White paper.”, Fireeye Inc., 2015.

[5] J. J. Giszczak és D. A. Paluzzi: „Ass or Fail? Data Privacy and Cybersecurity Risks in Higher Education”, McDonald Hopkins, 2016.
[6] Verizon: „Educational Services,” 2022. [Online]. Elérhető: https://www.verizon.com/business/resources/reports/dbir/2022/data-breaches-in-education/. [Hozzáférés dátuma: 2022.04.03.].
[7] M. Z. Zalat, S. M. Hamed, A. B. Bolbol: „The experiences, challenges, and acceptance of e-learning as a tool for teaching during the COVID-19 pandemic among university medical staff”, *PLoS One*, 16. kötet 1. szám, 1-12. o.
[8] *1163/2020. (IV. 21.) Korm. határozat Magyarország Nemzeti Biztonsági Stratégiájáról*, 2020.
[9] *1139/2013. (III. 21.) Korm. határozat Magyarország Nemzeti Kiberbiztonsági Stratégiájáról*, 2013.
[10] *Cyber security and defence European Parliament resolution of 22 November 2012 on Cyber Security and Defence (2012/2096(INI))*, 2012.
[11] *Stratégiai Koncepció az Észak-atlanti Szerzıdés Szervezete tagállamainak védelméért és biztonságáért*.
[12] NATO: *Defending the networks - The NATO Policy on Cyber Defence*, 2011.
[13] D. Appelman: „California Requires Disclosure of Database Security Breaches”, Usenix, 2004.
[14] „Australian Government Department of Home Affairs,” 2020. 11. [Online]. Elérhető: https://www.homeaffairs.gov.au/reports-and-pubs/files/exposure-draft-bill/exposure-draft-security-legislation-amendment-critical-infrastructure-bill-2020-explanatory-document.pdf. [Hozzáférés dátuma: 2022.01.11.].
[15] *2011. évi CCIV. törvény a nemzeti felsőoktatásról*, 2011.
[16] *2012. évi C. törvény a Büntető Törvénykönyvről*, 2012.
[17] „Az Európai Parlament és a Tanács (EU) 2016/679 Rendelete,” 2016.04.27. [Online]. Elérhető: https://eur-lex.europa.eu/legal-content/HU/TXT/?uri=uriserv:OJ.L_.2016.119.01.0001.01.HUN&toc=OJ:L:2016:119:FULL119%3AFULL#d1e1459-1-1. [Hozzáférés dátuma: 2022.01.10.].
[18] National Institute of Standards and Technology, „Framework for Improving Critical Infrastructure Cybersecurity,” 2018.04.16. [Online]. Elérhető: https://nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP.04162018.pdf. [Hozzáférés dátuma: 2019.15.23.].
[19] P. J. Ballard: „Measuring Performance Excellence: Key Performance Indicators for Institutions Accepted into the Academic Quality Improvement Program (AQIP),” 2013.
[20] E. K. Kwaa-Aidoo és M. Agbeko: „An Analysis of Information System Security of a Ghanaian University”, *International Journal of Information Security Science*, 7. kötet, 1. szám, 90-99. o., 2017.
[21] *Rendszeres szociális ösztöndíjakkal kapcsolatos adatkezelés a Budapesti Műszaki és Gazdaságtudományi Egyetemen.*, NAIH/2020/54.
[22] *Állásfoglalás a koronavírus elleni védettség tényének felsőoktatási intézmény általi megismerhetőségéről, nyilvántarthatóságáról kollégiumi elhelyezés és egyetemi rendezvények kapcsán.*, NAIH-6298-2/2021.
[23] I. G. Butnaru, V. Nita, A. Anichiti és G. Brînză: „The Effectiveness of Online Education during Covid 19 Pandemic—A Comparative Analysis between the Perceptions of Academic Students and High School Students from Romania”, *Sustainability*, 13. kötet, 9. szám 1-20. o., 2021.
[24] L. W. Loo: „Student Hacking into University’s Learning Management System to Save His Grades: A Cautionary Tale,” Singapore Management University, Singapore, 2016.
[25] „Unit-Department for ICT and Joint Services in Higher Education and Research,” Direktoratet for IKT og fellestjenester i høyere tdanning og forskning, Norway, 2019.

CC BY

[26] G. Vámosi: „Ezerhétszáz hallgató adatait vesztette el a veszprémi egyetem," 2008.12.10. [Online]. Elérhető: https://www.origo.hu/techbazis/20081210-1717-hallgato-adatait-vesztette-el-a-veszpremi-egyetem.html. [Hozzáférés dátuma: 2022.01.10.].

[27] „Zsarolóvírus-támadás érte a Pázmányt, leállt a Neptun," HVG, 2020.04.24. [Online]. Elérhető: https://hvg.hu/tudomany/20200424_pazmany_peter_katolikus_egyetem_zsarolovirus_neptun_tanulmanyi_rendszer_szakdolgozat_leadasi_hatarido. [Hozzáférés dátuma: 2022.10.01.].

[28] Nemzeti Adatvédelmi és Információszabadság Hatóság, „Közérdekű adatigénylés", 2018. 12.08. [Online]. Elérhető: https://kimittud.hu/request/12018/response/17739/attach/3/NAIH%202019%20741.pdf. [Hozzáférés dátuma: 2022.10.12.].

 DOI: 10.31915/NWS.2023.15

A digitális gyűjtésrekonstrukció lehetőségei: az *Ethiofolk projekt*

The Possibilities of Digital Reconstruction of Fieldwork: the Ethiofolk Project

Bolya Mátyás
Liszt Ferenc Zeneművészeti Egyetem (Budapest)
Népzene Tanszék Népzenei Kutatócsoport
bolya.matyas@zeneakademia.hu
ORCID: 0000-0002-6145-663X

Mátyás Bolya
Liszt Ferenc Academy of Music, Budapest
Folk Music Research Group
associate professor, Head of Folk Music Department

Abstract

In June, 1965, two young researchers arrived in Addis Ababa at the invitation of Emperor Haile Selassie. The purpose of György Martin (folk dance researcher) and Bálint Sárosi's (folk music researcher) journey was to examine and explore traditional Ethiopian folklore. They were members of the Folk Music Research Group of the Hungarian Academy of Sciences, whose head was Kodály at that time. From their home institution they had received internationally renowned knowledge and expertise in folk music research, thus they wished to be among the first to explore Ethiopian folklore. Thus, one of the most exciting and productive expeditions of Hungarian folklore research to Africa began.

As virtually nothing was available about Ethiopian folklore in Hungary at that time, their journey amounted to an academic leap of faith. At the beginning they had no idea of the richness of the archaic dance and music culture that they would encounter. Without any knowledge of the place and the material that awaited them in Ethiopia, their only support were the 70 years of experience crystalized in the methodology of Hungarian folk music research and the tools of contemporary documentation. While, some cultural exchange between the two countries followed their journey to Ethiopia for a few years, the collection's material slowly became forgotten.

During their journey they kept detailed notes and records, but also made audio and video recordings, photographs, and bought instruments. They returned home all together with approximately 3200 meters of silent video recordings, 30 strips of audio tape and 1000 photographs.

Processing the Ethiopian collection meant a new challenge for the team, since the collection itself took place more than five decades ago. We had to learn and understand a methodology that relied on the technology of the time and transfer it to a modern software environment. After digitalization we created a data structure and based on the available records and notes we made a full-scale collection reconstruction, fine-tuning the data and creating cross-references. Thus, we got a meta-data structure that could be placed to the software environment, developed by the Polyphony Project, which is capable of fulfilling online publication purposes as well as assisting research. Behind the scenes of a website that is accessible to everybody, there is a diverse database system that complies with the most rigorous of scientific standards and handles significantly more considerations that what is visible from the displayed elements.

CC BY

How much more is a digital reconstruction of fieldwork than the digitization of analog material? How can the information that can be extracted be maximized five decades later? How can all this be linked to a digital archive concept? The article will seek answers to these questions.

Keywords: Etiópia, népzene, digitális archívum, digitális gyűjtésrekonstrukció, terepmunka, Ethiopia, folk music, digital archive, digital reconstruction of fieldwork, fieldwork

1. Bevezetés[1]

1965 júniusában, Hailé Szelasszié császár személyes meghívására két fiatalember érkezett Etiópia fővárosába, Addis Ababa-ba. Martin György néptánckutató és Sárosi Bálint népzenekutató utazásának célja a hagyományos etióp folklór vizsgálata volt, amely munkát a Kodály vezette MTA Népzenekutató Csoport tagjaként, a világszerte elismert magyar néptánc- és népzenekutatás eszköztárával felvértezve, elsőként kívántak elvégezni. Így kezdődött a magyar folklórkutatás egyik legizgalmasabb és legtermékenyebb afrikai kutatóútja. Mivel a korabeli Magyarországon szinte semmilyen háttéranyag nem volt elérhető az etióp folklórról, útjuk valódi tudományos vakrepülésnek számított. Akkor még nem sejtették, hogy milyen gazdag és archaikus tánc- és zenekultúrával fognak találkozni. Az út során a jegyzetek mellett hang- és filmfelvételeket, fényképeket készítettek, valamint hangszereket is vásároltak. Összesen körülbelül 3200 méter (6 óra) némafilmfelvétellel, 30 tekercs (40 óra) magnószalaggal és 1100 fényképpel tértek haza.[2]



1. kép. Sárosi Bálint gyűjtés közben 1965. június 23-án Dessziében. Fotó: Martin György. ZTI_NZ_24311

1 A cikk a Veszprémben megrendezett NETWORKSHOP 2023 országos informatikai konferencián 2023. április 13-án elhangzott előadás leírt és szerkesztett változata. Az előadáshoz tartozó prezentáció megtekinthető itt: https://prezi.com/view/T3a97Po8UzeEcR4gr07z/

2 Nem volt technikai lehetőség a kamera és a magnetofon összekötésére, szinkronizált, közösen vezérelt működtetésére, így csak párhuzamos felvételek készültek. Az UNESCO segítségével 1965-ben beszerzett modernebb kamera képes lett volna erre, végül azonban ezt a drága eszközt – vélhetően a sok bizonytalan körülmény és a kezelési tapasztalatlanság miatt – a kutatók nem vitték magukkal Etiópiába.

Az utazást követően néhány évig még voltak kulturális kapcsolatok a két ország között, a gyűjtés anyaga azonban szép lassan feledésbe merült. A magyarországi rendszerváltozás több külképviselet megszüntetését hozta magával, erre a sorsra jutott az etióp intézmény is. A nagykövetség újranyitása csak a 2011-ben meghirdetett külpolitikai nyitás után 2016 tavaszán – két és fél évtizednyi szünet után – valósulhatott meg.[3] Végül a Magyarországon megőrzött archív anyag digitalizált változata 2019-ben a *Kontinenseken átívelő hidak az emberiség szellemi kulturális örökségének megőrzéséért* program keretein belül kerülhetett vissza Etiópiába.[4] A program szakmai vezetői a hosszabb távú megőrzés és az *open access* elvek alapján a digitális publikálás mellett döntöttek.

Vajon mennyivel több egy digitális gyűjtésrekonstrukció az analóg alapanyag digitalizálásánál? Hogyan maximalizálható a kinyerhető információ öt évtizeddel később? Mindez hogyan kapcsolódhat egy digitális archívum koncepcióhoz? Írásomban ezekre a kérdésekre keresem a választ. A cikk emellett tisztelgés a 2022-ben elhunyt Sárosi Bálint munkássága előtt.[5] Fontos megemlíteni azt is, hogy 2022-ben kutatócsoport alakult a Zeneakadémia Népzene Tanszékén, amelynek célkitűzése Sárosi ott őrzött hagyatékának feldolgozása és közzététele.

2. Kulturális kontextus

Az adatbázis-építés technikai és elvi kérdései előtt fontos röviden összefoglalni azt a kulturális hátteret – beleértve a tudományos és diplomáciai eseményeket is – amelyben értelmezhetővé válnak az etióp terepmunka mozgatórugói, jelentősége, nehézségei. Meggyőződésem, hogy ezek ismerete elengedhetetlen a hiteles és hatékony digitális modellezéshez.

A második világháború után új folyamatok indultak meg az éledező magyar népzenekutatásban. A bartóki útmutatás nyomán a kutatás egyre tágabb horizontot jelölt ki magának: a magyar nyelvterületek után a szomszéd népek,[6] majd az európai népek zenéjének vizsgálata következett.[7] A rokon népek zenefolklórjának megismerése már jelezte a nemzetközi nyitást.[8] 1964-ben nagyszabású népzenei konferenciát rendezett Budapesten a Nemzetközi Népzenei Tanács (IFMC), amelynek elnöke ebben az időben Kodály Zoltán volt.[9] Egy hónappal a konferencia után érkezett Budapestre I. Hailé Szelasszié, Etiópia császára. A látogatás politikai súlyát jelzi, hogy a császárt Dobi István,

---

3 Marsai Viktor. „A magyar-etióp diplomáciai kapcsolatok felvétele." *Külügyi Szemle*. 18/4 (2019). 48–66. 48. Hozzáférés: 2023.06.20. https://kki.hu/wp-content/uploads/2020/04/03-Marsai.pdf

4 A hivatalos honlap megnyitása mellett három vezető kulturális intézmény a digitális alapanyagot is átvehette az eseményen. „Afrika Magyarországon, Magyarország Afrikában." *Szellemi Kulturális Örökség Igazgatóság honlapja*. 2019. június 14. Hozzáférés: 2023.06.20. http://szellemikulturalisorokseg.hu/index0.php?name=hir_190614_szko_atado_addis_ababa

5 Bolya Mátyás. „Búcsú Sárosi Bálittól." *folkMAGazin*. 2022/4. 6–7. Hozzáférés: 2023.06.20. http://lapozo.folkmagazin.hu/mag22_4/?page=6

6 Bartók Béla. *Népzenénk és a szomszéd népek népzenéje*. Budapest: Somló Béla könyvkiadó, 1934. Hozzáférés: 2023.06.20. http://real-eod.mtak.hu/2635/1/14991.pdf

7 Az *Európai Dallamtár* az összehasonlító kutatások segédeszközeként jött létre a Zenetudományi Intézet Régi Zenetörténet Osztály gondozásában, Rajeczky Benjamin kezdeményezésére, Vargyas Lajos vezetésével az 1960-as évek elején.

8 Pálóczy Krisztina. *Egzotikus hangszerek és zene Magyarországon. Magyar kutatók a Kárpát-medencén túl.* Doktori disszertáció, Jyväskylä University Digital Repository, 2012. Hozzáférés: 2023.06.20. https://jyx.jyu.fi/bitstream/handle/123456789/37543/978-951-39-4670-8.pdf?sequence=1

9 The Present Volume Contains the Papers Read at the International Folk Music Council (IFMC) Conference Held in Budapest in August 1964. *Studia Musicologica*. 7/1–4 (1965). Hozzáférés: 2023.06.20. https://www.jstor.org/stable/i237242

CC BY

az Elnöki Tanács elnöke mellett Kádár János, a kormány elnöke is személyesen fogadta. A bevezetőben említett meghívás is ekkor történt.[10]

Érdekesség, hogy még 2019-ben is lehetett az eredeti anyagot kiegészítő adatokat gyűjteni a terepen. Az archívum etiópiai átadása után Both Miklóssal,[11] aki nagy szerepet vállalt a digitális felület létrehozásában is, bejártuk Sárosi és Martin útjának egy részét.[12] Az út során értettük meg az egykori gyűjtési pontok kijelölésének logikáját, a felvételek helyének kiválasztását, valamint pontosíthattunk számos leíró adatot, bemutatva a helyieknek a korabeli film- és hangfelvételeket. Világossá vált az is, hogy milyen fontos, ha egy kutatói életművet gazdag tereptapasztalat hitelesít.



2. kép. Fiatal helybeli azonosítja az 1965-ös felvétel szereplőit. Akszum, 2019. június 19. Fotó: Bolya Mátyás

10 Bolya Mátyás. *A kinyitott időkapszula: Etióp folklór 1965-ből. Kutatástörténet és digitális gyűjtésrekonstrukció magyar kutatók nyomán*. Polyphony, 2019. Hozzáférés: 2023.06.20. https://www.ethiofolk.com/hu/publications

11 Both Miklós jelentős erőfeszítéseket tett a korszerű, nemzetközileg is elismert kulturális adatbázisok építése terén. Legfontosabb munkái közé tartozik a *Folk_ME. Folk Music Education for Future Generations.* Creative folk music educational toolkit (Polyphony, 2021) www.folk-me.com és a *Polyphony Project: Internetes népzenei archívum.* (Polyphony, 2018) www.polyphonyproject.com Online archiválási rendszer ukrán népzenei gyűjtések feldolgozására és publikálására.

12 540 kilométert tettünk meg alkalmi járművekkel Etiópia északi részén a Makale, Akszum, Siré, Gondar útvonalon.



3. kép. Both Miklós és Bolya Mátyás gyűjtés közben. Gondar, 2019. június 23. Fotó: Kukár Manó

3. Virtuális gyűjtésrekonstrukció[13]

3.1 Digitális archívum koncepció

A digitális archívum koncepció elemeiről már volt szó egy korábbi publikációmban.[14] Az *Ethiofolk projekt* – hasznosítva ezeket az eredményeket – egy jól körülhatárolható gyűjtési egységen keresztül mutatja be a virtuális gyűjtésrekonstukcióban rejlő lehetőségeket. A legfontosabb elemek a következők: az internet elterjedésével az archívumok új funkciója – az archiválás mellett – a szolgáltatás lett. Ezzel új célcsoportokat lehet megszólítani a különböző szakterületeken.[15] Az adatfeldolgozás összetett folyamat, amelynek csak kis része kutatói feladat. A cél olyan szuperadatbázisok létrehozása, amelyek képesek kiváltani az elszigetelt, egymással nem kommunikáló, gyakran korszerűtlen szoftverkörnyezetben működő adatbázisokat. Bizonyos, tudományos szempontból arra érdemes gyűjteményi egységeket pedig ki lehet emelni, részletesebben feldolgozni és publikálni úgy, hogy az alapadatok a mindenkori szuperadatbázisból származnak. Erre példa az Ethiofolk projekt is. Egy ilyen platform a tudomány területén sokfunkciós eszközzé válik:

---

13 A részletes technikai leírások itt olvashatók: Bolya Mátyás: *A kinyitott időkapszula: Etióp folklór 1965-ből. Kutatástörténet és digitális gyűjtésrekonstrukció magyar kutatók nyomán*. Polyphony, 2019. 33–38. Hozzáférés: 2023.06.20. https://www.ethiofolk.com/hu/publications

14 Bolya Mátyás. „A BTK Zenetudományi Intézet digitális archívum koncepciója az oktatás és a tudomány szolgálatában." In: Tick József, Kokas Károly, Holl András (szerk.): *Online térben az online térért. Networkshop 30. országos online konferencia*. 2021. április 6–9. Hungarnet, 2021. 133–142. Hozzáférés: 2023.06.20. https://doi.org/10.31915/NWS.2021.13

15 Tudomány, oktatás, előadóművészet, civil szféra.

CC BY

Archiválisszempontból a leltárt, a feldolgozást, az adminisztrációt és a kutatószolgálatot segítő felület, kutatói szempontból pedig publikációs lehetőség és egyúttal a publikálást megelőző kutatást támogató digitális környezet. Nagy gyűjtemények belső összefüggéseinek feltárása digitális gyűjteménykezelés nélkül lehetetlen. Az adatok áttekintése után következik a klasszikus kutatói munka, vagyis a kulturális kontextus dekódolása, mintázatok keresése, rendszeralkotás. Ebben a környezetben olyan belső összefüggések tárulnak fel, amelyek a kétdimenziós táblázatokban láthatatlanok maradnak és rendkívül inspirálóak a kutatók számára.[16]

3.2 Adatbázis-építés

A több mint öt évtizeddel ezelőtt történt gyűjtés feldolgozása új feladat elé állította a projekt szakembereit. Meg kellett értenünk a korabeli technikákat alkalmazó gyűjtési módszertant, és átemelni modern szoftverkörnyezetbe. A digitalizálást követően kialakítottuk a leíró adatok struktúráját, majd a rendelkezésre álló gyűjtési jegyzőkönyvek alapján teljes körű gyűjtésrekonstrukciót végeztünk, pontosítva az adatokat, illetve kereszthivatkozásokat létrehozva.

3.3 A hagyományos gyűjteménykezelés dilemmái

Egy gyűjtés során sokféle dokumentumtípus keletkezhet. Ezeket a dokumentumokat különböző koncepció szerint tárolják az archívumokban, a legjellemzőbb – gyakorlati okok miatt – az analóg hordozók típusa szerinti csoportosítás. Ebből következik, hogy az egy gyűjtéshez tartozó egységek szétszórva találhatók a raktárakban, így a belső összefüggések csak aránytalanul nagy munka árán tárhatók fel, és sok esetben rejtve maradnak. A digitális feldolgozás éppen ezért kiemelt jelentőségű, hiszen lehetőséget ad a gyűjteményi egységek virtuális egyesítésére, rekonstrukciójára; akár intézményeken átívelő összefogás keretében. Azonban az analóg és a digitális világ átmeneti területe nem csupán alkotótér, hanem egyúttal ütközőzóna is. A 20. században kiforrott gyűjteménykezelési technikák, kutatói szokások és publikációs hagyományok csapnak össze a digitális technika által támogatott *open access* törekvésekkel. A továbbiakban részletezek néhány ilyen pontot:

- A digitális feldolgozás egy igen alapos fizikai revíziót is jelent. Az évtizedek óta nem bolygatott egységek esetén rendszeresen derülnek ki nyilvántartási hiányok és következetlenségek, esetleg fizikai hiányok is. A gyűjteményért felelős szakemberek ebben szakmai kompetenciájuk megkérdőjelezését látják, amely munkahelyi feszültségekhez vezethet.

- Egy kulturális adatbázis építése jellemzően csapatmunka. Az ilyen feladathoz szükséges munkacsoportok összetétele azonban nem mindig egyeztethető össze egy klasszikus kutatócsoport szervezési gyakorlatával. Gyakran sérti a hagyományosan értelmezett kompetenciahatárokat annak felmérése, hogy mely munkafázisokhoz kell valóban kutatói erőforrás. A tapasztalatok szerint ez az erőforrás a legértékesebb – és egyúttal a legszűkösebb is –, tehát a határidők miatt felhasználását jól meg kell tervezni. Jelen projektünkben a népzene- és néptánckutatók mellett archívumi-, adatbázis- és IT-szakemberek is dolgoztak. A megfelelő munkaszervezésnek köszönhetően néhány

16 Bolya Mátyás. „A BTK Zenetudományi Intézet digitális archívum koncepciója az oktatás és a tudomány szolgálatában.” In: Tick József, Kokas Károly, Holl András (szerk.): *Online térben az online térért. Networkshop 30. országos online konferencia.* 2021. április 6–9. Hungarnet, 2021. 133–142. 137. Hozzáférés: 2023.06.20. https://doi.org/10.31915/NWS.2021.13

hónap alatt sikerült elkészíteni az adatbázist. Kimondható, hogy csak kutatókból álló munkacsoport nem lenne képes ilyen összetett feladat végrehajtására.

- A különböző szakterületek találkozása miatt a kutatás funkciója és tárgya is eltérhet a korábbi értelmezésektől:

  A gyűjtemények mellett kialakuló adatbázisok már nem csupán arra használhatók, hogy néhány adatukat kiragadva cikkek szülessenek belőlük, hanem [...] a jövőben maga az elkészített adatbázis válhat publikáció tárgyává, szélesre tárva a kapukat a benne feltárt adattömegre kíváncsiak előtt. [...] A korunk kínálta technikai lehetőségek – ha okosan élünk velük – segítenek megvalósítani az eredeti szándékot: nagy mennyiségű, hiteles adat tudományos rendben való feltárását oly módon, hogy az mindenfajta kutatás számára hozzáférhető és átlátható legyen.[17]

- A virtuális gyűjtésrekonstrukció során fontos cél a dokumentumokból kinyerhető információ maximalizálása; ez jóval túlmutat a digitális archiváláson. Az analóg eredeti nagyfelbontású digitális másolatán gyakran olyan utómunkát kellett elvégezni, amely összeegyeztethetetlen a hagyományos archívumi protokollal. Esetünkben ez a hangfelvételek zajszűrését és sebességkorrekcióját, a filmek képstabilizálását, kontrasztok helyreállítását, színrestaurálását és sebességkorrekcióját, valamint az alulexponált fényképek korrekcióját jelentette.

3.4 Az Ethiofolk honlap[18]

A projekt végeredménye egy honlap, amelyen regisztráció nélkül érhető el minden közzétett adat, vagyis az 1965-ös gyűjtés teljes dokumentációja.[19] A felületet – ahogy az impresszumban is látható – az *open access* elveit követő önálló publikációnak tekintjük. Azonban jóval több annál. A korábban említett metaadat-struktúrát a javított hang- és vizuális fájlokkal együtt feltöltöttük a *Polyphony Project* által fejlesztett szoftverkörnyezetbe, amely az online publikációs funkció mellett összehangolt kutatói feladatok kiszolgálására is alkalmas. A mindenki által látogatható honlap mögött egy szerteágazó, a megjelenített elemeknél jóval szélesebb szempontrendszert kezelő, a legszigorúbb tudományos szabványokkal is kompatibilis adatbázismotor dolgozik.

A honlap központi eleme a *Gyűjtési egységek* menüpont. A gyűjtési egység kifejezés a folklórgyűjtők terminológiájában használatos, az egy helyen, egy időben, egy kutató vagy kutatócsoport által készített helyszíni dokumentáció összességét jelenti. A teljes dokumentációból itt az egy-egy faluban készített hang- és filmfelvételeket, valamint fényképeket mutatjuk be. Azonnal láthatók a leglényegesebb adatok: dátum, hely, nemzetiség, illetve a különböző dokumentumtípusok mennyisége. Fontos filológiai háttérmunka volt kialakítani a helynevek és a nemzetiségek neveinek következetes írásmódját; mindezt három

17 Sebő Ferenc. „Népzene és számítógép. Egy új írásbeliség filológiai problémái." In: Papp Márta (szerk.): *Zenetudományi tanulmányok. Kroó György tiszteletére.* (Budapest: Magyar Zenetudományi és Zenekritikai Társaság, 1996). 254–274. 264.

18 Bolya Mátyás, Both Miklós et al. *Ethiofolk: Online etióp népzene- és néptáncarchívum*. MTA BTK Zenetudományi Intézet / Polyphony Project, 2019. Hozzáférés: 2023.06.20. www.ethiofolk.com

19 A teljesség itt népzene- és néptánckutatói szempontból értendő. Nem kerültek fel az oldalra a tisztázatlan helyszíni jegyzőkönyvek, a levelezések, illetve az Országos Levéltárban őrzött kormányzati dokumentumok.

CC BY

nyelven. Az aloldalhoz egy térkép is tartozik, amely a legkorszerűbb formában mutatja be a gyűjtési helyeket az adatbázisban tárolt geokódok segítségével.

A filmek, a fotók és a hangfelvételek önálló menüpontból is elérhetők, itt kártyák jelenítik meg az egységeket jól áttekinthető rendben. A filmeket és a hangfelvételeket olyan platformokra töltöttük fel, amelyek működése és fejlesztése – a sok milliós felhasználói körnek köszönhetően – biztosított, illetve lehetőséget ad tetszőleges időponthoz egyedi hivatkozást rendelni.[20] Ennek segítségével az eredeti felvételeket csak virtuálisan szegmentáltuk, megőrizve az eredeti sorrendeket és modellezve a fizikai hordozókat.



4. kép. Az Ethiofolk honlap

A projekt egyik legeredetibb fejlesztése a némafilmek és a hangfelvételek virtuális szinkronizálása volt. Ez nem valódi, hanem látszólagos szinkronizálás. A feldolgozás során szembesültünk azzal, hogy a hangfelvételek és a filmek valós szinkronizálása nem megoldható. A gyűjtés alatt párhuzamos rögzítés és nem szinkronizált rögzítés történt, ezért csak a jegyzőkönyvek jegyzeteire és saját elemző megfigyeléseinkre támaszkodhattunk, miután az Etiópiába szállított kamera és a magnetofon nem volt alkalmas a szinkronizált, közösen vezérelt működtetésre. Mindez annyit jelent, hogy egy képzeletbeli referencia időegyenesen – információhiány miatt – nem lehetett egyértelműen elhelyezni a felvett anyag szegmenseit. Az igen bonyolult kereszthivatkozási rendszert kézi erővel – konkrét időkódok helyett – értelmezési tartományok meghatározásával sikerült felépíteni.

A rendszer alapját a hangfelvételek jelentik. Igyekeztünk a hangfelvételek szegmensei által kimetszett negyven órányi múltbeli időszelet koordinátarendszerébe a lehető legpontosabban elhelyezni a körülbelül hat órányi táncfilmet. A gyakorlatban a hangfelvétel lejátszása közben tetszőleges időpontban indított filmeket végtelenítve láthatjuk. Az eredmény magáért beszél: a megfelelő tempójú film- és hangrészletek együttesét az emberi agy a pontos szinkron hiánya ellenére is jól értelmezhető audiovizuális élménnyé egészíti ki.

20 SoundCloud, Youtube.

4. Összefoglalás

Utólag értékelve Martin György és Sárosi Bálint útját, elmondhatjuk, hogy a magyar népzenekutatás nemzetközi viszonylatban is egyedülálló szellemi és tárgyi eszköztárával rendkívül értékes áttekintést tudtak nyújtani Etiópia zenei örökségéről, néhány hét alatt olyan hatalmas anyagot összegyűjtve, hogy az a mai kutatásnak is bőven ad témát. Megtisztelő lehetőség ezt a munkát folytatni, együttműködésben etióp kollégákkal, kölcsönösen gazdagítva mindkét ország népzene- és néptánckutatási tapasztalatait. Az elkészült adatbázis pedig mintaként szolgálhat más kiemelt gyűjteményi egységek hasonló szemléletű feldolgozására is.

Forrásjegyzék

„Afrika Magyarországon, Magyarország Afrikában." *Szellemi Kulturális Örökség Igazgatóság honlapja*. 2019. június 14. Hozzáférés: 2023.06.20. http://szellemikulturalisorokseg.hu/index0.php?name=hir_190614_szko_atado_addis_ababa

Bartók Béla. *Népzenénk és a szomszéd népek népzenéje*. Budapest: Somló Béla könyvkiadó, 1934. Hozzáférés: 2023.06.20. http://real-eod.mtak.hu/2635/1/14991.pdf

Bolya Mátyás, Both Miklós et al. *Ethiofolk: Online etióp népzene- és néptáncarchívum*. MTA BTK Zenetudományi Intézet / Polyphony Project, 2019. Hozzáférés: 2023.06.20. www.ethiofolk.com

Bolya Mátyás. „A BTK Zenetudományi Intézet digitális archívum koncepciója az oktatás és a tudomány szolgálatában." In: Tick József, Kokas Károly, Holl András (szerk.): *Online térben az online térért. Networkshop 30. országos online konferencia*. 2021. április 6–9. Hungarnet, 2021. 133–142. Hozzáférés: 2023.06.20. https://doi.org/10.31915/NWS.2021.13

Bolya Mátyás. „Búcsú Sárosi Bálinttól." *folkMAGazin*. 2022/4. 6–7. Hozzáférés: 2023.06.20. http://lapozo.folkmagazin.hu/mag22_4/?page=6

Bolya Mátyás. *A kinyitott időkapszula: Etióp folklór 1965-ből. Kutatástörténet és digitális gyűjtésrekonstrukció magyar kutatók nyomán*. Polyphony, 2019. Hozzáférés: 2023.06.20. https://www.ethiofolk.com/hu/publications

Marsai Viktor. „A magyar-etióp diplomáciai kapcsolatok felvétele." *Külügyi Szemle*. 18/4 (2019). 48–66. 48. Hozzáférés: 2023.06.20. https://kki.hu/wp-content/uploads/2020/04/03-Marsai.pdf

Pálóczy Krisztina. *Egzotikus hangszerek és zene Magyarországon. Magyar kutatók a Kárpát-medencén túl*. Doktori disszertáció, Jyväskylä University Digital Repository, 2012. Hozzáférés: 2023.06.20. https://jyx.jyu.fi/bitstream/handle/123456789/37543/978-951-39-4670-8.pdf?sequence=1

Sebő Ferenc. „Népzene és számítógép. Egy új írásbeliség filológiai problémái." In: Papp Márta (szerk.): *Zenetudományi tanulmányok. Kroó György tiszteletére*. (Budapest: Magyar Zenetudományi és Zenekritikai Társaság, 1996). 254–274. 264.

The Present Volume Contains the Papers Read at the International Folk Music Council (IFMC) Conference Held in Budapest in August 1964. *Studia Musicologica*. 7/1–4 (1965). Hozzáférés: 2023.06.20. https://www.jstor.org/stable/i237242

CC BY

DOI: 10.31915/NWS.2023.16

A Kolozsvári Állami Magyar Színház jelmezterveinek digitalizációja és felvitele az ITIdata adatbázisba

Digitisation of the costume designs of the Hungarian State Theatre of Cluj-Napoca and their introduction into the ITIdata database

Dobás Kata
Bölcsészettudományi Kutatóközpont, Irodalomtudományi Intézet
dobas.kata@abtk.hu
ORCID: 0009-0009-7632-8276

Sidó Zsuzsa
ELTE Digitális Örökség Nemzeti Laboratórium
sido.zsuzsa@btk.elte.hu
ORCID: 0000-0002-3916-1675

Szabó-Reznek Eszter[1*]
Bölcsészettudományi Kutatóközpont, Irodalomtudományi Intézet
szabo.eszter@abtk.hu
ORCID: 0000-0002-9030-3130

Absztrakt

A Digitális Örökség Nemzeti Laboratórium Danube-AI alprojektjének részeként 2022-ben került sor a Kolozsvári Állami Magyar Színház Dokumentációs Tárában őrzött jelmeztervek nyilvántartásba vételére, digitalizálására, adatbázisba rendezésére és közzétételére. A tanulmány erről a több partner együttműködésében megvalósult részprojektről és tanulságairól számol be. Az adatbázisba a Kolozsvári Állami Magyar Színház 1959–1980 közötti 94 bemutatójához kapcsolódó jelmezterveinek képeit, metaadatait, illetve az egyes előadásokhoz tartozó alapinformációkat (például a szereposztást) vittük fel. Az adatok nemcsak a színpadi vizualitás kulisszái mögé engednek betekintést, hanem jelentős jelmeztervezői életművekbe is. Az ITIdata elsősorban irodalomtudományos kutatásoknak ad helyet, éppen ezért egy színháztörténeti projekt befogadása számos kérdés és kihívás elé állította a kutatókat. A specifikáció során olyan szempontokat is figyelembe kellett vennünk, amelyekkel eddig nem találkoztunk. A fő kérdésünk az volt, hogy milyen módon illeszthetnénk be a jelmeztervek rekordjait az eddigi struktúrákba, illetve milyen új tulajdonságok és entitások szükségesek az optimális megvalósuláshoz. A félautomatikus adatfelvitel mellett döntöttünk (QuickStatements), ehhez azonban számos előkészítő munkálatot kellett megvalósítani. A tanulmány a teljes munkafolyamatról, illetve annak tanulságairól is szól.

Abstract

As part of the Danube-AI program of the National Laboratory for Digital Heritage, in 2022 the costume designs of the Hungarian State Theatre of Cluj-Napoca were inventoried,

1 A szerző a tanulmány megírása idején az MTA BTK Lendület Magyar Irodalom Politikai Gazdaságtana Kutatócsoport (34080 LP 2019-10/2019) támogatásában részesült.

digitised, organised in a database and published. The paper reports about this project, which was carried out in cooperation with several partners. The database includes images and metadata of the costume designs of 94 performances of the Hungarian State Theatre of Cluj-Napoca between 1959 and 1980, as well as basic information (e.g. casting) related to each performance. The data provide insight not only behind the scenes of stage visuality, but also into the life's work of significant costume designers. ITIdata is primarily a semantic database designed for literary studies, which is why hosting a theatre history project presented a number of questions and challenges. The data specification also had to take into account aspects that had not been encountered before. Our main question was how to integrate the costume design records into the existing structures, and what new features and entities were needed for optimal realisation. We opted for semi-automatic data entry (QuickStatements), but this required a number of preparatory steps. The paper also describes the overall workflow and the lessons learned.

Kulcsszavak: kulturális örökség, színháztörténet, digitalizálás, jelmezterv
Keywords: cultural heritage, theater history, digitization, costume design

A tanulmány egy, a Digitális Örökség Nemzeti Laboratórium (DH-LAB) keretein belül zajló projekt első szakaszának bemutatásáról szól, nevezetesen a Kolozsvári Állami Magyar Színház Dokumentációs Tárában őrzött jelmeztervek digitalizálásáról, az adatfeldolgozásról és közzétételéről.

Az idei és egy korábbi Networkshopon is elhangzott az együttműködés fontossága a kulturális örökség feldolgozása és közzététele, a mesterségesintelligencia-alapú eszközök hasznosításának vonatkozásában. Az elmúlt időszakban az intézményi együttműködés témája több kerekasztal-beszélgetésen, konferencián is hangsúlyosan megjelent a közgyűjteményi digitalizáció, a metaadatok megosztása, általában a mesterséges intelligencia kulturális örökségi használata szempontjából. A most bemutatásra kerülő projektet ebből a szempontból egy mintaprojektnek tekintjük, hiszen több intézmény összefogásával valósult meg (benne határon túli gyűjteménnyel), az eredményei szabadon elérhetőek, az adatok pedig egy szemantikus adatbázisba kerülnek integrálásra.

A tanulmányban a projekt infrastrukturális, stratégiai ismertetése után a gyűjteményi állomány színháztörténeti relevanciájáról, majd az adatstruktúra kialakításáról és a rekordoknak az ITIdata adatbázisba való beviteléről lesz szó.

1. A Digitális Örökség Nemzeti Laboratórium Danube-AI programja

A Digitális Örökség Nemzeti Laboratórium (DH-LAB) Danube-AI[2*] programjának célkitűzése, hogy mesterséges intelligencia alapú jó gyakorlatok kialakításával feldolgozza a veszélyeztetett, vagy korlátozottan hozzáférhető (akár határon túli) gyűjteményeket, digitális és történelmi forrásokat. A DH-LAB az elmúlt évben több együttműködési megállapodást kötött különböző köz- és magángyűjteményekkel, amelyek keretén belül több pilot projekt zajlik. A born digital örökséghez kapcsolódóan a konferencián már hallhattak egy másik folyamatban levő munkáról: Kántor Lajos irodalomtörténész e-mail hagyatékának feldolgozásáról[3*], és reményeink szerint hamarosan beszámolhatunk a Móra Ferenc Múzeumban őrzött Móra

2 https://dh-lab.hu/danube-ai/

3 Lásd jelen kötetben: Alföldi István – Szemigán Dorottya Henrietta – Palkó Gábor – Fellegi Zsófia: Kutatói e-mail hagyaték archiválása és feldolgozása c. tanulmányt

CC BY

levelek kézírás-felismertetéséről a DH-LAB HTR eszközével. Az együttműködések sorában van a Kolozsvári Állami Magyar Színház is. Ezekkel a jól körülhatárolható, viszonylag kevés erőforrást igénylő esettanulmányokkal a Danube-AI alprojekt amellett, hogy bővíti a magyar kulturális örökség digitálisan elérhető állományát, hozzájárul a DH-LAB MI alapú eszközeinek teszteléséhez, magyar nyelvű szövegkorpuszainak bővítéséhez és nem mellékesen egyfajta missziós tevékenységet is végez. A többnyire alulfinanszírozott és humánerőforrással küszködő kisebb közgyűjteményekben kevés lehetőség van a szakmai továbbképzésekre; ezekben a pilot projektekben a gyűjteményekkel együttműködve tudunk dolgozni, őket mentorálni, az eszközöket testreszabni, a munkafolyamatokat az egyéni igények szerint rugalmasan összeállítani. A folyamat során tulajdonképpen folyamatos visszajelzéseket kapunk, amelyekkel finomhangolhatók az eszközök. Nem titkolt szándék, hogy ezeken a projekteken keresztül terjeszteni kívánjuk a digitális bölcsészet eszköztárát és a nyílt tudományosság szemléletét.

2. A munkafolyamat

A DH-LAB Danube-AI alprojektjének részeként 2022-ben került sor a Kolozsvári Állami Magyar Színház Dokumentációs Tárában[4*] őrzött jelmeztervek nyilvántartásba vételére, leírására, digitalizálására, adatbázisba rendezésére és közzétételére. A megvalósításban kulcsszerepe volt a DH-LAB konzorcium partnerének, a Magyar Nemzeti Levéltárnak, valamint közreműködött a Forum Hungaricum Nonprofit Kft. is.

A most bemutatásra kerülő jelmeztervek tehát egy kis szeletét képezik a gyűjteménynek és az elkövetkező időszakban folytatjuk a munkát a fotógyűjteménnyel, a színműdossziékkal és a kéziratgyűjteménnyel.

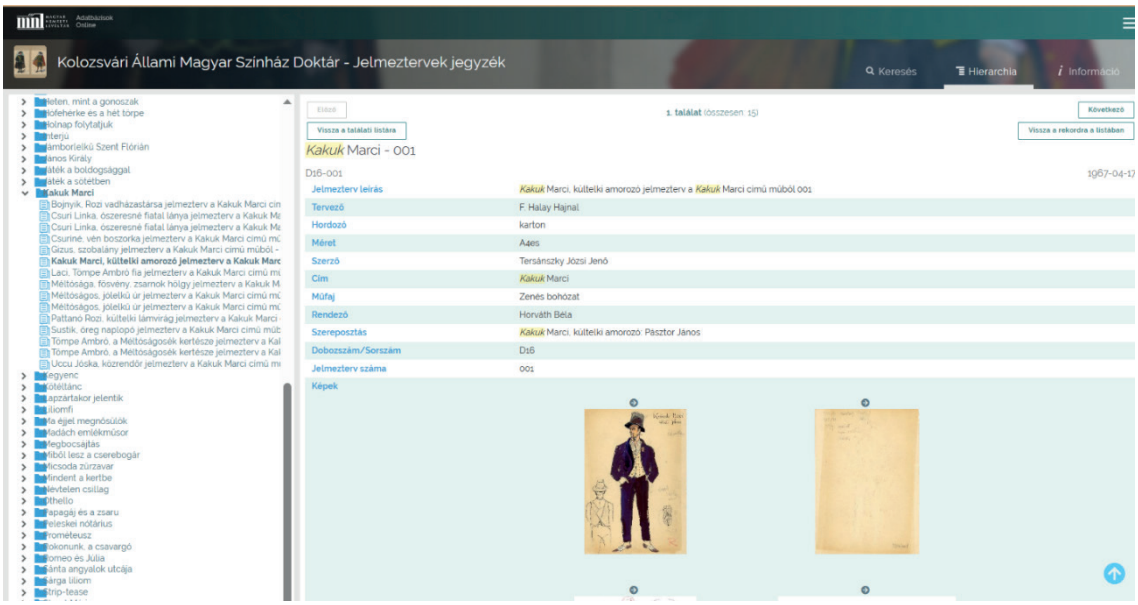
A papíron, kartonon levő jelmeztervek leltározása, számozása (1725 db), darabszintű jegyzékbe vétele és alapadatainak strukturált leírása volt az első lépés. A Forum Hungaricum közreműködésével ezeket, illetve két kötetet (az 1906–1928 között lejátszott darabok helyszínkóddal ellátott jegyzékét és a Színművek leltárkönyvét) Budapestre szállítottunk a Magyar Nemzeti Levéltárhoz. Az MNL műhelyében Hegedűs István irányításával Tauber Diana digitalizálta a jelmezterveket a kötetekkel együtt. Ezzel egyidőben az MNL restaurálta a rossz állapotban levő egyik kötetet és a jelmeztervek tárolására új savmentes tárolódobozokat készíttetett. A digitalizálás mellett ezáltal egy fontos állományvédelmi beavatkozás is történt. A digitalizálás után elkezdődött az adatok ellenőrzése, esetenkénti bővítése, a jegyzék egységesítése, korrigálása.

Következő lépés a közzététel volt. Ezt 2022. december 8-án, Janovics Jenő kolozsvári színész, rendező, színházigazgató és a magyar filmgyártás úttörő alakja születésének 150. évfordulójának alkalmára időzítettük. A jelmeztervek a Magyar Nemzeti Levéltár adatbazisokonline.hu felületén[5*] érhetők el nagy felbontásban. Ugyanakkor folyamatban

4 A Kolozsvári Állami Magyar Színház Dokumentációs Tárának létrehozója Dr. Salat-Zakariás Erzsébet. Tulajdonképpen neki köszönhető, hogy létezik, megőrződött ez a dokumentum és tárgyegyüttes és, hogy mindezt gyűjteménynek lehet nevezni. A Dokumentációs Tár a színház művészeti archívuma, a színház életére vonatkozó különféle dokumentumokat, fényképeket és emléktárgyakat, valamint a színházi élethez kapcsolódó szakkönyveket, folyóiratokat tartalmazza. Kiemelnénk Kocsis Tünde jelenlegi gyűjteménykezelő szerepét a projektben, aki részmunkaidőben, nagyon sok lelkesedéssel és lelkiismeretesen gondozza a gyűjteményt.

5 https://dh-lab.hu/a-kolozsvari-allami-magyar-szinhaz-jelmezterveinek-digitalizalasa/ ; https://adatbazisokonline.mnl.gov.hu/adatbazis/kolozsvari-allami-magyar-szinhaz-doktar-jelmeztervek-jegyzek/informacio

van a digitális állomány feltöltése a Magyar Nemzeti Digitális Archívumba, a MaNDA DB-be is, ahonnan szintén szabadon felhasználhatók.[6*] Végül pedig, amiről a tanulmány második felében számolunk be: integráltuk a jelmeztervek metaadatait a Bölcsészettudományi Kutatóközpont Irodalomtudományi Intézetének szemantikus adatbázisába (itidata.abtk.hu) és folyamatban van az adatok Europeanába való aggregálása is. Az elsőre redundánsnak tűnő többszörös közzététel azért is indokolt, mert különbözőképpen kapcsolódik ez az algyűjtemény az MNL-hez a Hungarika állománya kapcsán és a MaNDA DB-hez, az ott meglevő egyéb színháztörténeti állományhoz. Emellett más-más felhasználói kör látogatja a portálokat, így több csatornán, szélesebb körben hozzáférhető az eddig ismeretlen gyűjtemény.



1. ábra: Képernyőkép a Kakuk Marci c. előadás egy jelmezéről az adatbázisokonline.hu-n. 2023.07.27. Forrás: https://adatbazisokonline.mnl.gov.hu/adatbazis/kolozsvari-allami-magyar-szinhaz-doktar-jelmeztervek-jegyzek/informacio

A mesterséges intelligencia az utóbbi években történő hatalmas térhódítása és hasznossága mellett meg kell emlékeznünk a gyűjteményfeldolgozási munka egy másik oldaláról is. Ezzel a projektismertetővel szeretnénk rávilágítani a láthatatlan munkára: a manuális előkészítésre, feldolgozásra, az adatstruktúra kidolgozására (amelynek csak a végső eredménye látható, a számtalan próbálkozás, tesztelés általában kevésbé látszik). Habár apad a feldolgozatlan gyűjtemények száma, még mindig sok, a kulturális örökség témakörébe tartozó forrás nem érhető el digitálisan, sőt nincs leltárba véve; sokan ismerjük a raktárak mélyéről előbukkanó műtárgyak, dokumentumok történetét.

