Digitális bölcsészet – elméleti kutatások, gyakorlati eredmények

A 2021. évi Networkshop konferencia mottója az „Online térben – az online térért”, egyik kiemelt témája pedig az ebben a körben először színre lépő digitális bölcsészet volt. A száznál több elhangzott előadás több mint egy tizede a digitális bölcsészet körében zajló kutatásokról, valamint azok gyakorlati hasznosításáról szólt. Az előadók zöme az egyetemekről és a kulturális örökségvédelmi intézményekből érkezett, de a vállalkozói szféra is képviseltette magát az új technológiákról szóló beszámolók között.

A *Hungarnet Egyesület* által 2021. április 6−9. között szervezett Networkshop konferenciának az ELTE volt a házigazdája. A felsőoktatás, a köznevelés, a közgyűjtemények és a tudományos kutatás nélkülözhetetlen informatikai hátterét biztosító számítógép-hálózatok és -alkalmazások terén zajló új eredményekről számot adó Networkshop tekinthet vissza a legnagyobb múltra a magyarországi konferenciák között. Az idei a harmincadik, jubileumi szakmai találkozó volt, amely a Hungarnet Egyesület, az *Innovációs és Technológiai Minisztérium* (ITM), valamint a *Digitális Jólét Program* együttműködésével valósult meg, a lebonyolításhoz szükséges a számítógépes hátteret az *ELTE Informatikai Igazgatósága biztosította*.

A Közgyűjteményi Digitalizálási Stratégia értékelése és a jövő tervei

A beszámoló élére a *Petőfi Irodalmi Múzeum* (PIM) *Digitális Bölcsészeti Központjának* (DBK) igazgatója, *Bánki Zsolt* „Küszöbérték – tartalom- és szolgáltatásfejlesztések radar felett” című plenáris előadása kívánkozik. Bánki a közgyűjteményeket közvetlenül érintő szakmapolitikai kérdésekre fókuszált, átfogó helyzetképet rajzolt föl a digitális bölcsészet számára is létfontosságú hátteret biztosító könyvtári, múzeumi és levéltári digitalizálásól és tartalomszolgáltatásról, majd pedig ismertette a közeljövőre vonatkozó terveket.

Prezentációjának első részében az előadó egy egészen új szemszögből vette górcső alá az intézményrendszer megújulásának fontos kritériumait. Ezt követően a kulturális örökségvédelmi intézmények digitalizálási tevékenységét és a *Közgyűjteményi Digitalizálási Stratégia* (KDS) keretei között elért eddigi eredményeket tekintette át; az előadás második felében pedig a 2024-ig terjedő időszak kiemelt céljait ismertette.

Az egyik meghatározó intézmény igazgatói pozícióját betöltő, a szakterületen sok éve dolgozó szakemberként Bánki Zsolt joggal-okkal tette föl a kérdést: a közgyűjtemények állományára épülő online adatbázisok közül vajon melyik szolgáltatás éri el a küszöbértéket, amennyiben azt a bizonyos küszöböt ott jelöljük ki, ismeri-e, használja-e széles körben a nagyközönség az adott szolgáltatást? Nem kétséges, hogy az *Arcanum Digitális Tudástár* (ADT) megüti ezt a színvonalat; nemcsak az adatbázis méretei és összetétele folytán, hanem jó PR-tevékenysége okán is. A legnépszerűbb internetes portálok gyakran az ADT alapján közölnek cikkeket, múltbeli érdekességeket, és rendre beszámolnak a szolgáltatást érintő újdonságokról – vagyis az Arcanum zászlóshajójának, az ADT-nek nemcsak a használati statisztikája figyelemreméltó, de a marketingje és a sajtója is jó. A többiek közül a MEK és a DIA viszonylag széles ismertségnek

örvend, talán ezek az adatbázisok még a küszöbértéken felül vannak, de jó, ha tudatában vagyunk: a *Hungaricana*, a *MAPIRE*, a *MOKKA*, az *ODR* az *MTVA Archívum*, az *AdatbázisokOnline* stb. nincsenek benne a köztudatban, és – részben emiatt – túl kevesen használják ezeket a szolgáltatásokat. Fölmerülhet például az a kérdés: vajon hányan tudnak az *Országos Széchényi Könyvtár*, az OSZK webaratási projektjéről, annak eredményeiről? [1]

Annak idején nagy reménnyel kecsegtetett a kormány által elfogadott KDS, melynek elsődleges célja a magyar közgyűjteményekben őrzött dokumentumállomány mintegy 50%-áról elektronikus másolatok készítése és azok publikálása. A valóság ezzel szemben a következő: a megvalósítás egy év késéssel indult, ráadásul az előirányzotthoz képest jelentősen csökkentett forrásokkal. Ennek ellenére elmondhatjuk, hogy a féleségeket tekintve az előírt teljes vertikum megvalósult, de jelentősen csökkentett mennyiségben, és ugyancsak nagy arányú elmaradás tapasztalható az infrastrukturális fejlesztésekben.

Néhány adattal érzékeltetve az elmúlt évek terveit és azok megvalósulását: az előirányzotthoz képest a KDS 2019-ben 49%, 2020-ban 28%, 2021-ben 13% forrással gazdálkodhatott. A késést leszámítva a tervekhez képest az első három évben összességében 38% forrást lehetett felhasználni a stratégiai célok megvalósítására. Ebben a szakaszban a stratégiai célok nehézkesen valósultak meg, a projekt jelentősen alulfinanszírozott volt, az eredményei szinte láthatatlanok, és a tartalom értékét tekintve is vannak kérdőjelek. Mindez azt bizonyítja, hogy a KDS első ciklusának végrehajtása nem volt kellően hatékony, az egységes elvek szerinti digitalizálás sem igazán valósult meg.

Nem lehet említés nélkül hagyni a koronavírus-járványt. Az oktatáshoz hasonlóan a közgyűjteményi szféra is azt tapasztalta, hogy a pandémia katalizálta a digitális tartalomszolgáltatásokat: soha ekkora igény és érdeklődés nem mutatkozott a szakszerű, mégis közérthető üzenetekre épülő, új típusú digitális tartalomszolgáltatás iránt. Most vált igazán nyilvánvalóvá, hogy az oktatás és a turizmus kiválóan tudja hasznosítani a közgyűjtemények által elérhetővé tett digitális tartalmakat.