3. Színháztörténeti kontextusok

A láthatóság/láthatatlanság kérdése színháztörténeti szempontból is releváns probléma, ez implikálja a kulturális centrum és kulturális periféria/félperiféria kérdését is, azaz a nem fővárosi, budapesti – illetve, ha a 20. században vizsgálódunk, a nem a mai Magyarország területén lévő – magyar színházak helyét a magyar színháztörténetben. A magyar színháztörténetírás nagyelbeszéléseiben ezek a színházak, így a kolozsvári is, sokáig perifériku helyet foglaltak el, a klasszikus színháztörténeti monográfiák, kézikönyvek csak röviden

6 https://mandadb.hu/tart/kereses?HNDTYPE=SEARCH&name=doc&page=1&_clearfacets=true&_clearfilters=true&fld_compound=&fac_organization=217

CC BY

foglalkoztak velük,[7*] jelentőségük alárendelődött a fővárosi intézményeknek, a centrum felől nézve láthatatlanok maradtak, vagy alig és csak torzítva váltak láthatóvá. Szükség volt egy lépték- és paradigmaváltásra, a posztmodern történelemszemlélet fordulatára ahhoz, hogy az elbeszélések pluralitása, valamint a regionális és a lokális nézőpont emancipálódjon, ezáltal a korábbi perifériák láthatóvá váljanak.

Míg a kolozsvári színház kéziratos és nyomtatott forrásokban gazdag 19. századi története inkább a kutatók fókuszába került,[8*] az első világháború utáni időszak már jóval kevésbé feltérképezett. Jelen kutatásunk az intézmény 20. századi történetének értelmezéséhez kíván hozzájárulni, szorosabban az 1959–1980 közötti időszakot, illetve az ebben az időszakban bemutatott előadásokhoz készült jelmezterveket helyezve középpontjába. A Kolozsvári Állami Magyar Színház Dokumentációs Tárában őrzött korpuszból eddig 116 előadás jelmeztervének (ami több mint 1720 rajzot, illusztrációt jelent) feldolgozása kezdődött el (időközben még előkerült néhány, további előadásokhoz készült terv). A színháztörténetírás gyakran választ repertoárelemzést, vagy egy-egy kiemelt rendezői korszaknak, egy-egy híres színészi életpályának a feltárását. Mindezek számunka is fontosak: a kutatás jól körülhatárolható korpuszába, a jelmeztervekben a repertoár is kódolva van, általa pedig a rendezőkről és a színészekről is fontos adatok kerülnek felszínre. Ezen túlmutatva pedig a színház olyan, többnyire láthatatlan szereplőit is meg tudjuk mutatni, mint a jelmeztervezőt, akinek a munkája, kissé ironikus módon, a színházi világ nagyon is látható részéhez tartozik. F. Halay Hajnal, Schranz Kunovics Edit, Bârsan Éva, Sütő Éva, Mihai Nemeș, Cs. Erdős Tibor, Kozma Elza, Jules Perahim, Mircea Marosin, Carmencita Brojboiu, Bodor Mária, Constantin Russu, Tóth Szűcs Ilona, Paul Bortnovszkij – ilyen nevekkel találkozunk a tervek alkotói között. A vizuális anyaghoz rengetek egyéb adat tartozik, itt nemcsak arra gondolunk, hogy milyen hordozóra és milyen technikával készültek – ezeket is rögzítettük –, hanem arra is, amit az implikál, hogy ezek a tervek előadásokhoz készültek, s egy előadás egy mű színpadi változata, így az értelmezés, ahogy az ágrajz is, kinyílik a szövegkönyv, szereposztás, rendező, bemutató időpontja felé is.

A tanulmányban egy példán, Tersánszky Józsi Jenő *Kakuk Marci*jának Örkény István által készített adaptációján fogjuk megmutatni azt, ahogy az ITIdatában dolgozzuk fel a jelmezterveket.

---

7 Ennek alapművei: Kerényi Ferenc, szerk., Magyar színháztörténet 1790–1873. Budapest, Akadémiai Kiadó, 1990; Gajdó Tamás, szerk., Magyar színháztörténet 1873–1890. Budapest, Magyar Könyvklub – Országos Színháztörténeti Múzeum és Intézet, 2011.

8 Itt most csupán néhány példát emelünk ki a színház 19. századi történetével foglalkozó tanulmányok közül. Bartha Katalin Ágnes, Shakespeare Erdélyben. XIX. századi magyar nyelvű recepció. Budapest, Argumentum, 2010; Bartha Katalin Ágnes: Színházi professzió és presztízs Kolozsváron a 19. század utolsó harmadában. Erdélyi Múzeum, LXXVII. 2015. 3. 46-78.; Egyed Emese (szerk.) Theátrumi könyvecske. Színházi zsebkönyvek és szerepük a régió színházi kultúrájában. Kolozsvár, Scientia, 2002; Szabó-Reznek Eszter, Száz év előtt – száz év után. Az erdélyi hivatásos színjátszás 1892-es centenáriuma. Doktori disszertáció, kézirat, Szegedi Tudományegyetem, 2020.



2. ábra: Képernyőkép a Kakuk Marci c. előadás egy jelmezéről a mandadb.hu-n. 2023.07.27.
Forrás: https://mandadb.hu/tetel/855629/Kakuk_Marci_a_kultelki_amorozo

Az 1967. április 17-én bemutatott előadáshoz F. Halay Hajnal tervezett jelmezeket, aki, miután 1955-ben diplomát szerzett a kolozsvári képzőművészeti akadémián, a színházban kezdett el dolgozni 1970-ig, amely idő alatt több mint 70 előadás jelmeztervezője volt. A jelenleg digitalizált korpuszban 30 előadáshoz készített jelmeztervét találjuk az 1958 és 1970 közötti időszakból, valamint egy 1959-es Szophoklész-előadásnak a segédruhatervezője Cs. Erdős Tibor alatt – és ezzel a számmal ő vezeti a legtöbb előadáshoz tervezett jelmez listáját a mostani állás szerint. Egy színháztörténész számára érdekesek lehetnek a jelmeztervek széljegyzetei vagy a verzóra írt feljegyzések. Ezek sokszor a textilek mennyiségére és árára vonatkoznak (sőt, időnként textilmintákat is ragasztottak a tervekre), máskor arról kapunk (akár gazdaság- és intézménytörténeti szempontból is értelmezhető) adatokat, hogy egy-egy előadáshoz mennyi új ruhára volt szükség és mennyit tudtak a raktárból, a már meglévő ruhákkal megoldani. Arra is találunk példát, hogy egy-egy színész saját, civil ruhatárából hoz egy-egy darabot az előadásba (sajátos párhuzama ez a 19. századi színjátszás működésével, ahol a színészek maguk voltak felelősek a jelmezeikért).

Ez a megközelítés tehát túlmutat a szokványos repertoárelemzéseken, a műsorrendet a jelmeztervek szűrőjén át vizsgálja. Később az adatstruktúra bővíthető lesz további vizuális (színpadtervek, előadásképek) és írott források adataival, mint például előadáskritikák (azaz a recepciótörténet), interjúk vagy a szövegkönyvek – amiből sokszor van rendező-, ügyelő-, súgó- és színészpéldány is, a *Kakuk Marci* esetében például a Pattanó Rozit játszó Bereczky Júlia példányát őrzi a Dokumentációs Tár, ezekből az alapmű és annak színpadi változata közötti viszony is kirajzolódhat. Az adatbázis segítségével és a források felfejtésével, elemzésével eddig példa nélküli komplex képet kaphatunk az 1960–70-es évek kolozsvári színházának egyes előadásairól, vizuális kultúrájáról, a színészi foglalkoztatottság gyakoriságáról, műfaji mutatókról, előadások sikerességéről, népszerűségéről.

CC BY

4. A jelmeztervek rekordjainak specifikációja és feltöltése az ITIdata adatbázisba

4.1. A jelmeztervek specifikációja

A projekt végső fázisában az volt a célunk, hogy a Kolozsvári Állami Magyar Színház jelmeztervei felkerüljenek az ITIdata wikibase alapú szemantikus adatbázisba.[9*] Ehhez első lépésben a jelmeztervek metaadatait kellett összesítenünk Excel táblázatokban. Ezt követően a metaadatok összehangolását végeztük el: ezek egy részéhez már meglévő tulajdonságokat (properties) tudtunk rendelni (gyűjtemény, szerző entitás, műfaj, hordozó stb.), másik részéhez pedig létre kellett hoznunk új tulajdonságokat (Kolozsvári Állami Magyar Színház jelmezterveinek egyedi azonosítója, jelmeztervező entitás, rendező entitás, szereposztás stb.). A tulajdonságokat minden esetben kétféle szempontrendszer érvényesítésével hoztuk létre: 1. megfeleljen az adott projekt metaadat-specifikációjának; 2. más ITIdatában dolgozó projektek is hasznosítani tudják.

4.2. A feltöltés menete

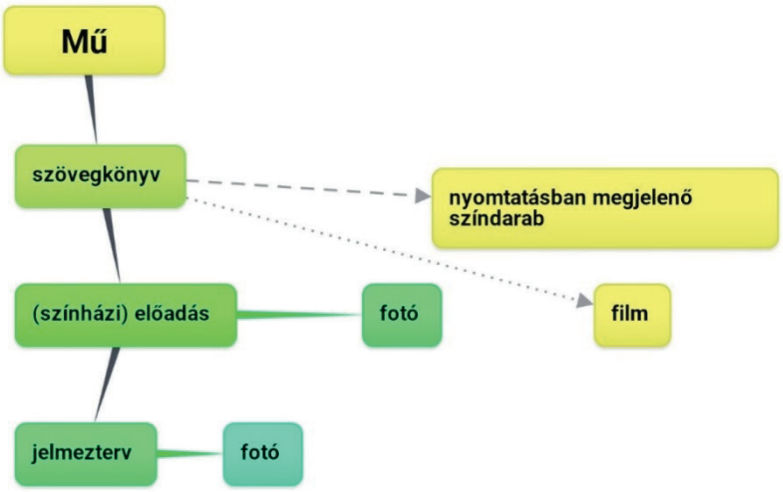
A specifikációt követően az adatok feltöltését végeztük el. A félautomatikus adatfelvitel mellett döntöttünk (QuickStatements), ezért az Excel táblázatban szereplő metaadatokat a megfelelő módon előkészítettük, formáztuk, hogy az adatfelvitel problémamentes legyen. A munka több napon keresztül tartott, végül 1722 rekord került fel az ITIdata adatbázisba. A projekt rövid leírást tartalmazó oldala, valamint az összes rekord listája elérhető az ITIdata nyitóoldalán keresztül.[10*]

4.3. Utólagos adatgazdagítás

A jelmeztervek rekordjainak metaadatai az első feltöltési fázisban hibridnek voltak tekinthetők. A jelmezterveket leíró tulajdonságok (hordozó, méret, jelmeztervező entitás) keveredtek a színházi előadás (rendező entitás, bemutató dátuma), valamint a szövegkönyv (szerző entitás) adataival. Ezért úgy döntöttünk, hogy létrehozunk egy szemantikus hálózatot, és a már meglévő rekordokat további rekordokra bontjuk, és ezek között kapcsolatokat hozunk létre. A színháztörténeti adatok, valamint az ITIdata irodalmi művekre vonatkozó struktúrájának fényében a következő rekordtípusokat hoztuk létre: 1. mű típusú, 2. szövegkönyv, 3. színházi előadás, 4. nyomtatásban megjelenő színdarab, 5. további specifikus rekordok (film). Első próbarekordjaink a *Kakuk Marci* című színházi előadáshoz tartozó jelmeztervekhez kapcsolódott. A színházi előadás Tersánszky Józsi Jenő négy regényének cselekményét használta fel, ezért a mű típusú tételünk, amely alapesetben egy rekord lenne, és a mű összes periodikában, kötetben megjelenését, valamint a kéziratos anyagot és digitális kiadást is tartalmazza, ez esetben négy különböző rekord lett. Létrehoztuk tehát az alábbi mű típusú rekordokat: Kakuk Marci ifjúsága,[11*] Kakuk Marci a zendülők között,[12*] Kakuk Marci kortesúton[13*] és Kakuk Marci vadászkalandjai.[14*] A regényekből készített Örkény István szövegkönyvet,[15*] a szövegkönyvből pedig létrejött a színházi előadás,

9 https://itidata.abtk.hu/wiki/Main_Page
10 https://itidata.abtk.hu/wiki/A_Kolozsv%C3%A1ri_%C3%81llami_Magyar_Sz%C3%ADnh%C3%A1z_jelmeztervei
11 https://itidata.abtk.hu/wiki/Item:Q329103
12 https://itidata.abtk.hu/wiki/Item:Q329535
13 https://itidata.abtk.hu/wiki/Item:Q308143
14 https://itidata.abtk.hu/wiki/Item:Q508
15 https://itidata.abtk.hu/wiki/Item:Q328182

amelynek szintén önálló rekordja lett az adatbázisban a hozzá tartozó tulajdonságokkal együtt (szereposztás, bemutató dátuma, rendező entitás).[16*] Az előadáshoz rendeltük hozzá kapcsolódó rekordokként a jelmezterveket. A vizsgált esetünk külön érdekessége, hogy a színdarab jóval később, 2016-ban megjelent nyomtatott formában is,[17*] illetve a művekből készült egy magyar játékfilm is 1973-ban.[18*] Az összes rekord között létrehoztuk a szemantikus kapcsolatokat, amelyeket az alábbi ábra szemléltet:



3. ábra: A jelmeztervekhez kapcsolódó szemantikus adatstruktúra az ITIdata-ban. Forrás: saját szerkesztés

4.4. Véglegesítés és további tervek

A rekordok közötti szemantikus kapcsolatok létrehozásához irodalomtörténeti és színháztörténeti kutatás is szükséges. Ez a későbbiekben is szükséges lesz, a szemantikus hálózat létrehozása tehát nem lesz automatikus, hiszen a rekordok közötti összefüggések minden esetben egyediek. A létrejövő rekordokat természetesen más ITIdatát használó projektek is tudják hasznosítani és kiegészíteni. Hosszú távú célunk, hogy az összes jelmeztervet elhelyezzük a fent vázolt szemantikus hálózatban.

Felhasznált irodalom

Bartha Katalin Ágnes. *Shakespeare Erdélyben. XIX. századi magyar nyelvű recepció*. Budapest, Argumentum, 2010.

Bartha Katalin Ágnes. Színházi professzió és presztízs Kolozsváron a 19. század utolsó harmadában = *Erdélyi Múzeum*, LXXVII. 2015. 3. 46-78.

Egyed Emese (szerk.) *Theátrumi könyvecske. Színházi zsebkönyvek és szerepük a régió színházi kultúrájában*. Kolozsvár, Scientia, 2002.

Gajdó Tamás (szerk.) *Magyar színháztörténet 1873–1890*. Budapest, Magyar Könyvklub – Országos Színháztörténeti Múzeum és Intézet, 2011.

Kerényi Ferenc (szerk.) *Magyar színháztörténet 1790–1873*. Budapest, Akadémiai Kiadó, 1990.

Szabó-Reznek Eszter. *Száz év előtt – száz év után. Az erdélyi hivatásos színjátszás 1892-es centenáriuma*. Doktori disszertáció, kézirat, Szegedi Tudományegyetem, 2020.

16 https://itidata.abtk.hu/wiki/Item:Q327322
17 https://itidata.abtk.hu/wiki/Item:Q308144
18 https://itidata.abtk.hu/wiki/Item:Q336193

CC BY

DOI: 10.31915/NWS.2023.17

H5P-ben létrehozható interaktív és adaptív tananyagok

Interactive and Adaptive Learning Materials Created in H5P

Köpösdi Zsuzsa
Debreceni Egyetem, Multimédia és E-learning Technikai Központ
koposdi.zsuzsa@metk.unideb.hu
ORCID: 0000-0002-2185-4887

Absztrakt

A H5P tartalomfejlesztő keretrendszer nyílt forráskódú, ingyenes és használata nem igényel speciális informatikai szakértelmet. A H5P tartalomfejlesztőnek köszönhetően viszonylag egyszerűen és gyorsan létrehozhatók interaktív és adaptív tananyagok. Az előadás és a publikáció célja, hogy a H5P keretrendszer és az ilyen típusú tananyagelemek használata tovább erősödjön a hazai oktatási folyamatokban.

Kulcsszavak: e-learning, interaktív tananyagok, adaptív tananyagok, H5P, Moodle, interaktív videó, elágazó forgatókönyv

Abstract

H5P framework is open source, user-friendly, free, and requires no special IT expertise. With H5P, we can create interactive and adaptive learning materials relatively quickly and easily. With the presentation and publication my goal is to further strengthen the use of H5P framework in the Hungarian educational processes.

Bevezető

Az előadás és a publikáció célja, hogy – a Networkshop 2022 konferencián elhangzott *Multimédiás, interaktív és adaptív tananyagok létrehozásának lehetőségei H5P keretrendszerrel* című H5P bevezető előadás folytatásaként – további H5P tananyagtípusokat mutasson be részletesen, illetve, gyakorlati megoldásokat mutasson arra, hogyan tudunk viszonylag egyszerűen és gyorsan H5P tananyagelemeket létrehozni, vagy hagyományos tananyagelemekből is gazdagabb és előremutató interaktív tananyagot fejleszteni. Ezentúl, konkrét példán keresztül bemutatásra kerül, hogy a H5P tananyagok milyen egyszerűen újrahasznosíthatók, azaz, hogy a H5P fájlok különböző portálok között milyen egyszerűen és gyorsan költöztethetők, szerkeszthetők és felhasználhatók. Természetesen, a legfontosabb cél az, hogy ezekkel az ismeretekkel az oktatók és a tartalomfejlesztők bátrabban használják a H5P tartalomfejlesztő eszközt és egyre szélesebb körben alkalmazzák e-learning kurzusaikban a H5P tananyagmodulokat – hiszen ezek a tartalmak a tanulási folyamatokat hatékonyan tudják támogatni és ezekkel az eszközökkel a tanulók figyelmét megragadó tananyagok hozhatók létre. Az előadásban rövid videókon keresztül kerültek bemutatásra az adott H5P tananyagmodulok, illetve a H5P fájlok letöltése, majd importálás utáni szerkesztése, újrahasznosítása is. Ezek a videók megtekinthetők az előadásról készült felvételen.

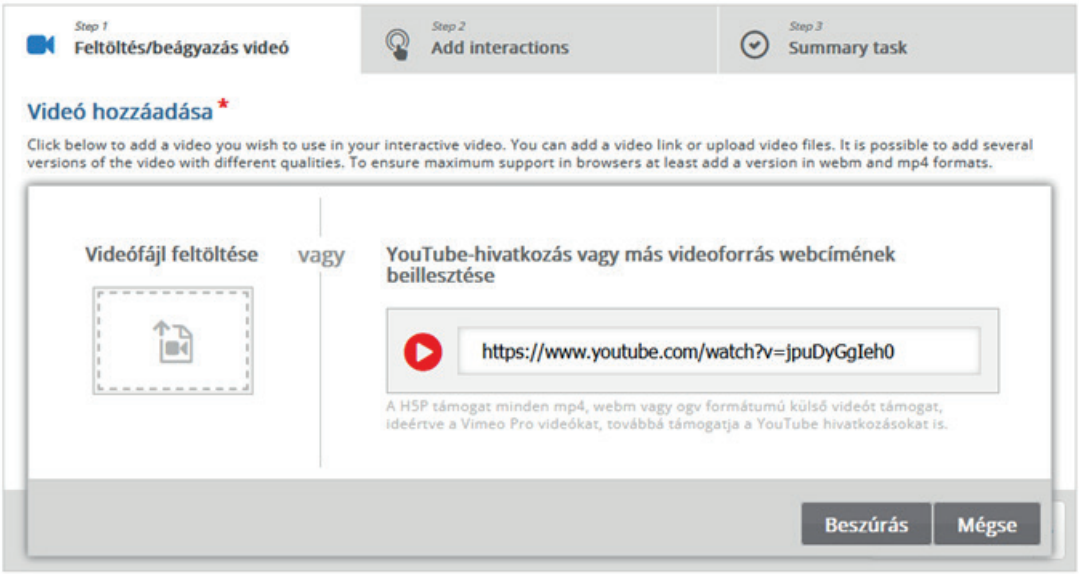
H5P tananyagmodulok, H5P-ben létrehozható tananyagtípusok

A H5P tartalomfejlesztő használatához nem szükséges speciális informatikai szakértelem, könnyen kezelhető, felhasználóbarát és ingyenes – segítségével viszonylag gyorsan és egyszerűen hozhatunk létre interaktív, illetve adaptív tananyagokat. A H5P abból a szempontból is felhasználóbarát, hogy a program integráltan működik a Moodle és a Canvas LMS rendszerekben is.[1] A H5P-ben létrehozott tananyagokat pedig szintén nagyon egyszerűen – .h5p kiterjesztésű fájlban – letölthetjük, majd a számunkra megfelelő e-learning rendszerbe importálhatjuk és tovább is szerkeszthetjük. Az ilyen szintű újrafelhasználhatóság nagyban növeli a tananyagkészítés hatékonyságát.

A H5P hivatalos honlapján megtalálható mindegyik jelenleg létező tananyagmodul, tananyagtípus, kérdéstípus, feladattípus – részletes leírással és letölthető példákkal.[2]

A publikáció megírásának időpontjában a honlapon 53 tananyagmodul található. Érdemes lehet ezeket egyenként megismerni és az e-learning kurzusunk céljainak leginkább megfelelő modulokat alkalmazni. Mivel az összes modul megismerése nagyon sok időt vehet igénybe, ezért ebben a publikációban is szeretném néhány kiemelkedő jelentőségű modulra felhívni a figyelmet.

Az **interaktív videó** funkcióit, létrehozásának lépéseit és alkalmazásának előnyeit a NWS 2022-es előadásban részletesen bemutattam. Itt kiegészítésként egy fontos lehetőségre szeretném felhívni a figyelmet az interaktív videó tananyagmodulban: YouTube videót is használhatunk az interaktív videó alapjaként. Azaz, nem feltétlenül szükséges saját videót felhasználni és azt feltölteni alapvideóként, hanem erre a célra a YouTube-on elérhető videókat is alkalmazhatunk. Ehhez az interaktív videó tananyagmodul szerkesztőfelületén a videó feltöltése lépésnél egyszerűen csak be kell másolnunk a megfelelő mezőbe az adott videó YouTube linkjét – ez a lépés látható a következő ábrán.



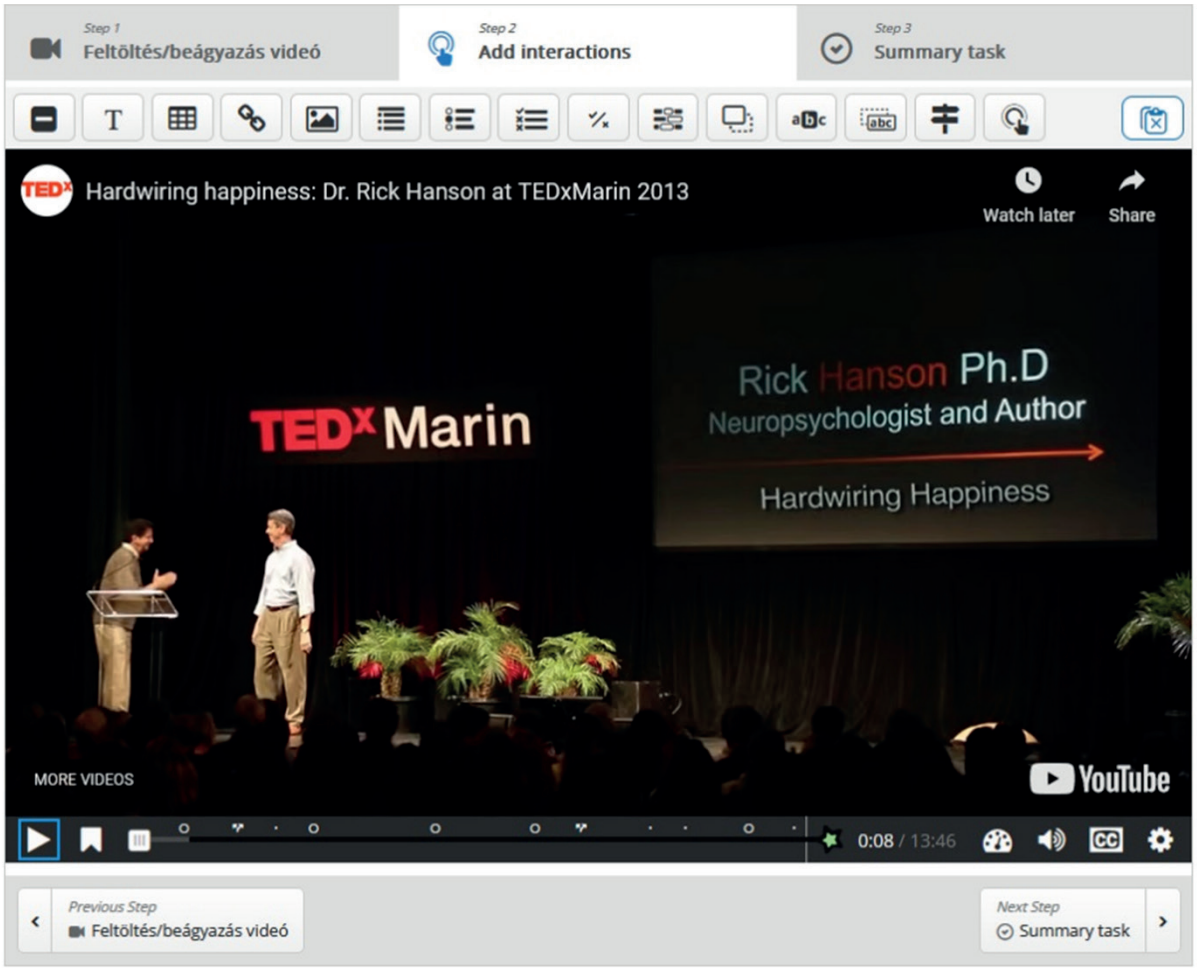
1. ábra Interaktív videó létrehozása YouTube videó felhasználásával

1 H5P - MoodleDocs. https://docs.moodle.org/39/en/H5P Hozzáférés: 2023. június 19.

2 Examples and Downloads | H5P. https://h5p.org/content-types-and-applications Hozzáférés: 2023. június 18..

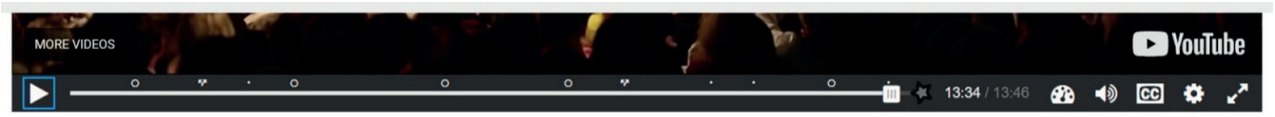
CC BY

A YouTube videó alapú interaktív videó előnye, hogy nem kell erőforrásokat befektetni saját videók leforgatásába, hátránya viszont, hogy ha elérhetetlenné válik az adott videó, akkor az természetesen hatással lesz az e-learning tananyagunkra is. A YouTube videó linkjének bemásolása után, ahogy a szerkesztőfelületen a *tevékenységek hozzáadása* lépéséhez érünk, már minden lépés és funkció azonos, mint a saját videó feltöltése esetén. A következő ábrán egy YouTube videó alapú interaktív videó szerkesztőfelülete látható, a *tevékenységek hozzáadása* funkcióban.



2. ábra YouTube videó alapú interaktív videó szerkesztése

A következő ábrán pedig egy elkészült YouTube alapú interaktív videó képernyőkép részlete látható: az alsó sáv, ahol a videóhoz hozzáadott tevékenységeket jelölő ikonok láthatók.



3. ábra YouTube videó alapú interaktív videó tevékenységeket jelölő sávja

Az **elágazó forgatókönyv** (Branching Scenario) megnevezésű tananyagmodul is szélesebb körben alkalmazásra érdemes modul. Ez a tananyagelem a döntéshozatal gyakorlására vagy gyakoroltatására kifejezetten jól alkalmazható. A való életből vett szituációkat, szimulációs gyakorlatokat lehet a hallgatókkal begyakoroltatni. A hallgatók azt is meg tudják tapasztalni, hogy a döntésüknek abban a pillanatban következménye van. Nagyon jól használható

tehát orvosok, szakápolók, mentőtisztek, mentálhigiénés szakemberek, pszichológusok képzésében – de természetesen sok más területen is.

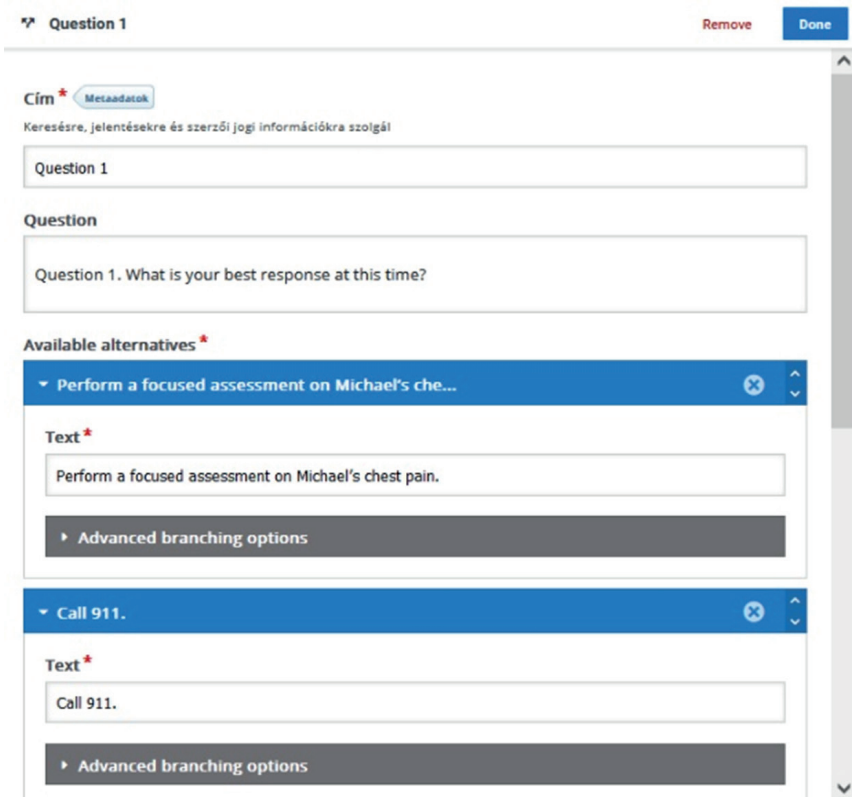
Az elágazó forgatókönyv modul segítségével meg tudunk határozni adott helyzetet, élethelyzetet, döntési helyzetet, és oktatóként információkat adhatunk erről a kiinduló helyzetről a hallgató számára. Ezeket az információkat átadhatjuk videó, interaktív videó, interaktív prezentáció, szöveges vagy képi tartalmak segítségével. A hallgató végig tud haladni ezeken az információs elemeken, majd eljut az első, általunk meghatározott döntési ponthoz. Az előadásban bemutatott példában egy középkorú férfi számol be bizonyos testi tüneteiről, amelyek szívinfarktus gyanúját vetik fel. A hallgató ebben a bemutatott tananyagban dönthet arról, hogy az adott szituációban, adott tünetek és adott információk alapján mentőt hív-e a beteghez, vagy néhány vizsgálatot önállóan elvégez és megkezd bizonyos szintű terápiát, stb. Adott válaszok adott útvonalon vezetik végig a hallgatót, ahol további információkat kaphat a döntésének következményeiről, és újabb döntési pontokon kell választania a lehetőségek közül. Így megtapasztalja, hogy döntéseinek következményei vannak, illetve, gyakorolhatja, hogy milyen tényezőket kell figyelembe vennie, esetleg milyen protokollt kell követnie, stb. adott helyzetekben. Döntései után azonnali visszajelzést kap, illetve, esetenként kaphat segítő információkat is arról, hogy adott döntés miért bizonyul adott helyzetben kevésbé jó döntésnek és melyik válaszlehetőség bizonyulna jobbnak. Oktatóként vagy tananyagfejlesztőként tehát meghatározom a kiinduló helyzetet, amelyről információkat adok a hallgatóknak (videó, képi, audio vagy szöveges tartalmak felhasználásával), döntési pontokat határozok meg (kérdések és válaszlehetőségek létrehozása), és megtervezem a lehetséges útvonalakat. A következő két ábra az elágazó forgatókönyv modul egy-egy szerkesztőfelületét mutatja. Az első ábrán az adott elágazó forgatókönyv típusú tananyag szerkezetének egy részlete látható, amely a különböző elemekből, illetve a döntési pontokból és válaszlehetőségekből áll össze.



4. ábra: Elágazó forgatókönyv típusú tananyag szerkezetének egy részlete

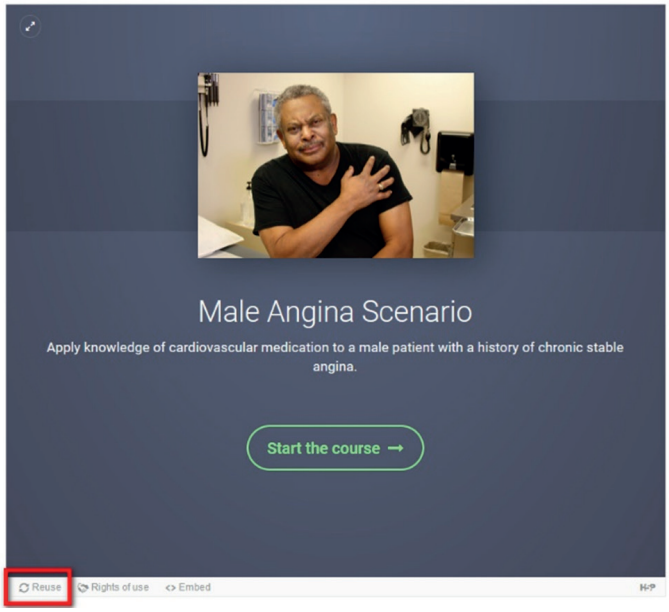
CC BY

A következő ábra pedig az egyik döntési pont szerkezetét és szerkesztői felületét mutatja:



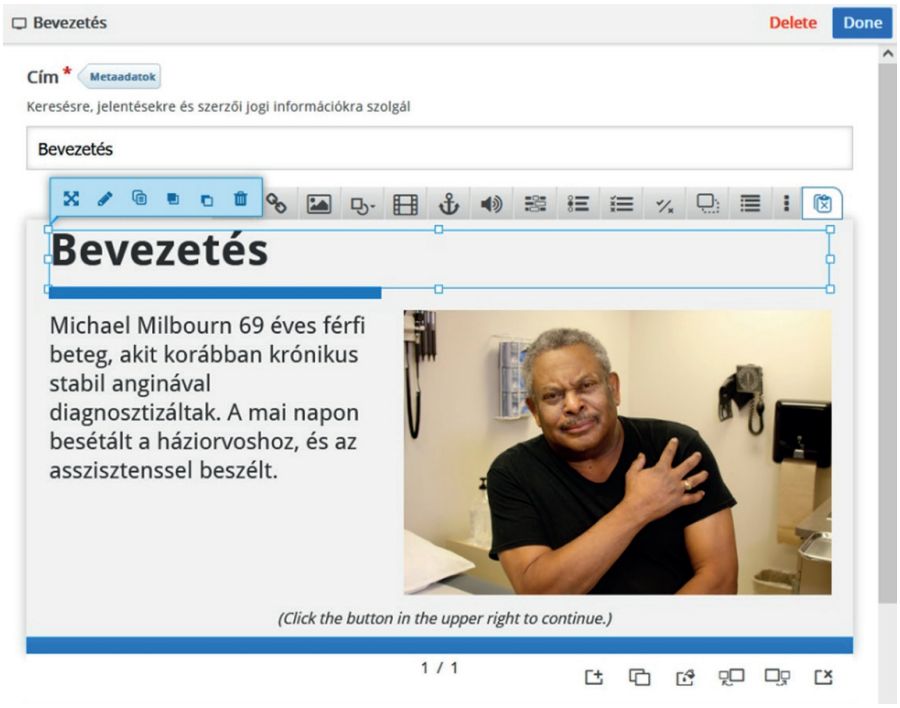
5. ábra: Elágazó forgatókönyv tananyag egyik döntési pontjának szerkesztői felülete

A H5P tananyagok egyszerű és gyors újrafelhasználhatósága is ezen az elágazó forgatókönyv típusú tananyagon került bemutatásra. Az elágazó forgatókönyv már haladó H5P felhasználói szintet feltételez, de az elérhető H5P repozitóriumokból, vagy egyes egyetemek tananyag-repozitóriumából könnyen letölthetünk már teljesen kész tananyagot – természetesen abban az esetben, ha az adott tananyag felhasználási joga ezt megengedi. A letöltést általában az adott tananyag alatt megtalálható *Reuse* gombra kattintva tudjuk elindítani. A következő ábra mutatja ezt a lépést.



6. ábra: H5P tananyagelem letöltése

Letöltés után importálhatjuk az adott H5P fájlt a Moodle rendszerünkben az általunk kiválasztott saját e-learning kurzusunk *Tartalombank*jába. Itt tudjuk szerkeszteni az adott tananyagot: egyrészt lefordíthatjuk a számunkra megfelelő nyelvre az adott elemeket, tartalmakat, másrészt a tananyag egyes elemeit törölhetjük, újakat is hozzáadhatunk, vagy a sorrenden is változtathatunk. A *Tartalombank*ba való betöltés után tulajdonképpen ugyanúgy szerkeszthető az adott H5P tananyag, mintha saját magunk hoztuk volna létre azt. Az elhangzott előadásban rövid videó részletek is láthatók erről a folyamatról, amely megtekinthető az előadásról készült felvételen. A következő ábra egy képernyőképet mutat be ebből a folyamatból.



7. ábra: Importált H5P tananyagelem fordítási folyamata

Az előadásban bemutatásra kerültek még a következő H5P tananyagmodulok:

- interaktív kép (Image Hotspot)
- tanulókártyák (Dialog Cards)
- képek párosítása (Image Pair)
- memóriajáték (Memory Game)

Az adott H5P modulokban a Debreceni Egyetem oktatói által létrehozott és aktívan használt tananyagelemek kerültek bemutatásra: rövid videó részletek segítségével az adott modul felhasználása, másrészt az adott modulok szerkesztői felületei. Ezek részletes bemutatása túlmutat jelen publikáció keretein.

Összegzés

Az előadás és a publikáció legfontosabb célja, hogy minél szélesebb körben váljon ismerté a H5P tartalomfejlesztő és minél többen alkalmazzanak H5P tananyagelemeket e-learning kurzusaikban, hiszen a H5P tartalomfejlesztő keretrendszerrel speciális informatikai tudás nélkül is egyszerűen és viszonylag gyorsan hozható létre multimédiás elemeket is tartalmazó interaktív és adaptív tananyag. Mivel kollégáimmal együtt mindennapi munkánk szerves részének tartjuk az előremutató és hatékony tananyagtípusok terjedésének elősegítését, emiatt a tavaly elindított H5P workshopok mellett elkezdtük egy olyan tananyag létrehozását is egyetemünk oktatói számára, amely a H5P tartalomfejlesztő eszköz használatát mutatja be, illetve jó gyakorlatokat, követendő példákat is tartalmaz.

IV.02.03
DOI: 10.31915/NWS.2023.18

Komplex kutatástámogató szolgáltatási portfólió az SZTE Klebelsberg Könyvtárban

Complex research support service portfolio at the SZTE Klebelsberg Library

Fülöp Tiffany*
tiffany.fulop@ek.szte.hu
ORCID: 0000-0002-2219-8455

Molnár Tamás*
tamas.molnar@ek.szte.hu
ORCID: 0000-0003-2571-3595

Hoczopán Szabolcs*
szabolcs.hoczopan@ek.szte.hu
ORCID: 0000-0002-7892-9974
* *SZTE Klebelsberg Könyvtár*

Absztrakt

Az elmúlt hét évben az SZTE Klebelsberg Könyvtárban egyre szélesebb körben igyekszünk felmérni azokat a kutatói, szerzői igényeket, melyeket könyvtárunk anyagi és szakmai lehetőségein belül ki tudunk szolgálni. Ezekre az igényekre épül a "Szerzői Eszköztár" nevű kutatástámogató portfóliónk, melynek keretében egyes szerzőknek és szerkesztőségeknek is nyújtunk szolgáltatásokat, melyek jelenlegi köre évek során alakult ki és jelenleg is bővül. A szerzőknek szánt szolgáltatási csokrunk végigköveti a kézirat életútját. Segítséget nyújtunk a megfelelő folyóirat kiválasztásában, szerzői jogi ügyintézésben, előzetes lektorálásban, beküldés előtti hasonlóság ellenőrzésben, kutatási adatkezelésben, intézményi Open Access publikálási támogatásban, valamint publikálási tréningek folyamatos szervezésével és közvetítésével segítjük a fiatal szerzőket.

kulcsszavak: Open Access, Lektorálás, hasonlóság ellenőrzés, szerzői jog, folyóirat ajánló

Abstract

In the past seven years, the SZTE Klebelsberg Library has been trying to assess the needs of researchers and authors in an increasingly broader range of fields, which we can serve within the financial and professional possibilities of our library. These needs are the basis of our research support portfolio, the "Author Toolkit", which also provides services to individual authors and editorial offices.

Our suite of services for authors follows the lifecycle of a manuscript: journal finder, copyright support, proofreading, similarity check, research data management, institutional Open Access support, publication trainings.

Keywords: Open Access, Editorial Review, Similarity Check, copyright, journal finder

Az SZTE Klebelsberg Könyvtár kutatástámogató szolgáltatási portfóliója évek alatt alakult ki, reflektálva az egyetem nemzetközi folyóiratokban publikáló szerzőinek felmerülő igényeire. Egy olyan rétegről beszélünk, akik meglehetősen ritkán járnak be a könyvtárba, az online szolgáltatásokat, adatbázisokat használják, így klasszikus könyvtári körülmények között nem találkozhattunk a szükségleteikkel, melyekre a képzett kollégáink és lehetőségeink révén sok esetben megoldást tudunk kínálni.

Az áttörést a korábban részletesen ismertetett, Open Access publikálás körüli ügyintézés hozta meg[1]. Kollégáinkkal ugyanis kínosan ügyelünk arra, hogy a publikációs költségek rendezése véletlenül se mehessen tévútra, vagy vezessen felesleges számlázáshoz a Read and Publish és membership szerződések esetén (a szerzői feleslegesen kikért és elfelejtett APC számlák nem ritkán behajtási eljárásba torkollanak). Ennek keretén belül a levelező szerzőkkel megbeszéljük, minden az egyetem által támogatott Open Access kézirat beküldésének pontos, kiadó specifikus útját és technikai feltételeit[2]. Amennyiben a publikációs kvóta merül ki valamelyik szerződésünk esetében[3], az előzetes pályázatnak és konzultációnak köszönhetően fel tudjuk hívni a szerzők figyelmét, hogy a szerződés jelen pillanatban nem tudja fedezni a publikációs költségeit, lehetőleg ne küldje be a kéziratot az új szerződés megkötéséig, várjon egy hónapot, mire újra élni fog a szerződésünk.

A kéziratok ilyen jellegű elő- és utógondozása során alkalmunk volt kapcsolatot kiépíteni a könyvtárba nem járó kutatókkal és felmérni a publikálással kapcsolatos igényeiket.

Tovább bővítette az információ forrásaink és kapcsolataink körét, amikor az Egyetem Beszerzési Igazgatóságától sikerült megszerezzük az egyetemi egységek Open Access publikálási számláinak ügyintézését. A klasszikus beszerzéseket végző egyetemi kollégáknak gondot okozott az APC[4] számlák nem éppen klasszikus életútja. Nem kapcsolódik hozzá megrendelés, szerződés, három árajánlat stb. A könyvtár kollégái addigra már mind a "klasszikus" beszerzésekben, mind az OA számlákkal komoly tapasztalata volt, nem jelentett problémát a feladat ellátása. Kialakított munkamenetünk a korábbi sokszorosára gyorsította az ügyintézés sebességét, ami nem elhanyagolható tényező a számla kifizetésétől függően megjelenő publikációk esetében[5]. Járulékos haszna az új szolgáltatásnak az volt, hogy még több szerzővel teremthettünk közvetlen kapcsolatot és még több publikálási információt gyűjthettünk, melyeket kamatoztatni tudtunk a szolgáltatásaink fejlesztése során.

Az eligazító beszélgetések során derült ki számunkra, hogy szerzőinknek nagy szüksége lenne kéziratuk előzetes hasonlóság ellenőrzésére, hogy a kellemetlen meglepetések ne a szerkesztőség részéről érje őket. A CrossRef ügynökséggel kötött DOI keretszerződésünk tartalmazta a kiadók által is használt iThenticate hasonlóság ellenőrző eszköz használatát is, amire addig nem is fordítottunk figyelmet. A rendszer minimális éves díjért és keresésenként

---

1 Muzs Krisztina ; Molnár Tamás ; Hoczopán Szabolcs: Open Access pályázati rendszer technikai megvalósítása és a szerzők támogatása a Szegedi Tudományegyetemen In: Tick, József; Kokas, Károly; Holl, András (szerk.) NETWORKSHOP 2019 konferenciakötet Budapest, Magyarország : Hungarnet 197 p. pp. 114-120. https://doi.org/10.31915/NWS.2019.15

2 egyetemi emailcím használata, egységes affiliáció, feltöltési kód, kiadói rendszerben történő

3 Gaálné, Kalydy Dóra: A kiadókkal kötött Read and Publish szerződések, és a nyílt hozzáférésű publikálás hazai lehetőségei. In: Gaálné, Kalydy Dóra (szerk.) Open Science : Nyílt tudomány magyar szemmel Budapest, Magyarország : Magyar Tudományos Akadémia Könyvtár és Információs Központ (2021) 58. p https://doi.org/10.36820/MTAKIK.KOZL.2021.OpenS.3

4 Article Processing Charge

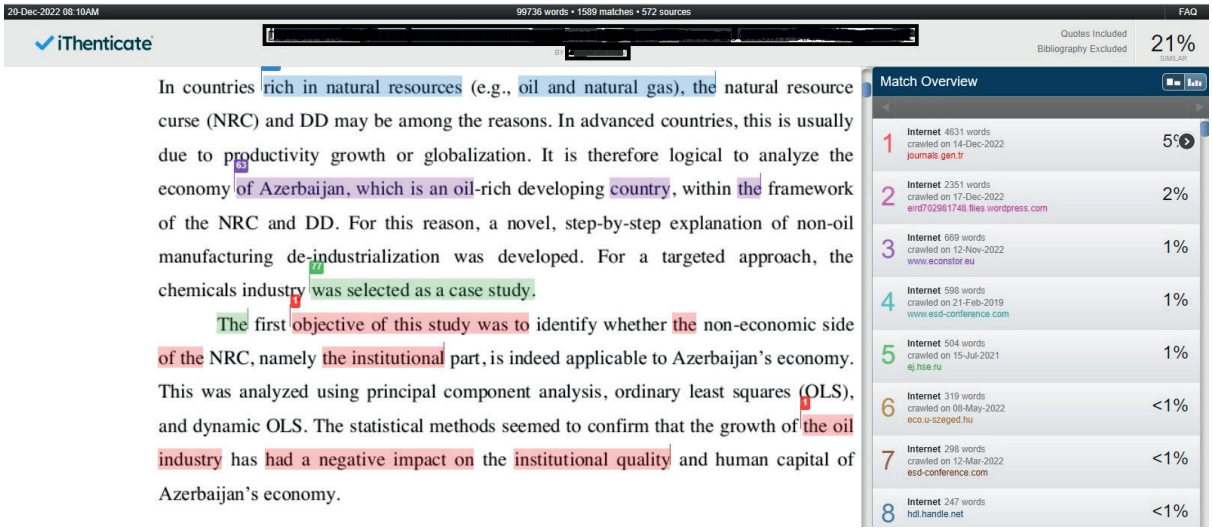
5 URL: http://szerzoknek.ek.szte.hu/tanszeki-open-access-szamla-ugyintezes/

CC BY

fix, szintén minimális összegért volt használatba vehető. Ráadásul a konstrukciónak köszönhetően a költségek jól követhetőek és tervezhetőek voltak.