A magyarországi mesterséges intelligencia (MI) stratégiai célja a magyar nyelv értékeinek megőrzése, illetve a nemzetgazdaság versenyképességének a biztosítása. A globalizáció egyre inkább veszélyezteti azokat a nyelveket, amelyek mögött nincs megfelelő technológiai támogatás – éppen ezért az MI-stratégiában kiemelt kutatási-fejlesztési irány a magyar nyelvre vonatkozó nyelvtechnológiai fejlesztés. A *Nemzeti Adatvagyon Ügynökség* (NAVÜ) által fejlesztendő, mesterséges intelligenciával támogatott szolgáltatások költséghatékonyabbá teszik a közgyűjtemények digitalizációját és strukturált üzemeltetését, továbbá nagymértékben javítják a mélységi keresések szövegértelmezéséhez szükséges szolgáltatások minőségét.

A KDS következő szakaszára, a 2021−2024 időszakra a PIM DBK elkészített egy cselekvési tervet, amelynek célja a kormánystratégia dinamizálása. A kormány elé került előterjesztést két minisztérium: az ITM és Emmi[2] a sajátjaként jegyzi, a résztvevők pedig a PIM DBK, a NAVÜ és a *Digitális Jólét Program* (DJP). Az említett szervezetek által elkészített dokumentum címe: „Előterjesztés a kormány részére a KDS megvalósítása érdekében a 2021−2024. között szükséges akciótervi intézkedésekről, valamint a hazai digitális tartalomipar helyzetének erősítését célzó piacfejlesztési koncepcióról és az abban foglalt intézkedések végrehajtásával kapcsolatos feladatokról.”

Fontos észrevenni, hogy az előterjesztés értelmezésében a KDS kiszabadult a közgyűjteményi szférából; már nemcsak kulturális projektként van definiálva, hanem a tartalomiparral van összekapcsolva – és ez a KDS szempontjából rendkívül fontos, emellett roppant szerencsés megoldás. Sikerült eddig nem létező, új szempontokra rátalálni és azokat kiemelni, amelyek révén az eddig egymástól távolinak látszó területek egymást erősíthetik.

1 A rövidítések feloldása: MEK – Magyar Elektronikus Könyvtár, DIA – Digitális Irodalmi Akadémia, MAPIRE, MOKKA – Magyar Országos Közös Katalógus, ODR – Országos Dokumentumellátó Rendszer, MTVA – Médiaszolgáltatás-támogató és Vagyonkezelő Alap, OSZK – Országos Széchényi Könyvtár

2 Innovációs és Technológiai Minisztérium, Emberi Erőforrások Minisztériuma

A KDS alapelvei és céljai lehetővé teszik a szolgáltatásorientált, hatékony, felhasználóbarát szemlélet érvényesülését. Az átfogó koncepció és a megalapozott elvek mentén újragondolt digitalizálás új szerepbe, a tartalomszolgáltató szerepébe helyezi a közgyűjteményeket. Itt kell megemlíteni a NAVÜ közgyűjteményi digitalizációval kapcsolatos, egyik elsődleges feladatát: a *Nemzeti Adattár Projekt* (NAP), vagyis a nyilvántartási és közgyűjteményi keresőrendszer kiépítését.

Előadása második részében Bánki Zsolt a kormányelőterjesztésben szereplő cselekvési terv kulcsfogalmait és fókuszát, illetve a 2021−2024 közötti időszakhoz kapcsolódó konkrét projektterveket ismertette. A KDS-ben összesen 52 projektterv szerepel, ezek közül 4 kiemelt projektként lett definiálva. Ez utóbbiak biztosítják a közgyűjteményi tartalmak szabad hozzáférését, illetve az egymás között megosztott intézményi tudás széles körű hasznosítását. A kiemelt projektek:

- az Arcanum Digitális Tudástár állami megvásárlása és ingyenes hozzáférhetővé tétele;
- a *Nemzeti Kulturális tartalomszolgáltató és Elektronikus Könyvtár* (NEKTEK) megvalósítása. Az OSZK-hoz köthető, könyvtári fókuszú projekt célja, hogy a megfelelő eljárások betartásával, közös jogkezelő szervezeten keresztül, ingyenesen szolgáltathasson jogvédett tartalmakat, amely leegyszerűsítve a sokak által ismert, ún. norvég modell hazai megfelelője lenne. A szolgáltatás modellje elkészült, a pilot projekt jelenleg fut az OSZK-ban. 1990-től kezdve, napjaink felé haladva, nagy tömegben zajlik majd a digitalizáció és a tartalmak ingyenes hozzáférhetőségének biztosítása;
- a digitalizált közgyűjteményi tudás- és adatvagyon hasznosítása, melynek lényege: a közgyűjtemények és a DJP égisze alatt létrejött *Digitális Pedagógiai Módszertani Központ* együttműködésében olyan típusú adatgazdagítás zajlik, amely a közgyűjteményi digitalizált tartalmakat közvetlenül hasznosíthatóvá teszi a közoktatás és köznevelés, a szakképzés, a kompetenciafejlesztés és felnőttképzés számára;
- Nemzeti Értéktár Projekt, amelyet a PIM DBK számos közgyűjtemény bevonásával, a *Magyar Turisztikai Ügynökséggel* karöltve valósít meg.

További két jelentős, ún. összközgyűjteményi projekt megvalósítása is szerepel a KDS második szakaszában. A *Nemzeti Névtér* fejlesztése az OSZK-ban kezdődött, jelenleg a DBK-ban folytatódik. A cél nem változott: nemcsak a teljes közgyűjteményi szféra, hanem bárki számára hasznosíthatóvá kívánják tenni a Nemzeti Névtér adatállományát. A *Digitális Bölcsészeti Platform* (dHUpla) nyílt eszközrendszert és platformot hoz létre a szövegekkel foglalkozó kutatók számára.