Mivel az előzetes igényfelmérést tulajdonképpen informálisan elvégeztük, nem volt meglepetés, hogy a szolgáltatás azonnal nagyon népszerű lett és tömegével küldték a szerzők a véglegesnek szánt kéziratait ellenőrzésre. Mivel a rendszerhez biztonsági és anyagi okokból nem akartunk a fél egyetemnek hozzáférést adni, a szerzők egy százalékos összesítőt, grafikus és szöveg alapú exportált eredményt kapnak kézhez. A rendszer beállítási lehetőségeit maximálisan kihasználva az átfedési eredmények közül kizártunk minden olyan elemet, mely torzíthatta volna az eredményt (pl. bibliográfia). Ha a szerző nem kapott megnyugtató eredményt (és itt soha nem valós plágiumról beszélünk, csak összejönnek az egymástól független hasonlóságok) lehetőséget biztosítunk számukra, hogy finomítsanak a kéziratukon, majd a javított kéziratukon újra lefuttatjuk az elemzést.

Gyakran kérdezik a kollégáinkat, hogy a kapott százalékos eredmény soknak, vagy kevésnek számít a szerkesztőségben? Mindig hangsúlyozzuk, hogy az eredmények tételes vizsgálatától nem lehet eltekinteni, a százalékos összesítő csak egy szám. Ennek ellenére a 30%-ot meghaladó hasonlóság esetén már óvatosságra intjük őket. Különösen sok problémát okoz fiatal szerzőinknek a DOI-val ellátott saját doktori dolgozatuk visszaköszönése a kézirat plágiumellenőrzésében[6].



1. ábra: Hasonlóság ellenőrzés eredményei

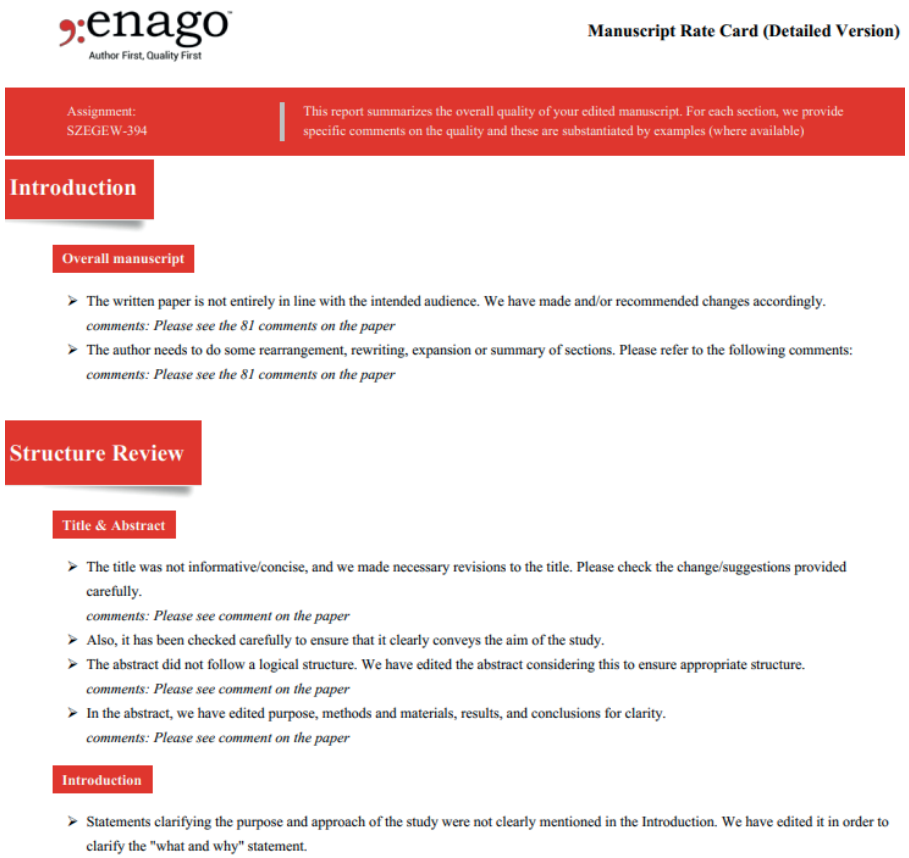
A szerzőkkel történt folyamatos egyeztetések és a hozzánk tévedt számlák alapján egyértelmű volt, hogy szükség van egy a kéziratokat előzetesen ellenőrző, lektoráló szolgáltatásra. Tájékozódva a szolgáltatói lehetőség között és ajánlatokat bekérve, hamar arra jutottunk, hogy központi ügyintézéssel és források bevonásával sokkal gyorsabban és gazdaságosabban, több szerző számára lehetséges a lektorálás biztosítása, mint ha mindenki önállóan intézkedik. Azonban az is világossá vált nagyon hamar, hogy így is nagyon meg kell rostálni a jelentkezőket, mert a szolgáltatásra masszív túljelentkezés volt, ami finanszírozhatatlan lett volna. Alapvetően a Web of Science - Scopus indexált folyóiratokban megjelentetni szándékozott kéziratok ellenőrzését vállaljuk, hiszen például egy könyv lektorálása akár két éves pénzügyi keretünket is felemésztené[7]. A szolgáltatás keretében nyelvi és nyelvtani

6 URL: http://szerzoknek.ek.szte.hu/plagiumkereso/
7 URL: http://szerzoknek.ek.szte.hu/lektoralas/

ellenőrzést, stílust és következetességet, technikai megfelelést, formázást, logikai és tartalmi ellenőrzést, valamint stilisztikai javítást végeznek.

Tekintve, hogy a szolgáltatás minőségének megítélése már messze túl van a lehetőségeinken, ezért a szolgáltatást igénybe vevő kutatókat kérjük meg, hogy minősítsék az aktuális ellenőrzést és így járuljanak hozzá a szolgáltatás jobbá tételéhez. Mivel a megbízott cég is külső lektorokkal dolgozik, az óvatosság nem fölösleges. Bár többnyire a szerzők meg vannak elégedve a minőséggel, extrém esetben arról kaptunk visszajelzéseket, hogy az átdolgozás kifejezetten rontott a kézirat minőségén. Természetesen alkalmi megoldásként rögtön ingyenes újralektorálást kértünk a szerző számára. Ellenőrizve az érintett kéziratokat, hamar kiderült, hogy mind egy témakörbe tartozott és ugyanaz a lektor dolgozott számunkra. Az egyértelmű szerzői visszajelzéseknek köszönhetően az anonim lektort is képesek voltunk azonosítani és jelezni az ügynökségnek, hogy vele a továbbiakban nem akarunk együtt dolgozni.

A szolgáltatás kihasználtsága, ahogy az várható is volt, folyamatosan maximumon, valójában azon felül van. Csak nagyon komoly tervezéssel és logisztikával lehet az igényeket kielégíteni és minden igyekezetünk ellenére a szolgáltatás időről időre kényszerűen szünetel.



enago
Author First, Quality First

Manuscript Rate Card (Detailed Version)

Assignment: SZEGEW-394

This report summarizes the overall quality of your edited manuscript. For each section, we provide specific comments on the quality and these are substantiated by examples (where available)

Introduction

Overall manuscript

- The written paper is not entirely in line with the intended audience. We have made and/or recommended changes accordingly.
  comments: Please see the 81 comments on the paper
- The author needs to do some rearrangement, rewriting, expansion or summary of sections. Please refer to the following comments:
  comments: Please see the 81 comments on the paper

Structure Review

Title & Abstract

- The title was not informative/concise, and we made necessary revisions to the title. Please check the change/suggestions provided carefully.
  comments: Please see comment on the paper
- Also, it has been checked carefully to ensure that it clearly conveys the aim of the study.
- The abstract did not follow a logical structure. We have edited the abstract considering this to ensure appropriate structure.
  comments: Please see comment on the paper
- In the abstract, we have edited purpose, methods and materials, results, and conclusions for clarity.
  comments: Please see comment on the paper

Introduction

- Statements clarifying the purpose and approach of the study were not clearly mentioned in the Introduction. We have edited it in order to clarify the "what and why" statement.

2. ábra: Kézirat bíráló lap

A legnagyobb igény a lektorálási szolgáltatás keretén belül a nyelvi, nyelvtani ellenőrzésre volna, és piaci viszonylatokban bármilyen ésszerű is a megállapodásunk a lektoráló céggel, költségvonzata miatt az igények töredékét sem tudjuk fedezni. Másik, költséghatékonyabb megoldásra volt szükségünk a SZTE-n születő kéziratok tömeges nyelvi ellenőrzéséhez. Egy M.I. alapú nyelvi alkalmazásra esett a választásunk, mely az előfizetés időtartama alatt korlátlan mennyiségű szöveg javítását képes elvégezni. Az alkalmazás háttér adatbázisát kifejezetten tudományos publikációkkal töltötték fel, ezek szolgálnak mintául az alkalmazásnak.

CC BY

A lemerült OA publikációs kvóták értelemszerűen frusztrációt okoztak az egyetem szerzőinek, különösen ha már kiválasztották a folyóiratot. Ezt a frusztrációt próbáltuk levezeti azzal, hogy a szerzőknek alternatív folyóiratokat ajánlottunk, olyanokat, melyek scope-ja tökéletesen megfelelt az adott kéziratnak és egyéb tudománymetriai jellemzői megfelelőek a szerző számára. Az OA kvóták stabilizálódtak az elmúlt években, de meglepetésünkre a szerzők továbbra is igényelték segítségünket a megfelelő folyóirat kiválasztásában, így 2020-ban szolgáltatássá szerveztük ezt a feladatot is, Folyóirat-ajánló néven[8].

Alapvetően egy kényelmi szolgáltatásról van szó (ezt természetesen szigorúan szakmai alapokon kell érteni, s arra utal, hogy a szerzők maguk is el tudnák végezni a folyóirat keresést, viszont a könyvtár rengeteg időt spórol meg nekik, segítve döntésüket.), mely az SZTE kutatóit segíti abban, hogy a már kész kéziratikat a megfelelő helyen tudják megjelentetni. Ehhez nem kell mást tenniük, mint, hogy küldjék el a kézirat címét, absztraktját, kulcsszavait, valamint a személyes, publikálással kapcsolatos preferenciákat, egy megadott kapcsolattartási e-mail címre.

A kutatók a szolgáltatás kimeneteként az absztrakt alapján figyelembe vehető folyóiratok viszonylag bő listáját kapják meg, címmel, az SJR-ből és a JCR-ból származó kvartilis értékkel, kiadói adatokkal, valamint a nyílt megjelenés egyetem által történő támogatásának megjelölésével, a JCR-ból származó Article influence score értékek szerint sorrendbe téve.

A folyóiratlista maga több folyóiratkereső szolgáltató által eredményként megadott találati listák metszetéből jön létre. A folyóiratkereső szolgáltatások között egyenlő arányban szerepelnek a kiadói webhelyekről származó és az általános célú lekérdezések. Így viszonylag pártatlan és kiegyensúlyozott szolgáltatást tud nyújtani a könyvtár a szerző általi végső döntéshez.

A szolgáltatást évente körülbelül harminc alkalommal veszik igénybe. Jellemzően olyanok, akiket az első vagy második alkalommal visszautasított már folyóirat. Ennek oka változatos lehet, általában arra hivatkoznak a szerkesztők, hogy a cikk nem illik a folyóirat profiljába. Olyanok is keresik e szolgáltatást, akik még a PhD képzésben vesznek részt, s a doktori képzés azon szakaszában vannak, amikor viszonylag gyorsan kell publikálniuk, azonban még nincs teljes rálátásuk az adott tudományterület teljes szakirodalmára.

Folyóirat	SJR	JCR	támogatható OA?	Kiadó	Article influence score
Electrochemical Energy Reviews	Q1	Q1	igen	Springer	5.772
Environmental Science & Technology	Q1	Q1	igen	ACS	1.926
Resources, Conservation and Recycling	Q1	Q1	igen	Elsevier	1.889
Applied Energy	Q1	Q1	igen	Elsevier	1.870
Chemical Engineering Journal	Q1	Q1	igen	Elsevier	1.758
ACS applied materials & interfaces	Q1	Q1	igen	ACS	1.608
Advanced Materials Technologies	Q1	Q1	igen	Wiley	1.586
ChemSusChem	Q1	Q1	igen	Wiley	1.566
Journal of hazardous materials	Q1	Q1	igen	Elsevier	1.515
Science of the Total Environment	Q1	Q1	igen	Elsevier	1.397
Journal of Industrial Ecology	Q1	Q2	igen	Wiley	1.393
Journal of Cleaner Production	Q1	Q1	igen	Elsevier	1.376
Waste Management	Q1	Q1	igen	Elsevier	1.224
Journal of environmental management	Q1	Q1	igen	Elsevier	1.114

3. ábra: Folyóirat ajánló eredményei

Az Open Access ügyintézés során következőnek a szerzői jogi ismeretek területén fedeztünk fel hiányosságokat a szerzőink tudásában. Gyakran a publikáció elfogadását követően a licenc választó formanyomtatvány kitöltése közben hívtak fel minket érdeklődve, hogy melyik

8 URL: http://szerzoknek.ek.szte.hu/folyoirat-ajanlo/

licencet válasszák. Melyik, mire jó? A CC-BY licencekkel kapcsolatban gyakran tartottunk kutatóinknak gyors telefonos képzéseket. Ezt követően megérkeztek az első olyan esetek, amikor a szerkesztőség azért tartotta vissza a kéziratot, mert a szerzők a más publikációkból újrafelhasznált ábrákra, képekre stb. nem kérték ki az újrafelhasználási engedélyeket a kiadóktól. A hagyományos publikálási modellben ez a lépés nem megkerülhető. A könyvtár munkatársai ezen a területen sok tapasztalatot gyűjtöttek korábban, több EFOP tananyagfejlesztési projekt keretében, újrafelhasználási engedélyek százainak megszerzésével[9]. Az ilyen esetek ismét arra sarkalltak minket, hogy az újrafelhasználási engedélyek beszerzését szolgáltatás szinten, professzionálisan végezzük el az SZTE szerzői számára[10]. A cél, hogy kiadói, szerzői jogi ismereteinket kamatoztatva, ingyenesen szerezzük be az engedélyeket.

Sokáig komoly bizonytalanságot okoztak szerzőink körében az úgynevezett rosszindulatú, predátor kiadók. Számos visszajelzés érkezett hozzánk szerzőinktől, jobb esetben csak gyanús megkeresésekről, rosszabb esetben már beküldött kéziratokról, legrosszabb esetben kifizetett számláról. Hála a már addigra Open Access vonalon jól bejáratot kapcsolatrendszerünknek a kutatók természetesnek érezték, hogy hozzánk fordulnak a predátor gyanús ügyekben. Időről időre nyilvánvalóan megkörnyékezték az egyetemet a predátor kiadók, a bejelentésekből még a nevükkel is tisztában voltunk. Mégis okulva Jeffrey Beall példáján, közvetlen riasztást nem mertünk kiadni honlapunkon és kommunikációs csatornáinkon keresztül, miszerint XY predátor kiadó éppen az SZTE szerzőire vadászik.

Ehelyett a hangzatos “predátor folyóirat azonosítás” szolgáltatásunkat hirdettük meg[11] és osztottuk meg minden létező csatornánkon keresztül. A szolgáltatás leírásába megadtunk minden információt, amivel a szerzők maguk is meg tudják vizsgálni a gyanús megkereséseket, de természetesen a szolgálati email címre bejövő megkeresések alapján, mi is ellenőrizzük a kérdéses kiadót és folyóiratot. Bár mostanra, évekkel később már nincs nagy forgalma a szolgáltatásnak, de a célját betöltötte. Elültette az egészséges gyanút a szerzőkben, hogy óvatosan kell kezelni a megkereséseket, nem szabad készpénznek venni a folyóiratoknak magukról tett állításait. Odáig jutottunk kutatóinkkal, hogy sok esetben már a határozott gyanúval keresnek meg minket és valójában tőlünk már csak megerősítést várnak. A predátor vadászat sikertörténetnek tekinthető.

Keressük a helyünket a kutatási adatkezelés világában is, mely számos új kihívást rejteget magában a felsőoktatási könyvtárak számára[12]. A Debreceni Egyetemi Könyvtár mintáját követve egy átfogó kérdőívet állítottunk össze, hogy felmérjük az SZTE kutatóinak elvárásait és igényeit az Open Science és különösképpen a kutatási adatkezelés terén[13]. A kérdőívre meglepően sok válasz érkezett, amit önmagában is jó jelnek számított, hiszen mutatta, hogy a kutatóinkat foglalkoztatja a téma és segítséget várnának a Könyvtártól.

---

9 Hoczopán, Szabolcs: Szerzői jog a gyakorlatban. Tananyag. 2019. http://dtk.tankonyvtar.hu/xmlui/handle/123456789/13213

10 URL: http://szerzoknek.ek.szte.hu/copyright-ugyintezes/

11 URL: http://szerzoknek.ek.szte.hu/predator-folyoirat-azonositas/

12 Lencsés, Ákos: Kutatási adatok könyvtári kezelése. TUDOMÁNYOS ÉS MŰSZAKI TÁJÉKOZTATÁS 68 : 11 pp. 663-670. , 8 p. (2021) https://doi.org/10.3311/tmt.13087

13 Zeller Rozália, Hoczopán Szabolcs, Nagy, Gyula: Kutatási adatkezelést támogató szolgáltatások előkészítése a Szegedi Tudományegyetemen kérdőív és válaszok [Data set]. Zenodo. https://doi.org/10.5281/zenodo.5166625

CC BY

A kérdőív válaszait kiértékeltük[14] és azok alapján három témakörben (melyekben kompetenseknek éreztük magunkat) indítottunk segítségnyújtást RDM vonalon[15]. Ezek az adatrepozitórium ajánló, kutatási adatmenedzsment tanácsadás és az adatkezelési terv konzultáció. Az RDM területén nagyon az utunk elején járunk, sokat kell képezzük még magunkat és kutatóinkat is. A szolgáltatások igénybevétele általában pályázat benyújtás előtti időszakban fut fel, míg az év további részében nem sok megkeresésünk akad.

Terveink szerint szolgáltatásaink köre folyamatosan bővülni fog a jövőben is. Továbbra is követjük az SZTE szerzőinek felmerülő igényeit és mérlegeljük, hogyan tudnánk segítségükre lenni.

Felhasznált irodalom

Muzs Krisztina ; Molnár Tamás ; Hoczopán Szabolcs: Open Access pályázati rendszer technikai megvalósítása és a szerzők támogatása a Szegedi Tudományegyetemen In: Tick, József; Kokas, Károly; Holl, András (szerk.) NETWORKSHOP 2019 konferenciakötet Budapest, Magyarország : Hungarnet 197 p. pp. 114-120. https://doi.org/10.31915/NWS.2019.15

Gaálné, Kalydy Dóra: A kiadókkal kötött Read and Publish szerződések, és a nyílt hozzáférésű publikálás hazai lehetőségei. In: Gaálné, Kalydy Dóra (szerk.) Open Science : Nyílt tudomány magyar szemmel Budapest, Magyarország : Magyar Tudományos Akadémia Könyvtár és Információs Központ (2021) 58. p https://doi.org/10.36820/MTAKIK.KOZL.2021.OpenS.3

Zeller Rozália, Hoczopán Szabolcs, Nagy Gyula: Kutatási adatkezelést támogató szolgáltatások előkészítése a Szegedi Tudományegyetemen. Tudományos Műszaki Tájékoztatás, 68. évf. 7. sz. 68 : 9 pp. 576-586. , 11 p. (2021) https://tmt.omikk.bme.hu/tmt/article/view/13120

Lencsés, Ákos: Kutatási adatok könyvtári kezelése. TUDOMÁNYOS ÉS MŰSZAKI TÁJÉKOZTATÁS 68 : 11 pp. 663-670. , 8 p. (2021) https://doi.org/10.3311/tmt.13087

14 Zeller Rozália, Hoczopán Szabolcs, Nagy Gyula: Kutatási adatkezelést támogató szolgáltatások előkészítése a Szegedi Tudományegyetemen. Tudományos Műszaki Tájékoztatás, 68. évf. 7. sz. 68 : 9 pp. 576-586. , 11 p. (2021) https://tmt.omikk.bme.hu/tmt/article/view/13120

15 URL: http://szerzoknek.ek.szte.hu/szolgaltatasaink

 DOI: 10.31915/NWS.2023.19

Az Open Science könyvtári vonatkozásai

Vass Johanna
Ökológiai Kutatóközpont
vass.johanna@ecolres.hu

Absztrakt

A kutatási folyamatnak csak egy része, mondhatni a nyilvánosságnak szánt végterméke a publikáció. A megelőzően összegyűjtött kutatási adatoknak gyakran a nyilvántartása is elmarad. Az Open Science, vagy *nyílt tudomány mozgalom* ezen a gyakorlaton kíván változtatni. Ugyanakkor kimutatható, hogy a magyarországi kutatóintézetek körében már vannak kialakult gyakorlatok a kutatási adatok repozitóriumokba töltésére. Erre vonatkozóan a jövőben érdemes kiterjedt, szisztematikus felméréseket végezni. A kutatási adatok menedzselése idő- és erőforrásigényes feladat, a könyvtárak számos területen hasznos szerepet játszhatnak a megfelelő adatkezelési gyakorlat kialakításában. A könyvtárosok lehetséges szerepvállalását illetően azonban tudatosítani kell az ezzel kapcsolatos kompetenciák meglétét a kutatói társadalomban, és aktívan keresni kell az együttműködés lehetőségét.

Kulcsszavak: Open Science, nyílt tudomány, könyvtárak, könyvtárosok, gyűjtés, rendszerezés, Fehér Könyv, ajánlások

Abstract

The publication is only one part of the research process, research data collected beforehand are often not kept in the register. The Open Science movement wants to change this practice. At the same time, it can be shown that Hungarian research institutes already have developed practices for uploading research data into repositories. In this regard, it is worth conducting extensive, systematic surveys in the future. Managing research data is a time- and resource-consuming task, and libraries can play a useful role in the development of appropriate data management practices in many areas. However, regarding the possible role of librarians, it is necessary to be aware of the existence of relevant competences in the research community and to actively seek the possibility of cooperation.

Keywords: Open Science, libraries, librarians, collection, systematization, White Book, recommendations

Bevezetés

A kutatási folyamat eredményeit bemutató publikációkat megelőzi az adatfelvétel (pl. természettudományi területen mintavételekkel), adatgyűjtés (pl. néprajzi területen fényképek, interjúk készítése stb.), a kísérletek sora, kísérleti eredmények változatos formában történő rögzítése, dokumentálása stb. Ez az a része a kutatásnak, amely nem hogy nyilvánosságot nem szokott kapni, de sokszor a nyilvántartása is elmarad, sőt a fellelhetősége is bizonytalanná válik egy idő után. [1]

1 Kovács László: Adatrepozitóriumok. Bevezetés [ppt]. ELKH Cloud, 2021. április 22.

CC BY

Mi minősül kutatási adatnak?

A 2022. 10. 28-án benyújtott új törvényjavaslat – *Egyes törvények közadatokkal összefüggő módosításáról; T/1786.* – szerint: Digitális formátumú, nyilvánosságra hozott, közfinanszírozott tudományos tevékenység keretében keletkezik. Lényegében minden, ami a számítógépen található és nem kereskedelmi program: mérési adatok, laborjegyzőkönyvek, számításokhoz fejlesztett programok (saját), kapcsolattartás más kutatókkal stb.[2]

Az Openscience.hu oldalon ennél árnyaltabb meghatározás olvasható. A kutatási adatok a tudományos közösség által létrehozott, rögzített, elfogadott és megőrzött tényadatok, amelyek a kutatási eredmények hitelességét támasztják alá. Létrejöhetnek megfigyelések, kísérletek, szimulációk eredményeképpen – egy konkrét kutatás számára előállítva – vagy korábban gyűjtött adatok összegyűjtésével, válogatásával, feldolgozásával.[3]

A problémát a „digitális objektum" meghatározás jelenti. Kérdés, hogy ez a meghatározás önmagában megállja-e a helyét, hiszen még az elmúlt évtizedekben (mondjunk elmúlt fél évszázadot) keletkeztek kéziratos, cédulákon, gépelve, analóg hangrögzítéssel, rajzolt térképeken ábrázolt stb. kutatási adatok. Ha a „digitális" egy alapvető kritérium a meghatározásban, akkor az analóg módon született kutatási adatok digitalizálására plusz erőforrásokat kell kalkulálni.

Az Eötvös Loránd Kutatási Hálózat Adatrepozitórium Projekt harmadik munkacsoportja 2022-ben felmérést készített a kutatási hálózatban keletkezett adatokról.[4] Ebből az összeállításból kiderül, hogy nemcsak tartalmilag, műfajilag, hanem a fájlformátumokat, illetve a méretet tekintve is rendkívül változatos állományok kezelését kell megtervezni.

Az Open Science és a könyvtárak

A szakkönyvtárak és az egyetemi könyvtárak tevékenységi körében egyre hangsúlyosabb a kutatástámogatás, illetve az e-science körébe tartozó tevékenységek – így az elektronikus tartalmak szolgáltatásának biztosítása; az Open Access publikálással kapcsolatos ügyintézés; tájékoztatás; valamint a tudományos kibocsátás nyilvántartása és mérése – határozzák meg a napi gyakorlatot.[5]

A hagyományos könyvtári szolgáltatások – a maguk tereivel, a fizikai állományaikkal, hagyományos szolgáltatásaikkal, mint nyitvatartás, reprográfia stb. – szignifikánsan veszítettek korábbi jelentőségükből, vagy háttérbe szorultak. Ezek a tevékenységek évek óta csökkenő tendenciát mutatnak, vagy éppen nem jellemzőek egy-egy kutatóhelyen. Kitörési pontokként olyan interfész szakemberekre van szükség, akik mind a két oldallal tudnak kommunikálni – jelen esetben a kutatókkal és az informatikusokkal. Megjegyzendő azonban, hogy az e-science szolgáltatások, illetve a nyílt hozzáféréssel kapcsolatos adatkezelési tevékenységek még nem feltétlenül kapcsolódnak akár a könyvtár fogalmához, akár a könyvtáros személyéhez.

---

2 A törvényjavaslat elérhető az alábbi linken: https://www.parlament.hu/irom42/01786/01786.pdf
Az itt idézett formában olvasható összefoglalás forrása: Szilágyi Edit: Az ELKH Adatrepozitórium projekt – motiváció [előadás és ppt]

3 Kutatási adatkezelés. Elérés: https://openscience.hu/kutatasi-adatok/ [2023. június 20.]

4 Meiszterics Enikő (TK): Kutatási adatok és adatkezelési gyakorlatok az ELKH-ban

5 Karácsony Gyöngyi: A könyvtárak szerepe a publikációs folyamatban

A közelmúltban megjelent egyik átfogó, a nyílt tudomány hazai helyzetével foglalkozó tanulmánykötetben[6] Holl András mintegy tizenhárom olyan területet sorol fel, amelyekben a könyvtárak hasznos szerepet játszhatnak az adatkezelési gyakorlat kialakításában, kezdve a szakirodalom feltárásától az adatkezelési szabályok megalkotásán keresztül a munkafolyamatok kialakításáig a kutatási adatok kezelésében.[7] „A könyvtárak memória-intézmények, alapfeladatuk, hogy információkat kezeljenek, tegyenek hozzáférhetővé hosszú távon […], gyakorta a könyvtár az egyetlen megőrző szervezeti egység. […] A könyvtáros-kultúra része a kutatók kiszolgálása, a könyvtárosoknak nagy gyakorlata van a metaadatok terén […]". [8]

A könyvtárosok számos olyan készséggel rendelkeznek, amelyeket a hagyományos dokumentumok, állományok kezelése során sajátítottak el, de amelyek az új (adatintenzív) környezetben is jól hasznosíthatók: megtalálni, összegyűjteni, rendszerezni, megtisztítani, szervezni, elemezni, prezentálni, nem utolsó sorban pedig feldolgozni és metaadatolni.[9] „Az egyik ok a szórvány adatok könyvtári kezelésére az, hogy a tárolt információk előhozásának fontos alapelve, hogy az információt valamiféle hálózatban, kontextusban kell elhelyezni. Ilyen hálózat (valójában többszörös hálózat) a bibliográfiai háló. A könyvtárak és könyvtári jellegű szervezetek jól kezelik ezeket a hálózatokat",[10] éppen ezért lehet hatékony a részvételük a kutatási adatok rendszerezésében, nyilvántartásában.

Amennyiben a könyvtárosok valóban be akarnak lépni erre a területre, hasznos érv lehet a kutatók felé, hogy a kutatási adatok feltöltése, metaadatolása, gondozása időigényes munka, elveszi a kutató idejét a publikálástól. A jelenlegi teljesítményértékelési rendszerben a kutatói értékelések középpontjában a publikációs tevékenység áll, az adatgondozás nem hoz annyi pontot a kutatónak, nem segíti az előmenetelét. Az MTMT-ben lehetőség van ugyan a „kutatási adat" felvételére, de ezek beszámítása a publikációs tevékenységbe nem rendezett (pl. mennyi pontot ér, melyik évre számolható el egy több évig tartó projekt esetén, a hivatkozásokat hogyan kell kezelni stb.).[11] Megfelelő kooperáció esetén a könyvtáros itt léphet be a folyamatba.

Mindehhez természetesen azt is hozzá kell tenni, hogy az önmagában nem elég, ha mi magunk igényt formálunk erre a szerepre és feladatra, a kutatók számára ma még nem egyértelmű, hogy a könyvtáros ezekben a folyamatokban partner lehet. Hasznos lenne a jövőben felmérni, vajon hány kutatóhelyen tervezik bevonni a könyvtárosokat a kutatási adatok kezelésével kapcsolatos feladatok ellátásába.

A nyílt hozzáférés kultúrája és gyakorlata

Elterjedt vélekedés, hogy az Open Science mozgalom egyik feladata a nyílt hozzáférés kultúrájának elterjesztése a kutatók körében. A következőkben egy „mikrovizsgálat" megfigyeléseivel szeretném árnyalni ezt a képet.

---

6 Open Science. Nyílt tudomány magyar szemmel. Szerk. Gaálné Kalydy Dóra

7 Holl András: A tudományos szakkönyvtárak és a nyílt tudomány (Open Science)

8 Holl András i.m. p. 21.

9 Koltay Tibor: Új könyvtári feladatok az adatintenzív kutatás korában

10 Holl András i.m. p. 22.

11 Lencsés Ákos: Bevezetés a nyílt tudományba. [ppt]

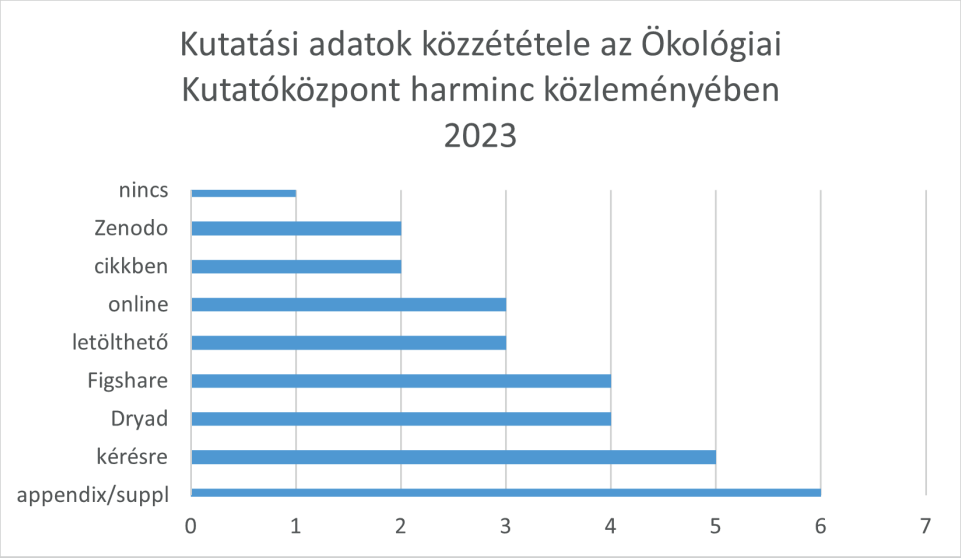
CC BY

Az Ökológiai Kutatóközpont 2023. évi publikációi közül találomra kiválasztottam harminc közleményt[12], és megvizsgáltam, milyen képet mutatnak a kutatási adatok közzétételét illetően.

A publikációk belső szerkezeti tagolásakor „Data availability", „Supplementary material", „Open Research" stb. fejezetcímek alatt találhatók a kutatási adatokra vonatkozó részletek. A véletlenszerűen választott harminc cikk között mindössze egyetlenegy volt, amelynél nem találtam kutatási adatot, ezt egyébként a publikáció szakirodalmat áttekintő jellege is indokolhatja. A fennmaradó huszonkilenc közleményben azonban változatos megoldások találhatók a kutatási adatok közzétételére.

Tíz esetben töltötték fel – vagy jelezték, hogy a cikk elfogadása után fel fogják tölteni – a kutatási adatokat repozitóriumba. A mintaanyagban a Dryad és a Figshare szolgáltatások fele-fele arányban (4-4 közlemény) voltak reprezentálva, míg a Zenodot két publikáció szerzői választották.

Hat közleménynél a kutatási adatok Supplement vagy Appendix fejezetcím alatt érhetők el, három esetben pedig a publikációt követően valamely (többnyire a folyóirat szerverére mutató) link vezet el a kutatási adatokhoz. Három publikációnál találtam letölthető (.doc vagy .xls) állományt, két esetben pedig magában a cikkben helyezték el kattintható, nagyítható, letölthető módon az ábrákat, táblázatokat. (Ezek a megoldások azonban nyilván inkább tükrözik az adott szerkesztőségek gyakorlatát, mintsem a kutatók elhatározását.)
Öt közleményben a szerzők azt a megoldást választották, hogy megkeresés után, csak az azt kérőknek bocsátják rendelkezésre a kutatási adatokat. A kéréshez kötés egyébként megjelent más típusok mellett is, két olyan cikknél, ahol Appendix is volt, valamint egy esetben a kutatási adatokhoz megadott linken túl is volt ilyen jelzés.



1. ábra Kutatási adatok közzététele az Ökológiai Kutatóközpont harminc közleményében
Forrás: saját szerkesztés

Ez a rövid vizsgálat arra enged következtetni, hogy a kutatási adatok nyílt hozzáférésű kezelése már része a hazai kutatói gyakorlatnak. A pályázatoknak egyre inkább kötelező eleme az adatkezelési terv elkészítése, ahol nyilatkozni kell az adatok tervezett repozitóriumba való feltöltéséről is. Nem a nulláról kiindulva kell tehát megteremteni ezt a kultúrát, azonban

12 A harminc közlemény bibliográfiáját önálló függelékként mellékelem.

gyakorlati téren kijelölhetünk néhány olyan területet, amelynek mentén az adatkultúra fejlesztését a közeljövőben érdemes elmélyíteni.

Célkitűzések

Említettük, hogy a könyvtárosok a gyűjtés és rendszerezés, továbbá a metaadatolás terén bírnak elmélyült tapasztalatokkal. Célszerű lenne **összegyűjteni** a jelenleg támogatott kutatócsoportok projektjeit, az azokhoz kapcsolódó publikációkat, valamint kutatási adatokat, és ezeket különböző szempontok szerint tipizálni (repozitóriumba kerültek-e; van-e azonosítójuk; vannak-e járulékos információk: pl. szoftver, verzió stb.) Az ELKH támogatásával aktuálisan 116 kutatócsoport működik; ha csak az ezekhez kapcsolódó kutatási adatok archiválását tűzzük ki célul, máris hatalmas feladat körvonalazódik.

Meg kell fogalmazni egy célkitűzést a teljességet illetően is: például a létrejövő hazai adatrepozitórium (ARP) tartalmazza-e a korábban a külföldi repozitóriumokba feltöltött adatokat? Van-e olyan szándék, amely retrospektíve is összegyűjtené a kutatási adatokat, vagy csak a jelenleg futó projektekkel foglalkozzunk? Továbbá nem lehet elég korán összegyűjteni és rendszerezni az eddigi információkat, szakirodalmat, ezekből szakbibliográfiát, vagy ajánló bibliográfiát készíteni.

A fentebb bemutatott vizsgálatot a kutatási adatok kezelésének gyakorlatára érdemes teljesebb adatsoron is elvégezni, előbb egy-egy kutatóközpont vonatkozásában, majd esetleg a humántudományi és természettudományi területek összehasonlításával is. Az adatok **elemzésével** teljesebb képet nyerhetünk a nyílt adatkezelés elterjedtségéről, gyakorlatáról, és arról is, hogy mely területeken, és milyen eszközökkel érdemes erősíteni a kialakult vagy kialakulóban lévő adatműveltséget.

Az adatműveltség kialakításának részeként célul lehet tűzni a hazai kutatótársadalom részére **ajánlások készítését** az adatrepozitóriumok tervezéséhez (ún. „Fehér könyvet", ha tetszik), például a DataScite Metadata Schema mintájára. Ennek keretébe illesztve kerülhet sor – akár szakterületenként – a metaadat-ajánlások kidolgozására is.

A fentebb megfogalmazott célkitűzések megvalósítása időigényes feladat, időben fel kell mérni, hogy az adatkezelés folyamata és erőforrás igénye tervezhető legyen.

Összefoglalás

Vajon „a könyvtárosok valós igényeket elégítenek-e ki azzal, hogy kutatási adatokkal kapcsolatos szolgáltatásokat nyújtanak, vagy csak új szerepeket keresnek maguknak, mivel csökkent a gyűjteményük fontossága?"[13] A kérdésbe rejtett állítás mind a két fele igaz lehet, azzal a kiegészítéssel, hogy a könyvtárosoknak új szerepeket *kell* keresniük maguknak, miután azok a tevékenységi formák, amelyekben korábban segítették a felhasználóikat – jelen esetben a kutatókat – már kevésbé kihasználtak. Szakmai tapasztalataikat azonban más, rokon területen és a jelenkor kihívásainak megfelelően ugyanúgy a kutatók szolgálatába tudják állítani.

A Koltay Tibor által feltett kérdést viszont érdemes kiegészíteni azzal, hogy vajon a fenti feladatvállalásban érintettek, vagyis a kutatóintézetek menedzsmentje, vagy maguk a kutatók tisztában vannak-e azzal, hogy a könyvtáros partner lehet az adatkezelés folyamatában?

13 Koltay Tibor: Új könyvtári feladatok az adatintenzív kutatás korában

CC BY

A könyvtárosi szerepvállalás lehetőségének felismerését mindenképpen erősíteni kell, mert a feladat az erőforrások bővítését kívánja, a könyvtárosok pedig képzettségüknél és tapasztalataiknál fogva megfelelő társak lehetnek a nyílt adatkezelés gyakorlatának elterjesztésében.

Irodalomjegyzék

ELKH Adatrepozitórium fejlesztése | TK Kutatási Dokumentációs Központ. Forrás: https://kdk.tk.hu/hirek/2022/01/elkh-adatrepozitorium-fejlesztese [2022. november 15.]

ELKH Adatrepozitórium Platform (science-research-data.hu). Forrás: https://science-research-data.hu/ [2022. november 15.]

Holl András: A tudományos szakkönyvtárak és a nyílt tudomány (Open Science). In: Open Science: Nyílt tudomány magyar szemmel. Szerk. Gaálné Kalydy Dóra. MTA KIK, 2021. p. 11-52.

Karácsony Gyöngyi: A könyvtárak szerepe a publikációs folyamatban. In: Könyvtári Figyelő, 2008, 1. pp. 10. Forrás: http://epa.oszk.hu/00100/00143/00066/pdf/EPA00143_konyvtari_figyelo_2008_1_009-021.pdf [2022. július 13.]

Koltay Tibor: Új könyvtári feladatok az adatintenzív kutatás korában = Könyvtári Figyelő 2019/2. p. 211-217. Forrás: http://epa.oszk.hu/00100/00143/00356/pdf/EPA00143_konyvtari_figyelo_2019_02_211-217.pdf [2022. november 15.]

Kovács László: Adatrepozitóriumok. Bevezetés [ppt]. ELKH Cloud, 2021. április 22. Hozzáférés: https://science-cloud.hu/eloadasok/adatrepozitoriumok-bevezeto [2022. november 15.]

Kutatási adatok kezelése – MTA Open Access (mtak.hu). Forrás: https://openaccess.mtak.hu/kutatasi-adatok-kezelese/ [2022. november 15.]

Kutatásiadat-archiválási pilot-projektek bemutatása | Open Science. Forrás: https://openscience.hu/events/kutatasiadat-archivalasi-pilot-projektek-bemutatasa/ [2022. november 15.]

Lencsés Ákos: Bevezetés a nyílt tudományba. [ppt] Nyílt Tudomány Workshopok, 2022. november 15.

Meiszterics Enikő (TK): Kutatási adatok és adatkezelési gyakorlatok az ELKH-ban Elérés: https://science-research-data.hu/eloadasok/kutatasi-adatok-magyarorszagon-hozzaferes-gondozas-megosztas/kutatasi-adatok-es [2023. június 20.]

Open Science: Nyílt tudomány magyar szemmel. Szerk. Gaálné Kalydy Dóra. MTA KIK, 2021

Szilágyi Edit: Az ELKH Adatrepozitórium projekt – motiváció [előadás és ppt]. Elérés: https://science-research-data.hu/eloadasok/kutatasi-adatok-magyarorszagon-hozzaferes-gondozas-megosztas/az-elkh-adatrepozitorium-0 [2023. június 20.]

 DOI: 10.31915/NWS.2023.20

A digitális oktatás módszertana a gyakorlatban

Digital Education Methodology in Practice

Antal Péter
Eszterházy Károly Katolikus Egyetem, Digitális Kultúra Tanszék
antal.peter@uni-eszterhazy.hu

Czeglédi László
Eszterházy Károly Katolikus Egyetem, Humáninformatika Tanszék
czegledi.laszlo@uni-eszterhazy.hu

Absztrakt

A digitális eszközök és a digitális környezet aktív, kreatív pedagógiai alkalmazása mára a pedagógus pálya szakmai rugalmasságának egyik fontos fokmérőjévé vált. A digitális oktatás módszertani megújulása nagymértékben hozzájárul számos pedagógiai siker eléréséhez, azonban nem árt tisztázni, hogy a digitális eszközök segítségével milyen pedagógiai célokat kívánunk és tudunk elérni. Az utóbbi néhány év tapasztalatai, különösen a Covid időszak, megmutatták azokat a személyi, technológiai, módszertani és szemléletbeli hiányosságokat, melyek rámutattak a digitális oktatás anomáliáira. Egyik ilyen valós probléma, hogy a pedagógusok által használt tartalomkezelő és tanuláskövető rendszerek még az egyes iskolai tantestületek esetében sem egységesek, sok esetben hiányzik a tanárok közötti alkotó kommunikáció. A kérdés: valóban tudjuk-e növelni a hatékonyságot, vagyis az alkalmazott technológia és módszertan a lehető legkisebb idő- és energiabefektetéssel képes-e a legnagyobb pedagógiai „hozamot" eredményezni. Ehhez azonban szükség van a reális kép vizsgálatára, a módszerek, az alkalmazott programok, és az infrastruktúra szempontjából.
Előadásunkban a fenti kérdéseket körbejárva, egy a pedagógusok körében végzett vizsgálat eredményeit és következtetéseit mutatom be.
Kulcsszavak: digitális oktatás, módszertan, digitális átállás

Abstract

The active, creative pedagogical use of digital tools and the digital environment has become an important measure of professional flexibility in the teaching profession. The methodological innovation of digital education is a major contributor to many pedagogical successes, but it is important to be clear about the pedagogical goals that digital tools are intended and capable of achieving. The experience of the last few years, especially in the Covid period, has shown the personal, technological, methodological and attitudinal shortcomings that have highlighted the anomalies of digital education. One of these real problems is that the content management and learning tracking systems used by teachers are not uniform even across school staff, and in many cases there is a lack of creative communication between teachers. The question is, whether we can really increase efficiency, i.e. whether the technology and methodology used can deliver the greatest pedagogical 'yield' with the least investment of time and energy. This requires, however, an examination of the real picture in terms of methods, programmes and infrastructure.
In our presentation, we will explore these issues and present the results and conclusions of a survey of teachers.
Keywords: digital education, methodology, digital transition

CC BY

1. Bevezetés

A digitális eszközök és a digitális környezet aktív, kreatív pedagógiai alkalmazása mára a pedagógus pálya szakmai rugalmasságának egyik fontos fokmérőjévé vált. Az oktatás sok területén rengeteg tapasztalat van az IKT alkalmazásával és hasznosságával kapcsolatban. A számítógépes alkalmazások előnyei között a szerzők legtöbbször külön kiemelik gyakorlati tapasztalatokat megerősítve, hogy az IKT eszközök használatának már önmagában is jelentős motiváló ereje lehet. [1]

Mivel az oktatásba kerülő gyerekek, egyre magabiztosabban mozognak a digitális térben, sokszor abba a hibába esünk, hogy azt gondoljuk, az információs és kommunikációs technológiák, pontosabban a digitális eszközök és módszerek rendszere, képes helyettesíteni a pedagógiai tervezést, nem pedig eszközként szolgálni ki azt.

Ruben Puentedura a 2000-es évek végére készítette el a digitális technológiai integráció pedagógiai modelljének egy változatát a SAMR modellt. [6]

Eszerint a technika kétféleképpen jelenhet meg az iskolában, vagy bővíti, vagy átalakítja a tanítás menetét. Mindegyiknek két szintje képzelhető el: a bővítés esetében a helyettesítés és a kiterjesztés, az átalakításnál a módosítás és az újraértelmezés.

Másként fogalmazva, a digitális átállás nyomán létrejövő pedagógiai modell (digitális pedagógia) azt kell, hogy jelentse, hogy a digitális eszközök segítenek abban, hogy képesek legyünk korábban nem, vagy csak sokkal körülményesebben megoldható feladatokat elvégezni, problémákat megoldani, az új tudásforma segítségével. [7]

A bevezetés során azonban fel kell mérni az országos és helyi környezet lehetőségeit (számítógéppelrendelkezőkszáma,internethasználatstb.),akülönbözőfolyamatok–társadalmi és technológiai egyaránt – irányát és jellemzőit, valamint a virtuális tér infrastruktúrájának kialakítására rendelkezésre álló szellemi, technikai és anyagi erőforrásokat. Emellett meg kell vizsgálni az eszközhasználat módszertanának elméleti és gyakorlati kérdéseit nem csak a jó gyakorlatok tekintetében, hanem a helyi viszonyokra való alkalmazhatóság oldaláról is. [4] Ebből fakadóan nagyon eltérő és helyspecifikus eredmények születnek a kutatások során.

A Covid időszak, igazán jó tesztnek bizonyult az IKT használat szempontjából, hiszen rámutatott azokra a személyi, technológiai, módszertani és szemléletbéli problémákra, melyek a digitális oktatást jellemzik napjainkban. Az egyik ilyen valós probléma, hogy a pedagógusok által használt tartalomkezelő és tanuláskövető rendszerek még az egyes iskolai közösségek esetében sem egységesek, sok esetben hiányzik a tanárok közötti alkotó kommunikáció. A másik fontos kérdés az attitűdváltozás minősége, hiszen az eszközhasználati rutin, magabiztosság az alkalmazás aránya, és főleg minősége erősen összefügg.

2. A felmérés célja

A Covid óta eltelt két évben teljesen átértékelődött a digitális alapú oktatás szerepe és megítélése. A kérdőíves kutatásban arra voltunk kíváncsiak, történt-e valamilyen előremozdulás, szemléletváltás a digitális kompetenciák fejlődése terén, növekedett-e, a hatékonyság az alkalmazott technológia és módszertan terén. A másik cél, az egyetemi tanárképzésben folyamatosan törekszünk az IKT képzéseket megújítani, olyan új technológiák, szoftverek bevonásával, melyek segítik a korszerű tanárképzés fenntartását

és a digitális átállás megvalósítását. Ennek érdekében szerettük volna megtudni, melyek a legtöbbet használt alkalmazások, módszerek a napi pedagógiai gyakorlatban.



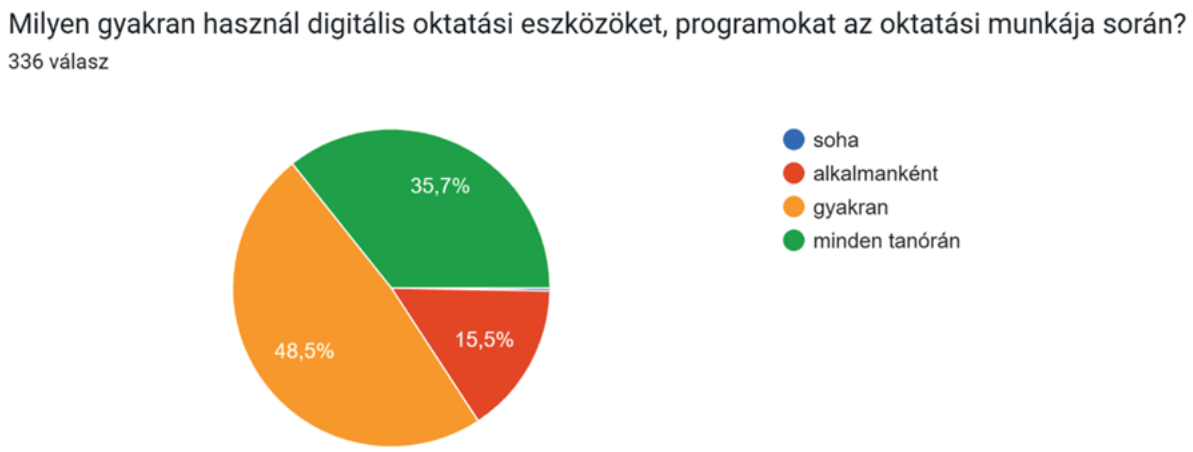
1. ábra. A kutatás célok meghatározása

3. A felmérés részletei

A céloknak megfelelően, egy kvantitatív módszertanú, országos szintű, kérdőíves felmérést végeztünk általános és középiskolai pedagógusi körben, eszközhasználat, attitűd, kommunikáció, alkalmazások használata, és pedagógiai módszertan témakörében. A kitöltők elsősorban dunántúli, és északkelet-magyarországi, zömében (74%) általános iskolában tanító kollégák közül kerültek ki. A kérdőívet összesen 336 válaszoló töltötte ki.

4. Eredmények

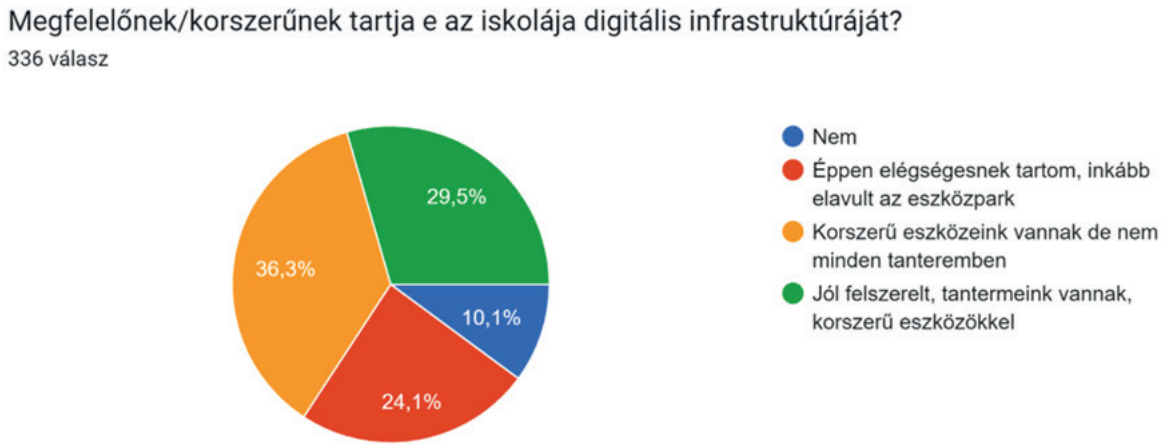
A koreloszlásra vonatkozó kérdés nem okozott meglepetést, hiszen a kitöltők 77%-a volt 45 év feletti, ami ebben az esetben is jól mutatja az aktív pedagógustársadalom elöregedését. Arra kérdésre, hogy milyen gyakran használ digitális eszközöket, biztató válaszok érkeztek a korábbi kutatások eredményeihez képest, hiszen a kitöltők 83%-a gyakran vagy minden tanórán használ digitális oktatási eszközöket.[3]



2. ábra. Digitális eszközhasználat arányai a pedagógusok körében

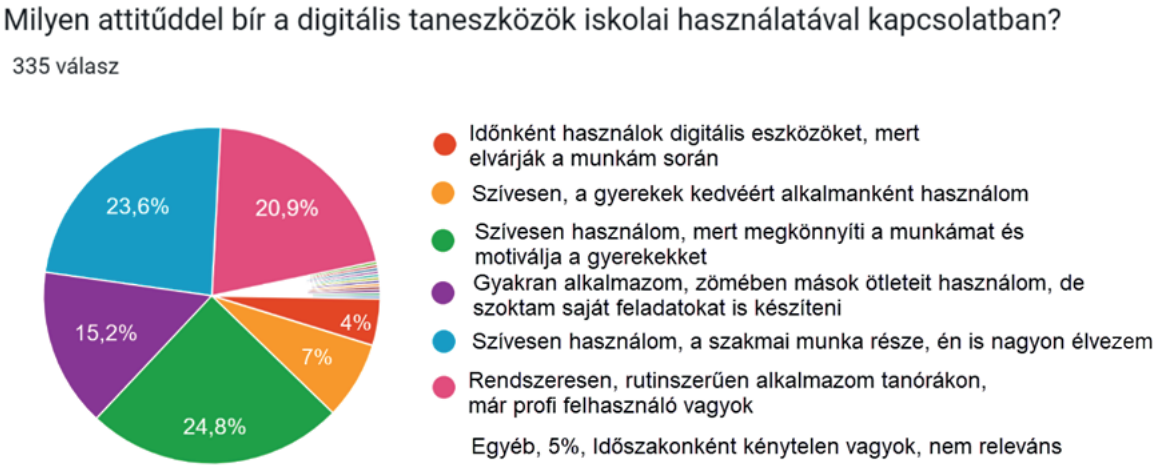
CC BY

A megkérdezettek az iskolai infrastruktúra tekintetében nem voltak maradéktalanul elégedettek, közel 35% korszerűtlennek, illetve éppen elégségesnek tartja iskolája digitális felszereltségét, ami elkeserítőnek mondható az utóbbi évek fejlesztéseinek tükrében. Csak az iskolák 29,5%-a rendelkezik a kollégák szerint, jól felszerelt, korszerű eszközökkel.



3. ábra. Infrastrukturális helyzet

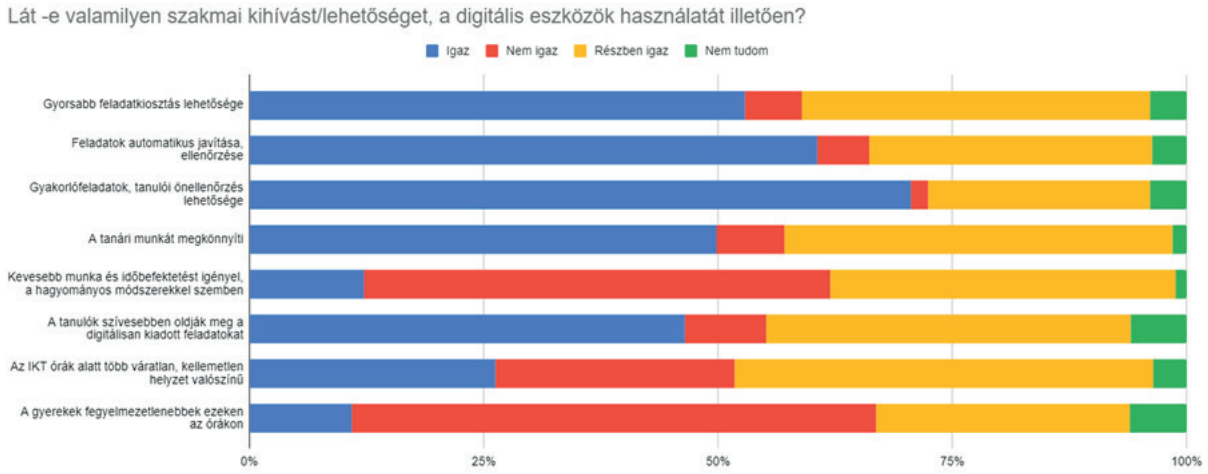
A digitális eszközök használatával kapcsolatos attitűdre vonatkozó kérdésre adott válaszok már nagyobb szórást mutattak, de inkább pozitív irányúak, ami előre mozdulást jelent a korábbi kutatások eredményeihez képest.[5] A rendszeresen, rutinszerűen és szívesen digitális eszközöket használó kollégák aránya 44,5%-os, a kevésbé motiváltak aránya (a gyerekek kedvéért használom) 7%, míg a digitális módszereket elutasítók vagy kényszerből használók aránya kb. 10%.



4. ábra. Az attitűd vizsgálat eredményei

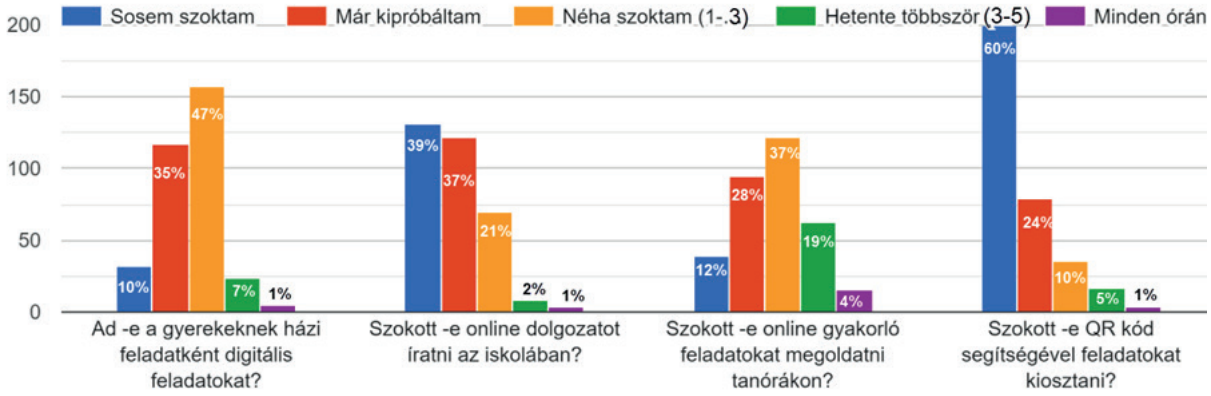
A következőkben arra kérdeztünk rá miként használják, és hogyan ítélik meg a digitális eszközöket. A válaszok javarészt bizonytalanságot, vagy rutintalanságot tükröznek, hiszen zömében 50% körüli igen válaszokat kaptunk azokra a kérdésekre is, amelyek egyértelműen és bizonyítottan pozitívak. Amit pozitívan értékeltek a digitális feladatok használatával kapcsolatban, a gyorsabb feladatkiosztás lehetősége, a feladatok automatikus javíthatósága, gyakorlófeladatok adása és az önellenőrzés lehetősége, de csak a megkérdezettek fele értett azzal egyet, hogy a tanári munkát az IKT használat megkönnyíti. A további kérdések esetében

(kevesebb munkát igényel a digitális órákra történő felkészülés), a tanulók szívesebben oldják meg a digitális feladatokat, illetve a tanórai fegyelem megítélése tekintetében is pozitív válaszok születtek. Egy kérdésben, a digitális órákon előforduló váratlan helyzetek esetében ítélték meg negatívan a helyzetet, miszerint gyakoribbak a műszaki problémák a digitális eszközökkel támogatott órákon.



5. ábra. A digitális eszközhasználat megítélése

A következőkben az általunk legnépszerűbbnek tartott digitális módszerek használatára kérdeztünk rá. Az eredmény szerint, házi feladatokat és gyakorló feladatokat adnak legtöbben a tanulóknak, bár így is mindkét esetben a megkérdezettek 10-12%-a elutasította ezeket a lehetőségeket. Az online dolgozat alkalmazása és a QR-kód használatát illetően, már kevésbé kedvező a helyzet, például a megkérdetettek 60%-a még ki sem próbálta, mint feladatmegosztási lehetőséget. Tanulság, hogy az egyetemi kurzusokon nagyobb figyelmet kell fordítani, a QR technológia módszertani bemutatására.

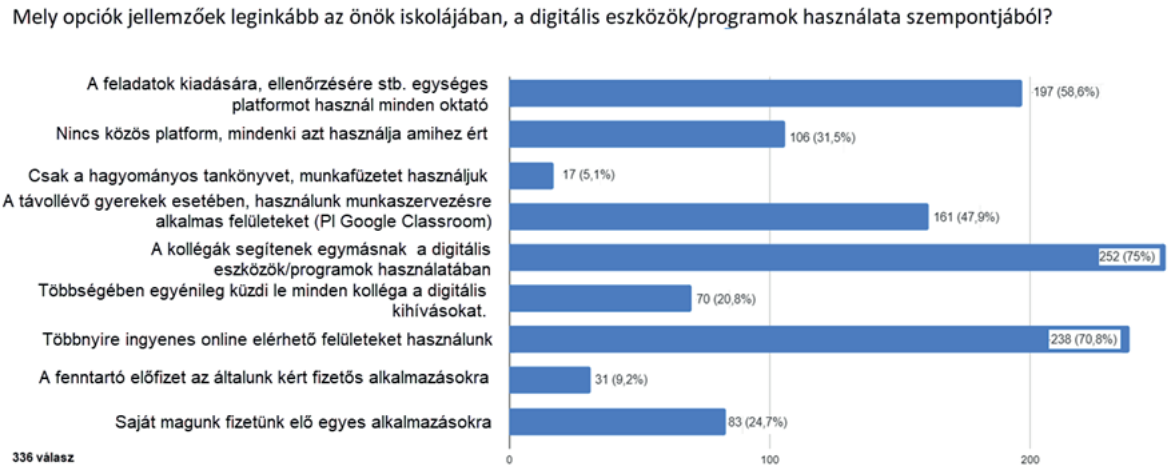


6. ábra. A digitális módszerek használata

A következőkben arra kérdeztünk rá, mennyire használnak egységes platformokat az adott iskolában dolgozó tanárok, milyen a kollaboráció szintje közöttük a mindennapi munkában. Itt is teljesen vegyesen alakult a kép, például a megkérdezettek csak 56,8%-a esetében használnak egységes digitális programokat és eszközöket az oktatók. 31% válaszolta azt, hogy semmilyen közös platform nincs, mindenki azt használja, amihez ért. További negatívum, hogy 20% azt jelölte meg, hogy egyénileg sajátítják el a használt alkalmazásokat. Itt is

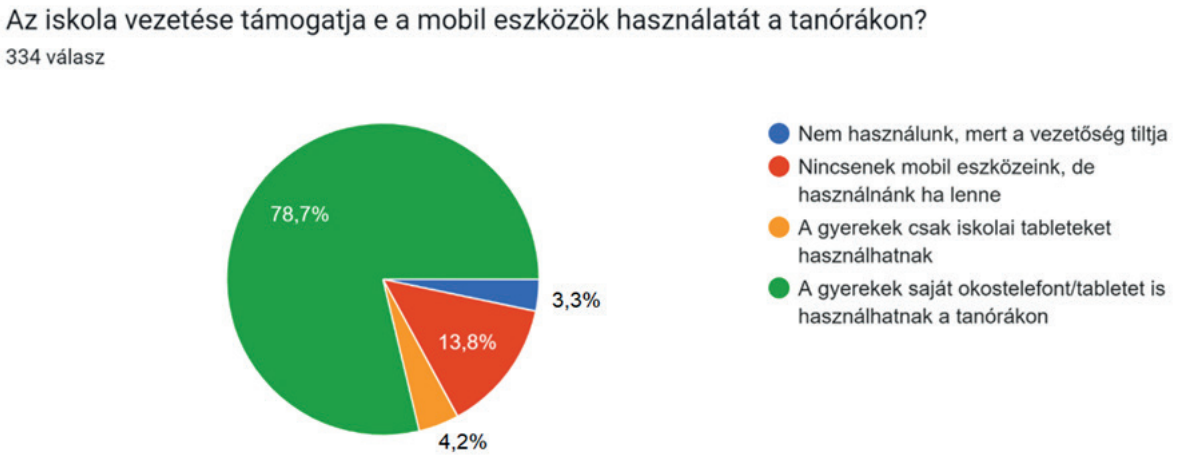
CC BY

érdemes lenne mélyebben megvizsgálni az okokat, miért nem tudnak, vagy akarnak segíteni egymásnak a digitális kompetenciák fejlesztésében. A megkérdezettek zömében ingyenes programokat használnak (70,8%), csak 31%-uk esetében fizet elő valamilyen digitális alkalmazásra a fenntartó, közel 25% pedig saját zsebből finanszírozza az előfizetéseket. Ezek az adatok mindenképpen elszomorítóak, hiszen a fenntartónak biztosítani kellene a feltételeket a korszerű oktatáshoz és ösztönözni a használatukat.



7. ábra. A digitális módszerek használatának körülményei

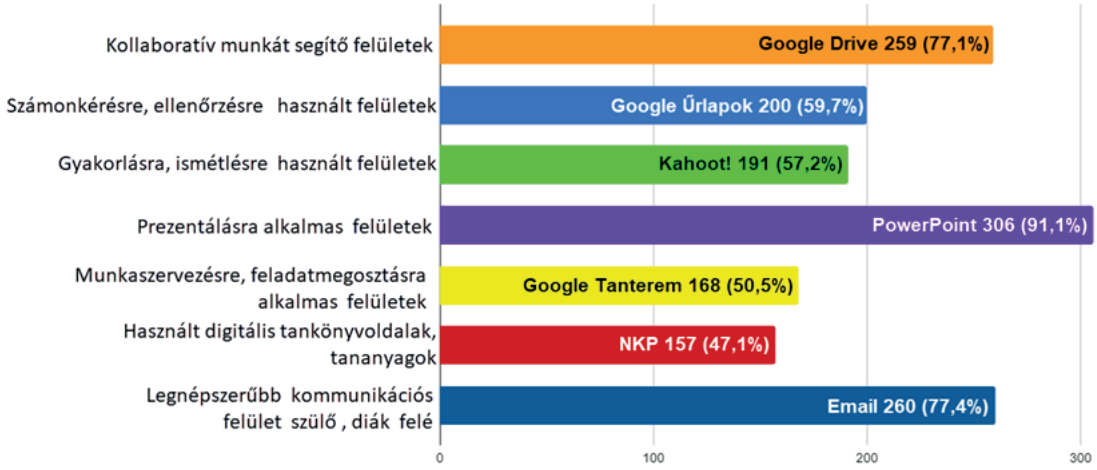
A mobil eszközök tanórai használatával kapcsolatos kérdés pozitív eredményeket hozott a korábbi évek felméréseihez képest.[2] Korábban majdnem teljes volt az elutasítottság, elsősorban a mobiltelefonokkal kapcsolatban mára ez szinte megfordult, majdnem 79%-ban használhatják a gyerekek a mobiljaikat a BYOD jegyében, és csak 3,3% adta azt a választ, hogy a vezetőség tiltja.



8. ábra. Mobil eszközök használatának arányai tanórákon

A kategóriánként leggyakrabban használt alkalmazások nem változtak az utóbbi években, elmondható, hogy a fejlesztők megtették a megfelelő lépéseket a felületek frissítése érdekében. Természetesen ettől jóval bővebb a paletta, például kommunikációs felületek között a Facebook és Messenger alig maradt le. A számonkérésre használt felületek közül a Redmenta népszerűsége csökkent, miután megszűnt az ingyenessége. Több kategóriában

a Google programjai nyertek, elsősorban ingyenességük és egyszerű használatuk miatt. A digitális tankönyvek esetében az NKP oldala a legnépszerűbb a megkérdezettek között, de még mindig előkelő helyen vannak a Mozaik Kiadó digitális felületei is.



9. ábra. A legnépszerűbb alkalmazások kategóriánként

5. Konklúziók

A digitális eszközök elfogadottsága és a tudatos kreatív használatuk iránti igény nőtt, szükségességük és szerepük elvitathatatlan. Sajnos a tanárok közötti kollaboráció a digitális kompetencia fejlesztésében elmarad a várttól, annak ellenére, hogy sokaknak nehézséget okoz és több időt igényel a digitális feladatok készítése. A kérdőív elemzéséből kiderül, hogy egyre kevesebben mellőzik a digitális feladatok alkalmazását a tanórákon, elsősorban gyakorlásra házi feladatként használják ki a digitális platformok lehetőségeit, viszont módszertani szempontból erősíteni kellene a munkájukat. Sajnos nagyon sokan még nem kérnek segítséget, vagy nincs kitől kérjenek, mindenképpen szükség lenne a tantestületeken belüli nyitottságra, illetve több ingyenes, vagy a fenntartó által támogatott digitális kompetencia fejlesztő tanfolyam szervezésére.

Az alkalmazható platformok száma jelentősen bővült az elmúlt években, viszont nagyon sok eddig ingyenesen használható program előfizetéses lett. Ezzel szemben a fenntartók csak csekély százaléka fizet elő központilag programokra, így sokan kényszerülnek saját maguk finanszírozni az előfizetéseket.

Legpozitívabb előremozdulás a mobil eszközök, elsősorban a mobiltelefonok használata és iskolai alkalmazásuk terén történt, néhány éve az iskolák teljesen elutasították ezek használatát.

Irodalom:

[1] ANTAL P.: Digitalization and Sports: ICT-related challenges in physical education teacher training. In: Abonyi-Tóth et al. (ed.): New Methods and Technologies in Education, Research and Practice. Proceedings of XXXIII. DidMatTech 2020 Conference. Budapest, ELTE Faculty of Informatics. URL: http://didmattech.inf.elte.hu/wp-content/uploads/2020/09/Didmattech2020_Proceedings_XXXIII_v20200921.pdf (Letöltés: 2022. 12. 10.)

CC BY

[2] ANTAL P.:Mobil eszközök alkalmazásának lehetőségei az oktatásban, trendek, lehetőségek, koncepciók, In: Forgó, Sándor (szerk.) Az információközvetítő szakmák újmédia-kompetenciái, az újmédia lehetőségei, Eger, Magyarország : Líceum Kiadó (2017) 152 p. pp. 103-124. , 22 p.

[3] ANTAL P., CZEGLÉDI L.: A távolléti oktatás tanulságai 2020-ban: felsőoktatás (EKE), közoktatás, iskolai könyvtárak In: Zagyváné Szűcs, Ida; K. Nagy, Emese (szerk.) Kihívások és megoldások a XXI. század pedagógiájában : válogatás a Pedagógiai Szakbizottság tagjainak a munkáiból Miskolc, Magyarország, Eger, Magyarország : Magyar Tudományos Akadémia Miskolci Területi Bizottsága, Eszterházy Károly Katolikus Egyetem Líceum Kiadó (2021) 269 p. pp. 67-88. , 22 p.

[4] CZEGLÉDI L.:. A felsőoktatás informatizálása, különös tekintettel a technikai eszközök integrációjára. In: Kis-Tóth Lajos (szerk.), Elektronikus tanulási környezetek kialakítása 1., Eger, Líceum K., pp. 10-31.

[5] FEKETE IMRE: A magyar közoktatásban tanító pedagógusok tapasztalatai a digitális munkarendű oktatásról IKT tudásszintjük tükrében: egy kevert módszertanú kutatáseredményei a Covid-19 idején In: Magyar Pedagógia, 120. évf. 4. szám 299–325. (2020) https://doi.org/10.17670/MPed.2020.4.299 (Letöltés: 2023. 02. 10.)

[6] RUBEN R. PUENTEDURA: Transformation, Technology, and Education. (2006) Online at: http://hippasus.com/resources/tte/ letöltés ideje: 2023. február 10.

[7] VAJNA TAMÁS: A magyar oktatási rendszer hegymenetben futott neki a digitális átállásnak, és meg is látszott az eredménye https://qubit.hu/2021/07/13/a-magyar-oktatasi-rendszer-hegymenetben-futott-neki-a-digitalis-atallasnak-es-meg-is-latszott-az-eredmenye letöltés ideje: 2023. március 21.

 DOI: 10.31915/NWS.2023.21

A szuperszámítástechnika mint európai stratégiai ágazat

Máray Tamás
ORCID: 0009-0007-8032-939X
maray@sztaki.hu

Absztrakt

A szuperszámítástechnika a tudomány és innováció megkerülhetetlen eszközévé vált. Ezért világszerte kiemelt figyelem övezi, és a fejlett országokban komoly programok keretében fejlesztik az infrastruktúrát, bővítik a felhasználást. Bár Európa ezen a területen is lemaradt Amerikához és Ázsiához képest, az utóbbi években megindult egy erőteljes felzárkózás. A cikk bemutatja, hogy hol tart ez a globális versengés ma, és mi a helyzete ebben hazánknak.
Kulcsszavak: HPC, szuperszámítógép, PRACE, EuroHPC, kvantum számítógép

Abstract

Supercomputing has become an indispensable tool for science and innovation. It has therefore attracted worldwide attention, with major programmes in developed countries to develop infrastructure and expand its use. Although Europe has lagged behind America and Asia in this area, it has started to catch up strongly in recent years. The article describes where this global competition is today and where Hungary stands.
Keywords: HPC, supercomputer, PRACE, EuroHPC, quantum computer

A szuperszámítástechnika (HPC[1]) a szuperszámítógépek építésének és alkalmazásának összefoglaló neve. Bár a HPC nem új, tulajdonképpen egyidős az elektronikus számítógépek történetével (hiszen mindig az adott kor legnagyobb teljesítményű számítógépeit tekinthetjük szuperszámítógépnek), mégis, mint a számítástechnika egyik különleges részterülete, az utóbbi 2 évtizedben kapott egészen kiemelt figyelmet és a jelentősége ugrásszerűen megnőtt. Sok oka van ennek, az adatforradalomtól kezdve a digitalizáció terjedésén át a tudományos módszerek és a high-tech iparágak gyorsuló fejlődéséig. Mindenesetre az széles körben nyilvánvalóvá és bizonyítottá is vált, hogy az innovációs- és versenyképesség szintje és a szuperszámítástechnika alkalmazásának mértéke között direkt összefüggés van. Miután ezt sok fejlett gazdaságban felismerték, megindult az egyre gyorsuló verseny a világ vezető országai között a szuperszámítógépek fejlesztése és alkalmazása terén. „*To compete, you must compute!*” hangzik a jól ismert szlogen, melynek érvényessége ma már nem kérdés. Bár a szuperszámítógépek mind a mai napig elsősorban kutatási infrastruktúrát jelentenek, a legkülönfélébb ipari és gazdasági termelőfolyamatok és szolgáltatások produkciós fázisaiban is jelen vannak már. Nemzetközileg is versenyképes tudományos/kutatási tevékenység alig képzelhető el szuperszámítógépes háttér nélkül, hiszen az egyre pontosabb és összetettebb szimulációk, adatbányászat vagy akár mesterséges intelligencia alapú eljárások más módon nem kezelhetők hatékonyan. A szuperszámítógépek segítségével számtalan olyan kérdés vizsgálható ami másképp nem lenne lehetséges. Természetesen, aki a legújabb tudományos eredményeket birtokolja, az az innovációt és fejlődést is meghatározza és így versenyelőnyre tesz szert. Ezért olyan komoly az erőfeszítés a világ összes fejlett régiójában a szuperszámítástechnika területén.

1 HPC – High Performance Computing

CC BY

Alkalmazás oldalról napjainkban a szuperszámítástechnika fejlesztésének egyik legfontosabb ösztönzője a mesterséges intelligencia. Technológiai szempontból pedig a fejlődést a tranzisztorok további méret csökkenése (már a 3nm-es tartományról van szó!), a GPU-k intenzív alkalmazása és az új hűtési módszerek teszik lehetővé (talán meglepő, de a szuperszámítógépek fejlődésének egyik legnagyobb technológiai kihívása a hatékony hűtési megoldások megtalálása). És közben egyre fontosabbá válnak a még mindig kísérletinek számító kvantum informatikai módszerek is. A kvantum számítógépek jelentik a HPC fejlődésének egyik legígéretesebb irányát.

Mára tehát nagyon éles, globális verseny alakult ki a világ legfejlettebb és legerősebb gazdasági/tudományos régiói között a szuperszámítógépek fejlesztését, építését és felhasználását illetően. A három pólus Észak-Amerika, Távol-Kelet és Európa. Ez a felsorolás a jelenlegi helyezések sorrendjét is tükrözi.

A világban üzembe helyezett szuperszámítógépek rangsorolására 1993-ban létrejött *top500.org* lista féléves frissítésekkel publikálja az 500 legnagyobb teljesítményű gép főbb jellemzőit, szemléletesen mutatva a trendeket és azt, hogy melyik ország vagy régió éppen hol áll ezen a területen. Szimbolikus, hogy az aktuális lista első három helyezettje éppen a fenti sorrendben egy-egy amerikai, ázsiai és európai szuperszámítógép.

Rank	System	Cores	Rmax (PFlop/s)	Rpeak (PFlop/s)	Power (kW)
1	**Frontier** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE DOE/SC/Oak Ridge National Laboratory United States	8,730,112	1,102.00	1,685.65	21,100
2	**Supercomputer Fugaku** - Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D, Fujitsu RIKEN Center for Computational Science Japan	7,630,848	442.01	537.21	29,899
3	**LUMI** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE EuroHPC/CSC Finland	2,220,288	309.10	428.70	6,016

1. ábra A 2022 novemberi top500 lista első három helyezettje (forrás: top500.org)

Bár Európa a régiók versenyében még mindig csak a 3., a lemaradásának mértéke pár évvel ezelőtt sokkal jelentősebb volt. Ez javult az utóbbi időszakban azáltal, hogy az EU döntéshozói felismerték az ágazat stratégiai jelentőségét, és komoly, össz-európai programot indítottak a felzárkóztatásra. Ennek első eredményei már láthatók. Az európai szintű összefogásra azért van szükség, mert a HPC rendkívül erőforrás igényes, ezért az egyes európai országoknak önállóan kevés esélyük van felvenni a versenyt pl. az USA-val vagy Kínával.

Észak-Amerika

A globális vezető szerep hagyományosan az Amerikai Egyesült Államoké, amely több száz szuperszámítógéppel, a világ teljes HPC kapacitásának legalább a felével rendelkezik. Az USA a hardver és szoftver technológia fejlesztésében, a HPC gyártók tekintetében, a

kiépített infrastruktúrában és a HPC alkalmazás mértékében is az élen jár. A Fehér Ház által még 2015-ben elfogadott NSCI (National Strategic Computing Initiative) [1][2] új lendületet adott a fejlődésnek és kitűzte az akkor még szinte utópisztikusnak tűnő célt, az exaflop határ áttörését. 1 Eflop/s = 10^{18} matematikai művelet másodpercenként. Nehéz felfogni, hogy ez milyen óriási szám. Akkora, hogy a Földön élő 8 milliárd ember mindegyikére másodpercenként több, mint 137 millió számítási művelet jut. A legnagyobb technológiai kihívást az jelentette, hogy hogyan lehet a fajlagos energiaigényt olyan mértékben leszorítani, hogy még egy ekkora gép is táplálható és hűthető maradjon. Bár az üzemterv szerint az exaflop határt 2020-ra kellett volna elérni, végül csak 2022-re sikerült, amikor üzembe helyezték az Oak Ridge National Laboratory-ban a világ első, 1 Eflop/s feletti teljesítményű szuperszámítógépét, a Frontiert. (2. ábra)



2. ábra: Frontier (Forrás: Oak Ridge National Laboratory, U.S. Dept. of Energy)

A Cray technológiát használó, közvetlen folyadék hűtésű, HPE-Cray által épített Frontier csúcs szuperszámítógép mért teljesítménye: 1,1 Eflop/s.

Elméleti teljesítmény	R_{peak} > 1,5 Eflop/s
Mért teljesítmény	R_{max} = 1,102 Eflop/s
CPU	9472 db AMD Epyc Milan (606,708 core)
GPU	37888 db Radeon Instinct MI250X (8,335,360 core)
Interconnect technológia	Slingshot
Memória	9,2 Pbyte
Háttértár	700 Pbyte
Operációs rendszer	HPE Cray OS (Linux)
Fizikai helyszükséglet	680 m^2
Energiaigény	21 MW (Green500 no.1)

1. táblázat: Frontier technikai paraméterei

CC BY

A Frontier az amerikai energiaügyi tárca egyik kiemelt kutatóintézetében kapott helyet, de felhasználása széleskörű és általános célú, a tudományos kutatás minden területét támogatja (anyagtudomány, űrkutatás, biotechnológia, klímakutatás, szoftver technológia, mesterséges intelligencia, lézerfizika, plazmafizika, stb.)

A Frontier teljesítménye elképesztő, de ez sem elegendő. A tudománynak még sokkal többre van szüksége. Így Amerikában hamarosan (a tervek szerint még 2023-ban) átadásra kerül további két, még nagyobb szuperszámítógép, az Aurora és az El Capitan, melyek külön-külön is meg fogják haladni a 2 Eflop/s teljesítményt. Mindeközben a fejlesztés és a gyártás erősen konszolidálódott, a gyártók száma lecsökkent. Az IBM kiszállt a hagyományos szuperszámítógépek piacáról, az SGI és a Cray a HPE-be olvadt.

A hagyományos szuperszámítógépek mellett ugyanakkor a kvantum számítógépek fejlesztése is gyors ütemben zajlik, ezen a területen az IBM vezető szerepet játszik. A kvantum informatika területén Kanada is nagyon jelentős szereplő.

Ázsia

Ázsia két vezető szuperszámítógép hatalma Japán és Kína. Természetesen más országokban (pl. Dél-Korea vagy India) is gyors a fejlődés, de a globális versenyben Japán és Kína a meghatározók. Japánban nagy hagyománya van a HPC-nek és az ország többször is elfoglalta a TOP500 lista első helyezését saját gyártmányú rendszereivel. Jelenleg a vezető Japán gép, a Fujitsu által épített Fugaku (442 Pflop/s) a TOP500 második helyezett, de a listán további 30 japán rendszer is található.

Kína csak az elmúlt két évtizedben kapcsolódott a versenybe, de a fejlődés nagyon erős. Érdekesség: ugyan abban az évben került Kína először a TOP500 listára (egy IBM gép beszerzésével) amikor Magyarország is: 2000-ben. Mára Kína birtokolja a TOP500 szuperszámítógépek kb. felét, a kapacitás kb. 1/3-át. Kína nemcsak alkalmazza a HPC technológiát, hanem fejleszti is, így a kínai szuperszámítógép gyártók (Inspur, Sugon, Lenovo, Huawei) az amerikaiak legerősebb versenyfelei. Kína legnagyobb szuperszámítógépe a, Sunway TaihuLight a legfrissebb TOP500 listán jelenleg a 7. helyet foglalja el, de a gyártás évében (2016) az első helyen állt. Érdekesség, hogy ez a szuperszámítógép kínai fejlesztésű processzorra épül. Operációs rendszere azonban szintén Linux alapú. Kína is törekszik az exaflop tartomány elérésére, és várható, hogy ez hamarosan meg is történik.

Európa

Nem meglepő módon az európai országok között a gazdaságilag legerősebbek a legaktívabbak a HPC alkalmazásában. Németország, az Egyesült Királyság, Franciaország, Olaszország, Spanyolország, Svájc vezetik a listát. Bár az 1980-as 90-es években Európa nem állt rosszul a globális versenyben, 2000 után fokozatosan lemaradt. Míg Észak-Amerikában és Ázsiában korán felismerték a HPC növekvő fontosságát, mindez Európában csak késve következett be, ráadásul az európai országok külön-külön akkora invesztíciót sem tudtak végrehajtani. Az EU az első jelentősebb, európai összefogásra építő HPC programot (DEISA) 2005-ben indította, majd ezt követte 25 ország együttműködésével a PRACE (Partnership for Advanced Computing in Europe) [3] 2010-től, amelyekhez kapcsolódva számos HPC K+F projekt is indult. Bár e programok sikeresek voltak az egyre gyorsuló globális versenytársakhoz való felzárkózásra nem bizonyultak elegendőnek.

A tanulságokat leszűrve az EU 2019-ben stratégiai fontosságú fejlesztési területnek nyilvánította a szuperszámítástechnikát [4] és 1 mrd EUR induló forrással létrehozta az EuroHPC Joint Undertaking (EuroHPC Közös Vállalkozás) nevű programot. [5] Ez deklaráltan egy felzárkóztatási program, széles körű európai összefogással, a tagországok saját forrásainak mobilizálásával, a teljes HPC ökoszisztéma felölelésével (infrastruktúra beruházások, hardver és szoftver fejlesztés, alkalmazások, szolgáltatások, képzés), az európai HPC „ipar" megerősítésével. Az EuroHPC is célul tűzte ki az exaflop tartomány elérését, de az ehhez vezető úton először petaflop és pre-exa kategóriájú gépeket épít. Az EuroHPC-hez 28 ország csatlakozott, köztük Magyarország is. Az EuroHPC első, petaflops kategóriájú szuperszámítógépei Szlovéniában, Luxemburgban, Csehországban, Bulgáriában és Portugáliában épültek. E gépek teljesítménye 4,5 - 13 Pflop/s közé esik, így a TOP500 felső ötödében találhatók. Az első – és egyben legnagyobb - EuroHPC pre-exa kategóriájú szuperszámítógép, a Lumi, Finnországban került telepítésre, 10 ország összefogásával. A Lumit 2022-ben adták át, jelenleg a világ 3. legerősebb szuperszámítógépe. A Lumi is HPE/Cray gyártmányú, és technológiája szinte teljes egészében egyezik a Frontier technológiájával. A gép messze északon, olyan környezetben épült fel, ahol egész évben megújuló energiával és természetes hűtéssel (free cooling) működtethető.



4. ábra: Lumi (Forrás: CSC, Finnország)

Elméleti teljesítmény	R_{peak} > 420 Pflop/s
Mért teljesítmény	R_{max} = 309,1 Pflop/s
CPU	5632 db AMD Epyc
GPU	10240 db Radeon Instinct MI250X
Interconnect technológia	Slingshot
Háttértár	117 Pbyte
Operációs rendszer	HPE Cray OS (Linux)
Fizikai helyszükséglet	150 m^2
Energiaigény	8,5 MW (Green500 no.1)

2. táblázat: a Lumi technikai paraméterei

CC BY

A Lumi után 2022-ben átadásra került az olaszországi Bolognában az EuroHPC második pre-exa szuperszámítógépe is, a 174 Pflop/s teljesítményű Leonardo. A Leonardot az európai Atos/Bull cég építette, Intel processzorokat és Nvidia GPU gyorsítókártyákat tartalmaz. A Leonardo a világ 4. legerősebb szuperszámítógépe.

A Leonardo után a 3. európai pre-exa szuperszámítógép a 2023-ban Barcelonában épülő MareNostrum 5 lesz.

Az új tervezési ciklusban már 8 mrdEUR forrásból gazdálkodó EuroHPC program időközben az első európai exaflop szuperszámítógép (Jupiter) megépítését is bejelentette, ami Németországban, a Jülichi Szuperszámítógép Központban (JSC) valósul meg 2024-re. És további, legalább ~20 Pflops/s teljesítményű kisebb rendszerek is épülnek, Görögországban, Írországban és Lengyelországban.

Az EuroHPC program a kvantum technológiára kiemelt figyelmet fordít. Komoly forrásokkal támogatja kvantum számítógépek beszerzését és a technológia fejlesztését.

A fejlődés tehát Európában is nagyon felgyorsult. Az előretörés a TOP500 listán is szembeötlő. Fontos, hogy az EuroHPC nemcsak az infrastruktúra fejlődésében hoz látványos eredményeket, hanem a felhasználás, az alkalmazás és algoritmus fejlesztések, illetve a szakember képzés területén is előre lép. Érdekesség, hogy Európa saját processzor fejlesztésbe is kezdett, hogy csökkentse technológiai függőségét. A 2015-ben indult EPI (European Processor Initiative) célkitűzése, hogy olyan processzort fejlesszen ki, amely a jövő európai szuperszámítógépeinek motorja lehet. A fejlesztés ARM és RISC-V alapokra épít. A megvalósításra létrehozott vállalkozás (SiPearl) éppen 2023-ra ígéri az első működőképes változat megjelenését.

Magyarország

Magyarországon az első valódi, tudományos célú szuperszámítógépet 2001-ben adták át az NIIF Program keretében. A gép 60 Gflops/s teljesítményével rögtön a TOP500 listára került és hamar népszerűvé vált a kutatók között. A gépet és utódait az egyetemeken és kutató intézetekben sok száz tudományos kutatási projektben használták, sok esetben kiemelkedő sikerrel, nemzetközileg is elismert eredményeket produkálva. Az infrastruktúra az NIIF Program keretében folyamatosan megújult és fejlődött technológiában és kapacitásban egyaránt.



5. ábra: Az első hazai szuperszámítógép (forrás: a szerző)

2015-ben az NIIF már 8 szuperszámítógépet működtetett, 0,5 Pflop/s aggregált kapacitással, köztük az első, GPU gyorsítókkal ellátott Leo nevű gépet, amely a Debreceni Egyetem campusán épített új, korszerű NIIF HPC központban kapott elhelyezést és 2022 év végéig a legnagyobb hazai szuperszámítógép volt.

Az NIIF jogutódja a KIFÜ által üzembe helyezett Komondor 2023-ban váltotta a kiöregedő Leot. 3,09 Pflop/s teljesítménnyel ez a rendszer a TOP500 lista 199. helyére került. Bár e gép teljesítményét tekintve a világ legnagyobb szuperszámítógépének alig 1/300-ada, technológiájában hasonlít ahhoz. Szintén a HPE/Cray építette, melegvizes hűtésű, AMD Epyc processzorokat és Nvidia A100 GPU-kat, valamint Slingshot interconnect technológiát tartalmaz.

2022-ben Magyarország elnyerte az EuroHPC 35%-os pénzügyi támogatását egy új, 20 Pflop/s nagyságú szuperszámítógép építésére és üzemeltetésére, a projekt azonban máig nem kezdődött el. Enélkül azonban a hazai tudományos-kutatási tevékenység nemzetközi szinten versenyhátrányba kerül, és Magyarország kimarad sok olyan lehetőségből – köztük pl. a kvantumszámítástechnika korai alkalmazása – amelyeket az EU jelentős forrásokkal támogat és amelyek a HPC területén előttünk járó környező országok (pl. Csehország, Lengyelország, Szlovénia, sőt Bulgária is) számára nyitva állnak. Mindez a magyar innovációs és versenyképességet hátrányosan érinti.

CC BY

Felhasznált irodalom:

[1] Creating a National Strategic Computing Initiative: lekérdezve: 2023.02.28. https://www.whitehouse.gov/the-press-office/2015/07/29/executive-order-creating-national-strategic-computing-initiative
[2] Advancing U.S. Leadership in High-Performance Computing: lekérdezve: 2023.02.28. https://www.whitehouse.gov/blog/2015/07/29/advancing-us-leadership-high-performance-computing
[3] Partnership for advanced computing in Europe: lekérdezve: 2023.02.28. https://prace-ri.eu/
[4] The European strategy for High Performance Computing: lekérdezve: 2023.02.28. https://digital-strategy.ec.europa.eu/en/library/european-strategy-high-performance-computing
[5] The European High Performance Computing Joint Undertaking: lekérdezve: 2023.02.28. https://eurohpc-ju.europa.eu/index_en

 DOI: 10.31915/NWS.2023.22

Szoftveres Cutter-keresés az SZTE Klebelsberg Könyvtárban

Cutter Search Software in the SZTE Klebelsberg Library

Frankó Máté
SZTE Klebelsberg Könyvtár
mate.franko@ek.szte.hu
ORCID: 0009-0009-3782-6571

Zeller Rozália
SZTE Klebelsberg Könyvtár
rozalia.zeller@ek.szte.hu
ORCID: 0000-0003-2501-8760

Absztrakt
A Cutter-számok a könyvtári dokumentumok betűrend szerinti elhelyezéséhez és rendszerezéséhez használt alfanumerikus kódok. Más közgyűjteményekhez hasonlóan az SZTE Klebelsberg Könyvtárban is - a dokumentumban foglalt mű témájára utaló szakjelek mellett – Cutter jelzeteket használunk a szabadpolcos állományunk sorrendezéséhez. Heti rendszerességgel mintegy 300-400 könyv Cutter jelzettel való ellátása a Könyvtári raktározási táblázatok c. segédlet használatával meglehetősen időigényes feladatot jelentett kollégáinknak. Ezért 2022-ben a munkafolyamat fokozott automatizálása mellett döntöttünk. Célunk az volt, hogy létrehozzunk egy olyan elektronikus felületet, amely a szerző, a cím vagy más bibliográfiai adatok megadása után, további emberi beavatkozás nélkül megmutatja, hogy az adott karaktersorozathoz milyen Cutter-szám tartozik.

Első lépésben az eredeti táblázatot kellett a kereső automatizmusok számára használható - koherens és a programnyelvek számára is értelmezhető - formátumra hoznunk, úgy, hogy az eredeti adatszerkezet ne sérüljön. Javítottuk az eredeti táblázat digitalizálása során keletkezett hibákat, átalakítottuk a speciális magyar karaktereket, komplementáltuk a hiányzó vagy hiányos zárótaggal rendelkező cuttereket, egyúttal igyekeztünk megoldást találni a szóközös és szóköz nélküli betűsorok egységesítésére. A táblát MS Excel környezetben, VBA-makrókkal konvertáltuk kódsorokká. A kész felület alatt egy PHP-alkalmazás működik. Előadásunkban a fejlesztés egyes munkafolyamatait, az alkalmazott módszereket, az elkészült rendszer működését és a tervezett fejlesztési irányokat kívántuk bemutatni.
Kulcsszavak: Cutter számok, Cutter táblázat, betűrendi jel, szoftveres keresés, VBA makró

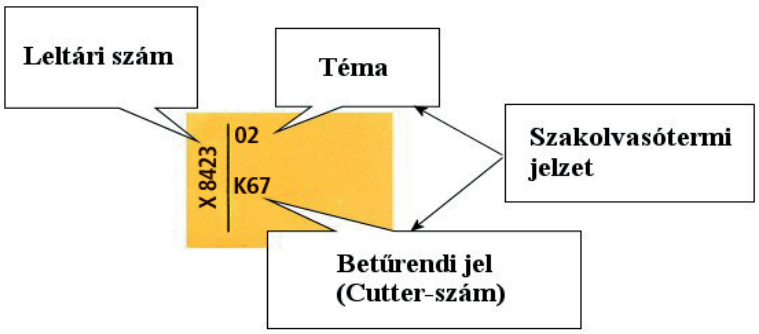
Abstract
Cutter numbers are alphanumeric codes used to alphabetise and organise library documents. Like in many other libraries, we use the Cutter classification system - in addition to the Universal Decimal Classification system - to organise our open shelf holdings in the SZTE Klebelsberg Library. It was a time-consuming task for our colleagues to assign Cutter numbers to about 300-400 books per week using the Alphabetic Order Table. Therefore, in 2022 we decided to increase the automation of the workflow. Our goal was to create an electronic interface that, after entering the author, title or other bibliographic data, would show the Cutter number associated with a given character set without further human intervention.

CC BY

As a first step, the Alphabetic Order Table had to be converted into a format that could be used by search automation - coherent and interpretable by programming languages - without damaging the original data structure. We corrected the errors that had occurred during the digitisation of the original table, converted the special Hungarian characters, complemented the truncated cutters, and tried to find a solution to unify the spaced and unspaced strings. The table was converted into lines of code in MS Excel using VBA macros. A PHP application runs under the finished interface. In our presentation we wanted to show the workflows of the development, the methods used, the functioning of the completed system and the planned directions of development.
Keywords: Cutter classification system, Cutter numbers, Alphabetic Order Table, search software, VBA macro

1. Bevezető

A nyomtatott dokumentumok fizikai helyét a legkülönfélébb szakjelek, jelzetek, szakcsoportszámok jelölik világszerte a könyvtárakban. Ezek a betűkből és számokból összeállított azonosító jelek a szakemberek számára információkkal szolgálnak a megjelölt publikáció témájára, típusára vagy bibliográfiai adataira vonatkozóan. Használatuk megkönnyíti a dokumentumok tematikus rendszerezését és sorrendezését, illetve biztosítja azok gyors visszakereshetőségét akár százezres vagy milliós példányszámú közgyűjteményi állományok esetén is.