Az intézménycsoportokat egyesítő ún. gyűjtőprojektek résztvevői az *Informatikai és Könyvtári Szövetség* (IKSZ), a *Vidéki Múzeumok Szövetsége* (VMSZ) és a *Magyar Nemzeti Levéltár* (MNL) – ebbe a körbe az említett szervezetek aggregációs projektjei tartoznak.

A fent említett kiemelt, országos projektek önmagukban nem jelentenek garanciát arra, hogy a közgyűjtemények tartósan meghatározó jelentőségű tartalomszolgáltatóvá váljanak, éppen ezért az átfogó projektekre építve az ország egészét lefedő, kisebb fejlesztésekre is szükség van. A 43 intézményi projekt azokra a területekre koncentrál, amelyek a legnagyobb hatást képesek elérni az oktatás, a turizmus, a kreatív ipar és a közgyűjteményi adatvagyonban rejlő értékek bemutatása révén. A kiválasztás elsődleges alapelvei: a projekt járuljon hozzá az ún. *Fehér könyv*[3] módszertani elveinek implementálásához, lehetőség szerint kapcsolódjon az MI-alkalmazásokhoz, valamint – a közgyűjtemények adatvagyonát nemzeti szinten értelmezve – mérhető hatást generáljon a közigazgatási, termékfejlesztési és szolgáltatástámogatási területeken. Az ide tartozó fejlesztések közös jellemzője: a felhasználók számára új, eddig nem ismert, hiteles és valós tartalmakat tesznek közzé. A megvalósító szervezetek között megtalálhatók a kultúrstratégiai intézmények, az országos múzeumok, a legjelentősebb szakmai szervezetek, a DJP Kft., a Magyar Turisztikai Ügynökség, az MTVA Archívum, valamint az evangélikus egyház is.[1]

3 Fehér könyv: módszertani útmutató a közgyűjteményi kulturális örökség digitalizálásához és közzétételéhez. Budapest: Emberi Erőforrások Minisztériuma, 2019. https://ommik.hu/media/attachments/2019/12/09/fehr_knyv.pdf

Digitális szövegkiadások, automatikus kézírásfelismertetés

Az alábbiakban a konferencia digitális bölcsészeti tárgyú előadásait tematikailag csoportosítva tekintjük át. Első helyre egy általánosabb érvényű téma kívánkozik: a PIM DBK munkatársai, *Mihály Eszter* és *Cséve Anna* arról beszéltek, milyen nehézségekkel, illetve milyen lehetőségekkel kell számolniuk a digitális szövegkiadások terén – és az ő tapasztalataik minden bizonnyal más közgyűjtemények gyakorlatára is érvényesek lesznek, ha hasonló feladatokat vállalnak föl.

A könyvtárak, múzeumok, levéltárak naponta szembesülnek az általuk felhalmozott analóg és digitális adatvagyon digitalizálására és hozzáférhetővé tételére vonatkozó, egyre fokozódó igényekkel, melynek eredményeként már nemcsak megtekinteni, illetve keresni lehet a digitalizált állományokat, de gépi feldolgozásra is alkalmassá lehet azokat tenni, miáltal hozzáférhetővé válik a magyar kulturális örökség eddig elérhetetlen rétege.

Az elmúlt években kialakult, általánossá váló gyakorlat szerint a szöveges dokumentumok digitalizálása során kétrétegű PDF-ek készültek – a fakszimile kép képezi az egyik, a számítógépes formátumra kódolt, kereshető szöveg a másik réteget.

Az elmúlt évek során néhány intézmény – köztük a PIM DBK – megkezdte a kézírással lejegyzett szövegek, kéziratok tartalmának számítógépes feldolgozását, textológiai-filológiai gondozását, az eredmények publikálását is. A feldolgozás során szabványos TEI XML-fájlokat állítanak elő, amely kiterjedt metaadatkészletet használ, lehetővé teszi annotációk csatolását, továbbá számos megjelenítési módra ad lehetőséget. A TEI XML alkalmazásával magasabb szintű forráskiadás, illetve egy még magasabb filológiai szint, a kritikai kiadás is elérhetővé válik.

Miért jó a digitális szövegkiadás?
- rugalmas, változó, javítható,
- nincs terjedelmi határ,
- annotálható,
- összeköthető egyéb tudástárakkal – például a névtérrel,
- több formában publikálható,
- a keresésen kívül szűrési, illetve adatvizualizációs lehetőség is van,
- új módszerekkel kutatható és
- természetesen archiválásra is alkalmas.

A PIM-ben komoly feladatot jelentett a digitális bölcsészeti eszközök integrációja a közgyűjteményi infrastruktúrába – és ugyanez visszafelé! Össze kellett egyeztetni a gyűjteményi, illetve a digitális bölcsészeti szemléletet, biztosítani kellett a humán erőforrást.

Pilot projektként *Kiss József* levelezésének a feldolgozását jelölték ki, ennek során ki kellett dolgozni a közös elvi és gyakorlati alapokat a szkennelés szabályairól, a névkonvencióról, valamint a feladatok nyilvántartásáról. Ki kellett alakítani továbbá a content management környezetet, majd az ütemezés és a workflow megtervezése, az eszközök kiválasztása, valamint a szerepek kiosztása következett. (Kiss József kéziratos levelezésének feldolgozásáról a következő előadás kapcsán lesz szó.)

A TEI XML-szerkesztést a DBK munkatársai „2.0 publikációnak" nevezték el, jelezve, hogy ez már egy másik szint, amely magasabb minőséget jelent. A szerkesztőprogram az Oxygen, amelyben framework-öket alakítanak ki a projektek számára. A korábbiakhoz képest itt jóval kiterjedtebb textológiai-filológiai jelölésrendszert lehet használni, mód van az adatok gazdagítására, bővített metaadatolásra, be lehet kapcsolni külső adatbázisokat, névtereket, bibliográfiákat az entitások azonosításához, illetve annotációkkal lehet ellátni a szöveget. Fontos továbbá, hogy a forráskiadásokban a text-image linking módszert használják, vagyis összekötik a képet a szöveggel.