1. ábra. Az SZTE Klebelsberg Könyvtár szabadpolcos állományában használt címke

A Szegedi Tudományegyetem Klebelsberg Kuno Könyvtárának szakolvasótermeiben[1] elhelyezett szabadpolcos állomány azonosítására egy kétosztatú jelölési rendszert használunk. A dokumentumok szakterületi besorolásánál egyszerűsített ETO-számokat tartalmazó szakjelzetekkel, míg a kötetek ábécé szerinti rendezésénél általában a befoglalt mű szerzőjére vagy címére utaló Cutter-számokkal dolgozunk.

2. Cutter-számok és használatuk az SZTE Klebelsberg Könyvtárban

A Cutter-szám egy nyomtatott nagybetűből és egy kétjegyű számból álló alfanumerikus kód. Minden számhoz tartozik egy kezdő és záró érték. A Cutter-számok a két végpont közé eső, illetve azokkal megegyező karaktersorozatokat jelölik. A Cutter-számok hivatalos listáját – a hozzájuk tartozó kezdő és záró tagokkal – a Könyvtári Intézet (KI) által kiadott Könyvtári raktározási táblázatok[2] című segédkönyv tartalmazza.

1 Az SZTE Klebelsberg Könyvtár olvasói tereinek bemutatása. http://www.ek.szte.hu/olvasoi-terek/ Hozzáférés: 2023.06.15.

2 Könyvtári raktározási táblázatok. 9. átdolg. javított kiadás. Budapest: Könyvtári Intézet, 2001. https://www.ki.oszk.hu/dokumentumtar/konyvtari-raktarozasi-tablazatok Hozzáférés: 2023.06.15

Például a B14-es cutterhez a „Baim” és a „Bakor” karaktersorok tartoznak. Követve a megadott szabályt minden olyan nevet vagy könyvcímet, amely betűrend szerint a „Baim” után következik, de a „Bakor” szó előtt szerepel az ábécében, a B14-es Cutter-számmal kell jelölni. A B14-es szám alá kerül egyebek közt a „Bajorország” kifejezés és „Baka István” neve is.

A Cutter-számok kiosztását az SZTE Klebelsberg Könyvtárban az egyes tudományterületekért felelős szakreferensek, valamint a Feldolgozó Osztály munkatársai végzik. Korábban a Könyvtári raktározási táblázatok nyomtatott, illetve annak digitalizált és OCR-ezett elektronikus változatából dolgoztak a kollégák, de heti 300-400 dokumentum jelzetelése és átirányítása mellett meglehetősen problémás és időigényes feladat volt az egyes címekhez és szerzőkhöz tartozó Cutter-számok lekeresése.

Néhány évvel ezelőtt a probléma megoldása érdekében készült egy egyszerű keresőprogram, de az sem váltotta be igazán a hozzá fűzött reményeket. Az alkalmazásban kizárólag a kezdő és záró tagokra lehetett keresni, így teljes szavak, nevek keresésére nem volt alkalmas, csak a potenciális lehetőségek körét tudta szűkíteni.

2022-ben a könyvtár érintett vezetői a folyamat fokozott automatizálása mellett döntöttek. A cél az volt, hogy létrehozzunk egy olyan online felületet, amely a szerző, a cím vagy más bibliográfiai adat megadása után automatikusan megmutatja, hogy az adott karaktersorozathoz milyen Cutter-szám tartozik.

3. Adattisztítás a Cutter-táblázatban

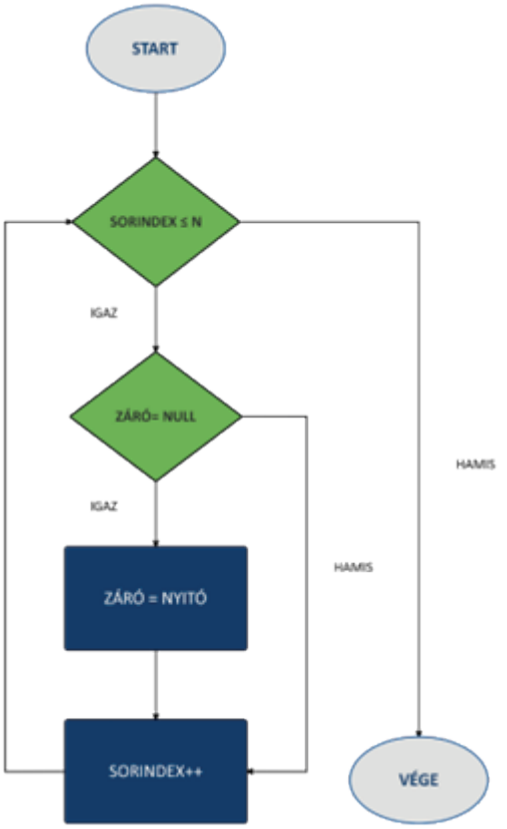
Első lépésben a Cutter-táblázatot kellett koherens és a programnyelvek számára is értelmezhető formára hozni – természetesen úgy, hogy az eredeti adattartalom ne sérüljön. Át kellett írnunk a forrásállományban az egyszerűsítő jelöléseket és a számítógépes automatizálás folyamatát megnehezítő speciális karaktereket. A táblázat egyes sorait Microsoft Excelben, VBA makrók segítségével tisztítottuk le.
A Cutter-tábla belső felépítésének, logikai szerkezetének feltérképezése során két olyan, a főszabálytól eltérő típuskivételt találtunk, amelyek a számítógépes feldolgozás során problémát okozhattak volna.

3.1. Hiányzó és hiányos záró tagok

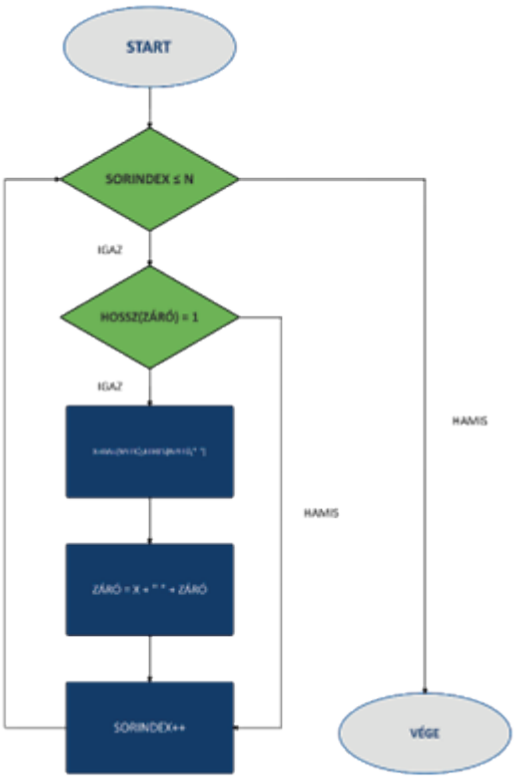
Az első csoportba a hiányzó záró taggal rögzített sorok tartoznak. Ilyen például az A25-ös kódszámhoz tartozó rekord, ahol is az „Ady” nyitó tag mögött nincs záró érték. Könnyen belátható, hogy itt – és az összes hasonló felépítésű sorban – a nyitó és a záró tag megegyezik, vagyis az A25-ös Cutter-szám az „Ady”-tól „Ady”-ig terjedő karaktersorozatokat jelöli, csak az utóbbit külön nem tüntették fel a segédlet készítői. Ahhoz, hogy a későbbiekben a keresőprogramnak legyen mit-mivel összehasonlítania, utólagosan pótolnunk kellett a hiányzó záró tagokat. Ehhez lefuttattunk egy szkriptet a tábla összes során, utasítva a táblázatkezelőt, hogy ott, ahol a záró érték helyén üres cellát talál, írja be a nyitó tagot.

A második kivételhalmazt a hiányos záró taggal felvitt sorok alkotják. Példaként említhetjük a K77-es Cutter-számot, ahol is a „Kovács E” előtag mögött csak egy „G” betűt tüntettek fel a szerkesztők. Ebben az esetben is egy formai egyszerűsítésről van szó: a „G” betű a „Kovács G” karaktersorozatot hivatott jelölni, a keresőprogram helyes működéséhez azonban a teljes záró tagnak szerepelnie kell a táblázatban. Ezeknél a soroknál szoftveresen kimásoltuk a szóközt megelőző karaktereket a nyitó tagból és beillesztettük azokat a záró tag elé.

CC BY



2. ábra. A program működési elve hiányzó záró tag esetén



3. ábra. A program működési elve hiányos záró tag esetén

3.2. Speciális karakterek

Két fontos könyvtári-nyelvtani szabályt kellett még átültetnünk a keresőprogramhoz készített új táblázatba. Az egyik, hogy a könyvtári ábécé nem tesz különbséget a rövid és a hosszú magánhangzók, valamint az „sz” kivételével az egy és a többjegyű betűk között. Ennek megfelelően az „á” betűs szavakat az „a”-hoz, a „cs” betűs szavakat a „c”-hez sorolja. A másik, hogy a szóközök minden más karaktert megelőznek, így például „Abonyi Andor” neve az „Abonyiak” kifejezés elé kerül a betűrendben, függetlenül attól, hogy az „Abonyiak” szót lezáró „k” karakter az „Andor” név „n” betűje előtt van az ábécében.

Figyelembe véve a fenti szabályokat létrehoztunk egy speciális ábécét, ahol minden betűt egy egyjegyű szám vagy az angol ábécé egyik betűje jelöl. Látható, hogy minden betű és szám mögé beillesztettünk egy felkiáltójelet is. Erre a helyes adattípus-deklaráció miatt volt szükség. Például „Baja” város neve az átírás után „1090”-ként jelenne meg a táblázatban. Ha nincsenek felkiáltójelek, a program számként és nem szöveges adatként értelmezné ezt a kifejezést.

Kódolt ábécé									
A | a	0!	É | é	4!	L | l	B!	P | p	G!	Ü | ü	M!
Á | á	0!	F | f	5!	Ly | ly	B!Q!	Q | q	H!	Ű | ű	M!
B | b	1!	G | g	6!	M | m	C!	R | r	I!	V | v	N!
C | c	2!	Gy | gy	6!Q!	N | n	D!	S | s	J!	W | w	O!
Cs | cs	2!J!	H | h	7!	Ny | ny	D!Q!	Sz | sz	J!R!	X | x	P!
D | d	3!	I | i	8!	O | o	E!	T | t	K!	Y | y	Q!
Dz | dz	3!R!	Í | í	8!	Ó | ó	E!	Ty | ty	K!Q!	Z | z	R!
Dzs | dzs	3!R!J!	J | j	9!	Ö | ö	F!	U | u	L!	Zs | zs	R!J!
E | e	4!	K | k	A!	Ő | ő	F!	Ú | ú	L!	[Szóköz]	!

4. ábra. A Cutter kereső szoftver számára előkészített speciális ABC

4. A Cutter-kereső működése

Az átalakított Cutter-táblázat egy XML fájlként került be a keresőprogramba. Az állomány az eredeti és a kódolt nyitó illetve záró tagokat is tartalmazza. A kódolt változat a kereséshez kell, a felhasználói interfész találati listájában viszont az eredeti alakban jelenik meg az eredmény.

```
<cutter>
  <number>R49</number>          Cutter-szám
  <firsttag>Régi</firsttag>     Nyitó tag
  <endtag>Regn</endtag>         Záró tag
  <xfirsttag>I!4!6!8!</xfirsttag>   Átalakított nyitó tag
  <xendtag>I!4!6!D!</xendtag>       Átalakított záró tag
</cutter>
```

5. ábra. Részlet az XML fájllá alakított Cutter-táblázatból

A kereső első lépésben törli a felesleges karaktereket (például a szövegben hagyott írásjeleket vagy sortöréseket) a felhasználói inputból, majd átalakítja a megadott karaktersort a Cutter-tábla transzformálása során alkalmazott szabályok szerint.

Ezt követően a program végigfut a Cutter-táblázaton, és szükség esetén feltölti plusz karakterekkel a nyitó és záró tagokat. Ha a nyitó és záró tagok rövidebbek, mint a felhasználó által megadott kifejezés, a program hibás eredményt adhat. A nyitó tagok mögé szóközöket (azaz a kódtábla szerint felkiáltójeleket), a záró értékek mögé pedig „z” betűket (azaz „R” betűket) ír a program, így a művelet a betűrendet nem módosítja.

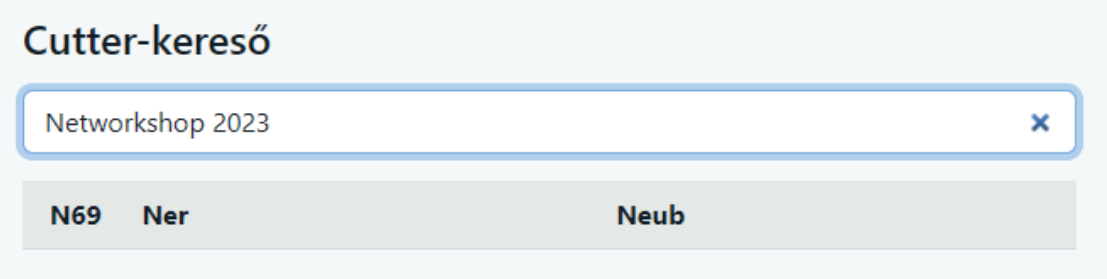
Zárásként megvizsgálja, hogy melyik sorban igaz az, hogy az input nagyobb vagy egyenlő, mint a nyitó tag és kisebb vagy egyenlő, mint a záró. Végül az „IGAZ” értékkel visszatérő sorokat megjeleníti a beviteli szövegmező alatt egy táblázatban.

CC BY

Fontos megjegyeznünk, hogy bár a program fejlesztése során egy szerver oldali programnyelvet (nevezetesen PHP-t) használtunk, a kezelőpanel valós időben kommunikál a kiszolgálóval. Három karakter megadása után automatikusan lefut a program. Ha a könyvtáros átírja a szövegdobozban megadott kifejezést, az alkalmazás azonnal módosítja a találati listát, ezért nincs szükség „Keresés" gombra.

5. Fejlesztési irányok

A kereső első verziója nyilvánosan is elérhető az SZTE Klebelsberg Könyvtár weboldalán[3], így már nem csak a saját kollégáink munkáját tudja megkönnyíteni.



6. ábra. A Cutter-kereső felülete az SZTE Klebelsberg Könyvtár weboldalán

Mindazonáltal van egy apró hiba, ami még javításra szorul a programban. A napi munkavégzés során derült ki, hogy egyes szóközös kifejezések - nevezetesen azok, ahol a szóközök jelöletlenül, a zárótagok végén vannak - az átalakításnál rejtve maradtak, így ezeknél a soroknál két találatot kap a felhasználó. Utólag összesen hét hibás sort azonosítottunk, ezek javítása még folyamatban van.

Ezen felül célszerű lenne elindulni a munkafolyamat teljes automatizálása felé. Hasznos lenne, ha a Cutter-keresés funkció beépülne az integrált könyvtári rendszerek katalogizáló moduljaiba, így a MARC-rekordokból kinyert információk alapján, további emberi beavatkozás nélkül, automatikusan generálhatóvá válna a megfelelő Cutter-szám. Ez a funkció egyes rendszerekben (például a LinLib[4]-ben vagy a TextLib[5]-ben) már elérhető, de a legtöbb IKR-ben még nem vált alapszolgáltatássá.

3 Cutter-kereső az az SZTE Klebelsberg Könyvtár weboldalán. http://www.cutter.bibl.u-szeged.hu Hozzáférés: 2023.06.15.

4 LinLib integrált könyvtári rendszer. http://www.linlib.hu/konyvtarirendszer.htm Hozzáférés: 2023.07.31.

5 TextLib. A Cutter mező kitöltése. https://www.textlib.hu/html/cutter.htm Hozzáférés: 2023.07.31.

 DOI: 10.31915/NWS.2023.23

Tudománymetriai műhely könyvtári környezetben

Zsiborács Judit
zsiboracs.judit@uni-pannon.hu

Dési Ádám Dániel
desi.adam.daniel@uni.pannon.hu

Nagy Attila Árpád
nagy.attila.arpad@uni-pannon.hu

Urbán Katalin
urban.katalin@uni-pannon.hu
Pannon Egyetem, Egyetemi Könyvtár és Tudásközpont

Absztrakt

A *Pannon Egyetem Könyvtár és Levéltár* 2021 végén *Egyetemi Könyvtár és Tudásközponttá* alakult. A szervezetfejlesztés során új szolgáltatások és tevékenységek kerültek feladatkörébe, többek között kutatástámogatási feladatok is. Az átalakulás keretében intézményünk szolgáltatási spektrumát számos tevékenységi körrel bővítettük. Ezek egyike a Tudománymetriai Műhely létrehozása volt.

A Műhely mindennapi munkájában jelen cikk négy szerzője vesz részt, akik szakmai hátterüket tekintve teljesen eltérő környezetből érkeztek, ennek megfelelően a Műhely munkájához is eltérő szempontokkal tudnak hozzájárulni.

Feladataink közé tartozik az Egyetem kiválósági profiljának gondozása, melynek egyaránt része a leíró tudománymetriai statisztikai adatok kezelése, hazai és nemzetközi felsőoktatási rangsorok követése és elemzése, valamint a saját kvalitatív kutatások lebonyolítása. Ezen kívül a tudományos teljesítmény monitorozása, a fenntartó által meghatározott minőségi teljesítménymutatók összegyűjtése, elemzések és jelentések készítése az Egyetem vezetői számára, a döntéshozatal előkészítése, alátámasztó adatok biztosítása.

A modellváltás következtében átalakuló felsőoktatási térben kiemelt figyelemmel kísérjük az intézményhez köthető tudományos publikációk trendjeit, valamint azok hatását a hivatkozások nyomon követésével. Számos kihívással is meg kellett küzdenünk a Műhely létrejötte óta: miképpen lehet integrálni a Műhely tevékenységét egy egyetemi könyvtár életébe, kihívások a kollaboráció területén, új kapcsolatrendszer kiépítése, akár egyetemi, akár országos szinten.

Abstract

At the end of 2021, the *Library and Archives* of the University of Pannonia was transformed into a *University Library and Knowledge Centre*. In the course of the organizational development, new services and activities were added to its responsibilities, including scientific metrics and research support tasks. As part of the transformation, we expanded the scale of our services with a number of new activities. One of these was the establishment of the „*Scientometrics Team*”.

CC BY

The four co-authors of this article create the Team. Coming from completely different environments and professional backgrounds, the Team members can contribute to the work with very different aspects.

Our tasks include monitoring the University's excellence profile, which includes the management of descriptive scientific statistical data, tracking and analysing national and international higher education rankings, and conducting our own qualitative research. In addition, we monitor the academic performance, collect the KPIs determined by the Ministry and prepare reports for the University Management to support data-driven decision-making.

In the rapidly changing HE ecosystem, we pay special attention to the trends of scientific publications linked to the University, as well as the scientific impact by monitoring the citations.

We have faced many challenges since the Team had been set up like the integration the new activities into the daily routine of our university library, or challenges of collaboration during the building a new network of contacts, either at the university or national level.

I. Szakkönyvtár egy modellváltó egyetemen

Az 1949-ben alapított Veszprémi Vegyipari Egyetem könyvtára egészen 1966-ig a Budapesti Műszaki Egyetem Központi Könyvtárának tagkönyvtáraként működött, elsődlegesen vegyipari, kémiai szakkönyvtárként, majd 1990 után, az Egyetem több karúvá válásával folyamatosan bővült gyűjtőköre. Mára egy széles spektrumú, közel 250 000 nyomtatott kiadványt és számos online elérhető, digitális szakirodalmi forrást tartalmazó állománya szolgálja ki az egyetemi polgárokat. 2009-ben egyetemünk alapító tagként csatlakozott a Magyar Tudományos Művek Tára kezdeményezéshez, és a 4-es szintű intézményi adminisztrátor feladatait a könyvtár munkatársa látta el. Az Egyetemhez köthető tudományos közlemények szisztematikus nyilvántartása és a szerzői profilok értő, naprakész gondozása mindig kiemelten fontos feladat volt szakkönyvtárunkban.

A 2020-ban bekövetkezett modellváltás azonban az egyetemi könyvtár számára is sok változást hozott. A *Pannon Egyetem Könyvtár és Levéltár* 2021 végén *Egyetemi Könyvtár és Tudásközponttá* alakult. A névváltás alapvető koncepcionális és strukturális változásokat takar. Egyetemünk 2022-ben csatlakozott a Nemzeti Open Science Állásfoglaláshoz, és a nyílt tudomány alapelveinek gyakorlati alkalmazása sok szálon kötődik a könyvtárunk tevékenységéhez.[1]

A könyvtári szaktájékoztatás megújítása, az integrált könyvtári rendszer fejlesztése, a gyűjtőkör bővítése és muzeális különgyűjtemény létrehozása egyaránt része az új stratégiának. Az egyetemi kiadó tevékenységének teljes újragondolása, illetve a levéltári szolgáltatások fejlesztése és átfogó digitalizálási program elindítása mind az újdonságok közé tartoznak. Ebbe az ambiciózus fejlesztési tervbe illeszkedik a Tudománymetriai Műhely létrehozása is.

II. Új típusú együttműködési modellek

Ma már világszerte jellemző, hogy az egyetemi szakkönyvtárak központi szerepet kapnak a felsőoktatási intézmények életében.[2] Vajon fel vagyunk készülve idehaza egy teljesen új

1 Urbán „Előtérben az Open Science: Az NKFIH előadása a 3. Nyílt Tudományos Fórumon"

2 Zemsky, Wegner „Changing Roles of Academic and Research Libraries"

típusú szolgáltatás színvonalas biztosítására? Rendelkezünk-e azzal a tudásbázissal, ami az új feladatok szakszerű ellátásához szükséges?[3]

A modellváltás következtében átalakuló felsőoktatási térben kiemelt figyelemmel kísérjük az intézményhez köthető tudományos publikációk trendjeit, valamint azok hatását a hivatkozások nyomon követésével.[4] Át kellett gondolnunk, hogyan lehet integrálni a Műhely tevékenységét egy egyetemi könyvtár életébe.

Az egyetemi szakkönyvtárban „hagyományosnak" tekintett oktatás- és kutatástámogató szakirodalmi tájékoztatás, szakdolgozati konzultációs és tanulás-kutatásmódszertani segítségnyújtás már jól bejáratott gyakorlatán túl, számos kihívással is szembesültünk az egyetemen belüli kollaboráció területén. Más szervezeti egységektől szerzünk adatokat, ugyanakkor mi is szolgáltatunk adatokat más szervezeti egységeknek, illetve az Egyetem és a fenntartó Alapítvány vezetőinek, ahogyan az 1. számú ábra illusztrálja. Ahhoz, hogy ez a kommunikáció hatékonyan működjön, teljesen új típusú kapcsolatrendszer kiépítésére volt/van szükség.



1. ábra: Új típusú együttműködések az egyetemen

Műhelyünk gyűjti, rendszerezi, kezeli és elemzi azokat az adatokat, melyeket az Egyetem más szervezeti egységei szolgáltatnak. A munkánkhoz szükséges adatokat a következő szervezeti egységektől gyűjtjük össze: Fejlesztési és Projektigazgatóság, Gazdasági Igazgatóság, Humán Erőforrás Igazgatóság, Karrier Iroda, Minőségfejlesztési Központ, Nemzetközi Főosztály. Az adatigénylés és adatszolgáltatás módszertana a kezdeti kihívásokat követően mostanra kialakult, megtaláltuk a kapcsolattartókat, és közösen létrehoztuk a szükséges nyomtatványokat és táblázatokat.

Az adatokat a munka következő fázisában ellenőrizzük, elemezzük, majd szolgáltatjuk tovább azon szervezeti egységek számára, amelyek az Intézmény egésze tekintetében a döntéshozatalra jogosultak: a Rektori Kabinet, a Kancellári Kabinet, a Szenátus, illetve a Pannon Egyetemért Alapítvány számára. Bizonyos szervezeti egységek tekintetében ez a folyamat kétirányú: például a karok és egyetemi központok, vagy az Oktatási Igazgatóság esetén.

3 Mohammad „Adapting to change in academic libraries"
4 Varga „A szakkönyvtári innováció lehetséges útjai"

CC BY

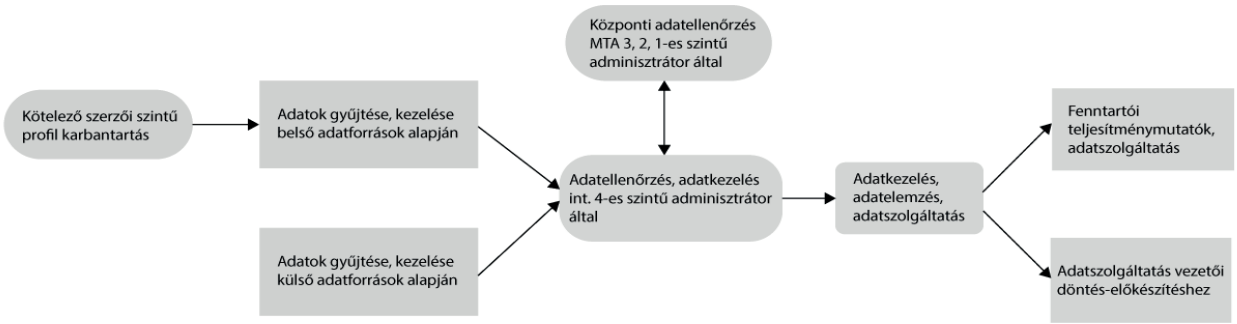
III. A Műhely napi feladatai

Feladataink közé tartozik az Egyetem kiválósági profiljának gondozása, melynek egyaránt része a leíró tudománymetriai statisztikai adatok kezelése, hazai és nemzetközi felsőoktatási rangsorok követése és elemzése, valamint a saját kvalitatív kutatások lebonyolítása. Célunk a tudományos teljesítmény monitorozása, a fenntartó által meghatározott minőségi teljesítménymutatók összegyűjtése, majd ezek alapján elemzések készítése az egyetem vezetői számára, a döntéshozatal előkészítése, alátámasztó adatok biztosítása.

A Műhely mindennapi munkájában négy kolléga vesz részt, akik szakmai hátterüket tekintve eltérő környezetből érkeztek, ennek megfelelően a Műhely munkájához is eltérő szempontokkal tudnak hozzájárulni a stratégia és módszertan, a kvalitatív elemzések és narratív szociológia, a kvantitatív adatelemzés és adatvizualizáció területén, valamint hozzánk tartoznak az MTMT intézményi 4-es szintű adminisztrátori feladatok. Mivel a Műhely elsődleges célja a Pannon Egyetem tudományos tevékenységének támogatása, igyekszünk ezt a feladatot a négy teljesen eltérő nézőpont szimbiózisával a lehető legszéleskörűbben kiszolgálni.

MTMT szerzői profilok gondozása

Az intézményi, 4-es szintű adminisztrátor fő feladata a Pannon Egyetemen az MTMT adatbázisba bekerülő hatalmas adatmennyiség ellenőrzése, illetve, az adatok kezelése, a 2. számú ábrán bemutatott munkafolyamat szerint. Munkája során az intézményi adminisztrátor folyamatos kapcsolatban áll a központi (MTMT 3, 2, 1-es szintű) adminisztrátorokkal.



2. ábra: MTMT szerzői profilok gondozása

Az adatgyűjtés, adatkezelés folyamata során megkülönböztetünk - az intézmény szempontjából - külső és belső adatforrásokat. Belső adatforrásokként tekintünk az intézmény szerzőire, valamint az egyes karokon dolgozó 5-ös, 6-os szintű kari, intézeti, tanszéki adminisztrátorokra, akikkel napi kapcsolatban állunk. Tőlük származnak azok az adatok, melyek bekerülnek az MTMT adatbázisba, és melyeket munkánk során ellenőrzünk.

Külső adatforrások alatt azokat az adatbázisokat értjük, amelyekből közvetlenül emelünk be adatokat a rendszerbe. Ezek közül a legfontosabbak a Web of Science, a Scopus, a Google Scholar, valamint a Dimensions adatbázisok.

Szükségesnek látjuk rendszeres belső képzések szervezését az 5-ös és 6-os szintű adminisztrátorok számára, annak érdekében, hogy ők is megismerjék az aktuális fókuszpontokat, tisztában legyenek a legfontosabb indikátorokkal. Az ilyen alkalmakon lehetőség nyílik a jó gyakorlatok kialakítására, cseréjére. Mindenképpen hangsúlyozni kell a

teljességre törekvés elvét (pl. önhivatkozások esetében), de nem szabad elfeledkezni arról sem, hogy melyek azok a prioritások (pl. közlemények esetében a tárgyév és a megelőző két év, hivatkozások esetében, a tárgyév és az azt megelőző öt éves időablak publikációi) amelyek előnyt élveznek az adatfelvitelkor.

Külső adatforrások kapcsán felmerülő feladatok

A Web of Science/InCites (Clarivate), mint külső adatforrás, igen hangsúlyos szerephez jutott a modellváltó egyetemek teljesítményének értékelése során a hivatkozások elvárt számára vonatkozó indikátor kapcsán. A Web of Science hivatkozáskezelő, szigorú minőségi szűrőket alkalmazó bibliográfiai adatbázis és az ennek adattömegére épülő InCites elemző rendszer a legrangosabb, szakmailag lektorált folyóiratok metaadatait tartalmazza, illetve ezek alapján teszi lehetővé elemzések készítését.

A Dimensions (Digital Science), mint külső adatforrás használata szintén a napi munkát könnyíti meg, segítségével az MTMT rendszerébe a Q1, Q2 közlemények szinte már a megjelenés pillanatában beemelhetők a DOI azonosítók segítségével. A Dimensions egy átfogó kutatási támogatásokkal foglalkozó adatbázis, amely teljesen más elven épül fel, mint a Web of Science vagy a Scopus, mivel a CrossRef rendszerén alapul: nem alkalmaz minőségi szűrőt, viszont mindent indexál, aminek DOI azonosítója van. Több témakörben, szélesebb időbeli és publikációs forráslefedettséget biztosít, mint a hagyományos bibliográfiai adatbázisok, továbbá lefedettségében közelebb áll az olyan ingyenes, aggregált adatforrásokhoz, mint a Lens, vagy a Google Scholar.

A Dimensions alkalmazása hasznos, mivel folyamatosan frissül, gyakran itt lehet megtalálni leghamarabb a Pannon Egyetemhez köthető publikációk adatait, beleértve az Open Access státuszt is. A fenntartói indikátor elvárások miatt a D1, Q1, Q2 SJR minősítésű publikációk MTMT-be történő minél előbbi bekerülése az elsődleges prioritás. Ez pedig leghamarabb a Dimensions adatbázisból valósulhat meg, mivel a Web of Science és a Scopus csak jóval később indexál, és nem is minden SJR értékelésű közleményt.

A hivatkozások tekintetében is számos aktuális feladatunk van, a jelenlegi hazai helyzetet vizsgálva. A hivatkozások száma az MTMT-ben gyakran kevesebb, mint a Scopusban. Ez a probléma minden modellváltó egyetemet érint, hiszen a közfinanszírozási szerződésekben elvárt tudományos teljesítménymutatók kizárólag a közhiteles adatforrásnak számító MTMT-ből tölthetők le. “Az MTMT-ben tárolt közleményadatok hitelességéért a szerzők és munkáltató intézményeik felelnek. Az adatbázisban az eredeti közleménnyel összevetett, hitelesített adatok már nem változtathatóak meg (a hibák javítására természetesen lehetőség van). A hiteles adatszolgáltatás egyik garanciája a nyilvánosság. Egyrészről a kutatók látják – és módosíthatják – saját publikációs adataikat, de látja azt mindenki más is. Látják (és módosíthatják) a társszerzők, látják az intézményi vezetők, a pályázatok bírálói, a fenntartók és finanszírozók, a <<versenytársak>> és a többi intézmény is.”[5]

Célszerűnek tűnik, hogy ezt a problémát az egyes intézmények valamilyen módon házon belül próbálják meg megoldani. Mindenképpen számolni kell azzal, hogy egy egyetem szintjén ez hatalmas adathalmazt jelent, így a szerzőnkénti és közleményenkénti, tehát manuális idézőimport nem járható út, vagy csak részleges megoldást nyújt a problémára. Szükségesnek tűnik valamilyen alternatív megoldást kidolgozni a munkafolyamat automatizálására, akár API alapú lekérdezések használatával.

5 Holl „A Magyar Tudományos Művek Tára – alapvető információk és működési alapelvek”

CC BY

Munkafolyamataink további lényeges eleme az összegyűjtött adatok elemzése, valamint szolgáltatása az egyetemi vezetők számára, mintegy döntéselőkészítési céllal. Az általunk lekeresett adatokat archiváljuk is, hiszen az MTMT adatbázisa nem alkalmas idősoros adatmegjelenítésre, miközben a döntéshozók számára az egyes trendek alakulása is meghatározó jelentőségű információval bírhat. A rendelkezésünkre álló adatokból statisztikákat készítünk.

Az Egyetem fenntartásának módját és feltételrendszerét tartalmazó közfinanszírozási szerződésben szerepelnek a Minisztérium által pontosan meghatározott minőségi mutatószámok, amelyek között sajátos helyet foglalnak el a tudományos teljesítményre és kiválóságra vonatkozó indikátorok. Az általunk kidolgozott, egyszerű, táblázatos formában megjelenített adatszolgáltatás segítségével az egyetemi vezetők kijelölhetik az Egyetem adott időszakra vonatkozó vállalásait, és összevethetik a célértékekkel, valamint kijelölhetik a beavatkozási pontokat.

Szerzői profilok gondozása a nemzetközi adatbázisokban

A Web of Science és Scopus hivatkozáskezelő adatbázisokban található nagy mennyiségű adattömeg mellett nem meglepő, hogy egyre több szerzőnek okoz problémát a szerzői profiljába keveredett idegen cikk, vagy az, hogy több néven szerepel az adatbázisokban. A bibliográfiai adatbázisok által alkalmazott algoritmusok a közleményekben szereplő metaadatok alapján társítják azokat a szerzői profilokhoz. Ha a rendszer nem tud hozzácsatolni egy publikációt egy szerzői profilhoz, úgy automatikusan új profilt hoz létre, emiatt fordulhat elő az, hogy egy szerző több profillal is rendelkezik. Ebben az esetben a szerzői profil nem a valóságnak megfelelően reprezentálja a szerző tudományos teljesítményét, a teljesítményértékelések pedig hibás eredményeket adhatnak.



3. ábra: Szerzői profiltisztítás a nemzetközi adatbázisokban

A 3. számú folyamatábra illusztrálja, hogy a szerzői profilok ellenőrzése során megvizsgáljuk, helyes-e az intézményi affiliáció, illetve nem rendelkezik-e a szerző több profillal, továbbá, hogy aktív-e és helyesen van-e vezetve a profil. Esetleges hiba esetén felhívjuk a szerző figyelmét a fennálló problémára, és igény szerint segítünk a javításban.[6] Célunk, hogy a Pannon Egyetem szerzőinek a nagy nemzetközi adatbázisokban megjelenő profiljai minél teljesebb képet mutassanak a szerzők tudományos munkásságáról.

6 Adams, Pendlebury, Potter, Rogers „Unpacking research profiles: Moving beyond metrics”

Kvalitatív kutatásunk

A statisztikai elemzések mellett törekszünk intézményünk teljes körű leírására, ennek érdekében indítottuk el kvalitatív kutatásunkat. A felmérésünk célja az Egyetem kiválósági profiljának kialakítása, melynek alapja az egyetemi polgárokkal felvett interjúk. A holisztikus szemlélet érdekében egyaránt törekszünk arra, hogy az aktív hallgatók mellett az Egyetem alumni közösségét, valamint a felsőoktatási felmérésekben méltatlanul hanyagolt nem-akadémiai területen dolgozó kollégáinkat is megszólítsuk.

A kutatásunk egyaránt mutat rá lehetséges fejlesztési pontokra az Egyetem vezetősége számára, valamint akadémiai oldalról szociológiai elemzésként tágabb értelemben vett szakmai párbeszédekre is lehetőséget nyújt.

IV. Tapasztalataink

Az elmúlt másfél év során szerzett tapasztalataink alapján elmondható, hogy sikeresen vettük az első akadályokat. Megértettük, hogy tevékenységünk nagy felelősséggel jár, és ez másfajta, eddig nem ismert elvárásokat is hozott magával.[7]

Úgy gondoljuk, hogy vannak még feladataink, úgy az adatforrások minél hatékonyabb használata, mint a munkamegosztás racionalizálása tekintetében. De a legfontosabb a kommunikáció: szeretnénk támogatni egyetemünk szerzőit annak felismerésében, hogy a megváltozott működési modell és finanszírozási struktúra által támasztott új típusú követelmények sikeres teljesítése érdekében elengedhetetlen, hogy mindenki a teljességre törekvően tartsa nyilván tudományos közleményeit és azok hatását a hivatkozások pontos és naprakész rögzítésével, hiszen ez nem csak egyéni, hanem intézményi érdek is egyúttal.

Előremutató lenne, hogyha a modellváltó egyetemek szakkönyvtárai együttműködnének ezen az új területen is, és a jó gyakorlatokat rendszeres workshopokon megosztva és elemezve segíthetnék egymást a minél magasabb szolgáltatási színvonal elérésében.[8]

Irodalomjegyzék:

Adams, J., Pendlebury, D., Potter, R., & Rogers, G. (2023). *Unpacking research profiles: Moving beyond metrics*. https://doi.org/10.14322/isi.grr.unpacking.research.profiles

Aslam, M. (2022). Adapting to change in academic libraries. *Global Knowledge, Memory and Communication*, *71*(8/9), 672–685. https://doi.org/10.1108/GKMC-04-2020-0053

Holl, A. (2021). A Magyar Tudományos Művek Tára – alapvető információk és működési alapelvek.*MagyarTudomány*,*182*(1),81–89.https://doi.org/10.1556/2065.182.2021.1.12

Kovácsné Koreny, Á. (2022). *Az ötlettől a megvalósításig*. Könyvtári Intézet.

Urbán, K. (2022). Előtérben az Open Science: Az NKFIH előadása a 3. Nyílt Tudományos Fórumon. *Tudományos És Műszaki Tájékoztatás*, *69*(3), 104–108. https://doi.org/10.3311/tmt.13151

Varga, K. (2008). A szakkönyvtári innováció lehetséges útjai. *Könyv és Nevelés*, *10*(4).

Zemsky, R., & Wegner, G. (2018). *Changing Roles of Academic and Research Libraries*. Association of College and Research Libraries American Library Association. https://www.ala.org/acrl/issues/value/changingroles

7 Kovácsné Koreny „Az ötlettől a megvalósításig”

8 Zemsky, Wegner „Changing Roles of Academic and Research Libraries”

V.04
DOI: 10.31915/NWS.2023.24

A Digitális Örökség Nemzeti Laboratórium webszolgáltatásai automatikus kézírás-felismertetéshez

Web services of the Digital Heritage National Laboratory for automatic handwriting recognition

Palkó Gábor
Eötvös Lóránd Tudományegyetem, Digitális Bölcsészet Tanszék
palko.gabor@btk.elte.hu

Szekrényes István
Eötvös Lóránd Tudományegyetem, Digitális Bölcsészet Tanszék
Debreceni Egyetem, Filozófia Intézet
szekrenyes.istvan@btk.elte.hu

Bobák Barbara
Bölcsészettudományi Kutatóközpont, Irodalomtudományi Intézet, DigiPhil
bobak.barbara@abtk.hu

Absztrakt

Fejlesztési projektünk célja, hogy a csak kézírásos formában elérhető gyűjtemények feldolgozásához egy olyan ingyenesen használható, nyílt hozzáférésű eszközökre épülő platformot biztosítson, amellyel az eredetileg képként tárolt anyagokból kereshető, digitális feldolgozásra valóban alkalmas dokumentumok hozhatók létre. A kézírás felismertetésére a TrOCR eszközhöz elérhető alapmodelleket finomhangoltunk magyar, illetve latin nyelvre a *Transcribus* szolgáltatásával összehasonlítható eredménnyel. Jelenleg három modell áll rendelkezésre: a 900 oldalnyi, többszerzős Arany János levelezésen és hivatali iratokon tanított modell magyar nyelvű szövegekhez, a *Rerum Ungaricarum Libri* korpusz 200 oldalán tanított modell latin nyelvű kódexekhez, a Magyar Nemzeti Levéltártól kapott 200 oldalnyi anyagon tanított modell pedig levéltári iratok feldolgozásához. A szolgáltatás egy webes interfészen és Rest API-n keresztül is igénybe vehető.
Kulcsszavak: kézírás-felismertetés, TrOCR, Kraken, Alto-XML

Abstract

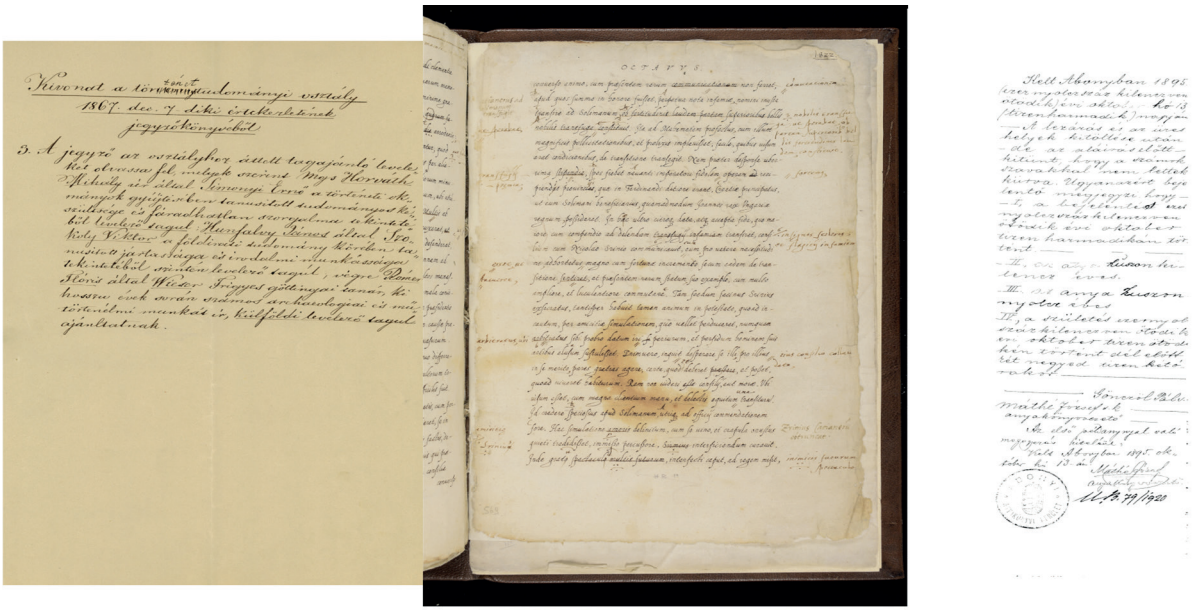
The aim of the development project is to provide a free and open-access platform for processing collections available only in handwritten form. This platform enables the creation of searchable and digitally processable documents from materials originally stored as images. We fine-tuned basic models available in the TrOCR tool for handwriting recognition in Hungarian and Latin languages, with results comparable to the *Transcribus* service. Currently, three models are available: one trained on a 900-page, multi-author collection of correspondence and official documents by Arany János for Hungarian texts, one trained on 200 pages of the *Rerum Ungaricarum Libri* corpus for Latin codices, and one trained on 200 pages of archival documents obtained from the Hungarian National Archives. The service can be accessed through a web interface and via a REST API.
Keywords: HTR, TrOCR, Kraken, Alto-XML

1. Bevezetés

A magyar közgyűjtemények egy tetemes hányada csak kézírásos formában érhető el. A korszerű NLP technológiák (lekérdezések, tartalomkivonatolás, entitásfelismerés, wikifikáció stb.) alkalmazásához az eredeti dokumentumok szkennelt képként történő tárolása önmagában nem elégséges. A szövegek tényleges digitalizálása manuálisan rendkívül időigényes feladat, a jelenleg elérhető, valóban hatékony HTR szolgáltatások pedig költségesek. Fejlesztési projektünk célja, hogy olyan, kizárólag nyílt hozzáférésű, ingyenesen használható eszközökre épülő alternatívát biztosítson, ami a *Transcribus* (Kahle et al, 2017) szolgáltatásaihoz hasonló, minél kevesebb kézi korrekciót igénylő minőségben teszi lehetővé a gyűjtemények digitalizálását, XML és kétrétegű PDF fájlok előállítását. A projekt négy fő részfeladatra tagolódott: (1) saját tanítóanyagok előállítása és előfeldolgozása, (2) a TrOCR keretrendszerhez elérhető transzformációs modellek finomhangolása magyar és latin nyelvű kéziratok felismertetéséhez, (3) a dokumentumok szövegsorokra szegmentálásának automatizálása (4) online alkalmazás implementálása, ami lehetővé teszi, hogy a láncba szervezett szegmentáló és felismerő modul saját weboldalunkon kívül egy REST API-n keresztül is elérhető és integrálható legyen. A fejlesztéseket a Magyar Nemzeti Levéltárral, az ELTE IK Mesterséges Intelligencia Tanszékével és a Monguz Információtechnológiai Kft-vel együttműködésben végeztük.

2. Tanítóanyagok előállítása

A HTR modelljeink betanításához 3 korpusz állt rendelkezésre. Ebből az egyiket a Nemzeti Levéltár bocsátotta rendelkezésre, amely 300 oldal anyakönyvi iratot tartalmazott két anyakönyvvezető kézírásával. A második 900 oldalnyi kéziratot tartalmazó korpusz Arany János levelezéséből és hivatali irataiból (Arany, 1859–77) állt össze, Arany János mellett 4 további szerző kézírásával (Bobák & Gábori Kovács, 2019). A harmadik korpusz Gian Michele Bruto (latinos alakban használt nevén Brutus) *Rerum Ungaricarum Libri* latin nyelvű munkájának 200 oldalából állt (Bobák & Kasza, 2019). A dokumentumképek szöveges átírása mindhárom esetben a *Transcribus* szolgáltatással, a korábban betanított, publikusan elérhető magyar nyelvű modellek segítségével történt. Az automatikus átírás hibáit kollégáink manuálisan



1. ábra: Minták a tanítóanyagokból. Balról jobbra: Arany János levelezés, Brutus, anyakönyvi iratok (MNL)

CC BY

javították (az anyakönyvi iratok esetében a feladatot a Magyar Nemzeti Levéltár munkatársai végezték). Az eredményeket Alto-XML formátumban exportáltuk, amely a szövegek mellett azok pozícióját, a szövegkép elrendezésére vonatkozó információkat is tárolja. Mivel a TrOCR modellek finomhangolása ezt megköveteli, az eredeti képeket az XML fájlokban tárolt adatok alapján egy Python szkripttel szó, illetve szókapcsolat-szintű egységekre szegmentáltuk. A tanítóanyag előkészítésének eredményeként tehát képszegmentumokat és egy tabuláris szerkezetű szövegfájlt kaptunk, ami a szövegrészleteket a kapcsolódó képfájllal összekötve reprezentálja a tanításhoz szükséges szöveg-kép párokat.

3. TrOCR modellek finomhangolása

A TrOCR egy 2021-ben publikált (Li et al, 2021), Microsoft által fejlesztett, ingyenesen használható keretrendszer,[1] amely az addigi CNN és RNN alapú OCR technológiákat egy képkódoló és szövegdekódoló részből álló transzformációs modellel váltotta fel (a dekódoló rész Bert-típusú nyelvi modellekkel is inicializálható). A rendszer használatához nagy méretű, szintetizált anyagokat is tartalmazó, angol nyelvű kéziráson előre-tanított alapmodellek érhetők el, amelyek saját szöveg-kép párokon tovább finomhangolhatóak. A finomhangoláshoz az előző szakaszban bemutatott korpuszokat használtuk fel. A tanítás egy A-100 GPU segítségével a kisebb korpuszokon néhány órát, de a 900 oldalas Arany János korpusz esetében sem vett fél napnál több időt igénybe. A HTR modellek pontosságát a tanítóanyagok 5 százalékán teszteltük. A tanítás eredményeit karakter-hibaarányban (CER: Character Error Rate) az 1. táblázat tartalmazza.