A Framework és a Git alapú funkciók együttese révén lehetővé válik az automatizált műveletek beépítése (transzformációk, azonosítókiosztás), a metaadatok beemelése a Huntékából, a metaadatok automatikus továbbítása, valamint a szerkesztőségi rendszerből a közvetlen publikáció.

Az egyes entitások azonosítása során számos kérdés fölmerült: hogyan jussanak el az új adatok a névtérbe? Mi legyen azokkal a nevekkel, amelyek az éles névtérbe nem kerülnek bele – vagy azért,

mert nincs elég adat, vagy azért, mert egy névnek csak az adott projektben van jelentősége? Tanulság: a névtérprojekttel szoros együttműködésben kell kidolgozni a menet közben fölmerülő kérdésekre a megoldást.[2]

A PIM DBK-munkatársai, *Szűcs Kata Ágnes* és *Mihály Eszter* bemutatták az automatikus kézírásfelismertetés működését. Amint előbb említettük, a PIM DBK egyik kiemelt projektje Kiss József[4] levelezésének feldolgozása és digitális forráskiadása. A kézírásos szöveg átírását számítógépes hordozóra gépeléssel is meg lehet oldani, de erre a célra kifejlesztettek egy jó minőségű szoftvert, a Transkribust – egy számos hasznos funkcióval rendelkező, felhasználóbarát eszközt, amely megkönnyíti a kéziratok feldolgozását, a szövegek átírását és megteremti a későbbi filológiai elemzések alapját. A DBK pilot projekt egyik célja a kézírás felismertetésére kidolgozott Transkribus szoftverben rejlő lehetőségek kiaknázása, a másik fontos célkitűzés pedig egy publikus, magyar nyelvű *Handwritten Text Recognition* (HTR) modell létrehozása és a kutatók rendelkezésére bocsátása. A Kiss József levelezésének feldolgozása során szerzett tanulságokat és tapasztalatokat később felhasználják más, eddig publikálatlan kéziratos hagyatékok feldolgozása terén.

Az előadók először azt az adminisztrációs felületet, a Trellot mutatták be, ahol a levelek tényleges feldolgozását lehet nyomon követni. Minden levélnek saját kártyája van, amelyen a rendszer minden lépést, minden adatot rögzít. A szkennelés során jönnek létre a fakszimile minőségű képfájlok. Az egyes dokumentumok metaadatait a Huntéka rendszerben rögzítik.

Az ingyenesen elérhető Transkribus szoftver elvileg lehetővé teszi az automatikus kézírásfelismerést. A beszkennelt kézírásos dokumentumot fel kell tölteni a szerverre, ezt követi a szöveg szegmentálása, utána az átírás,[5] majd a korrektúra, végül az ellenőrzés. A jóváhagyott szövegről kétrétegű PDF-, illetve TEI XML-outputok készülnek. A tervek szerint a PIM OPAC-felületén a kétrétegű, kereshető PDF jelenik meg.

A PIM DBK-ban zajló kézirat-digitalizálási pilot projektben egy automatikus kézírásfelismerő modell, az előbb már említett HTR is épül. A HTR a digitális bölcsészet egyik új, erőteljesen fejlődő ágazata, amely a mesterséges intelligencia, azon belül a neurális háló alapú technológia segítségével automatikusan írja át a kéziratok tartalmát számítógéppel olvasható szövegre.

A digitális fakszimilék és pontos átírásuk alapján a mesterséges intelligencia segítségével a HTR-t folyamatosan lehet tanítani, és így egyre pontosabban ismeri föl a tanulásba bevont, illetve a hasonló kézírási stílusokat. Mivel a modellek egymásba építhetők, egyre szélesebb körben válnak alkalmassá a kézírásfelismerésre.[3]

Digitális filológia és kritikai kiadások

Fellegi Zsófia, a DigiPhil projekt kutatója a digitális filológiai korpusz szövegstatisztikai és nyelvelemző vizsgálatairól tartott előadást. A *DigiPhil* – a Tudományos szövegkiadások, bibliográfiák és kutatási adatbázisok online tudástára – projekt 2012-ben indult a PIM és a *Bölcsészettudományi Kutatóközpont* (BTK) *Irodalomtudományi Intézet* (ITI) együttműködésében. Indulása óta a DigiPhil számos kritikai kiadás digitalizálását végezte el retrokonverzió révén (pl. Arany János Összes Művei), illetve kutatócsoportokkal együttműködve segíti born-digital kiadások elkészítését és publikálását (pl. Móricz Zsigmond levelezése 1892−1913). A projekt indulása óta közel tízezer, a TEI ajánlásának megfelelően elkészített XML-fájlból álló korpusz épült. A gyűjtemény jelentős részét XIX-XX. századi szerzők munkái teszik ki. A korpusz méretéből és a szerzők időbeli közelségéből adódik a lehetőség, hogy a DigiPhil által készített jelölőnyelvi átiratokon nyelvstatisztikai elemzéseket végezzenek.

A jelenleg zajló kutatás célja, hogy a rendelkezésre álló kritikai kiadásokon, szövegstatisztikai módszerekkel olyan, adott szerzőre jellemző mintázatokat

4 Kiss József, a XIX. századi író, költő és A Hét című irodalmi folyóirat alapító szerkesztője kiterjedt személyes és szakmai levelezést folytatott.

5 A nyomtatott szövegek digitalizálására szolgálnak az optikai karakterfelismerő OCR programok, a kéziratok szövegének digitalizálására a HTR szoftverek szolgálnak.

derítsenek föl, amelyek erőforrás hiányában korábban sokszor láthatatlanok maradtak a kutatók számára. Fény derülhet például arra, változik-e Móricz Zsigmond levélírási gyakorlata az egyes leveleken végzett javítások tükrében, vagy Kosztolányi utólagos módosításai mögött felsejlik-e valamilyen tendencia. A kutatás során kiemelt szempont a saját elemző algoritmusok készítésén túl a rendelkezésre álló eszközök és algoritmusok felhasználása, illetve azok hatékonyságának vizsgálata.