Modell rövid neve	Korpusz	Méret	CER (TrOCR)	CER (Transcribus)
Arany900	Arany János levelezés és hivatali iratok	900 oldal	5.86%	9.30%
Brutus	Rerum Ungaricarum Libri	200 oldal	2.03%	1.90%
MNL	A Magyar Nemzeti Levéltár anyakönyvi iratai	300 oldal	2.77%	N/A

1. táblázat: A finomhangolt TrOCR modellek pontossága

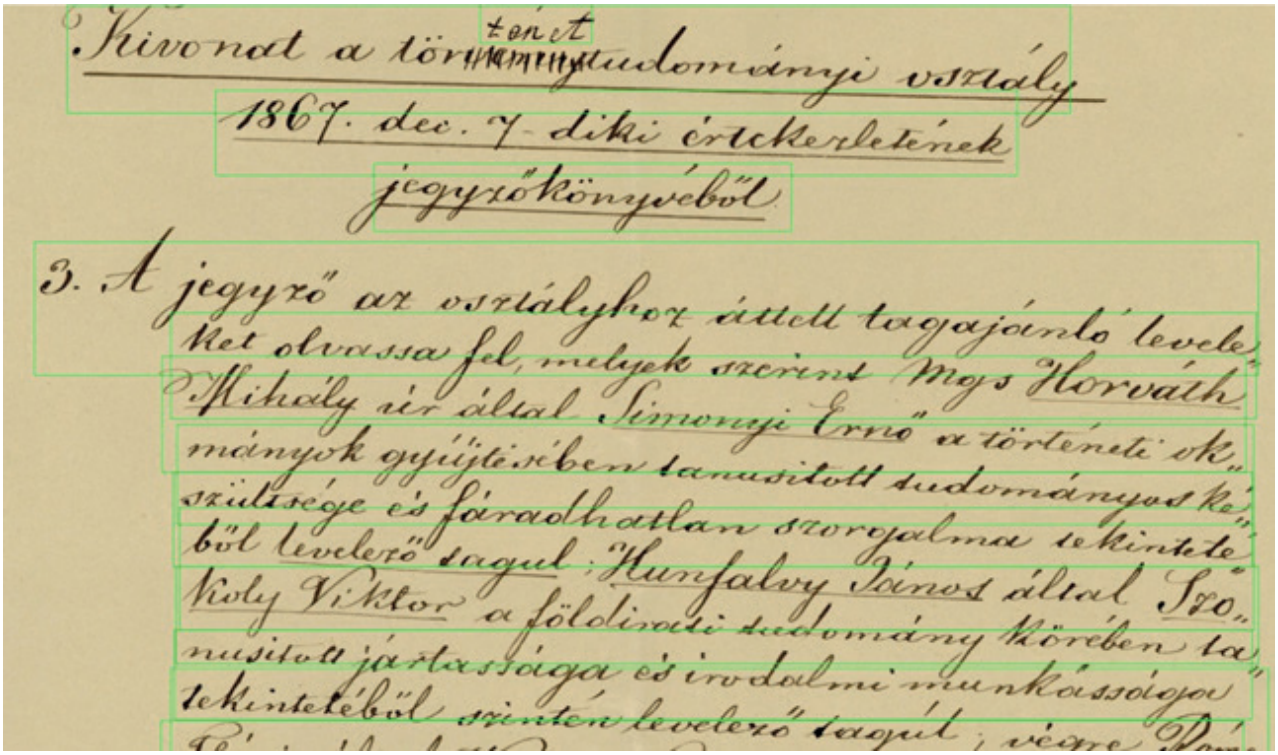
A legjobb eredményeket a *Rerum Ungaricarum Libri* korpuszon értük el, amely a felhasznált kézírás homogenitásának (egyetlen szerző) és gondos, szinte a nyomtatott szöveg minőségét megközelítő kivitelezésének könyvelhető el. Általános célokra a legnagyobb és a legtöbb szerző kézírását tartalmazó anyagon finomhangolt *Arany900* modellt használjuk.

4. A dokumentumok automatikus szegmentálása

Mivel a TrOCR modellek csak szövegsor-szintű képek feldolgozására képesek, a webszolgáltatások teljessé tételéhez szükséges volt olyan szövegkép-elemző és szegmentáló modulok integrálására, amelyekkel a bemenetként használt szövegoldalak kisebb egységekre darabolhatók. Az első feladathoz két, a használat során opcionálisan választható, nyílt

1 A TrOCR GitHub repozitóriuma: https://github.com/microsoft/unilm/tree/master/trocr

hozzáférésű eszközt, a *Transkribus*ban is használt P2Pala szövegblokk és alapvonal-detektálót,[2] illetve a Kraken HTR rendszer *blla* szegmentáló modulját[3] építettük be szolgáltatásainkba. A P2Pala PageXML formátumú kimenetet ad, ami szövegblokkok és az alapvonalak koordinátáit tartalmazza. A Kraken szegmentálója JSON formátumban adja vissza a szövegsorok poligonjait. Alkalmazásainkban ezeket a kimeneteket egy egységes objektum-formára konvertáljuk, ami a poligonok (a P2Pala esetében az átlagos alapvonal-távolság alapján számolunk egy becsült sormagasságot az alapvonalakra) befoglaló négyszögeinek koordinátáival reprezentálja a szövegsorokat (lásd 1. ábra).



2. ábra: A szövegszegmentálás grafikus kimenete

A TrOCR modellek a koordináták alapján kivágott képdarabokat kapják bemenetül. Az aktuális sor alá és fölé esetlegesen belógó karakterek elmaszkolása nem szükséges (a P2Pala esetében erre nem is lenne lehetőség) , a szövegfelismerést nem zavarják meg.

5. Webszolgáltatások

Webszolgáltatásaink a szövegszegmentáló és a szövegfelismerő modulon túlmenően több utófeldolgozó eljárást is tartalmaznak, amelyek a TrOCR szimpla szöveges kimenetéből Alto-XML dokumentumokat, kétrétegű PDF fájlokat generálnak, illetve elérhetővé teszik a szövegszegmentálás grafikus kimenetét is a sorokra-bontás pontosságának ellenőrzéséhez.

Az elsősorban demonstrációs céllal készült webes interfész[4] a Python *gradio* moduljával készült. Mint a 3. ábrán látható a bemeneti oldalon a felhasználó választhat az általunk készített három kézírás-felismerő modell (Arany900, Brutus, MNL) közül, illetve beállíthatja, hogy melyik integrált szövegkép-elemző modul (Kraken vagy a P2Pala) segítségével kívánja a szövegkép szegmentálását elvégezni. A kimeneti oldalon pedig a kívánt formátumban megtekintheti, letöltheti az eredményfájlt.

2 https://readcoop.eu/transkribus/docu/p2pala/
3 https://kraken.re/4.0/api_docs.html
4 http://mobydick.elte-dh.hu:42005/

CC BY



3. ábra: Pillanatkép a webes interfészről

A Python *flask* moduljával implementált Rest API a „/HTR/” végponton vár feltöltésre képfájlokat, a kimenetet pedig a kérésben definiált formátumban (szimpla szöveg, Alto-XML, kétrétegű PDF, képfájl, vagy ZIP archívum, ami mind a négyet tartalmazza) adja vissza sikeres feldolgozás után.[5] A kéréshez egy, a kérés fejlécében elhelyezett API kulcs elküldése szükséges. A kéréshez az alábbi paramétereket kell beállítani:

- „**model**”: a feldolgozáshoz használni kívánt TrOCR modell neve. Jelenleg elérhető modellek: „Arany900”, „Brutus”, „MNL”,
- „**segmenter**”: a szövegkép sorokra szegmentálásához használt modul kiválasztása. Jelegnek két szegmentáló érhető el: „kraken”, „P2Pala”,
- „**output_format**”: a kimenet formátuma. Elérhető formátumok: „txt”, „pdf”, „xml”, „zip”
- „**request_type**”: „sync”, vagy „async”.

A „sync” típus esetén az API a feldolgozás végéig nem szakítja meg a kapcsolatot a klienssel, az eredményfájlt pedig válaszként küldi vissza. Az „async” kéréseknél a kliens egy kérésazonosítót kap vissza, a feldolgozás állapotát és az eredményfájlt pedig külön végpontról lehet lekérdezni. Az API-ban konfigurálható, hogy összesen hány feldolgozási folyamat futhat párhuzamosan. A limiten felül érkező kérések várólistára kerülnek. A folyamatok állapotának lekérdezésére irányuló kérések („/HTR/processes/<kérésazonosító>/” végpont) az alábbi JSON formátumú válaszüzenet érkezik:

5 A Magyar Nemzeti Levéltár kérésére készült egy külön végpont („HTR-JSON”), ahol a bemenet nem képfájlként, hanem egy JSON dokumentumként kerül feltöltésre, amik a kliens oldalon előre szegmentált dokumentum egyes részeit base64 formátumban tartalmazza.

```
{
   „id”: „c7ac94da-0de7-11ee-ae05-00163edf70e9”,
   „receivedAt”: „2023-06-18 16:52:43”,
   „startedAt”: „2023-06-18 16:52:43”,
   „processingTime”: 29.06680178642273,
   „statusCode”: 2,
   „status”: „COMPLETED”,
   „input”: „1.jpg”,
   „output”: „1.xml”
}
```


Ha folyamat státusza „COMPLETED” értéket vesz fel, akkor az előre választott formátumú eredményfájl a „/HTR/processes/<kérésazonosító>/result” végpontról tölthető le. Egy szövegoldal átlagos feldolgozási ideje saját méréseink szerint átlagosan 40 másodperc.

A Rest API használatához kutatási célokra előzetes egyeztetés alapján tudunk hozzáférést és részletes dokumentációt biztosítani. A szolgáltatás az *Invenio* másodlagos interfészén keresztül is elérhető.[6]

További terveink

A TrOCR keretrendszerben eddig finomhangolt modelleket további korpuszok, illetve az ELTE IK Mesterséges Intelligencia Tanszékének közreműködésével előállított szintetizált anyagok segítségével tervezzük tovább bővíteni és fejleszteni. A szövegszegmentálás és szövegkép-elemzés hatékonyabb módszereinek feltérképezésén, fejlesztésén, a meglévő modellek kombinálásán, azok pontosságának összehasonlító kiértékelésén is dolgozunk.

Bibliográfia

Arany, János. „Hivatali iratok 2.: Akadémiai évek (1859–77)”. szerk., jegyz. Gergely Pál. Bp. Akadémiai K. (1964) (Arany János összes művei, 14.)

Bobák, Barbara és Gábori Kovács, József (2019) Kézírásfelismerés Arany János levelein. In: Networkshop 2019. HUNGARNET Egyesület, Budapest, pp. 38 – 44.

Bobák, Barbara és Kasza, Péter (2019) Az MI lehetőségei a kora újkori filológiában: Johannes Michael Brutus Rerum Ungaricarum libri kéziratának digitális kiadása. In: Networkshop 2022. HUNGARNET Egyesület, Budapest, pp. 154–160.

Kahle, P. and Colutto, S. and Hackl, G. and Mühlberger, G. (2017) „Transkribus - A Service Platform for Transcription, Recognition and Retrieval of Historical Documents,” 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Kyoto, Japan, 2017, pp. 19–24, doi: 10.1109/ICDAR.2017.307.

Li, Minghao and Lv, Tengchao and Cui, Lei and Lu, Yijuan and Florencio, Dinei and Zhang, Cha and Li, Zhoujun and Wei, Furu (2021) TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models. ArXiv, doi: 10.48550/arXiv.2109.10282

6 https://digitization.nemzetilab.qulto.eu

V.04
DOI: 10.31915/NWS.2023.25

Adatvizualizációs lehetőségek a bölcsészettudományban

Data visualisations in the Digital Humanities

Szűcs Kata Ágnes
Digitális Bölcsészeti Központ
Országos Széchényi Könyvtár
szucs.kata@oszk.hu

Abstract

In my paper, I introduce data visualisations based on textual content published on the Digital Humanities Platform (dHUpla) Creative site.

Showing the workflow how to create visualizations based on writers' correspondence helps to make information valuable to literary history available to users. In addition to the authors' correspondence, I will present the data visualisation based on the web archiving activities of the National Széchényi Library.

Presenting the Digital Humanities Centre's data visualisation toolkit can be effectively used by institutions in the GLAM sector.
Keywords: data visualisation, web archiving, text mining, correspondence, research, manuscript

1. Bevezetés

A Digitális Bölcsészeti Központ által fejlesztett Digital Humanities Platform[1] (dHUpla) egy online publikálási környezet, amely a közgyűjtemények szöveges forrásainak digitális közreadására szolgál. A kéziratos korpuszok mellett, a nyomtatott és born digital (digitálisan született) tartalmak közzététele is szerepel a szolgáltatásai között.

A dolgozat első részében a klasszikus kéziratos forrásokon alapuló adatvizualizációk elkészítéséről adok számot. Egy digitális környezetben készült forráskiadás során a szövegből adathalmaz lesz, amely megjelenítése újszerűen képes megmutatni egy-egy kéziratos korpuszt, jelen esetben írói levelezést. A digitálisan született tartalmak vizuális megjelenítése nemcsak segít a rendelkezésre álló információ értelmezésben, hanem láthatóvá tesz eddig rejtett tartalmakat is. A dolgozat második része a learatott webes tartalmak nyelvi annotálásán keresztül készült adatvizualizáció bemutatásával foglalkozik.

A dolgozatnak nem célja az ismertetett adatvizualizációk értelmezése, elemzése, inkább egy olyan gyakorlatot ismertet, amely lehetővé teszi a közgyűjtemények kutatástámogató szerepének kiterjesztését.

2. Írói levelezések

A dHUplán a Kreatív menüpont alatt publikált tartalmak egyik halmaza írói levelezéseken alapszik. A kéziratok adatainak és metaadatainak újszerű felhasználásával szeretnénk

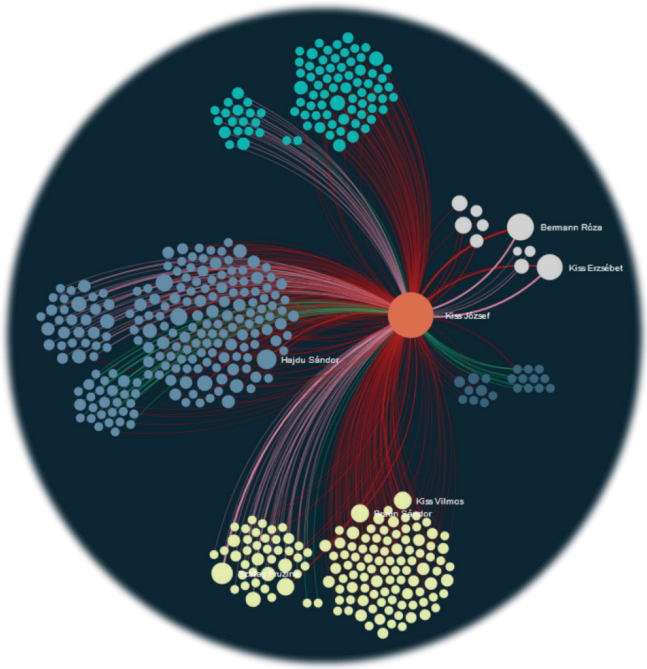
1 Digital Humanities Platform, https://dhupla.hu/ (hozzáférés: 2023. 09. 19.)

felhívni a kutatók és az érdeklődők figyelmét a közgyűjtemények kéziratanyagára. Az írói, értelmiségi levelezéseket digitális kiadás formájában publikáljuk folyamatosan. Az itt őrzött források metaadatai a célnak megfelelő adatstruktúrába rendezve kiválóan hasznosíthatók adatvizualizáció készítésére is.

Többféle adatvizualizációs eszközt is teszteltünk: kapcsolati hálók, térképes vizualizációk; kapcsolati hálók; szófelhők és stilometriai elemzések.[2] A legfrissebb tartalmak közé tartoznak a Széchényi Ferenc, a Vas István és a Kiss József kapcsolatait megjelenítő hálózatok.

A vizualizációkhoz a rendelkezésünkre álló, eddig feldolgozott kéziratanyagok adatait használtuk fel, így az elkészült munkák első fázisa lezártnak tekinthető. A kialakított digitális infrastruktúra azonban lehetővé teszi a folyamatos bővítést és javítást, így az újonnan bekerülő források adatai, további kutatások eredményei beépíthetők az adatvizualizációkba.

A továbbiakban a kapcsolati hálók közül egyet kiemelve szeretném bemutatni az elkészítése körüli munkafolyamatokat. Kiss József *A Hét* című folyóirat alapítója és szerkesztője, emellett költőként és íróként is működött a századfordulón. A levelezését feldolgozó adatvizualizáció elkészítésekor felhasználtuk a személyes és szakmai levelezését egyaránt.[3] Ez utóbbi a Nyugat folyóirat előtt formálódó modern magyar irodalom közegébe enged betekintést. A projekt első szakaszában a vizualizáció, a Petőfi Irodalmi Múzeumban őrzött levelezés alapján készült, és közös munka eredménye.[4]



1. ábra: Kiss József kapcsolati háló.[5]

2 Gephi https://gephi.org/, Stylo R csomagja, https://eadh.org/projects/stylo-r-package, Voayant Tools, https://voyant-tools.org/, Thing Link https://www.thinglink.com/, Genial.ly, https://genial.ly/.

3 Kiss József kapcsolati háló, https://dhupla.hu/s/dataviz/kissjozsef-kapcsolatihalo/ (hozzáférés, 2023. 06. 29.)

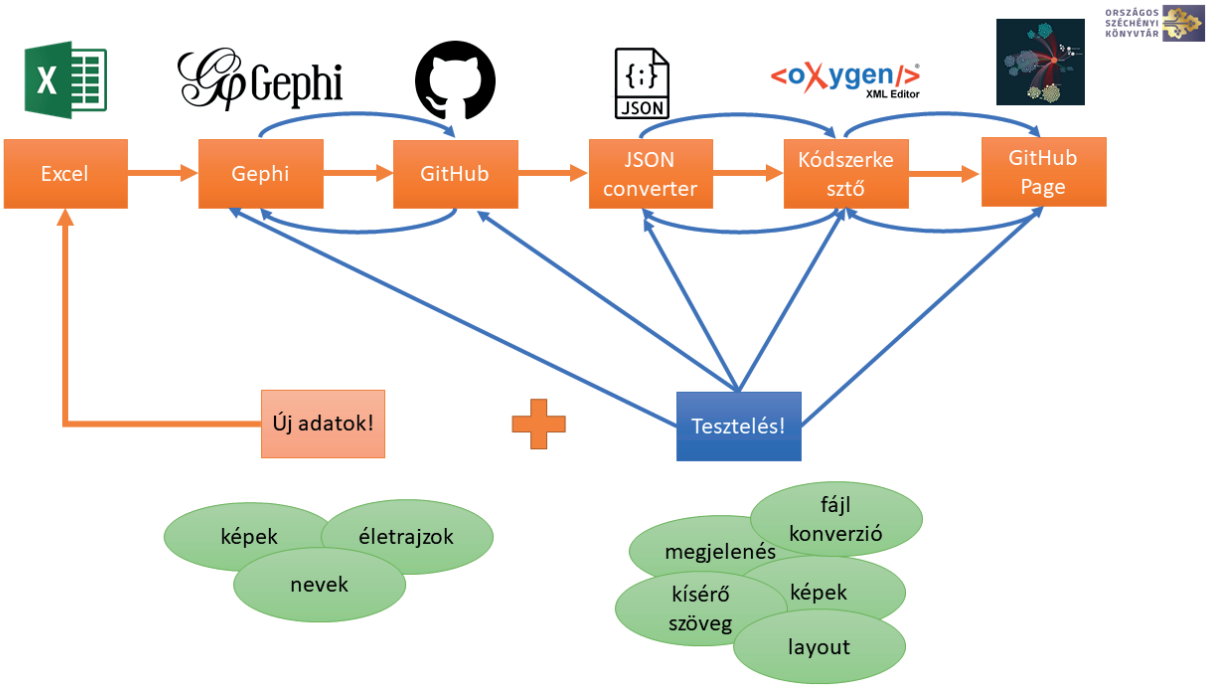
4 Ezúton is köszönöm Horváth Dánielnek.

5 A kész kapcsolati háló ismertetése a dHUplán olvasható. https://dhupla.hu/page/kreativ/kissjozsef-kapcsolatihalo (hozzáférés, 2023. 06. 29.)

CC BY

Kapcsolati háló a gráf típusú vizualizációk közé tartozik, amely a matematikai és a számítógéptudomány egyik alapfogalma; csomópontok (csúcsok) és a rajtuk értelmezett összeköttetések (élek) halmaza. A rendelkezésre álló adatokból egy egyetlen pontban összefutó, ún. fagráf hálózatot alakítottunk ki.[6]

Elkészítéséhez a Gephi[7] nyílt forráskódú, ingyenesen hozzáférhető szoftvert használtuk, a munkamenetet pedig az alábbi ábra szemlélteti.



2. ábra: Gephi Workflow

2.1 Adatok előállítása

Az első lépésben összegyűjtöttük, majd csomópontok (node) és élek (edge) szerint rendszereztük a gráfhoz felhasználandó adatokat két Excel táblázatban. A csomópontok jelen esetben a levélírók, amelyekhez több adat is kapcsolódik (egy rövid leírás, a születési és halálozási időpontok, milyen kapcsolat fűzte Kiss Józsefhez, a levelezőpartner neme vagy típusa és a levélváltások száma). Az élek ezzel szemben a csomópontok közötti kapcsolatot határozzák meg, ezen kívül pedig két további információt hordozhatnak: az egyik az él irányítottságára, a másik pedig a súlyozottságára vonatkozik. Ez a vizualizáció egy irányított gráf (bigráf), tehát különbséget tesz az „A-ból B-be", illetve „B-ből A-ba" menő élek között.

Az adatok egy része a Petőfi Irodalmi Múzeum online katalógusából[8] származik. A hiányosakat ellenőrzés és tisztítás után[9] kiegészítettük, továbbá olyan új adatokkal is gazdagítottuk, amelyeket a vizualizáción ábrázolni szerettünk volna, de a katalógusban nem voltak feltüntetve (pl. a levélíró neme és a kapcsolat jellege). Ugyanakkor érdemes figyelembe

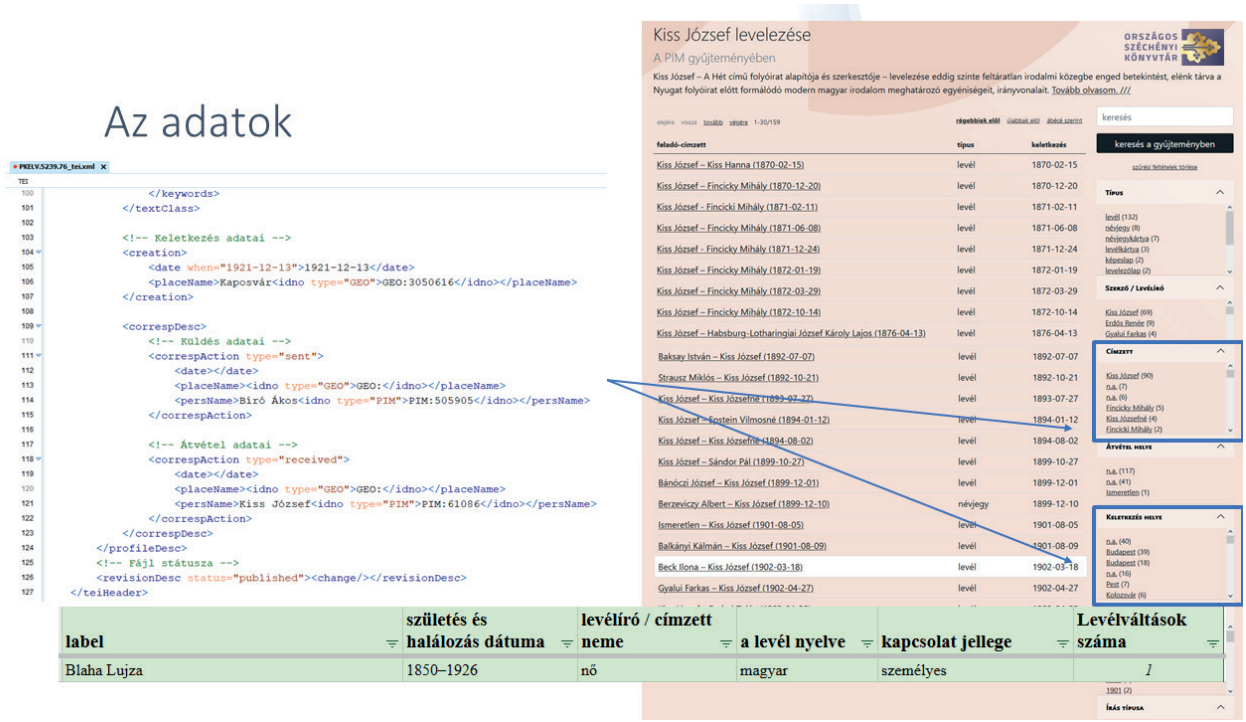
6 Ha egy gráf összefüggő és nem tartalmaz egy adott pontjába visszavezető utat (kört), akkor azt fának nevezzük. (pl.: weboldalak menüstruktúrája, vagy egy családfa is tekinthető fagráfnak.), Vö: Hajnal Péter: Gráfelmélet, Polygon, Szeged, 2003.

7 Gephi, The Open Graph Viz Platform, https://gephi.org/

8 Petőfi Irodalmi Múzeum, Gyűjtemények, https://opac.pim.hu/hu (hozzáférés: 2023.06.29.)

9 Ezúton is köszönöm Török Sándor Mátyásnak a korrektúrázásban vállalt áldozatos munkát.

venni, hogy egy digitális szövegkiadás esetében az ábrázolható adatok (pl. a levélíró és a címzett neve) a metaadatakkal ellátott TEI XML fájlokból is kinyerhetők (ld. alább).



3. ábra: Az adatok előállítása

2.2 Gephi lépések

Ezután az Excel adatbázist importáltuk a Gephi gráfvizualizációs szoftverbe. Ezen a felületen alakítottuk ki vizuálisan is a csomópontok és az élek rendszerét. A GUI-n különböző beállítások segítségével tudjuk megadni a gráf további jellemzőit. A Layout beállítása az elrendezési algoritmusok futtatását jelenti, ezek a számítások határozzák meg a gráf kinézetének alapját. Emellett lehetőség van a gráf vizuális elemeinek a finomhangolására is (pl. színek, távolságok, feliratok, betűtípusok megadása stb.).

A Gephi lehetővé teszi az adatelemzést, az objektumok közötti kapcsolatok mögöttes struktúráinak feltárását. A szoftver 2008-as eredeti megjelenést követően 2010-ben alapult meg a Gephi Consortium,[10] egy nemzetközi nonprofit vállalat, amely a működést és a verziók fejlesztését támogatja. Ehhez egy széles felhasználói közösség is hozzájárul különböző fórumokon, vitacsoportokban és blogbejegyzéseken keresztül.[11]

Ebben a széles társadalmi bázisban van a szoftver erőssége, ugyanis az alapbeállítások mellett a fejlesztők jóvoltából lehetőség van ingyenes bővítmények telepítésére, amelyek további lehetőséget nyújtanak a hálózatok megjelenítéséhez, elemzéséhez. A szoftverben létrehozott gráf alapesetben képként exportálható, viszont a SigmaExporter[12] nevű bővítmény segítségével a Gephiben megrajzolt gráfot minden paraméterével együtt exportálni lehet

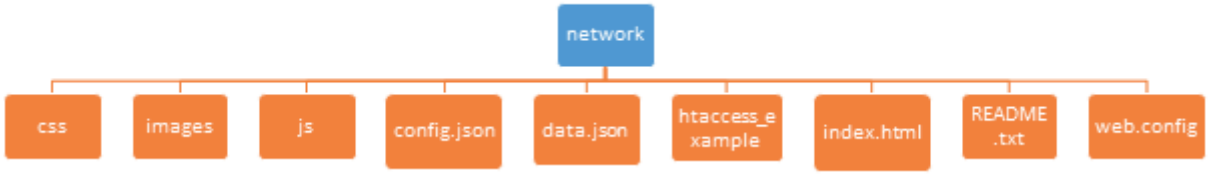
10 Announcing the Gephi Consortium, https://gephi.wordpress.com/2010/09/14/announcing-gephi-consortium/ (hozzáférés: 2023. 09. 19.)

11 Gepphi blog, https://gephi.wordpress.com/ (hozzáférés: 2023. 09. 19.), Learn how to code in Gephi, https://gephi.org/developers/ (hozzáférés: 2023. 09. 19.)

12 https://github.com/oxfordinternetinstitute/gephi-plugins/tree/sigmaexporter-plugin (hozzáférés: 2023.06.30)

CC BY

olyan formátumban, hogy az immár alkalmas a böngészőn keresztüli közzétételre, utána pedig lehetőséget ad a felhasználók számára az interaktív adatlekérdezésre.



4. ábra Az exportált mappa felépítése

Az online publikáláshoz szükséges egy GitHub[13] repozitórium létrehozására, amely a verziókezelés mellett lehetőséget biztosít a feltöltött tartalmakat a weben publikálni.[14] Az oldal kinézetét és az adatok strukturáltságát a SigmaExporter kimeneti fájljai határozzák meg, amelynek – lévén egy előre lefejlesztett oldal – projektenként eltérő lehet az alkalmazhatósága.

A Kiss József-adatvizualizáicó elkészítése kapcsán utólagos szerkesztésre is szükség volt, ez jobbára a színek korrekcióját, node-ok nagyságát, elírások javítását jelentette közvetlenül módosítva a forráskódban. A Széchényi-levelezésnél azonban arra is szükségünk volt, hogy az adatlapon minden node esetében egy bizonyos sorrendben jelenjenek meg az adatok. Ez a lépés ki is hagyható.

Az itt felvázolt folyamat egymásra épülő elemekből áll, de nem ilyen lineárisan formálódott a munkamenet. Az egyes lépések kialakítása a workflow végén és a köztes lépések között is sok teszteléssel járt. Nehézséget okozott továbbá az időközben beérkező új adatok elhelyezése, integrálása. Sokszor az eredeti adatbázist gazdagítva újra végig kellett menni a workflow lépésein.

2.3 Kapcsolati hálók – összegzés

Az adatbázisokban (katalógusok, könyvtári adatbázisok, táblázatok, tudástárak) tárolt adatok vizuális megjelenítése új kérdéseket, nézőpontokat vethet fel egy kutatásban. A levelezések alapján az adatvizualizációk segítségével rekonstruáltuk egy-egy fókuszba helyezett személy kapcsolati rendszerét, láthatóvá téve annak eddig ismeretlen jellemzőit, aspektusait.

A kéziratok feldolgozása során az adott korpusz adatai mellett egyéb adatbázisokban, gyűjteményekben lévő anyagokat is hasznosítunk. Ilyen például Móricz Zsigmond családfájának vizualizációja, melyhez a levelezés adatai mellett születési, házassági vagy halálozási anyakönyveket és egyéb nyilvántartásokat is felhasználtunk. Ennek alapján olyan rokoni kapcsolatok váltak egyértelműsíthetővé és mindenki számára elérhetővé, amelyek egyébként átláthatatlanok és nehezen hozzáférhetőek.[15]

13 GitHub fejlesztői környezet, https://github.com/about (hozzáférés: 2023. 09. 18.)

14 GitHub Docs, Creating a GitHub Pages site, https://docs.github.com/en/pages/getting-started-with-github-pages/creating-a-github-pages-site Vö: kiss-jozsef-levelezes, https://github.com/szucs-kata/kiss-jozsef-levelezes (hozzáférés: 2023. 09. 18.)

15 Vö.: Varga, Emese és Makkai, T. Csilla (2022) *„Ki a fenének kell collstok?" A digitális szöveg rejtett mértékegységei.* In: Valós térben - Az online térért: Networkshop 31: országos konferencia. 2022. április 20–22. Debreceni Egyetem. Kiadja a HUNGARNET Egyesület az MTA Könyvtár és Információs Központ közreműködésével, Budapest, pp. 204-210. ISBN 978-615-82243-0-7, https://doi.org/10.31915/NWS.2022.26

3. Vizualizáció az aratott webes tartalmakból

Az OSZK webarchívuma[16] azzal a céllal jött létre, hogy gyűjtse és hosszútávon megőrizze a nyilvánosan elérhető magyar webtartalmakat, amelyeket a magyar közönség számára szántak és részei a kulturális örökségnek, különös tekintettel a hungarikumokhoz tartozó elektronikus dokumentumokra. Webtérszintű és a tematikus aratások mellett a webarchiváló csoport eseményalapú gyűjtéseket is végez[17] a jelentősebb kulturális, politikai és sporteseményekről, egy adott témában hetente egy alkalommal.

2022. február 21-én kezdődött meg és azóta is tart az orosz-ukrán konfliktussal és a későbbiekben kibontakozó háborúval kapcsolatos cikkek és írások szelektív aratása, 75 magyarországi és határon túli hírportálról az oldalakon használt címkék vagy kategóriák alapján.[18]

Az esemény alapú aratások anyaga zárt archívumba kerül, tehát a learatott webes tartalmak teljes szövege jogi okok miatt nem szolgáltatható, de az adott eseménnyel kapcsolatban összegyűjtött és archivált weboldalak, webhelyrészek, illetve webhelyek címlistáját tartalmazó táblázatok megtekinthetők.[19] Az eddigi gyűjtések közül az orosz–ukrán konfliktushoz kapcsolódó részgyűjteményhez a SolrWayback[20] nevű szoftver segítségével készült egy publikus keresőfelület is.[21]

A Digitális Bölcsészeti Központ a hírportálok anyagából tematikus korpuszt épített. A továbbiakban azt az interaktív felületet mutatom be, amely egyedülálló módon láthatóvá tette a learatott webes tartalmak szókészletét. Egy korábbi tanulmányban már szó esett arról, hogy miként jött létre az infrastruktúra, erre itt nem térek ki.[22]

3.1 Tematikus aratás adatainak megjelenítése

Az orosz-ukrán konfliktus szókészletének vizualizációja egy demó projektnek tekinthető, amely új lehetőséget nyit meg a webarchivumok előtt.

A tematikus aratást követően a cikkek egy nyelvi elemzésen estek át az e-magyar Digitális Nyelvfeldolgozó Rendszer segítségével,[23] melynek a kimenete az adott heti aratás kétezer leggyakrabban használt szavát tartalmazta lemmatizált formában, tehát eltávolítva róla minden ragot és jelet, továbbá egy szófaját jelölő kóddal ellátva azt.

---

16 OSZK Webarchívum, https://webarchivum.oszk.hu/ (hozzáférés: 2023.09.14

17 Archívumtípusok, MIA WIKI, https://webarchivum.oszk.hu/mediawiki/index.php?title=Archívumtípusok

18 Böngészés: Orosz-ukrán konfliktus – 2022 https://webarchivum.oszk.hu/webarchivum/bongeszes/bongeszes-az-esemeny-alapu-gyujtemenyekben/bongeszes-orosz-ukran-konfliktus-2022/ (hozzáférés: 2023.06.30)

19 Böngészés az esemény-alapú részgyűjteményekben, https://webarchivum.oszk.hu/webarchivum/bongeszes/bongeszes-az-esemeny-alapu-gyujtemenyekben/

20 netarchivesuite/solrwayback, https://github.com/netarchivesuite/solrwayback (hozzáférés: 2023.09.18.)

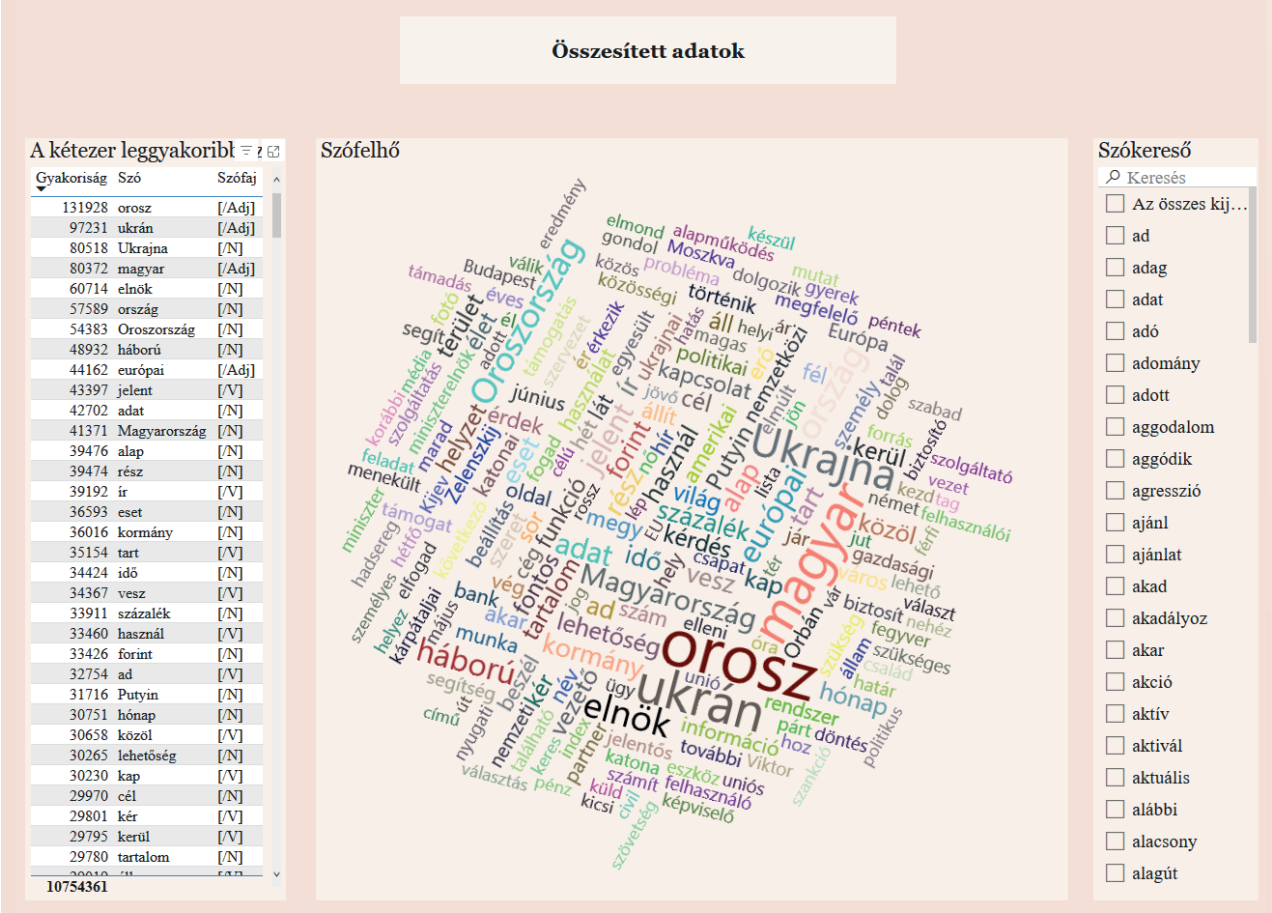
21 OSZK SolrWayback kereső, https://ukrajnapublic.webharvest.oszk.hu/solrwayback/ (hozzáférés: 2023.09.18.)

22 Bővebben a projekt szakmai hátteréről ld..: Kalcsó Gyula, Mihály Eszter, Szűcs Kata Ágnes, Korpuszépítés és -feldolgozás learatott webes tartalomból, 447–456., XIX. Magyar Számítógépes Nyelvészeti Konferencia Szeged, 2023. január 26–27. https://rgai.inf.u-szeged.hu/sites/rgai.inf.u-szeged.hu/files/mszny2023_0.pdf

23 e-magyar, https://e-magyar.hu/hu/intro (hozzáférés: 2023.09.18.)
nytud/emtsv, https://github.com/nytud/emtsv#e-magyar-text-processing-system-emtsv (hozzáférés: 2023.09.18.)

CC BY

Ezeket a szólistákat manuális tisztítás és szűrés után heti bontásában egyesével, és összesítve is meg tudtuk jeleníteni a Microsoft által üzemeltetett Power BI[24] ingyenesen is hozzáférhető szolgáltatását használva.



5. ábra: Magyar hírportálok orosz–ukrán háborús tartalmainak interaktív vizualizációja, első oldal[25]

A vizualizáció első oldalán egy szófelhő jeleníti meg a szavakat, kiegészítve az összesített szólistával és egy gyakoriságot mutató oszlopdiagrammal. Egy-egy szóra rá is lehet keresni, illetve ezen az oldalon jelenik meg az összesített statisztika is. Az oldalon található idővonal legkisebb értelmezhető alapegysége a hónap: ki lehet választani egy, vagy több összefüggő hónapot.

A második oldalon is látható egy szófelhő, amely összetétele aszerint változik, hogy mit választunk a heti szintű aratásokat felsoroló legördülő listából. Itt már lehetőség van szófaj szerint is szűkíteni a találatokat, illetve egy gyakoriságot mutató diagram után a szófajok előfordulásának arányát egy fatérkép mutatja.

A Power BI (Business Intelligence) lehetővé teszi a nagy mennyiségű adatfeldolgozást- és kezelést egy GUI felületen keresztül, emellett újszerű vizualizálási lehetőségeket kínál azok megjelenítésére és közzétételére (diagramok, szófelhők, térképes vizualizációk, stb).[26]

24 Microsoft Power BI, https://powerbi.microsoft.com/hu-hu/ (hozzáférés: 2023.09.18.)

25 Magyar hírportálok orosz–ukrán háborús tartalmainak interaktív vizualizációja, https://dhupla.hu/page/kreativ/ukrajna-hirek-szokeszlet-interaktiv (hozzáférés: 2023.09.18.)

26 Visualization types in Power BI, https://learn.microsoft.com/en-us/power-bi/visuals/power-bi-visualization-types-for-reports-and-q-and-a (hozzáférés: 2023.09.18.)

Emellett hatalmas előnyt jelent az interaktivitása, azaz, hogy a különböző adattáblákon szereplő információk összeköttetésben állnak egymással. Így, ha a felhasználó szűrést végez az egyiken, a másik eredményei is aszerint változnak. A Power BI továbbá rugalmasan kezeli a bemeneti forrásokat is (Excel, CSV, egész mappa, stb), ezáltal több projektnek is otthont adhat.

A webarchívumokban rejlő adatok kutathatósága és közzététele egy folyamatosan napirenden lévő kérdés, ezzel a demó projekttel bemutattunk egy lehetséges felhasználási módot, amelyet más gyűjteményekre, és más kérdések feltevésével is hasznosítani lehet kiindulópontként.

4. Kitekintés

Az adatvizualizáció a kutatás számára is hasznos nézőpontokat adhat, melyek hagyományos eszközökkel nem, vagy nem ennyire kézenfekvő módon lennének elérhetőek. A big data alapú kutatások továbbá arra is alkalmazhatóak, hogy cáfoljanak vagy megerősítsenek kutatási kérdéseket, felvetéseket.

Intézményi szempontból egy adatvizualizáció azokat az érdeklődőket is kiszolgálja, akik nem kutatási céllal látogatnak a dHUpla honlapjára, és nem utolsó sorban pedig kiaknázza a rendelkezésre álló adatokat.

Távlati céljaink közé tartozik az egyes írói levelezések vizualizációinak összekapcsolása, hálózattá alakítása (például Móricz Zsigmond és Kiss József kapcsolatai sok esetben metszik egymást), valamint a vizualizációk digitális szövegkiadásokkal való összekötése is. Amelyet a Móricz Zsigmond levelezéséből készült vizualizáció esetében részben megvalósítottunk.

Bibliográfia

Gephi, The Open Graph Viz Platform, https://gephi.org/ (hozzáférés: 2023.06.30)

Kiss József kapcsolati háló, https://dhupla.hu/page/kreativ/kissjozsef-kapcsolatihalo (hozzáférés, 2023. 06. 29.)

SigmaExporter plugin, https://github.com/oxfordinternetinstitute/gephi-plugins/tree/sigmaexporter-plugin (hozzáférés: 2023.06.30)

OSZK Webarchivum, https://webarchivum.oszk.hu/ (hozzáférés: 2023.06.30)

Böngészés: Orosz-ukrán konfliktus – 2022, https://webarchivum.oszk.hu/webarchivum/bongeszes/bongeszes-az-esemeny-alapu-gyujtemenyekben/bongeszes-orosz-ukran-konfliktus-2022/ (hozzáférés: 2023.06.30)

Hajnal Péter: Gráfelmélet, Polygon, Szeged, 2003.

Kalcsó Gyula, Mihály Eszter, Szűcs Kata Ágnes, Korpuszépítés és -feldolgozás learatott webes tartalomból, 447–456., XIX. Magyar Számítógépes Nyelvészeti Konferencia Szeged, 2023. január 26–27. https://rgai.inf.u-szeged.hu/sites/rgai.inf.u-szeged.hu/files/mszny2023_0.pdf

Varga, Emese és Makkai, T. Csilla (2022) „Ki a fenének kell collstok?" A digitális szöveg rejtett mértékegységei. In: Valós térben - Az online térért: Networkshop 31: országos konferencia. 2022. április 20–22. Debreceni Egyetem. Kiadja a HUNGARNET Egyesület az MTA Könyvtár és Információs Központ közreműködésével, Budapest, pp. 204-210. ISBN 978-615-82243-0-7, https://doi.org/10.31915/NWS.2022.26

V.05
DOI: 10.31915/NWS.2023.26

A BME Építészettörténeti és Műemléki Tanszék repozitóriuma

Leitgéb Mária
BME Építészettörténeti és Műemléki Tanszék - tudományos segédmunkatárs, könyvtáros
ELTE BTK Könyvtár- és Információtudományi Intézet – PhD hallgató
ORCID: 0000-0002-5423-7367
leitgeb.maria@epk.bme.hu
leitgeb.maria@btk.elte.hu

Absztrakt
A BME Építészmérnöki Kar Építészettörténeti és Műemléki Tanszékének Rajz- és Fotótára a műegyetemi építészoktatás kezdeteitől tartalmaz építészeti rajzokat és oktatási anyagokat. A korábban rendezett, metaadatokkal ellátott és digitalizált anyagok számára a DSpace keretrendszerben került kialakításra repozitórium, amelyben a Rajz- és Fotótár mellett további tematikus és dokumentumtípus szerinti gyűjteményeket hoztunk létre. A tanulmány a tervezés lépéseit, speciális szempontjait és feladatait mutatja be a Rajz- és Fotótár és a Diplomatervek példáján, valamint a hozzáférési szintek és jogosultságok meghatározásának módját a nem homogén gyűjtemények esetén.
Kulcsszavak: építészeti rajz, repozitórium, tervezés, megőrzés, hozzáférhetővé tétel

Abstract
The Drawing and Photo Collection of the Department of History of Architecture and Monument Preservation of BUTE contains architectural drawings and educational materials from the beginning of architecture education at the university. The repository for the previously organized, metadata-equipped and digitized materials has been created in the DSpace framework, in which, in addition to the Drawing and Photo Library, we have created additional thematic and document type-based collections. The study presents the steps, special aspects and tasks of planning of the repository on the example of the Drawing and Photo Collection and Diploma Plans, as well as the way of determining access levels and authorizations in the case of the nonhomogeneous collections.
Keywords: architectural drawing, repository, planning, preservation, accessibility

Történet és gyűjtemények

A magyarországi egyetemi szintű építészképzés 1870 októberében indult, amikor a Műegyetemen megkezdte működését a Steindl Imre vezette Műépítészet tanszék. 1887-re alakult ki az ókori, középkori és újkori építéstan tanszékekből álló felépítés, amely lényegében 1957-ig fennmaradt. A tanszékek egyesítése után jött létre az a szervezeti egység, amelynek jogutódja az Építészmérnöki Kar Építészettörténeti és Műemléki Tanszéke.[1] Bár a szervezeti struktúra, a tantárgyak neve és tartalma jelentősen változott az elmúlt 150 év folyamán, az építészképzésben mindenkor szerepet játszottak az építészettörténet, az építészetelmélet és a műemlékvédelem diszciplínák.[2]

1 Krähling 2021. 8–9.
2 Krähling 2021. 7.

A jogelőd építészettörténeti tanszékek állományát is részben tartalmazó tanszéki könyvtár mintegy 21 ezer kötet könyvet és 3 ezer egység folyóiratot tartalmaz, köztük az egykori Steindl-tanszék dokumentumait, valamint a 16-18. századi szakirodalmat reprezentáló muzeális könyveket.[3] A tanszéki Rajz- és Fotótár hallgatói rajzokat, műemléki felméréseket, professzori rajzokat és terveket, valamint fotókat és diákat őriz a műegyetemi építészoktatás kezdeteitől napjainkig.[4] A tanszéken számos, egyelőre feldolgozatlan dokumentumtípus is megtalálható: tudományos dokumentációk, felmérések; műemlékvédelmi szemináriumi- és szakdolgozatok; hagyatékból származó kéziratos anyagok, fényképek, diák; a tanszéki működéshez és a közösségi médiafelületekhez kapcsolódó nyomtatott és born digital plakátok, fotók. Ezek rendszerezése, archiválása jelenleg nem megoldott. Fontos feladat továbbá a tanszéken keletkezett diplomatervek feldolgozása és archiválása.