A genetikus kritika a szövegek genezisének, a mű keletkezési folyamatának rögzítésére és ennek bemutatására törekszik. Az előadó *Kosztolányi Dezső* művein keresztül mutatta be, hogyan készül a digitális kritikai kiadás.

Vizsgálati módszerek:

- szövegstatisztikai vizsgálatok (törlések, javítások aránya, mintázatok felismerése),
- nyelvi elemző (szófaji arányok változása az írásfolyamat során, mintázatok felismerése), amelyhez az e-magyar szoftvert alkalmazták.

A BTK ITI és az ELTE Digitális Bölcsészet Tanszék (ELTE.DH Tanszék) Stilometriai kutatócsoportjának a közeljövőre vonatkozó tervei:

- irodalmi szövegeken tanított vektortér modellek (pl. Jókai prózájának nyelvi világa) kialakítása.
- Vizsgálni fogják, egy vektortér segítségével az írói szótár és a szóhasználat alapján megjósolhatóvá válik-e egy kiolvashatatlan, vagy csak részben kiolvasható szöveghelyen a legnagyobb valószínűséggel szereplő szóalak?
- A tervek szerint a Babits kritikai kiadások egységes kódolással, az előre kialakított filológiai specifikáció alapján fognak készülni.

Az előadás végén kitekinthettünk a szemantikus hálózatok világába. A BTK dolgozik egy szoftver fejlesztésén, amely hálózati modellezéssel teszi kutathatóvá, megjeleníthetővé az irodalmi kapcsolatokat. A közeljövő tervei között *Arany János*, *Vörösmarty Mihály* és *Olahus* (Oláh Miklós) levelezésének digitális kiadása szerepel.[4]

A born digital anyagok feldolgozása

A PIM DBK-ban javában zajlik a dHUpla elnevezésű digitális bölcsészeti platform létrehozása, amelynek egyik fontos része a born digital anyagok kezelésének megtervezése. A born digital workflow kidolgozásáról *Kalcsó Gyula* tartott előadást.

Sürgető igény mutatkozik a born digital anyagok eljárásrendjének a kidolgozására: a PIM-ben már szép számmal vannak ilyen jellegű gyűjteményi elemek, de a jövőben várhatóan egyre több digitálisan létrejött tartalom kerül a gyűjteménybe, amelyek szakszerű kezeléséről gondoskodni kell. A másik fontos feladat a born digital anyagok kezelésére vonatkozó eljárásrend kidolgozása a közgyűjtemények számára.

Kalcsó Gyula a born digital anyagok feldolgozására a PIM DBK-ban tervezett workflow-t mutatta be. A born digital fájlok túlnyomó többsége 'digital exclusive' – azaz kizárólag digitálisan létező számítógépes állomány, nincs analóg megfelelője. Elsősorban nem az egyediség, hanem az archiválás okoz igazi nehézséget: az állományok megőrzése a mennyiség, a változatosság, az elavulás, az értelmezhetőség szempontjából egyaránt bonyolult, nagy feladat.

A Zotero repozitóriumban megtalálható a dHUpla born digital nyilvános csoportja, ahol mintegy 200 szakirodalmi forrás adatait gyűjtötték össze az e-mailek archiválásától a törvényszéki módszerek használatáig.[6]

Egy ausztrál szerző, *Somaya Langley* készítette el a 14 fázisból álló „Digital stewardship end-to-end workflow" modellt, amelyből a DBK-ban leginkább a digitális megőrzés feladatsorral kell foglalkozni. Létezik egy nemzetközileg elfogadott referenciamodell, amelyet a born digital archiválásban szinte mindenütt alkalmaznak, ez pedig az Open Archival Information System – OAIS-modell,[7] egy magas szintű elméleti modell a digitális környezetben keletkező adatok/információk hosszú távú megőrzésére.

6 pim-dbk-dh-born-digital. https://www.zotero.org/groups/2532329/pim-dbk-dh-born-digital/library

7 Open Archival Information System. ISO 14721:2012

Az OAIS alapvetően három ágenst különböztet meg: az adat létrehozóját, a kezelőjét és a felhasználóját. A digitális megőrzés kulcsfontosságú tevékenységei a bevitel/gyarapítás, a megőrzés/archiválás és a hozzáférés/szolgáltatás; ennek megfelelően az OAIS háromféle információs csomagot ír elő:

- átadás – Submission Information Package (SIP);
- archiválás – Archive Information Package (AIP);
- szolgáltatás – Dissemination Information Package (DIP).

A nemzetközi elterjedtség mellett az is az OAIS-modell alkalmazása mellett szól, hogy Magyarországon létezik már kidolgozott workflow, amelyet a Magyar Nemzeti Levéltár fejlesztett ki, levéltári archiválásra.

A hosszútávú megőrzésre szolgáló módszerek és eszközök ismertetése után az előadás két érdekes példával zárult, melynek során a hallgatóság megtudhatta, hogy egy írói hagyaték hosszú távú megőrzése során miért fontos gondoskodni az adatok integritásáról, az esetleges adatmódosítás kizárásáról, amelyhez gyakorlatilag ugyanazt a módszert kell használni, mint amellyel a rendőrség őrzi a bizonyítékok integritását. Ezt a megoldást hívják merevlemez-filológiának, amelynek segítségével tudományos szintű szövegkiadást is elő lehet állítani.[5]

Webarchiválás és internetes újságcikk-kereső

A webarchiválás világszerte az egyik legnagyobb kihívás a szakemberek számára, ugyanis az online hírportálok kiadói főként az új tartalmakra, és nem az archívumaikra koncentrálnak, holott a napjainkban megjelenő internetes közlemények egyben a jövő történeti dokumentumai, ezért azok összegyűjtése, megőrzése, feltárása ugyanolyan fontos, mint az analóg dokumentumoké.

A nemzeti kulturális örökség szerves részét képező online sajtó termékeinek megőrzése viszont sokkal nehezebb feladat, mint a nyomtatott kiadványoké. A webaratás önmagában hatalmas kihívás, de még inkább az, ha a learatott tartalmat cikk-szinten kereshetővé kívánják tenni. Ez utóbbira vállalkozott az ELTE vezetésével létrejött Digitális Örökség Nemzeti Laboratórium (DH-Lab) internetes újságcikk-kereső és archiváló szolgáltatása. A webaratásprojekt, amelyről *Sárközi-Lindner Zsófia* és *Indig Balázs* számolt be, az ELTE.DH Tanszék égisze alatt indult, és 2020 őszétől a DH-Lab keretei között folytatódik.

A fejlesztés elsődleges céljai: archiválás, a learatott webanyagok tisztítása, metaadatolása, repozitóriumba szervezése, kutatható formában való közzététele, és mindehhez a jogi alapok tisztázása.

A projekt keretei között kifejlesztett eszköz a webcrawler, egy saját fejlesztésű szabad szoftver, az archiválás formátuma a WARC[8]. Az adatok feltárása terén a webes erőforrásokkal kapcsolatos információk tárolására tervezett Schema.org metaadat-szabványt használják, amely kompatibilis a szemantikusweb-technológiával. A feldolgozás kimeneti formátuma a TEI XML. Az archivált anyagok feldolgozása során figyelmet fordítanak a deskriptív szövegekre, a multimédiás tartalmakra, a tipográfiai elemekre, megkülönböztetik egymástól a releváns szövegegységeket és a tagolókat, illetve megőrzik a külső-belső hivatkozási rendszert alkotó linkeket. Az előállított WARC- és TEI-állományokat repozitálják, e célra a Zenodo nyílt hozzáférésű repozitóriumát veszik igénybe, ahová feltöltik a kutatással kapcsolatos tanulmányokat, dokumentumokat, adatkészleteket, szoftvereket stb.[9]

A kereső felület technikai háttere: SQL adatbázis, PHP-lekérdezőfelület. A szerzői jogi szabályoknak megfelelően a cikk-kereső csak a metaadatokat jeleníti meg, hivatkozva az eredeti cikk URL-jére. A letöltött anyag csakis a szerzői jogi előírásoknak megfelelő korlátozásokkal érhető el.

Összefoglalva: a webaratási munkamenetre épülő cikk-kereső szolgáltatás a learatott cikkek metaadatait (szerző, cím, rovat, megjelenés időpontja,

8 A Web ARChive formátum eredetileg az Internet Archive által kidolgozott ARC továbbfejlesztett változata, 2009 óta nemzetközi szabvány – legújabb verziója az ISO 28500:2017.

9 ELTE Department of Digital Humanities Repository. https://zenodo.org/communities/elte-dh/ – A European OpenAIRE program keretében kifejlesztett, a CERN által működtetett Zenodo valamennyi feltöltött objektumhoz szabványos DOI-t biztosít.

az aratás ideje stb.), illetve a cikkek teljes szövegét is kereshetővé teszi. A DH-Lab műhelyében kidolgozott módszertan a lehető legteljesebb módon, szabványos adat- és metaadat-formátumokkal leírva rögzíti a hírportálok cikkeit, lehetővé téve ezáltal a cikkek repozitóriumba helyezését, verziókövetését, keresését.[6]

Két további új szolgáltatás: a Verskorpusz és a Regénykorpusz

Főként a magyar irodalommal hivatásszerűen foglalkozók érdeklődésére számíthat az a két új, a DH-Lab műhelyében kifejlesztett szolgáltatás, amely az adatbázisokban szereplő szépirodalmi művek sokszempontú elemzését, kvantitatív vizsgálatát teszi lehetővé. Mindkét fejlesztés az ELTE-n indult, és mindkettő egyik fontos hívószava a *Franco Moretti* által bevezetett, az általa az irodalomhoz való új viszonyulás kulcsának tartott „távoli olvasás" fogalom. A távoli olvasás kifejezést – némileg leegyszerűsítve – ma gyakran a számítógépes olvasás szinonimájaként használják.

A jubileumi Networkshop konferencia tiszteletére adták át a nagyközönség számára az ELTE digitális bölcsészeti műhelyében elkészült, Verskorpusz nevű szolgáltatást, amelyről *Horváth Péter*, *Kundráth Péter* és *Palkó Gábor* tartott előadást. A DH-Lab projektvezetője, Palkó Gábor bevezetőjében Umberto Eco szellemes megjegyzését említette, mely szerint több könyv van a világon, mint ahány óra rendelkezésünkre állana elolvasni őket, de az el nem olvasott könyvek is mély hatással lehetnek, sőt vannak is ránk. Eco megállapítása új megvilágításba kerül a XXI. század új digitális tudományosságának a fényében. Franco Moretti irodalomtörténész – Ecohoz mérhetően provokatív éllel – a távoli olvasásban látja az irodalomhoz való új viszonyulás kulcsát, amely az ő jóslata szerint alapjaiban változtatja meg az irodalomhoz fűződő viszonyunkat.

Kérdés, van-e létjogosultsága, hogy számítógépes versolvasásról beszélhessünk? Érdemes felidézni, hogy Magyarországon, a Szegedi Tudományegyetemen már 40 évvel ezelőtt kísérleteztek számítógépes versolvasással, vagyis tudományos célú versfeldolgozással. Ezt a – mind a mai napig folytatódó – *Horváth Iván* nevével fémjelzett kutatást a világon a legelsők között tartják számon

Az ELTE Verskorpusz számára a közvetlen inspirációt a cseh verskorpusz jelentette, amelyben közel 80 ezer lírai mű számítógépes elemzése kereshető meg. Az ELTE.DH Tanszék kutatói által fejlesztett Verskorpusz jelenleg a középiskolai irodalmi kánonban szereplő 45 költő összes versét tartalmazza; a korpusz forrása a MEK adatbázisa. A feldolgozás eredményeként az adatbázisban lévő 11 295 vershez 3 128 000 tokent[10] különítettek el.

Az ELTE Verskorpusz építésének első lépéseként gépileg annotálták a versek szerkezeti egységeit (cím, vers, verssor), majd az annotáció kézi ellenőrzését követően szintén gépileg annotálták a szavak grammatikai tulajdonságait (lemma[11], szófaj, morfoszintaktikai jellemzők), valamint a vershangzás formailag egyszerűbben megragadható jellemzőit (rímképlet, rímpár, időmértékes ritmus, alliterációk, a szavak fonológiai jellemzői).