Rajz- és Fotótár projekt

Az építészeti rajzokat tartalmazó, töredékesen fennmaradt, mégis hatalmas gyűjtemény feldolgozása 1990-ben kezdődött, a jogelőd tanszékek kiemelkedő professzoraira fókuszálva.[5] Hosszabb szünet után, OTKA támogatással 2014-ben kezdődött újra hallgatók bevonásával az anyag rendszerezése, katalogizálása, melynek során a feldolgozók a dokumentumok formai és tartalmi ismérveit egységes metaadat-struktúra alapján egy Excel-táblázatban rögzítették.[6] A munka eredményét reprezentáló, a rajztári dokumentumok metaadatait és bélyegképeket tartalmazó három nyomtatott kötet 2016-tól jelent meg Építészettörténeti Rajztár címmel,[7] de a .pdf formátumú digitális változatok sem biztosítják a hozzáférhetőséget, és a visszakeresés is nehézkes. A projekt következő fázisában NKA támogatással került sor a már feldolgozott rajzok, valamint diapozitívok digitalizálására.[8] 2021-ben, a tanszék alapításának 150. évfordulóján NKA pályázati támogatásból jött létre a tanszék és a képzés történetét bemutató kiállítás és kötet[9], megújult a tanszéki honlap,[10] és a Tudástár részeként lehetőség nyílt egy saját erőforrásokkal működtethető repozitórium fejlesztésére is.

A repozitórium tervezésének lépései, speciális szempontok és feladatok

A repozitóriumot a nyílt forráskódú DSpace keretrendszerben hoztuk létre, amelyben egy hierarchikus gyűjteménystruktúrába (kategóriák, alkategóriák, gyűjtemények) rendeződnek a tételek. A tervezés során tanulmányoztuk a hazai DSpace rendszerű repozitóriumok felépítését, elsősorban az ELTE Digitális Intézményi Tudástárat (EDIT).[11] A Rajz- és Fotótár gyűjteményszervezési és szolgáltatási modelljének kialakításához hasznosnak bizonyult az

3 Leitgéb 2021. 84.
4 Krähling–Fehér 2018. 93.
5 Krähling et al. 2015. 8.
6 Leitgéb 2021. 84–-85.
7 Krähling–Baku 2016., Krähling-Baku–Fehér 2017. Digitálisan elérhetőek a REAL-ban, az MTA Könyvtárának Repozitóriumában: http://real.mtak.hu/ .
8 Leitgéb 2021. 84–85.
9 Gy. Balogh et al. 2021.
10 A honlap elérhetősége: https://eptort.bme.hu/
11 Az EDIT elérhetősége: https://edit.elte.hu/

CC BY

Architekturmuseum der Technischen Universität Berlin in der Universitätsbibliothek digitális gyűjteményének tanulmányozása.[12]

A kiinduláskor egy gyűjteményt terveztünk a Rajz- és Fotótárban lévő anyag elhelyezésére, azonban érdemesnek látszott, hogy a tanszéken található, fentebb említett dokumentumok feltöltésének, tárolásának és szolgáltatásának a lehetőségét is megteremtsük. Így jött létre egy hibrid struktúra, ahol tematikus és dokumentumtípus szerinti gyűjtemények egyaránt helyet kaptak. Ezután szakmai szempontok és az infrastrukturális hatékonyság miatt a repozitóriumot kari szintűvé bővítettük, így a későbbiekben a többi tanszék számára is lehetővé válhat anyagaik archiválása és hozzáférhetővé tétele. A kategóriákat ebben a struktúrában a szervezeti egységek (tanszékek) alkotják, ezeken belül lesz mód igény szerint a gyűjtemények létrehozására. Az Építészettörténeti és Műemléki Tanszék gyűjteményei tehát egyfajta pilot-projektként szolgálhatnak majd a többi tanszék számára.[13]
A tervezés első lépése a gyűjtemények és jellemzőik meghatározása volt. Számba vettük, hogy hogy milyen gyűjteményeket érdemes létrehozni, mi jellemzi ezeket a tartalom, a dokumentumtípusok, a hozzáférés (szerzői jogi szempontok), a tervezett funkció (archiválás, hozzáférhetővé tétel) szempontjából.[14] Mindezek alapján a következő gyűjteményeket hoztuk létre:

Gyűjtemény	Gyűjtemény tartalmi jellemzői, funkciója	Gyűjtemény hozzáférhetősége
Diplomatervek	A tanszéken megvédett diplomatervek archívuma, elsődleges feladata a megőrzés.	Szerzői jogilag védettek, a jogszabályoknak megfelelően csak a tanszéki könyvtár másolásvédett számítógépén tekinthetők meg, kivéve azokat a dokumentumokat, melynek szolgáltatásához a szerző kifejezetten hozzájárult.
Hallgatói dolgozatok	Az azonos dokumentípust képviselő Műemlékvédelmi Szakmérnöki képzés során keletkezett szakdolgozatok és falukutatási dolgozatok, valamint a tanszéki érintettségű régebbi TDK dolgozatok.	
Disszertációk	Régebbi doktori disszertációk.	
Publikációk/preprint verziók	A nyílt tudomány szellemében a tanszéki tudományos kutatások láthatóságának növelése, szabad hozzáférés biztosítása. Nem teljességre törekvő.	Open Access hozzáférés
E-könyvtár	A tanszékhez kötődő szerzők born digital és digitalizált dokumentumainak gyűjteménye, digitalizálandó muzeális könyvek.	Open Access hozzáférés
Fotótár, Videótár, Plakáttár	A tanszék rendezvényeihez, oktatási eseményeihez kapcsolódó dokumentumok.	Szerzői jogi szempontból különböző státuszú anyagok, a hozzáférés differenciált.
Kéziratok	Jellemzően hagyatékokból származó autográf és gépirat.	Archiválás. Szerzői jogilag védett, csak a szabad felhasználás feltételeinek megfelelően szolgáltathatók.
Éptört 150	Tematikus gyűjtemény.	Szerzői jogi szempontból különböző státuszú, nem nyilvánosságra szánt anyagok is, a hozzáférés differenciált.

1. ábra: A létrehozott gyűjtemények és jellemzőik

12 2016-ban került megrendezésre tanszékünk szervezésében az „Építészeti rajztár" workshop, melyen az intézmény igazgatója, Dr. Hans-Dieter Nägelke is tartott előadást. A gyűjtemény elérhetősége: https://architekturmuseum.ub.tu-berlin.de/index.php?p=18 A jövőbeli fejlesztéseknél érdemes lehet további építészeti vonatkozású repozitóriumok gyakorlatának vizsgálata, pl. a Delft University of Technology építészeti témájú szakdolgozatokat is tartalmazó repozitóriuma (https://repository.tudelft.nl/content/about), vagy a Digitale Sammlung der Universitätbibliothek Stuttgart építészeti rajzainak gyűjteménye (https://digibus.ub.uni-stuttgart.de/viewer/architekturzeichnungen/).

13 A felsőoktatási intézményekben létrehozott intézményi repozitóriumokkal szembeni elvárásokkal kapcsolatban lsd. Drótos 2011. 310.

14 A tanulmány elkészítésének időpontjában nem volt információm arról, hogy mi a BME állásfoglalása a nemzeti felsőoktatásról 2011. évi CCIV. törvény 50. § (1)-ben foglaltakkal kapcsolatban.

Speciális tervezési szempontok a Rajz- és Fotótár és a Diplomatervek esetén

A metaadatok szempontjából speciális jellemzőkkel bíró Rajz- és Fotótár, illetve Diplomatervek gyűjteményeknél kialakítottuk dokumentumtípusonként a metaadatstruktúrát és ezzel párhuzamosan a feltöltési űrlapokat, amelyekkel a későbbiekben a tételenkénti feltöltés történik. A munka során számos különleges szempont merült fel, a tanulmány ezek közül mutat be néhány példát.

A *Rajz- és Fotótár* esetén szükségessé vált a korábbi feldolgozás során készült Excel-táblában található metaadatok pontosítása és Dublin Core-sémába való konvertálása. Ezek közül kiemelendő, hogy az Excel-táblában a rajz készítője szerepelt szerzőként, azonban az építészeti alkotások (tervrajz, épület) szerzője a tervező/építész, így a rajzok készítői a Készítő elnevezésű mezőbe kerültek (dc.creator). A későbbiekben a Tervező/Építész mezőben (dc.contributor.author) lehet megadni a szerzőt. Ez a megoldás a felhasználók felé is tükrözi a szerzői jogi viszonyokat. Meg kellett továbbá határozni, hogy mit tekintünk a dokumentum címének. A rajzokon szereplő feliratok az Excel-táblában a Megnevezés mezőben szerepeltek, ezek azonban sokszor általános megnevezést tartalmaznak, pl. „Homlokzati terv". Sok rajzon nem is szerepel felirat, így ez nem alkalmas a dokumentum megfelelő azonosítására. A dokumentum címének ezért a feldolgozók által megállapított, a rajz tartalmát ténylegesen tükröző elnevezést tekintjük, amely a Megnevezés mezőben (dc.title.image) szerepel, míg a felirat a Felirat mezőben (dc.title.caption). A tételek áttekintő adatainak megjelenítésekor jelenleg a Készítő és a Megnevezés jelenik meg, a későbbiekben a szerzők pótlásával a Tervező/Építész mező is megjeleníthető.

A földrajzi elhelyezkedésre utaló adatokat tartalmazó metaadatoknál, pl. Megye, Helység, Közölt közterület, Közölt Hrsz. ügyelni kellett arra, hogy a dokumentumokon szereplő helyszínek neve nem feltétlenül azonos a mai elnevezésekkel, valamint nem biztos, hogy a feldolgozók az egységesített nevet adták meg. Ezért a most betöltött tételeknél a földrajzi nevek egyelőre nem az egységesített földrajzi nevekre szolgáló DC-mezőbe kerültek, a későbbiekben kerülhet sor a tételek módosítására, és adatok megfelelő DC-mezőben való feltüntetésére.

Az építészeti rajzok ismérveinél számos speciális, a felhasználók számára informatív metaadat jelenik meg, pl. Hordozó, Technika, Méretarány/lépték, Méret, Állapot, az ezekhez a mezőkhöz tartozó DC-mezők meghatározása is feladat volt.

A kutatók számára az is fontos információ lehet, ha egy adat, pl. a méretarány nem szerepel rajzon. Ezért meghatároztuk azoknak a metaadatoknak a körét, amelyeknél ilyen esetben a „nincs adat" érték jelenik meg.

Mindezek figyelembevételével alakítottuk ki a feltöltéshez használandó űrlapokat, melyekbe bekerültek az eddigi feldolgozásoknál megadott metaadatok és új, releváns metaadatok is. Kijelöltük a kötelezően kitöltendő mezőket, ahol a feldolgozónak az adathiányt is jelölnie kell. Végiggondoltuk, melyek azok a mezők, amelyeknél a feldolgozók számára előre definiálhatunk értékeket, amelyet menüből választhatnak ki, gondolva arra, hogy nem könyvtárosok fognak feldolgozni.[15] Ugyancsak emiatt a földrajzi és a személynevek esetében a besorolási adatoknál a jövőben a feldolgozók számára elő kell írni a névterek alkalmazását.[16]

15 A gyarapítás céljából a szerzői önfeltöltés mellett hallgatókat kell bevonni, kettős felügyelet: építészi és könyvtárosi mellett. Lsd. Holl 2022. 359.

16 Az intézményi repozitóriumok és a besorolási adatok problémaköréről lsd. Köntös 2012. 264–265., 272.

CC BY

A *Diplomaterv* gyűjtemény előzmények nélküli, új fejlesztés, amelynek kialakítása során arra törekedtünk, hogy a metaadatolás és az űrlap a többi tanszék számára is megfelelő legyen. Az építészeti tervet tartalmazó diplomaterveknél speciális közreműködők vannak (a konzulens mellett szakági konzulensek), speciális adatok (Építészeti terv helyszíne/helység; Építészeti terv helyszíne/közterület; Építészeti terv helyszíne/helyrajzi szám) és speciális formai jellemzők (Eredeti hordozó, Eredeti méret, Eredeti terv digitális formátuma, Tartalom/tervlapok), amelyeknek létrehoztuk az egyedi mezőket a hozzájuk tartozó DC-mezőkkel. A képzési forma megjelenítésénél tükröződnek az utóbbi évek átalakulásai (BSc, MSc, Osztatlan képzés), Szak/szakirány/specializáció).

A repozitórium működésének kezdetei – a Rajz- és Fotótár anyagának betöltése

A rajzok esetében 96 dpi felbontású, vízjellel ellátott képek feltöltése mellett döntöttünk, amelyek alapvetően megfelelőek a kutatók számára. A tételeket szerzői jogi szempontból három csoportba soroltuk: a szabadon hozzáférhető, illetve a szerzői jogilag még védett művek mellett kialakítottunk egy harmadik csoportot, amelyeknél még szerzői jogi vizsgálat szükséges, s ezt a felhasználók felé is jelezzük. Ezután tömeges feltöltéssel betöltöttük a rajzokat és diákat tartalmazó, több mint 9 ezer tételt, és publikussá tettük a repozitóriumot. A szerzői jogilag védett rajzok megtekintésére a tanszéki könyvtár másolásvédelemmel ellátott számítógépén biztosítunk lehetőséget. Mivel a gyűjtemény minden tételéről készült nagyfelbontású digitális változat is, a szerzői jogilag már nem védett művekről kérésre (pl. publikálás, kiállítás, üzleti célú felhasználás) egyedi elbírálás alapján adunk digitális másolatot.

Hozzáférési szintek és jogosultságok meghatározása a további gyűjteményeknél

A Rajz- és Fotótár betöltése után került sor a további gyűjteményeknél a hozzáférési szintek és jogosultságok meghatározására. A DSpace-ben alapértelmezett szerepkörök (a *Központi adminisztrátor*, az *Anonymous* (bejelentkezés nélküli felhasználó) és a *Gyűjteményi adminisztrátor)* mellett két speciális szerepkört határoztunk meg. A *Feltöltők* csak feltöltést, szerkesztést végző személyek (demonstrátorok, könyvtáros gyakornokok), akik csak egyes, meghatározott gyűjteményhez férnek majd hozzá. Az *Oktatók* csoportjába felvett munkatársak feltölthetnek egyes gyűjteményekbe, illetve hozzáférhetnek egyes, szerzői jogilag nem védett, nem publikálásra szánt tételekhez.

A gyűjtemények jogosultságainak meghatározásánál fontos tényező volt, hogy egyes gyűjtemények nem homogének, a gyűjteményen belül különböző jellemzőkkel bíró dokumentumokat tartalmaznak. A gyűjteményeket a szerzői jogi státusz, a hozzáférhetőség, a jogosultságok és az elhelyezés célja szempontjából vizsgáltuk.

Szerzői jogi státusz	Hozzáférési szintek
Szerzői jogilag (már) nem védett Szerzői jogi védettség alatt	Teljes tétel: metaadatok + fájl Csak metadaatok Sem metaadatok, sem a fájl
Szereplők	**Elhelyezés célja**
Administrator: központi adminisztrátor Anonymous: felhasználók bejelentkezés nélkül Adminisztrátor: gyűjteményi adminisztrátor Feltöltők: csak feltöltés, szerkesztés Oktatók: feltöltés, egyes tételekhez hozzáférés	Archiválás + hozzáférhetőség biztosítása Archiválás + tájékoztatás a dokumentum metaadatairól (hozzáférés csak a szabad felhasználás követelményeinek megfelelően) Nem publikálásra szánt tartalmak archiválása (hozzáférés csak speciális csoportnak)

2. ábra: Hozzáférési szintek és jogosultságok meghatározásának szempontjai

Igyekeztünk az összes lehetséges kombinációt számba venni. Például egy szerzői jogilag védett kézirathoz csak az adminisztrátorok férnek hozzá, de az Anonymous felhasználók tájékozódhatnak a metaadatokról és a könyvtári terminálon meg is tekinthetik azt, vagy például egy szerzői jogilag nem védett, de belső használatra szánt fotónak még a metaadatait sem látják az Anonymous felhasználók, de az Oktatók hozzáférhetnek. Mindezek alapján két alaptípus határozható meg:

1. gyűjteménytípus	2. gyűjteménytípus
Szerzői jogilag nem védett, publikálásra szánt dokumentumokat tartalmaz (E-könyvek, Plakáttár, Publikációk) vagy szerzői jogilag jellemzően nem védett, de kisebb számban védett, illetve publikálásra szánt és nyilvánosságnak nem szánt anyagokat tartalmaz (Éptört 150, Fotótár, Videótár).	**Szerzői jogilag:** jellemzően szerzői jogilag védett, és kisebb számban nem védett dokumentumokat tartalmaz (Rajz- és Fotótár, Disszertációk, Hallgatói dolgozatok) vagy szerzői jogilag védett anyagokat tartalmaz (Diplomaterv, Kézirat)
Alapértelmezett hozzáférés: A gyűjtemény tételei mindenki számára teljességgel - metaadatok és fájl - hozzáférhetőek. Az alapértelmezett jogosultságot a tételek szintjén lehet módosítani, pl.: az Anonymous felhasználók csak a metaadatokat láthatják, vagy, hogy az Anonymous felhasználók a metaadatokat sem láthatják, de az Oktatók igen.	**Alapértelmezett hozzáférés:** a metaadatok és a fájlok csak a központi adminisztrátor és a gyűjteményi adminisztrátor(ok) számára láthatóak, de a gyűjtemény szintjén a metaadatok megtekintését engedélyezzük az Anonymous felhasználóknak. Az alapértelmezett jogosultságot a tételek szintjén lehet módosítani, pl. az Anonymous felhasználók is láthatják a fájlokat, vagy pl. az Anonymous felhasználók a metaadatokat sem láthatják, de az Oktatók igen.

3. ábra: Gyűjtemények alaptípusai a hozzáférhetőség és jogosultság szempontjából

A gyűjteményekbe való feltöltés előtt a típusba sorolásnak megfelelően kell megállapítani a hozzáférési szinteket és jogosultságokat, valamint a feltöltés folyamatát és az ehhez tartozó jogosultságokat. Ez utóbbinál minden esetben adminisztrátornak kell jóváhagynia és publikussá tennie a tételeket.

CC BY

A repozitóriumépítés további lépései

A hozzáférési szintek és jogosultságok meghatározása után kerülhet sor azok gyűjteményenkénti beállítására a konkrét közreműködők megadásával. Ezután következik a rendszer tesztelése, szükség esetén a beállítások módosítása, esetleges hibák javítása. A Rajz- és Fotótár esetén hosszútávú feladat a szerzői jogilag még felülvizsgálat alatt lévő tételek ellenőrzése, indokolt esetben hozzáférhetővé tétele, az egységes besorolási adatok (építész/tervező neve, illetve földrajzi név) pótlása.

A felvázolt protokoll mentén a következőkben tanszékenként meg lehet határozni az egyedi struktúrát, definiálni az egyes gyűjtemények jellemzőit (dokumentumtípus, szerzői jogi státusz, alapértelmezett hozzáférhetőség, jogosultságok és szerepek megadása), kinevezni a gyűjteményi adminisztrátorokat, meghatározni ki fogja az egyes gyűjteményeknél a feladatokat végezni (feltöltés, ellenőrzés, publikussá tétel), és létrehozni a jogosultságokat. Ezek után be kell tanítani a gyűjteményi adminisztrátorokat és más szereplőket. Fontos feladat még a működéshez szükséges dokumentumok előállítása (feltöltési szabályzat, elhelyezési nyilatkozat, útmutatók), különös tekintettel arra, hogy könyvtárosi ellenőrzés mellett, de nem könyvtárosok fogják a feladatokat elvégezni.

Irodalomjegyzék

Drótos 2011
Drótos László: Bölcsészek információs igényeinek megismerése egy intézményi repozitórium tervezésekor. *Tudományos és Műszaki Tájékoztatás* 58 (2011) 7. 308–311. https://tmt.omikk.bme.hu/tmt/article/download/856/876 Hozzáférés: 2023.06.02.

Gy. Balogh et al. 2021
Gy. Balogh Ágnes – Fehér Krisztina – Krähling János – Vukoszávlyev Zorán (szerk.): *150 év – 150 rajz. A BME Építészettörténeti és Műemléki Tanszék másfél évszázada.* BME Építészettörténeti és Műemléki Tanszék, Budapest 2021.

Holl 2022
Holl András: Repozitóriumok – különleges terület a könyvtárak világában. *Tudományos és Műszaki Tájékoztatás 69* (2022) 7. 358–365. https://doi.org/10.3311/tmt.13180. Hozzáférés: 2023.06.02.

Köntös 2012
Köntös Nelli: Szerzők nyomában. A könyvtári szabványok szerepe az intézményi publikációs adattárak névkezelési stratégiájában. *Könyvtári Figyelő* 22=58(2012)2.255–279. http://epa.oszk.hu/00100/00143/00083/pdf/EPA00143_konyvtari_figyelo_2012_2_255-279.pdf Hozzáférés: 2023.06.02.

Krähling 2021
Krähling János: Bevezető. In: *150 év – 150 rajz. A BME Építészettörténeti és Műemléki Tanszék másfél évszázada.* Szerk.: Gy. Balogh Ágnes et al. BME Építészettörténeti és Műemléki Tanszék, Budapest 2021. 7–10.

Krähling-Baku 2016
Krähling János – Baku Eszter (szerk.): Építészettörténeti Rajztár. Az Építészettörténeti és Műemléki Tanszék rajzgyűjteményének katalógusa. *1–2*. BME Építészmérnöki Kar, Budapest 2016.

Krähling-Baku–Fehér 2017
Krähling János – Baku Eszter – Fehér Krisztina (szerk.): Építészettörténeti Rajztár. Az Építészettörténeti és Műemléki Tanszék rajzgyűjteményének katalógusa. *3*. BME Építészmérnöki Kar, Budapest 2017.

Krähling–Fehér 2018
Krähling János – Fehér Krisztina: Kincsek az építészképzés 150 éves történetéből. *Műemlékvédelem* LXII (2018) 3–4. 93–239.

Krähling et al. 2015
János Krähling – Balázs Halmos – Katalin Marótzy – István Sajtos – Zorán Vukoszávlyev – Eszter Baku – Anna Józsa – Zsuzsanna Kiss – Krisztina Fehér – Gergő Kovács: Architectural drawing and education. Principles to the evaluation of the historic plan collection at Budapest University of Technology and Economics. *Architectura Hungariae* 14 (2015) 1. 7-18. http://real.mtak.hu/24827/1/AH_vol14_no1_pp7_18_Krahling_etal_u.pdf Hozzáférés: 2023.06.02.

Leitgéb 2021
Leitgéb Mária: A tanszéki könyvtár, rajz- és fotótár története. In: *150 év – 150 rajz. A BME Építészettörténeti és Műemléki Tanszék másfél évszázada*. Szerk.: Gy. Balogh Ágnes et al. BME Építészettörténeti és Műemléki Tanszék, Budapest: 2021. 82–85.

CC BY
VI.01 DOI: 10.31915/NWS.2023.27

Szerkesztői környezet TEI-alapú szövegkiadásokhoz

An Editor Framework for Digital Scholarly Editions in TEI

Mihály Eszter
Országos Széchényi Könyvtár, Digitális Bölcsészeti Központ (OSZK DBK)
mihaly.eszter@oszk.hu

Micsik András
Számítástechnikai és Automatizálási Kutatóintézet, Elosztott Rendszerek Osztály (SZTAKI DSD)
micsik@sztaki.hu

Absztrakt

Az OSZK Digitális Bölcsészeti Központ által fejlesztett új platform, a dHUpla (Digital Humanities Platform – dhupla.hu) elsősorban digitális szövegkiadások publikálására jött létre. Ennek hátterében a szövegkorpuszok előállításához egy teljes szerkesztőségi rendszer kialakítására volt szükség, amely egyrészt felhasználóbarát módon teszi lehetővé a szerkesztést, másrészt funkcióival támogatja a szövegek jelölőnyelvi kódolását. Ezt a felületet a SZTAKI és az OSZK közösen fejleszti egy XML szerkesztő bővítményeként. A szövegkódolás alapja a TEI (Text Encoding Initiative) szabványa, a keretrendszer e nemzetközi ajánlás bonyolult konstrukcióinak bevitelét, többek között kontextus-menüvel, beszúrható mintákkal, Schematron-validációval segíti. A fejlesztésben központi szerepet játszanak a szöveg feldolgozását támogató további eszközök is: névterekkel, adatbázisokkal való összekapcsolás, MI-alapú névelem-felismerés, valamint különböző automatizált műveletek, úgymint PDF- és képkonverzió, vagy adatvizualizáció támogatása. A manuálisan és a gép által végezhető részfolyamatok minden esetben kiegészítik egymást, megteremtve ezzel egy minőségi, a digitális közeg adta lehetőségeket kiaknázó szövegkiadási módszert.

Kulcsszavak: digitális szövegkiadás, TEI, XML szerkesztőfelület

Abstract

The Digital Humanities Centre (DBK) of the Hungarian National Library has developed a platform called dHUpla (Digital Humanities Platform - dhupla.hu) for publishing digitized text editions. For this, the creation of a complete editing environment was necessary to support editors and digital humanists and to help them in the input and validation of correct TEI XML encoding. This environment was implemented as an XML editor extension jointly by DBK and SZTAKI (Insititute for Computer Science and Control). The extension contains custom toolbars, templates, Schematron validation and a set of scripts to automate steps of the conversion to TEI XML. Scripts support named entity recognition and linking, PDF generation, extraction of data for visualizations and other task automations. These operations combine automated execution with manual supervision to reach high quality TEI production.

Keywords: digital scholarly editions, TEI, XML editor

Bevezetés

Az Országos Széchényi Könyvtár Digitális Bölcsészeti Központ (DBK) egyik elsődleges feladata egy korszerű online platform fejlesztése a közgyűjteményekben őrzött szöveges források kezelésére, amely egységes kutatói környezetet jelent az irodalomtudomány, a nyelvtudomány, és más humán tudományok számára. A dHUpla[1] (Digital Humanities Platform) 2021 óta üzemel (2021 decemberéig a Petőfi Irodalmi Múzeum, majd az Országos Széchényi Könyvtár szolgáltatásaként), és számos szövegkiadást tesz nyilvánosan elérhetővé, amelyhez entitástár, valamint kreatív tartalmak, vizualizációk is társulnak. E szolgáltatások hátterében a szövegkorpuszok előállításához egy teljes szerkesztőségi rendszer kialakítására volt szükség, amely egyrészt felhasználóbarát módon teszi lehetővé a szerkesztést, másrészt funkcióival támogatja a szövegek jelölőnyelvi kódolását. Ezt a felületet a SZTAKI és az OSZK közösen fejleszti egy XML-szerkesztő bővítményeként. Az alábbiakban röviden bemutatjuk a dHUpla funkcióit, majd rátérünk a szerkesztői felület részletes ismertetésére.

A dHUpla infrastruktúra

A dHUpla elsődleges célja, hogy a kutatók, olvasók számára egységes felületen, filológiai igényességgel hozzáférhetővé tegye a magyar kulturális örökség különböző intézményekben őrzött, eddig ismeretlen vagy méltatlanul elfeledett szöveges tartalmú, elsősorban kéziratos forrásait. Emellett olyan szerkesztőségi keretrendszert is kínál a tartalmak digitalizálásához, amely biztosítja a források egységes és színvonalas feldolgozását.

A platform alapjait és felépítését korábban már részletesen bemutattuk [1]. A rendszer rugalmas és moduláris, a tartalom előállítása többféle úton történhet. Az átírt szövegek a nemzetközileg támogatott TEI (Text Encoding Initiative) szabvány szerint annotált XML formátumban készülnek, amely a nemzetközi integráció mellett lehetővé teszi többek között a dokumentumok gépi feldolgozását és értelmezését (linked open data), szemantikus hálók kiépítését (semantic web), adatgazdagítást (data enrichment), illetve a távoli olvasás (distant reading) különböző aspektusait. A szolgáltatás kapcsolatot létesít különböző névterekkel, bibliográfiai forrásadatbázisokkal, illetve nyelvi elemző szoftverekkel. Mindezek segítségével a legkülönbözőbb elemzések, korpuszlekérdezések, adatvizualizációk válnak megvalósíthatóvá.

Az infrastruktúra középpontjában a git verziókövető szoftver áll, amely végül teljes mértékben kiváltotta egy XML-adatbázis használatának szükségességét, nagyban leegyszerűsítve a rendszer használatát, karbantartását és fejlesztését. A dHUpla git-ben lévő források (szöveg, programkód) alapján publikál, így a git repository-k birtokában bárhol újraépíthető a teljes dHUpla honlap. A projektek mind önálló git repository-ban vannak, a publikáláshoz szükséges transzformációt docker containerek végzik, minden egyes projekthez meg lehet adni saját ún. buildert, amelyekben tetszőleges programnyelvet lehet használni. A HTML tartalmon túl Apache Solr indexfájlt is előállítunk, amelynek segítségével a legkülönbözőbb facettás keresések is lehetővé válnak.

A publikációs felületen több módon lehetőség nyílik az átírt szöveg és az eredeti facsimile együttes vizsgálatára, a digitális objektumok különböző szempontú rendezésére, szűrések elvégzésére. Az egyes gyűjtemények megjelenésének, funkcióinak konfigurálása egyszerű szöveges fájlokban (yml) történik.

1 dhupla.hu

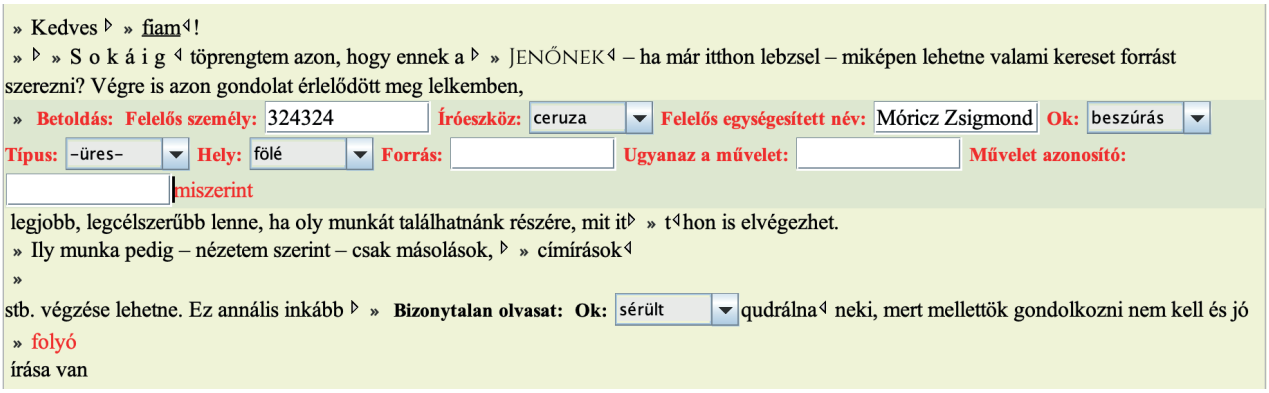
CC BY

A kéziratok átírása során ezenkívül olyan kézírásfelismerő-modell épül, amely folyamatosan bővülve alkalmassá válik a magyar nyelvű kézírások automatikus felismertetésére, azaz mesterséges intelligencián alapuló gépi feldolgozására (Handwritten Text Recognition).

A TEI szerkesztői felület

A dHUpla oldalára szánt szövegeket a DBK által készített TEI XML specifikációk szabályai szerint kell létrehozni, mivel a TEI igencsak rugalmas XML formátum, és több különböző ábrázolási mód is választható ugyanarra a szövegelemre. A TEI XML-ek részletes szerkesztésére az Oxygen XML Editor[2] programot használjuk. Az Oxygen XML Editor (röviden továbbiakban Oxygen) beépített TEI támogatással rendelkezik, de ez a már említett teljesen általános, korlátozások nélküli TEI-n alapul. Az Oxygenen belül lehetőség van ún. frameworkök kifejlesztésére, amely egy személyre szabott szerkesztői felületet tud nyújtani, eszköztárakkal, menükkel, specializált megjelenítéssel és ellenőrzési lehetőségekkel.

A frameworkben alkalmazott megoldásokból elsőként magát a szövegszerkesztő ablakot vesszük sorra. Itt háromféle nézetből lehet választani, ebből számunkra kettő a lényeges: az első magát az XML kódolást mutatja, ahol a rideg valóság tárul elénk, és ellenőrizhetjük, illetve változtathatjuk az XML-t. A szerzői nézet ezzel szemben egy CSS-sel barátságossá alakított, olvasható nézetet kínál, ahol az elemek szerkesztése vezetett módon lehetséges (1. ábra).

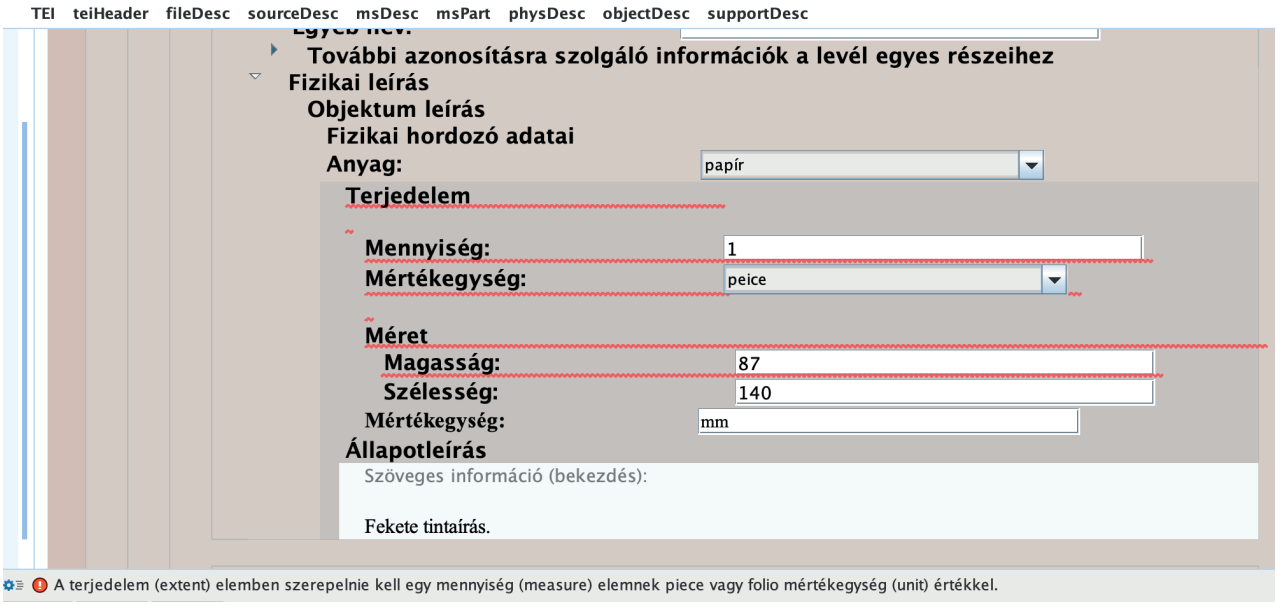


1. ábra: Az ún. szerzői nézetben segítséget kapunk a speciális elemek kitöltéséhez

Az 1. ábrán megfigyelhetjük, hogy a speciális szövegelemek előtt álló » jelre kattintással kinyitható és becsukható az elem tartalma. Az ábrán egy <add> és egy <unclear> elem attribútumait láthatjuk.

A metaadatok beírását részletesen kidolgozott űrlapok segítik (2. ábra). A piros aláhúzás hibás kitöltést jelez, a magyarázat a legalsó sorban olvasható. Az ellenőrzést Schematron[3] szabályokkal végezzük.

2 https://www.oxygenxml.com/xml_editor.html
3 https://www.schematron.com/



2. ábra: A TEI fejléc kitöltése és ellenőrzése

A 3. ábrán láthatóak a szerkesztők munkáját megkönnyítő eszköztárak. A felső sorban található menüvel a gyakori TEI elemeket lehet beilleszteni a szövegbe. A középső sor főleg a fejléc, vagyis a metaadatok kitöltéséhez nyújt segítséget. A legalsó sorból feldolgozási parancsokat lehet indítani. Ebben a sorban nem látszik az összes lehetséges művelet, van még keresés több más névtérben is, illetve PDF készítés mint menüpont.



3. ábra: Az általunk készített eszköztárak

Példa munkafolyamat

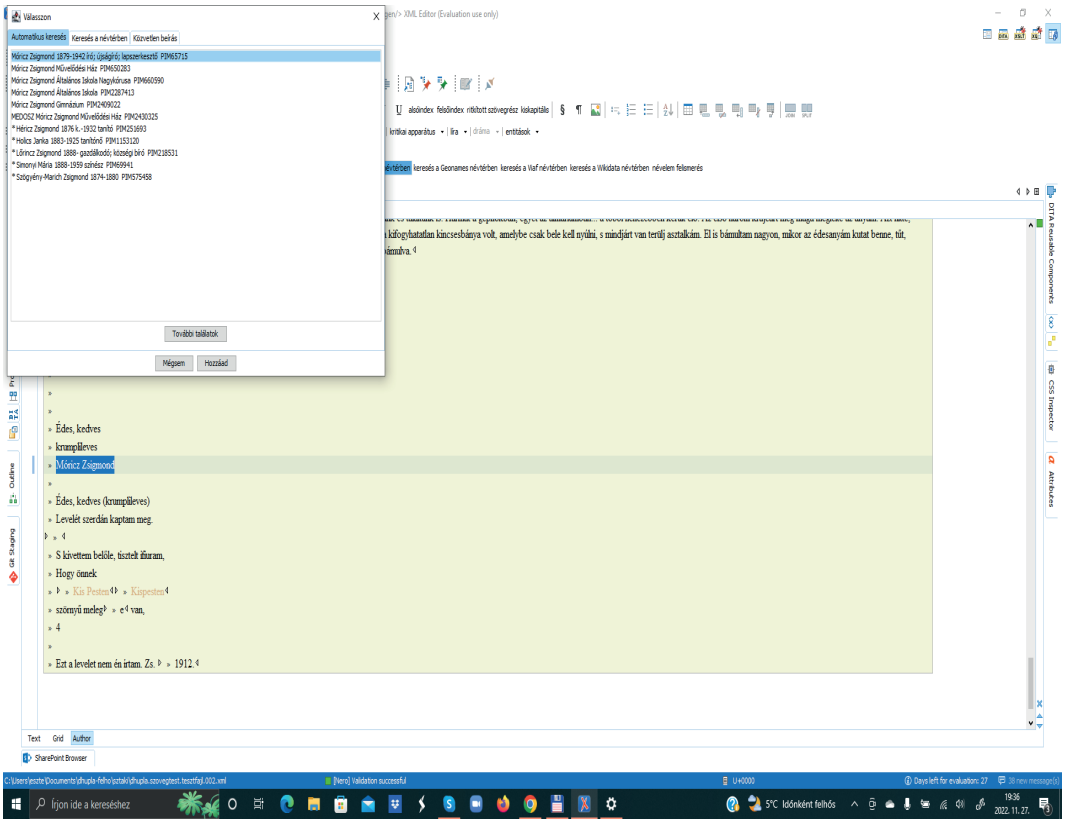
A kéziratok kézi és automatikus gépi átírásához, valamint az ún. text-image linking (kép és szöveg összekötése) elvégzéséhez a Transkribus programot használjuk, amely TEI XML formátumban is tud exportálni, azonban ez a formátum nem szabványos, és a DBK előírásainak sem felel meg. Az Oxygen eszköztárból indítható Transkribus export konverzió segítségével azonban megfelelő formára tudjuk hozni a TEI XML-t.

Ezután a szerkesztőben ellenőrizhetjük az XML-t, javíthatjuk a kapott Schematron hibajelzéseket, és egyéb finomításokat tudunk tenni a szövegjelölésekben. Ehhez a fázishoz kapcsolódik a TEI fejléc kitöltése, amelyhez egy teljes sablont tudunk egyetlen gombnyomással beilleszteni, majd a kapott űrlapokat kitölteni.

A szöveg feldolgozásában segít a megemlített névelemek (személyek, települések, szervezetek stb.) bejelölése. Erre saját fejlesztésű névelem-jelölőnket lehet használni, amely a HuSpaCy [2] nyelvi feldolgozó eszköz segítségével beilleszti a megfelelő TEI XML elemeket (pl. <persName>) a talált nevekhez. Ezek a nevek még azonosítatlanok, vagyis nincsenek egy gondozott névtérben a megfelelő névelemhez kötve. Erre a célra az eszköztár keresőgombokat tartalmaz, amely a kijelölt szöveget megkeresi az adott névtérben (pl. PIM[4], GeoNames, Wikidata), és a szerkesztő választásának megfelelően beírja az XML-be a névtér azonosítót (4. ábra).

4 https://opac-nevter.pim.hu/search

CC BY



4. ábra: Keresés a szövegben kiválasztott névre az Oxygen szerkesztőfelületén

Ha elkészült a TEI XML feldolgozás, az elemzési és terjesztési feladatokat is támogatja a framework. Az XML-ből például PDF-et tudunk előállítani, amely nem tartalmazza ugyan az XML-ben kódolt összes részletet, viszont jól olvasható és megosztható.

Dokumentumok csomagjaiból (pl. egy író levelezése) a Metaadatok kinyerése funkció segítségével táblázatokat kapunk, amelyek aztán különböző összesítések és vizualizációk alapjául szolgálhatnak. Az egyik ilyen vizualizációs eszköz hamarosan elkészül: a levelek feladását és kézhezvételét mutatja térképen.

Összefoglalás

A digitális szövegkiadások TEI alapú publikálása rengeteg új lehetőséget ad a digitális bölcsészeknek, amelyből párat jelen cikkben említettünk. Nehéz viszont odáig eljutni, hogy a szövegek jó minőségű TEI XML formátumba kerüljenek, mivel a folyamat sok szakértelmet és szakértői munkát igényel. Ennek a feldolgozási munkának a megkönnyítésére készítettük el a bemutatott Oxygen XML Editor framework megoldást, amely számos hasznos funkcióval segíti a szerkesztést, de ezen felül lehetőséget teremt a TEI-n belüli saját kódrendszerek megalkotására és az elkészített TEI dokumentumok e szerinti ellenőrzésére, és egy alap-minőség biztosítására a konverziók végén. A jelenlegi framework a levelezések XML kódolására fókuszál, de látjuk már a lehetőséget többféle „szakosodott” leírási mód (pl. drámák, versek, naplók) egyidejű támogatására is. A keretrendszer fejlesztése jelenleg a házon belüli alkalmazás és tesztelés fázisában van.

Irodalomjegyzék

[1] Mihály, Eszter (2022) *Mi az a dHUpla?: A Digitális Bölcsészeti Platform bemutatása.* In: Valós térben - Az online térért, Networkshop 31: országos konferencia. 2022. április 20–22. Debreceni Egyetem. Kiadja a HUNGARNET Egyesület az MTA Könyvtár és Információs Központ közreműködésével, Budapest, pp. 345–358. ISBN 978-615-82243-0-7 DOI: 10.31915/NWS.2022.44

[2] Orosz György, Szántó Zsolt, Berkecz Péter, Szabó, Gergő, Farkas Richárd (2022). HuSpaCy: an industrial-strength Hungarian natural language processing toolkit. In XVIII. Magyar Számítógépes Nyelvészeti Konferencia.

CC BY
VI.02 DOI: 10.31915/NWS.2023.28

A kis gömböc meséje – az ITIdata irodalomtudományos adatbázis fejlesztése 2022–2023-ban[1]

The tale of the roly-poly – development of the ITIdata literary database in 2022–2023

Dobás Kata
Bölcsészettudományi Kutatóközpont, Irodalomtudományi Intézet
dobas.kata@abtk.hu
ORCID: 0009-0009-7632-8276

Fellegi Zsófia
Bölcsészettudományi Kutatóközpont, Irodalomtudományi Intézet
fellegi.zsofia@abtk.hu
ORCID: 0000-0001-9199-1759

Palkó Gábor
Digitális Örökség Nemzeti Laboratórium
ELTE Digitális Bölcsészet Tanszék
Bölcsészettudományi Kutatóközpont, Irodalomtudományi Intézet
palko.gabor@btk.elte.hu
ORCID: 0000-0002-4394-8577

Absztrakt

A Wikibase szoftverrel működő, a Wikidata struktúráját részben követő ITIdata irodalomtudományos adatbázis specifikációját még 2022-ben úgy találtuk ki, hogy minél több típusú projekt befogadására alkalmas legyen. Az elmúlt évben számos kutatás csatlakozott az ITIdata-hoz, így tanulságos volt összegzést tartani az elmúlt egy év történéseiről.

Tanulmányunkban a következő témaköröket fogjuk említeni: milyen konkrét projektek csatlakoztak az adatbázishoz az első tesztprojektünk, a Kosztolányi-forrásjegyzéket követően; milyen ütemben követte az egyik kutatás a másikat; milyen elképzelésekkel érkeztek a kutatók az adatstruktúrát illetően és mi és hogyan valósult meg ebből, valamint milyen workflow-t alkalmaztunk az egyes esetekben. Az adatgazdagítás mikéntjéről is szót ejtünk: a félautomatikus (QuickStatements) és a nagy mennyiségű adatfelvitel tapasztalatairól egyaránt. Fontosnak tartottuk kiemelni, hogy az ITIdata adatstruktúrája hogyan változott meg a különböző adattípusoknak köszönhetően, milyen új tulajdonságok és entitások felvitelére volt szükség, illetve a jövőt illetően milyen lépésekre volt/lesz szükségünk, az adatfelviteli protokoll szigorításától kezdődően az egyes projektek elkülönítésén át az ellenőrző scriptek kidolgozásáig. Tanulmányunk végén azokat a nemzetközi, szintén wikibase szoftverrel működő adatbázisokat is áttekintem, amelyek eredményeit, tanulságait hasznosítani tudtuk az ITIdata fejlesztésekor.

Kulcsszavak: szemantikus web, Wikibase, adatgazdagítás, digitális filológia, workflow

1 Jelen tanulmány a Digitális Örökség Nemzeti Laboratórium támogatásával készült.

Abstract
The specification of ITIdata, a wikibase-based database that partially follows the structure of wikidata, was designed in 2022 to accommodate as many types of projects as possible. In the past year, a large number of research projects have joined ITIdata, so it was instructive to provide a summary of what has happened in the past year.

In our study, we will cover the following topics: what specific projects joined the database after our first test project, the Kosztolányi Resource Directory; the pace at which one research followed another; what ideas researchers came with about the data structure and what and how this was achieved, and what workflow was used in each case. We will also discuss how data was managed: both semi-automatic (QuickStatements) and experiences with large-scale data entry. We considered it important to highlight how the data structure of ITIdata has changed due to the different data types, what new properties and entities needed to be added, and what steps were/are needed for the future, from tightening the data upload protocol to the separation of each project and the development of control scripts. At the end of our study, I will also review the international databases, also using wikibase software, whose results and lessons learned we could use in the development of ITIdata.