Az első lépés bemenete: a MEK RTF-formátumú fájljai, az annotálás eszköze: XQuery szrikpt, a kimenet: TEI XML, amely tartalmazza a szerkezeti egységek annotációit; minden vers egy fájlt alkot. Ezt a lépést kézi ellenőrzés követte. A második lépés a tokenizálás, a lemmatizálás[12], valamint a szófaji és morfoszintaktikai annotálás. A bemenet az előzőleg létrehozott TEI XML-fájlok az annotációkkal, az annotáló eszköz az e-magyar szoftver Python szkriptbe ágyazva. A második lépés kimenete: TEI XML-fájlok a szerkezeti egységek, valamint a szavak lemmájának, szófajának és morfoszintaktikai jellemzőinek annotációival. A következő lépés a vershangzás jellemzőinek gépi annotálása, amelynek a bemenete a második lépésben létrehozott TEI XML-fájlok, annotáló eszköze a *hunpoem-analyzer-TEI*[13] elnevezésű, Python nyelvben írt program, kimenete pedig a második lépésben előállított TEI XML-fájlok,

10 A token egy szövegben előforduló bármely szó, szövegszó.
11 A lemma a szótő, a szavak szótári alakja.
12 A tokenizálás a szavak meghatározása, vagyis azok elválasztása a szóközöktől, írásjelektől stb. Lemmatizálás: a szótövek meghatározása a tokenekből
13 A versek hangzásjellemzőinek elemzésére szolgáló programot Horváth Péter írta a Verskorpusz számára.

kiegészítve a vershangzás jellemzőivel. A negyedik lépés a formátum átalakítása és az annotációk bővítése, amelynek input oldalán az előző lépésben létrehozott TEI XML-fájlok állnak, az annotáló eszköz egy XSLT-stíluslap, az output oldalon pedig a szerkezeti egységeknek, a szavak grammatikai tulajdonságainak és a vershangzás bizonyos jellemzőinek az annotációit tartalmazó XML fájlok állnak.

Az előadás végén a kutatók bemutatták a Verskorpusz online elérhető, SQL-alapú lekérdezőprogrammal működő keresőfelületének számos funkcióját. A hallgatóság meggyőződhetett arról, hogy a Verskorpusz a magyar költészettel foglalkozó irodalomtudományi és nyelvészeti vizsgálatokat segíti, de a közoktatásban is haszonnal alkalmazható.[7]

Bajzát Tímea Borbála, *Szemes Botond* és *Szlávich Eszter* „A magyar regény korpusza és a »távoli olvasás«" című előadásában a Verskorpuszhoz „párját", a DH-Lab szerverén szabadon elérhető Regénykorpuszt mutatta be.

A projekt előzménye, hogy az ELTE csatlakozott a *European Cooperation in Science and Technology* (COST) *Distant Reading for European Literary History* kutatási projektjéhez, melynek fő célja egy többnyelvű európai irodalmi szöveggyűjtemény, a European Literary Text Collection (ELTeC) létrehozása. Az ELTeC célja: források és módszerek fejlesztése az európai irodalomtörténet-írás modernizálása érdekében, módszere a Distant Reading (nagy korpuszok digitális technológiai elemzése). Ez utóbbi cél elérése érdekében európai irodalmi szövegkorpuszt építenek, amelybe minimum 10 európai nyelven mintegy 2500 regény kerül be, hogy lehetővé váljon az innovatív digitális szövegvizsgálati módszerek tesztelése, az összehasonlító vizsgálatok elvégzése. Az ELTeC ambíciója, hogy az irodalmi szövegkorpuszon végzett vizsgálatok eredményeinek alapján a hagyományos irodalomelméleti és irodalomtörténeti koncepciók és azok alapvető fogalmai (pl. kánon, stílus) újragondolhatóvá, újraértelmezhetővé váljanak.

Az ELTE.DH Tanszék regénykorpusza szervesen illeszkedik az *ELTeC* nemzetközi gyűjteményébe, ugyanis a korpuszba került első 100 regény ezen összeállítás magyar nyelvű alkorpuszát képezi; jelenleg 81 szerző 100 magyar nyelvű regényének digitalizált, annotált szövegtesteit tartalmazza. A szövegeket az e-magyar elemzőlánccal dolgozták föl, ezzel készült a lemmatizálás, valamint a morfológiai és a szófaji elemzés – ennek során 6 948 590 token jött létre. A szövegek feldolgozásának és tárolásának formátuma ebben az esetben is a strukturált szövegfeldolgozást és az online megosztást lehetővé tevő TEI XML. A szolgáltatás szabadon hozzáférhető, a keresőfelületen sok szempontú, részletes keresést lehet végezni.

Az előadók bemutatták, hogy a keresőfunkciók használata, illetve a statisztikai és nyelvészeti megközelítések együttes alkalmazása milyen sokféle szerepet tölthet be a szövegek értelmezésének folyamatában. Van mód például alkorpuszok létrehozására szerzők és/vagy műcímek, a keletkezés ideje, a mű terjedelme, illetve kanonikussága alapján. Egy másik keresési lehetőség a tokenek és tokenkapcsolatok, nyelvi szerkezetek, szóalakok, szótövek, szófajok stb. szerinti szűrés. Egyszerre több tokenre is lehet keresni, de megadható a tokenek távolsága és kapcsolata is. A keresőfelületen meg lehet adni a találatok megjelenítésére vonatkozó beállításokat, a keresés eredményét pedig el lehet menteni.

A Regénykorpuszban tárolt nagy mennyiségű szöveg kvantitatív vizsgálata és a ráépülő vizualizáció az egyes művek, de akár a történeti korszakok olyan jellemzőit képes láthatóvá tenni, amelyek a hagyományos olvasás révén vagy reflektálatlanok maradnak, vagy a mérések hiányában nehezen igazolhatók.[8]

Stilometria, szerzőazonosítás, szerzői ujjlenyomat

A stilometria a stílus statisztikai alapú vizsgálatát jelenti, amelynek segítségével a kutatók megállapítják egy adott szerző műveire, egyéni szóhasználatára jellemző nyelvi tényezőket. A szövegeket lexikális és más nyelvi jegyek alapján mérik és hasonlítják össze, miáltal lehetővé válik a szövegek közötti azonosságok, illetve különbségek meghatározása és értékelése.