Keywords: semantic database, Wikibase, digital philology, data management, workflow

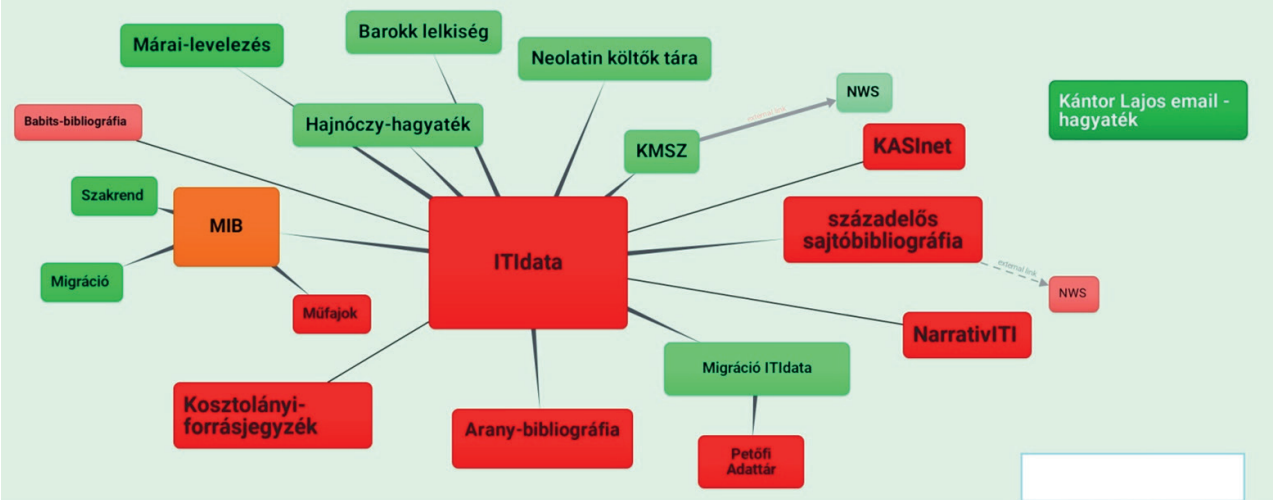
1. Bevezetés

A tavalyi év során, amikor az ITIdata, a Bölcsészettudományi Kutatóközpont Irodalomtudományi Intézetének Wikibase alapú szemantikus adatbázisának kezdeti lépéseiről számoltunk be, még csupán egyetlen projektünk volt, amelyik már látható eredményeket tudhatott magáénak, s amelynek nagyobb részben készen volt a specializációja is: a Kosztolányi Dezső-forrásjegyzék. Ez egy 10 ezer tételt magába foglaló bibliográfia, amelynek egy része már felkerült az adatbázisba. A 2022-es év során kiépítettük az adatbázis alapjait, létrehoztuk a legfontosabb tulajdonságokat (properties) és elemeket (items), mindezeket a Wikidata struktúrájával és felépítésével szinkronban tettük, a rekordokhoz és a tulajdonságokhoz minden esetben hozzákapcsoltuk a WikiData azonosítót (ha volt ilyen), megkönnyítendő egy majdani adatmigrációt. Természetesen egy-egy projekt megkívánhatja az eltéréseket, illetve léteznek ellentmondások a WikiData-n belül is, ettől függetlenül ezt a gyakorlatot kívánjuk folytatni a továbbiakban is. Jelenleg 280 ezer személynév rekord, 187 egyetem/kar, 2590 foglalkozás, 17850 településnév, összesen pedig 336 190 rekord szerepel az ITIdata-ban, amelynek száma napról napra növekszik.

2. Projektek

A 2023-as évre már 10 fölé nőtt az ITIdata-ban helyet kapó projektek száma. Ezekről az alábbi ábra ad áttekintést:

CC BY



1. ábra: Az ITIdata szemantikus adatbázisban található projektek buborékábrája
Forrás: saját szerkesztés

A projekteket az alábbiakban is felsoroljuk. A különböző kutatások mögött egy kutató (Hajnóczy-hagyaték, Századelős sajtóbibliográfia), vagy egy egész kutatócsoport munkája is áll (Neolatin Költők Tára, Kosztolányi-forrásjegyzék):

- Hajnóczy Péter hagyatéka (Ludmán Katalin)
- Magyar Irodalomtörténet Bibliográfiája (Irodalomtudományi Intézet, Bibliográfiai Osztály)
- Márai-levelezés (Mihályi Ödönnel folytatott levelezés, Ötvös Anna)
- Barokk irodalom lelkiség (Pázmány Péter Katolikus Egyetem, kutatócsoport)
- Neolatin Költők Tára (Irodalomtudományi Intézet, kutatócsoport)
- Arany János-bibliográfia (Gönczy Monika)
- Kolozsvári Állami Magyar Színház jelmeztervei (Szabó-Reznek Eszter, Sidó Zsuzsa)
- Századelős sajtóbibliográfia (Virágh András)
- Kosztolányi-forrásjegyzék (Dobás Kata)
- NarratívITI (Irodalomtudományi Intézet, Elméleti Osztály)
- KASInet (Irodalomtudományi Intézet, 20. századi Osztály)
- Babits-bibliográfia (Bucsics Katalin, Káli Anita)
- Petőfi-adattár (Jakab Éva)

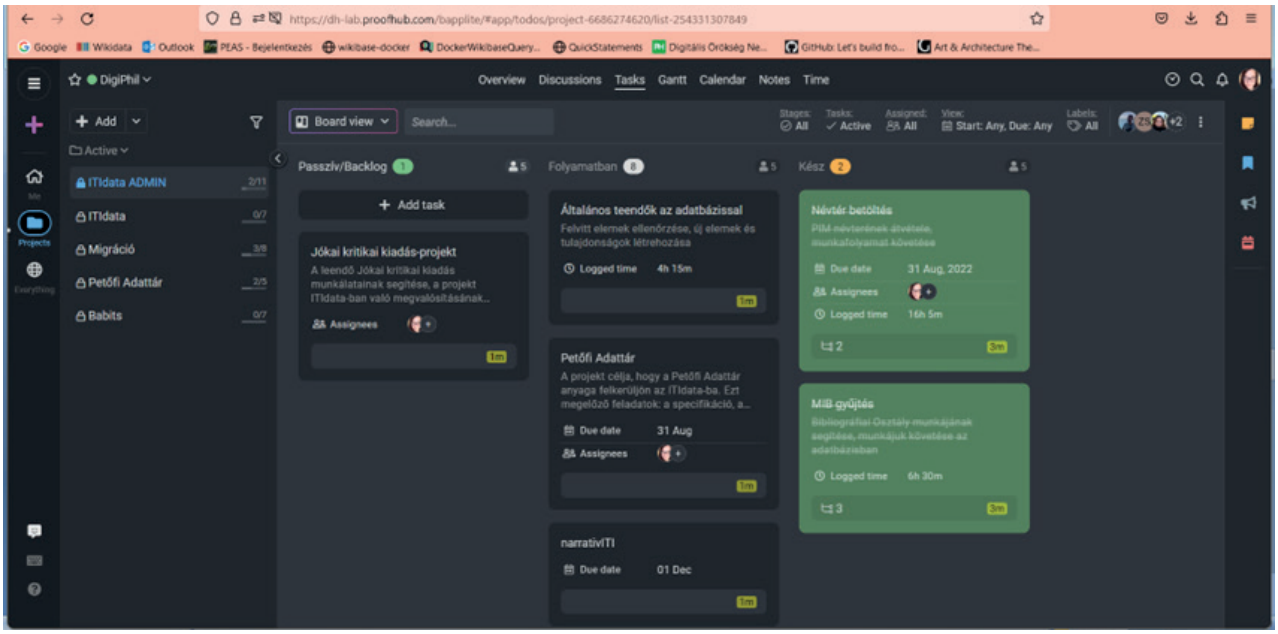
3. Workflow

3.1. Általános workflow

Mint az talán a fentiekből is érzékelhető, a projektek megnövekedett száma másfajta munkamenetet kívánt meg tőlünk. A munkaszervezést ezért a megbeszélések és az adminisztráció szintjén is átalakítottuk. Ezekről általában kevés szó esik, ezért különösen fontosnak tartjuk azt, hogy ezekről is írjunk, hiszen a munka sikerességét nagy részben ez garantálja. Az egyik sarokkő a megbeszéléseink átszervezése lett. Ezek az alábbiak szerint strukturálódnak:

- hétindító meetingek,
- vezetői meetingek,
- fontos döntésekre meeting,
- projektek meetingjei.

A projektek számának növekedésével párhuzamosan az adminisztrációnk is jelentős mértékben megnövekedett, ezért egyre fontosabb lett a megbeszélések és vezetői döntések dokumentációja, hogy az egyes kutatások, adatfelvitelek, feladatok státuszát követni tudjuk. Éppen ezért kiemelten fontossá vált, hogy minden megbeszélésről jegyzeteket készítsünk. A DigiPhil csoport, a Digitális Örökség Nemzeti Laboratóriummal együttműködésben a széles körben használt, ProofHub[2] elnevezésű projektmenedzsment-szolgáltatást használja. A DigiPhil a saját szerveren működő ProofHub-ban rögzíti a munkaidő ráfordítást és az egyes projektekhez kapcsolódó feladatokat, határidőket és a folyamatok előrehaladását. A megbeszélésekről készült jegyzeteket is itt tároljuk, hozzáférése az adott részprojekthez tartozó személyeknek van. A szolgáltatás használata jelentősen megkönnyíti a feladatok átláthatóságát, a munka visszaellenőrizhetőségét.



2. ábra: Képernyőkép az ITIdata projektek aktuális állását tartalmazó ProofHub oldalról, 2023. 08. 01. Forrás: https://dh-lab.proofhub.com/bapplite/#app/todos/project-6686274620/list-254331307849

3.2. Egyes projektekhez kapcsolódó workflow.

Mivel exponenciálisan nő az adatbázishoz csatlakozó kutatási projektek száma, 2023 elején új workflow-t dolgoztunk ki arra, hogy hogyan vezessünk be egy tetszőleges új projektet az ITIdata-ba. Az újonnan csatlakozó projektek két csoportba sorolhatók: amennyiben egy teljesen új projekt kiépítéséről van szó, és amennyiben egy már meglévő adatbázis migrációját hajtjuk végre.

Ha új projektet szeretnénk elindítani, akkor a specifikációt a meglévő metaadatok alapján készítjük el, a kutatókkal közösen. A mappinget követően elkészítjük a projekthez tartozó leírási útmutatót, megtörténik a betanítás, és elkezdődhet az adatfelvitel, ebben az esetben manuálisan. Amennyiben már meglévő adatbázist kell integrálnunk az ITIdata-ba, úgy szintén mapping az első lépés, majd a tulajdonságok megfeleltetése után következik az adatok migrációja. Ez az adatok mennyiségétől függően történhet automatikus átalakító és betöltő algoritmusok, vagy a Wikibase által biztosított QuickStatements modul használatával. Utóbbi esetében az adatokat táblázatban helyezzük el, és megfelelő formátumban előkészítjük őket, a táblázatból generált CSV formátumot az online felületen keresztül töltjük be. Ez a metódus

2 Bővebben a ProofHub-ról lásd: https://www.proofhub.com/ (Hozzáférés dátuma: 2023. 08. 01.)

CC BY

néhány ezer rekord betöltéséig rendkívül hatékonyan működik, azonban több tíz- és százezer rekord esetén python nyelven írt betöltő algoritmusokat készítünk, amelyek valamilyen szabványos adatcsere-formátumból (pl. CSV, Marc21 stb.) API protokollon keresztül töltik be a rekordokat. A szemantikus kapcsolatok döntő hányadát manuálisan hozzuk létre, illetve az ellenőrzés is így történik. Ez alól kivétel a Magyar Irodalomtörténet Bibliográfiájának több mint 130 ezer rekordja, ahol a Qulto Kft.-vel együttműködésben automatikusan zajlik a szemantikus kapcsolatok kialakítása, a MIB névtérre fektetése. Mivel az egyes projektek igényei eltérőek lehetnek, ezért minden projektnek saját specifikációja van, de ezek között természetesen nagy átfedések vannak.

4. Mit tegyünk, hogy a kis gömböc sorsára ne jussunk?

Ahhoz, hogy a rengeteg projekt ne feszítse szét sem szervezésben, sem belső struktúrákban a munkánkat, fontosnak tartottuk azt a kikötést megtenni, hogy az új kutatások legalább 80%-ban véglegesített specifikációval léphessenek be az ITIdata-ba. Eddig minden esetben szükség volt utólagos korrekcióra, kiegészítésre, erre a szemantikus adatbázis természetéből fakadóan van lehetőség, de az előkészítő munkálatoknak mindenképpen le kell fednie a specializáció nagy részét.

Kiemelten fontos tulajdonságként tekintünk a P1-es (osztály, amelynek példánya/instance of) tulajdonságra, amely az adott rekord besorolását első lépésben elvégzi (ember, irodalmi mű, egyetem stb). A Wikibase egyik legnagyobb előnye, szemben például egy könyvtári rendszerrel, hogy az egyes rekordok adatszerkezete rugalmasan alakítható ki, a kutatás során felmerülő új adatok bármikor hozzárendelhetők a rekordokhoz. Egy tulajdonság létrehozásakor meg kell határozni, hogy a tulajdonsághoz milyen érték kapcsolódhat, mint például entitás, dátum, szöveges érték vagy URL. Azt azonban külön kell meghatározni például az entitások esetében, hogy a kapcsolt entitás milyen további tulajdonságokkal kell, hogy rendelkezzen. Ezek az úgynevezett megkötések, amelyek nélkül olyan értelmetlen állítások is létrehozhatók egy személy entitás esetében, hogy például a 'neme' tulajdonsághoz a személy foglalkozását rendeljük. Ezzel összefüggésben beállítottunk megkötéseket az adatbázison belül. Például nem lehet születési idő tulajdonságot felvinni egy irodalmi műhöz. Ha ez mégis megtörténik, a rendszer egy felkiáltójelet tesz az adatsor mellé.

További fontos kikötés, hogy új tulajdonságot kizárólag a rendszer adminisztrátorai vihetnek fel, a megfelelő Wikidata tulajdonsággal harmonizálva azt.[3] Az új rekordokat a leírási útmutató alapján töltik fel a kutatók, ezek ellenőrzésére, lektorálására is lehetőség van a lap alján található „ellenőriztem" funkció használatával.

Az ITIdata-n belül lehetőség van a tulajdonságok felviteli sorrendjének beállítására. Így ha egy adatfelvitel során kimaradna egy tulajdonság, azt fel lehet venni utoljára is, a rendszer automatikusan besorolja azt a megfelelő helyre, mentés után.

A rendszer bonyolultsága miatt több, táblázatos formában tárolt listát vezetünk, ebből a két legfontosabb a tulajdonságok listája illetve a műfajok listája. Az utóbbi az irodalomtudományi adatbázisok esetében kifejezetten indokoltnak bizonyult. A műfaji hierarchiák kiépítettsége, a műfaji átfedések miatt, a használt rendszerünk átláthatóságát többek között ez a táblázat is garantálja.

3 Jelenleg 202 tulajdonság található az ITIdata-ban, amelynek teljes listája elérhető itt: https://itidata.abtk.hu/w/index.php?title=Special:ListProperties/&limit=250&offset=0 (Hozzáférés dátuma: 2023. 08. 01.)

Bevezettük az egyplatformos kommunikációt. A Bölcsészettudományi Kutatóközpont Microsoft Teams felületén a DigiPhil önálló felülettel rendelkezik, ezen belül minden projektnek létrehoztunk egy-egy önálló csatornát, ahová kizárólag azokat a tagokat vettük fel, akiket valamilyen mértékben érint a kérdéskör. Ez jelentős mértékben megkönnyíti a kommunikációnkat a tagokkal, sem az e-mailek, sem más chatprogramok nem bizonyultak elég hatékonynak és átláthatónak. Itt lehetőség van a munkafájlok, útmutatók, videós oktatási anyagok tárolására is, illetve arra is, hogy párhuzamosan szerkesszük a dokumentumokat.

Jelenleg olyan algoritmusok fejlesztésén dolgozunk, amelyek lehetővé teszik a tömeges adatellenőrzést, előre meghatározott szabályok mentén. A középtávú tervek között szerepel az automatikus adatszinkronizáció olyan külső rendszerekkel, mint a WikiData vagy az MTMT. Munkaszervezés szempontjából célunk, hogy a DigiPhil csoportban olyan felelősöket jelöljünk ki, akik az egyes kutatási projekttel való mindennapi kommunikációért és az adatok minőségéért felelnek, ezzel is gyorsítva a folyamatokat.

5. A szerkezeten innen, az adatokon túl

2022 második felében kezdtük meg két, az Irodalomtudományi Intézetben folyó, egymással összefüggő kutatási projekt, a NarrativITI és a KASInet ITIdatá-ba történő integrálását.[4] Előbbi a narratológia alapvető fogalmairól készít szócikkgyűjteményt, valamint válogatott bibliográfiát, utóbbi Kassák Lajos kapcsolatrendszerét dolgozza fel. Mivel mindkét projekt az adatokon túl narratív szövegeket is publikál, olyan megoldásra volt szükségünk, amely lehetővé teszi az adatok és a szövegek egy rendszerben történő leírását és tárolását. A szövegek közlésének és az adatgazdagítás specifikációjának kidolgozása során a Digitális Bölcsészet Tanszék közreműködésével készült *A tudományos tudás áramlásának mintázatai a Magyar Királyságban, 1770–1830* projekt gyakorlatát követtük.[5] A Wikibase rendszer lehetővé teszi a Wikipedia szócikkeihez hasonló oldalak létrehozását, felhasználóbarát vizuális felületet (Visual Editor) biztosít azok szerkesztéséhez, valamint támogatja a szövegek összekapcsolását más szövegekkel és magával a szemantikus adattérrel. A két kutatási projekt teljes anyaga így egy rendszerben tárolható és kereshető, a kutatók által felvitt szemantikus adatok pedig tovább gazdagítják az ITIdata állományát.

6. Inspirációk, külföldi példák, részvételek

A számos Wikibase alapú külföldi adatbázisból kettőt szeretnénk kiemelni: a német történészek által működtetett FactGridet[6] és a finn Lettersampoo-t.[7] Az előbbiben elsősorban a tulajdonságok jól kiépített struktúráját tartjuk követendő példának,[8] jelenleg is dolgozunk egy hasonló kiépítésén a saját felületünkön. Az utóbbinál pedig az átláthatóságot és a könnyen kereshetőséget emelnénk ki. Az ITIdata-n belül jelenleg is dolgozunk keresőspecifikációkon, ezért a finn adatbázist szintén jó példának tartjuk.

---

4 A NarrativITI honlapja: http://narratologia.btk.mta.hu/ (Hozzáférés dátuma: 2023. 08. 01.)

5 A Tudásáramlás projekt leírása és a közzétett forrásszövegek az alábbi linken érhetők el: https://eltedata.elte-dh.hu/wiki/Tud%C3%A1s%C3%A1raml%C3%A1s (Hozzáférés dátuma: 2023. 08. 01.)

6 A FactGrid elérhetősége: https://database.factgrid.de/wiki/Main_Page; https://blog.factgrid.de/ (Hozzáférés dátuma: 2023. 08. 01.)

7 A Lettersamppo elérhetősége: https://lettersampo.demo.seco.cs.aalto.fi/en (Hozzáférés dátuma: 2023. 08. 01.)

8 https://database.factgrid.de/wiki/FactGrid:Directory_of_Properties (Hozzáférés dátuma: 2023. 08. 01.)

CC BY

2023-tól tagjai vagyunk a DARIAH-EU szervezet BiblioData Working Groupjának,[9] amely számos együttműködést tesz lehetővé, illetve számos szakirodalomról, kiadványról első kézből értesülhetünk.[10]

7. Jövőbeli tervek

A DigiPhil csapat jövőbeli tervei között szerepel, hogy minden egyes projekthez készítsünk olyan belépési pontokat, ahol a felhasználók könnyen és egyszerűen tudnak az adott projekt rekordjai között keresni; a keresőspecifikáció fejlesztése folyamatban van. További tervünk olyan ellenőrző scriptek írása, amelyeket időről időre lefuttatunk az ITIdata rendszerén belül, hogy az anomáliák, duplumok szűrése, illetve a javítás egyszerűbbé váljon. Fontosnak tartjuk, hogy a rendszeren belül további megkötéseket hozzunk létre, ezzel is segítve az adatfelvitel menetét, az ITIdata használatát.

9 https://www.dariah.eu/activities/working-groups/bibliographical-data-bibliodata/ (Hozzáférés dátuma: 2023. 08. 01.)

10 A legfrissebb példa erre a Working Group által írt és összeállított kiadvány: *An Analysis of the Current Bibliographical Data Landscape in the Humanities. A Case for the Joint Bibliodata Agendas of Public Stakeholders*, https://zenodo.org/record/6559857 (Hozzáférés dátuma: 2023. 08. 01.)

 DOI: 10.31915/NWS.2023.29

Kutatói e-mail hagyaték archiválása és feldolgozása

Long-term preservation of e-mail heritage

Alföldi István
Digitális Örökség Nemzeti Laboratórium
alfi@poliphon.hu
ORCID: 0009-0002-9634-0482

Szemigán Dorottya Henrietta
Digitális Örökség Nemzeti Laboratórium
szemigan.dorottya@btk.elte.hu
ORCID: 0009-0007-5116-838X

Palkó Gábor
Digitális Örökség Nemzeti Laboratórium
Bölcsészettudományi Kutatóközpont, Irodalomtudományi Intézet
palko.gabor@btk.elte.hu
ORCID: 0000-0002-4394-8577

Fellegi Zsófia
Bölcsészettudományi Kutatóközpont, Irodalomtudományi Intézet
fellegi.zsofi@abtk.hu
ORCID: 0000-0001-9199-1759

Abstract

The management of born-digital content is becoming an increasingly relevant task in the different cultural fields, thus the long-term preservation of digitally generated material has been in the focus of the archivist community since the beginning of the new millennium. In the early twenty-first century, the Library of Congress pointed out its concerns about the fragility of the solely digitally created records. Since such material does not have an analog (physical) version, there is a greater risk of either getting lost, or not being available or accessible in their original form as historical records for future researchers.

Preserving digitally generated correspondence between individuals has been an important research area of digital archiving for many years. There are already good practices for archiving e-mails, being a priority especially in the corporate sector. However, merely preserving and making e-mails researchable is not the only issue, since in the case of the cultural heritage materials for instance a critical attitude and a curatorial practice is also essential in the archiving work. Furthermore, one of the reasons why a general solution to e-mail preservation still has not yet been found is that archiving e-mails at different stages of their lifecycle requires radically different solutions.

Our research project at the National Laboratory for Digital Heritage focuses on the long-term preservation of an e-mail heritage, more precisely the electronic correspondence of a Hungarian critic and literary historian from Cluj-Napoca (Transylvania), Lajos Kántor. We

CC BY

are investigating how e-mail archiving tools can be used in this specific field, how e-mails can be made visible and researchable as an integral part of the overall heritage, and how the experience gained in preserving and processing paper-based correspondence can be used in archiving and publishing e-mails.

Keywords: Born-digital, e-mail preservation, cultural heritage, long-term preservation

Bevezetés

A hálózatba kötött elektronikus levelezés robbanásszerű fejlődése az elmúlt 52 év egyik jelentős technikai és társadalomtörténeti fejleménye volt. A számítógép alapú levelezés kezdeti szakaszában, az 1960-as években kialakított time-sharing[1] révén lehetővé vált az elektronikus üzenetek küldése ugyanazon rendszer/számítógép felhasználói között, majd Ray Tomlinsonnak köszönhetően 1971-ben már ennek hálózatos kiterjesztése is megszületett. Tomlinson az MIT meghízásából az internet elődjének, az ARPHANet (Advanced Research Project Agency Network) hálózatának kiépítésében segédkezett, amelynek rendszerében később létre is hozta az első e-mail programot.[2] Tomlinson voltaképpen összekötötte a SNDMSG elektronikus levelezésre használt programját és az ARPHANet hálózatát egy kísérleti file-transfer program segítségével, emellett meghatározta a mára közismert cím-szintaxist[3] a @ szimbólum felhasználásával. Bár kezdetben csupán egy szűk szakmai réteg részére gyorsította meg jelentősen az információ továbbítását, mára egy mindenütt jelenlévő kommunikációs eszközzé vált, mely releváns forrásnak tekinthető a történeti vizsgálódások számára.

Az e-mail történetének több mint fél évszázada során számos átviteli protokollt és különböző szabványt határoztak meg, emellett megannyi felhasználói eszköz, e-mail kliens jött létre, melyek nagyfokú diverzitása nagyban megnehezíti az e-mailek archiválási folyamatát, amelyre jelenleg még nem áll rendelkezésre szabványosított megoldás. A problémát tovább bonyolítja, hogy az e-mailek rendkívül heterogén digitálisan keletkezett (born-digital) objektumok, melyek különböző fájlokat és adatokat tartalmazhatnak, a különböző szolgáltatók és kliensek pedig különféle sztenderdeket követhetnek. Mivel a formátumok idővel elavulhatnak, így ezen dokumentumok hozzáférhetősége megszűnik, fontos hosszútávú megőrzésük és integritásuk biztosítása.

A Digitális Örökség Nemzeti Laboratórium (DH-LAB) born digital alprojektje a digitálisan keletkezett kulturális örökség hosszútávú megőrzésének,[4] archiválásának és kereshetővé tételének módszertanát hivatott kidolgozni az adatvesztés minimalizálásának célkitűzésével.[5] Az aktuális pilot projektünk fókuszába a kutatói e-mail hagyatékok fent jelzett szempontok szerinti vizsgálatát állítottuk.

---

1 A time-sharing lehetőség voltaképpen a számítási erőforrások megosztott használatát jelentette a számítástechnikában. Mivel a számítási kapacitás és a (A time-sharing történetéhez lásd: J.A.N. Lee, "Claims to the Term 'Time-Sharing,'" *IEEE Annals of the History of Computing* 14, no. 1 (1992): 16–54, https://doi.org/10.1109/85.145316.)

2 Az e-mail fejlődésének technikai részleteiről lásd Craig Partridge, "The Technical Development of Internet Email," *IEEE Annals of the History of Computing* 30, no. 2 (April 2008): 3–29, https://doi.org/10.1109/MAHC.2008.32.

3 user@host

4 Az alprojektben résztvevő munkatársak: Alföldi István (born digital szakértő), Dobás Kata (ITIdata specifikáció), Fellegi Zsófia (digitális filológiai szakértő), Indig Balázs (Python programozás), Szemigán Dorottya Henrietta (born digital szakértő), Palkó Gábor (projektvezető).

5 Emellett hosszútávú terveink közé tartozik ezen anyagok közzététele, illetve oktatói anyagok létrehozása is.

A DH-LAB DANUBE AI programja a határon túli hungarikumok, a magyar nyelvű vagy vonatkozású kulturális örökség felkutatásával foglalkozik, valamint kiemelt célja együttműködések kialakítása a közép-európai régió kutatási és oktatási intézményeivel, a tudáscsere és a tudományos hálózatépítés. E program biztosította továbbá a lehetőséget, hogy a Minerva Művelődési Egyesülettel együttműködésben kutatási és archiválási célra megkapjuk Kántor Lajos elektronikus levelezését.[6]

Kántor Lajos kolozsvári születésű, meghatározó jelentőségű kulturális és közéleti szereplő, termékeny kritikus, szerkesztő volt. Élete során közel 70 önálló és általa szerkesztett kötet jelent meg. Haláláig a *Korunk* kolozsvári kulturális folyóirat főszerkesztője volt, emellett kiterjedt kultúraszervező tevékenységet folytatott Kolozsváron, többek között alapító tagja és elnöke is volt (2006-tól egészen 2017-ben bekövetkező haláláig) a Kolozsvár Társaságnak, mely a város magyar kisebbségének gazdasági, társadalmi és kulturális integrálását tűzte ki célul. Tekintélyes életműve, ezen belül nagy terjedelmű levelezése – elsősorban, de nem kizárólagosan a határon túli magyar irodalomtörténetírás számára – értékes forrás. Elsődlegesen a *Korunk* szerkesztői levelezése számtalan érdekes és eddig ismeretlen adatot tartalmazhat irodalom-, művészet- és társadalomtörténeti vonatkozásban is.

Kántor Lajos e-mail hagyatéka tulajdonképpen hibrid jellegű, hiszen az összesen 6 különböző e-mail fiók leveleit tartalmazó hagyatékrész mellett a korai, csupán papíron megőrzött, kinyomtatott leveleket is az e-mail hagyaték szerves részeként értelmeztük, ezeket digitalizáltuk.[7]

Az eredeti e-mailek biztosítása

Az első feladat a jogi kérdések tisztázása és a megállapodás aláírása után az örökösöktől és a kolozsvári Minerva Művelődési Egyesülettől megkapott elektronikus levelezés biztosítása volt. Az e-mailek átadása egy Google Gmail fiók hozzáférési adatainak megadásával történt. Kántor Lajos 2017-ben bekövetkezett halála óta a fiókot az örökösök kezelték.
Az e-mailek biztosítása a következő feladatokat jelenti:

- A Gmail fiók biztosítása
- MBOX mentés készítése
- Az MBOX állomány integritásának biztosítása

A Gmail fiókról fontos tudni, hogy ha egy ideig nem lépünk be, az e-mailek törlődnek. Ezt elkerülhetjük, ha beállítjuk a Beállítások \ Adatok és adatvédelem \ További lehetőségek \ Rendelkezzen digitális hagyatékáról szekció alatt, hogy huzamosabb inaktivitás esetén mi történjen az e-mailekkel és kit értesítsen a rendszer.

A Gmail fiókból a Google Takeout segítségével készítettünk MBOX mentést. Az MBOX egy széleskörűen támogatott sztenderd e-mail konténer formátum, amely több e-mail strukturált tárolását teszi lehetővé. A további feldolgozás során az MBOX állományokból indultunk ki, ezért fontos volt az állomány integritásának biztosítása. Ehhez a SHA (Secure Hash Algorithm)

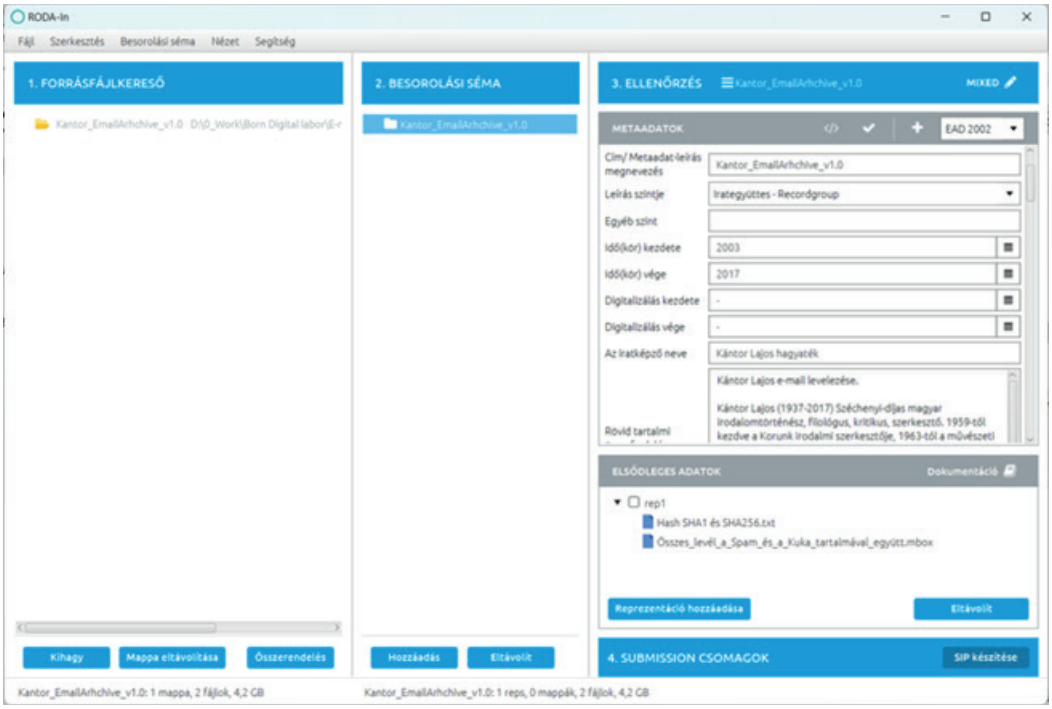
6 A Kántor Lajos hagyatékot alapvetően a Minerva Művelődési Egyesület gondozza, ennek egy részét, a hagyaték elektronikus levelezését bízták a DHLAB-ra.

7 A nyomtatott e-mailek szkennelését követően, a DH-LAB által fejlesztett másodlagos GUI felület mesterséges intelligencia alapú OCR (Optical Character Recognition) moduljának segítségével felismertettük nyomtatott levelek szövegét, így a továbbiakban a digitális hagyatékkal együtt tudtuk kezelni.

CC BY

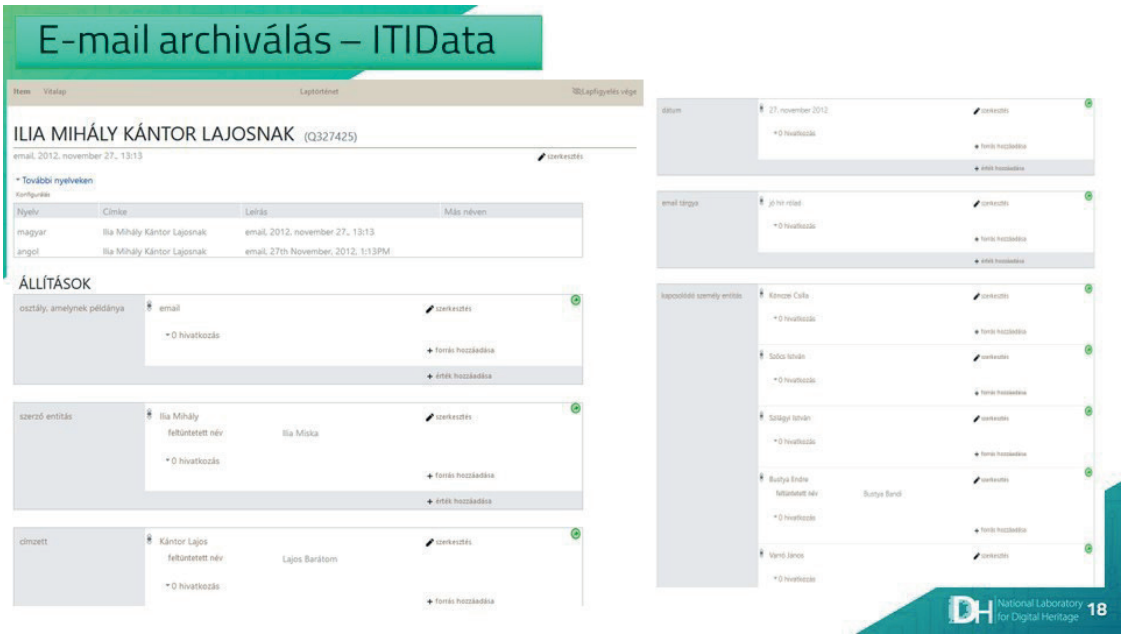
256 bites hash formátumát használtuk. Az algoritmus egy úgynevezett „visszafordíthatatlan és egyedi kivonatot” készít, amellyel a későbbiek során bitszinten ellenőrizhető, hogy nem történt változás az állományban. Az SHA-256 hash-t a Windows Power Schellbe beépített Get-FileHash eszközével hoztuk létre.

Az MBOX fájlból ezek után készítettünk egy E-ARK SIP formátumú beadási csomagot. A Born-digital alprojekt kimondott célja az Európai Bizottság E-ARK/eArchiving programja által készített és támogatott digitális archiválási specifikációk és eszközök tesztelése, használatuk magyarországi előmozdítása. A digitális archiválás elméleti alapjait lefektető Open Archival Information System (OAIS) referenciamodell definiálja a beadási csomagot, de nem rendelkezik annak pontos tartalmáról és struktúrájáról. Az E-ARK projekt során kidolgozott E-ARK SIP egy szabványos csomag formátum az archívumnak átadandó állományok és metaadatok kezelésére. Az E-ARK SIP csomagot a RODA-in SIP-készítő eszközzel állítottuk elő.



1. ábra: E-ARK SIP készítése

A csomagot végül a CERN nyílt adatrepozitórium szolgáltatásának (Zenodo.org) segítségével archiváltuk. Ezekkel az eszközökkel biztosítottuk a hagyaték hosszútávú szabványos megőrzését, és a továbbiakban a feldolgozásra koncentrálhattunk.



2. ábra: A Kántor hagyaték a Zenodo repozitóriumban

MBOX állományok feldolgozása

Az előbbiekben archivált MBOX konténer a feldolgozás nyers bemeneti állománya. Ebben megtalálható az összes levél, függetlenül annak fejlécadataitól és tartalmától. A feldolgozás előtt el kellett végezni a levelek szűrését. A nyers állomány összesen 11245 e-mailt tartalmazott.

A következő leveleket szűrtük ki a feldolgozás első lépéseként:

Szűrési feltétel	Levelek száma
Összes levél	11245
Érzékeny tartalmú levelek (családon belüli levelezés)	779
Levélszemét és egyéb lényegtelen e-mail	1710
Duplikátumok, tömeges levelek	0
Maradt	8756

3. ábra: Elektronikus levelezés szűrésének feltételei

Először az érzékeny tartalmú levelek kerültek kiszűrésre. Az örökösök megnevezték azokat a családtagokat, akik sem címzettként, sem feladóként nem szerepelhettek a publikálandó levelek között. Az összes többi levelet elvben publikálhatónak nyilvánították. A következő csoport a levélszemét volt. A levélszemetet vagy eleve el sem mentették, vagy később törölték, mert alig találtunk biztosan kéretlen leveleket. Viszont sok olyan körlevél és hírlevél volt, ami Kántor Lajos halála után keletkezett. Ezeket is kiszűrtük. Megtartottuk viszont a halála előtt keletkezett hasonló leveleket, mert olyan elvek definiálása lehetetlennek tűnt, amelyek mentén egyértelműen lehetett volna dönteni arról, hogy melyik levél nem releváns a hagyaték összefüggésében. A duplikátumok azonosítására sem áll rendelkezésre megkérdőjelezhetetlen módszer, ezért a válaszlevélben megismételt, vagy különböző feladótól többször megkapott leveleket sem szűrtük ki.

CC BY

A következő lépés a metaadatok (voltaképpen az e-mail fejléc [header] adatainak) kinyerése és elemzése volt. A legfontosabb e-mail metaadatokat a feldolgozás ezen korai szakaszában a Python programnyelven készült email.parser alkalmazással[8] nyertük ki és rendeztük táblázatos adatformátumba.

A csak a metaadatokat és a csatolmányok neveit tartalmazó táblázat segítségével különböző bontásokban (dátum, szerző, címzett, tárgy) lehet elemezni az e-maileket.

Papíralapú levelezés tapasztalatainak felhasználása

A pilot keretei között arra tettünk kísérletet, hogy a digitális kritikai levelezések kiadása során a DigiPhil projektben összegyűjtött tapasztalatokat összevessük az elektronikus levelezések publikálásának gyakorlati kérdéseivel. A levelek szemantikus adatgazdagítására az azokban található bibliográfiai és filológiai adatokat az Irodalomtudományi Intézet szemantikus web alapú adatbázisát, az ITIdata-t használtuk. Megvizsgáltuk annak a lehetőségét is, hogy a papír alapú levelezések digitális kiadásánál széles körben használatos jelölőnyelvi ajánlás, a Text Encoding Initiative alapján írjunk le e-maileket.

ITIdata

Elsőként az ITIdata hagyományos levelezések adatainak leírására készített specifikációját optimalizáltuk, figyelmet fordítva az e-mailek sajátosságaira. Az ITIdata rendszerében minden e-mailt különálló entitásként képeztünk le, a rekordok egyedi azonosítót kaptak, az entitáshoz olyan adatokat rendeltünk, mint a szerző, címzett, küldés időpontja, fájlformátum stb. A szerző és címzett esetében, az adatbázisban entitásként szereplő rekordok adott e-maillel való összekapcsolásán túl feltüntettük a szerző aláírását és a címzett megszólítását, ezek sokszor a beceneveiket reprezentálják. Az egyes e-mailek rekordjainál kapcsolódó rekordként feltüntettük továbbá a levélben említett személy- és helyneveket. Az ITIdata rendszerében új tulajdonságokat is létre kellett hoznunk a hagyaték adatainak kifinomult leírásának érdekében, ilyen például az e-mail tárgyának leírására szolgáló tulajdonság, ennek segítségével váltak sorrendezhetővé a levelek. Az e-mailekben említett művek bibliográfiai adatait szintén feltöltöttük a rendszerbe, azokat az e-mailt reprezentáló rekorddal összekapcsoltuk.

8 Lásd: https://docs.python.org/3/library/email.parser.html



4. ábra: ITIdata: egyedi e-mail entitás és kapcsolódó állításai

TEI jelölőnyelv és XML séma felhasználása

A DigiPhil szakmai irányításával készült a levelezés digitális szövegkiadását reprezentáló TEI XML átirat specifikációja. A DigiPhilnek kéziratos levelezés kritikai kiadásainak digitális közzétételében volt gyakorlata,[1] az e-mailek specifikációjának kidolgozásakor ezeket a tapasztalatokat vettük alapul. Ahogyan arról Fellegi Zsófia beszámolt,[2] a DigiPhil a korábbi gyakorlatához képest változtatott a TEI XML specifikációin, a metaadatokat az ITIdatá-ban írja le, a TEI XML fájlok meghatározott pontjain pedig az adatbázisra hivatkozik. A specifikáció kialakítása ennek az új gyakorlatnak a figyelembevételével történt.

A digitális kiadás címét, a korábbi, a hagyományos filológiai gyakorlatból eredeztethető módon, a levél szerzőjének, címzettjének és a levél megírásának dátumából hoztuk létre. A TEI XML fájl fejlécében, amely a levélre vonatkozó metaadatokat tartalmazza, a „correspAction" jelölő alatt tüntettük fel a feladó és a címzett nevét, a levél küldésének pontos idejét, ezeket összekapcsoltuk az ITIdata megfelelő entitásaival. Fontos kiemelni, hogy bár a DigiPhil gyakorlatában a metaadatok kifinomult leírása az ITIdatában történik, nem várható el a felhasználótól, hogy minden esetben együtt vizsgálja a szövegközlést és az adatbázist, így a leveleket reprezentáló TEI fájlokban is, karaktersorok szintjén közölnünk kell a releváns adatokat.

1 Ehhez részletesebben lásd Móricz Zsigmond levelezésének kritikai kiadásáról szóló beszámolót: Cséve Anna, Fellegi Zsófia, Kómár Éva: *Móricz Zsigmond levelezésének (1892–1913) digitális kritikai kiadása*, Digitális bölcsészet 1. sz. (2018.) 159-174 https://doi.org/10.31400/dh-hun.2018.1.227, valamint Fellegi Zsófia, Palkó Gábor: *Arany-kéziratok és kritikai kiadások közzététele az Arany János Emlékévben*, Helikon 66. Sz. (2020) 82-98.

2 Fellegi Zsófia: *A digitális filológia infrastruktúrái. A DigiPhil megújulásáról*, Valós térben - az online térért, NIIF Networkshop konferenciák; 31., szerk. Tick József, Kokas Károly, Holl András, MTA Könyvtár és Információs Központ (Budapest:2022) 338-344. Ld. Még: Fellegi, Zsófia, & Palkó, Gábor. (2023, June 14). Publishing Digital Text Editions on the Semantic Web. Open Repositories 2023 (OR2023), Stellenbosch, South Africa. Zenodo. https://doi.org/10.5281/zenodo.8091659

CC BY



5. ábra: TEI jelölőnyelv XML séma kódrészlet, kapcsolódó entitások kódolása, „idno" identifier használata

A szövegközlés betűhű, tehát karakterszinten követi az e-mailben található szöveget, azonban a levelek szövegében is elvégeztük az adatgazdagítást, tehát például a személyneveket jelölővel láttuk el és összekötöttük az ITIdatával.

Az egyéb technikai adatok *(paradata)* rögzítése még nem definiált. Egy lehetséges megoldásként javasoljuk a fájl méretének feltüntetését a ‚measure' elemben a @type=”quantity” és a @ unit=”kb” attribútum-értékekkel, emellett még a fájlformátum, illetve a hordozó fájltípus feltüntetése is szükséges lesz.

További tervek

Ahogy azt fentebb is jeleztük, a Born-digital alprojekt pilotjának fontos célja volt megvizsgálni, hogy a papíralapú levelezés feldolgozása és publikálása során szerzett tapasztalatokat hogyan lehet hasznosítani az elektronikus levelek feldolgozásában. A Digitális Örökség Nemzeti Laboratórium egy másik kiemelt projektje a webaratás. A labor folyamatosan végzi a kutatási és innovációs szempontból releváns webes források kiválasztását, aratását és az ehhez szükséges technológiák fejlesztését. A webaratás projekt célja az archiváláson kívül az összegyűjtött anyagok kutathatóvá tétele. A webaratás és az adatok publikálása saját fejlesztésű Python-nyelvű szoftverekkel történik. Az e-mailek feldolgozása közben jöttünk rá, hogy valószínűleg ez az eszközkészlet is jól hasznosítható az e-mailek további tartalmi feldolgozása és publikálása céljából.

A tervezett feldolgozás során először a teljes metaadat-készletet nyerjük ki egy saját fejlesztésű Python programmal. Az e-mailek elvben szabványos metaadatkészlete az évtizedek alatt felmerülő új igényekre válaszul úgy változott, hogy sokszor ugyanazt a metaadatot a különböző szolgáltatók és szoftverkészítők máshogy nevezik. Ez azt jelenti, hogy a továbblépés előtt az e-mail meta- és para-adatok gondos elemzése szükséges. A metaadatok elemzése és szűrése után a levelek szövegét tartalmazó levéltörzset kell szükség esetén dekódolni és további feldolgozásra alkalmas formára hozni. Ezt az állományt töltjük be végül egy relációs adatbázisba, illetve további tervünk, hogy egy e-mailek elemzésére testre szabott keresőalkalmazás segítségével meg tudjuk jeleníteni az adatokat, így a webaratás projekt keretei között létrehozott kereső és vizualizációs alkalmazáshoz hasonlóan statisztikák, kimutatások készülhetnek az e-mailekről is.

Bibliográfia

Balázs Indig, Árpád Knap, Zsófia Sárközi-Lindner, Mária Timári, Gábor Palkó. “The ELTE.DH Pilot Corpus – Creating a Handcrafted Gigaword Web Corpus with Metadata” In the Proceedings of the 12th Web as Corpus Workshop (WAC XII), pages 33-41 Marseille, France 2020.

Kirschenbaum, Matthew G., Richard Ovenden, Gabriela Redwine, and Rachel Donahue. *Digital Forensics and Born-Digital Content in Cultural Heritage Collections*. CLIR Publication, no. 149. Washington, D.C: Council on Library and Information Resources, 2010.

Lee, J.A.N. “Claims to the Term ‘Time-Sharing.’” *IEEE Annals of the History of Computing* 14, no. 1 (1992): 16–54. https://doi.org/10.1109/85.145316.

Partridge, Craig. “The Technical Development of Internet Email.” *IEEE Annals of the History of Computing* 30, no. 2 (April 2008): 3–29. https://doi.org/10.1109/MAHC.2008.32.

Prom, Christopher. “Preserving Email (2nd Edition).” Digital Preservation Coalition, May 2019. https://doi.org/10.7207/twr19-01.

Radicati, Sara. *Electronic Mail: An Introduction to the X.400 Message Handling Standards*. Uyless Black Series on Computer Communications. New York: McGraw-Hill, 1992.

Task Force on Technical Approaches to Email Archives, Andrew W. Mellon Foundation, and Digital Preservation Coalition, eds. *The Future of Email Archives: A Report from the Task Force on Technical Approaches for Email Archives, August 2018*. CLIR Publication, no. 175. Washington, DC: Council on Library and Information Resources, 2018.