Az itt következő két fejlesztés szintén a korábban már említett, a BTK ITI és az ELTE.DH Tanszék Stilometriai kutatócsoportja keretei között zajlik. A kutatások digitális bölcsészeti hátteréről, valamint a közeljövő terveiről Palkó Gábor elmondta, hogy a stilometriai kutatási eredményeket szeretnék nemzetközi színtéren is bemutatni, melyre legközelebb az *International Journal of Digital Humanities* szerzőazonosítással foglalkozó különszáma ad lehetőséget. A magyar nyelvre vonatkozó benchmark eredmények publikálása rendkívül fontos.

A hazai tudományos műhelyekben dolgozó stilometriai kutatók számára a DHLab infrastruktúrát, tárhelyet is tud biztosítani, megfelelő autentikációval pedig hozzáférést nyújt a korpuszokhoz. A DH-Lab közeli tervei között az egyik első helyen az oktatás szerepel: workshopokat, nyári egyetemet terveznek, eLearning tananyagok készülnek. A távlati célok közül Palkó Gábor a mesterséges intelligencia alapú, új technológiák integrálását, például a mélytanuló algoritmusok használatát emelte ki.

Kiss Margit, *Palkó Gábor* és *Szakács Béla Benedek*: „Szöveghasonlósági vizsgálatok automatizálása" című előadása szintén egy újonnan indult szolgáltatásról számolt be.

A stilometriai elemzésekhez jelentős számítási kapacitásra és komoly nyelvészeti-statisztikai háttértudásra is szükség van. Ez annak ellenére van így, hogy ma már elérhető néhány elemzőszoftver (Websty, JGAAP, Stylene stb.). A prezentáció első részében a hallgatóság megismerhette a stilometriai elemzés különféle alkalmazási területeit és eddigi eredményeit, különös tekintettel a magyarországi fejleményekre. A második részben az előadók azt a DHLab által üzemeltetett szolgáltatást mutatták be, amelyet az ELTE.DH Tanszék, a BTK ITI, illetve a *Budapesti Műszaki Egyetem Méréstechnika és Információs Rendszerek Tanszék* együttműködésével fejlesztettek ki. Az előadók a fejlesztésben részt vállaló három intézményt képviselték.

A DH-LAB szerverén futó szolgáltatás webes környezetbe ágyazva működik, így a felhasználók mentesülnek a szoftver telepítéséhez, illetve a számítási feladatok elvégzéséhez szükséges, igen jelentős gépi erőforrás biztosítása alól. A webes szoftverkörnyezet, illetve az elemzések automatizálása révén korszerű, felhasználóbarát eszköz áll a számítógépes szövegelemzési munkálatokat végzők rendelkezésére, és ez megkönnyíti a stilometriai elemzést végzők munkáját.

A stilometria nem új keletű, az 1850-es években már voltak ilyen jellegű kutatások. Ma különféle tudományterületekhez kapcsolódik, nyelvészet, irodalomtudomány, filológia, stilisztika, statisztika, informatika. Napjainkban nagy terjedelmű szövegkorpuszok vizsgálata, a szövegek stilisztikai jegyeinek a mérése, eredmények összevethetősége és értékelése zajlik. Nemcsak az irodalomtudomány és a nyelvészet, de a jogi, igazságügyi eljárások, az orvostudomány, a zenetudomány, a képzőművészet is alkalmazza.

Alkalmazási területek:
- idiolektus[14] vizsgálata,
- anonim vagy vitatott szerzőség,
- egyéni nyelvezet alakulása, formálódása,
- korszakolás szerzői életművekben vagy nyelvtörténeti korszakokban,
- csoporthoz tartozás vizsgálata,
- műfaji jelleg elemzése,
- nyelvi szempontból megmutatkozó hatás.

A digitális bölcsészetben nagyon sok tulajdonság alapján lehet szövegeket összehasonlítani. A stilometriai elemzés menete: a szövegekre jellemző tulajdonságok meghatározása, a stílusmarkerek (egyedi stíluselemek) megállapítása, amelynek legelterjedtebb módszere a MFW (Most Frequent Words – a leggyakoribb szavak) megkeresése, ezt követi a mondathossz, a szóhosszúság, a szókészlet gazdagsága, a leggyakoribb funkciószavak, a szó és karakter n-gramok vizsgálata. Ezek a vizsgálatok komoly elméleti hátteret és empirikus tudást igényelnek.

A közös projektben a 2016-ban kifejlesztett, R nyelven írt Stylo programcsomagot fejlesztették tovább, ennek átdolgozásából és kibővítéséből jött létre

14 Idiolektus: egyéni nyelvhasználat, egy adott személy nyelvhasználatára jellemző nyelvi vonások összessége.

a Shtylo. Az új fejlesztés nagy előnye, hogy a vizsgálandó korpusz URL-ről betölthető, a paraméterezés pedig elmenthető, és a későbbiekben ismét felhasználható korpuszok elemzésére. A Shtylo programhoz varázslót, illetve részletes elemzőt is kifejlesztettek, amelynek a működését az előadás során bemutatták.[9]

A stilometriai kutatások legújabb eredményeiről szólt a DH-Lab négy kutatójának, *Bajzát Tímea Borbálának*, *Nemeskey Dávidnak*, *Palkó Gábornak* és *Timári Máriának* az előadása a *Jókai Mór* prózájával kapcsolatos koncepcióról.

A számítógépes stilisztika területén közkeletű nézet szerint léteznek az egyéni nyelvhasználatra jellemző egyedi mintázatok, az ún. szerzői „ujjlenyomatok", amelyek felderítése a kvantitatív szöveghasonlósági vizsgálatokat alkalmassá teheti a szerzőazonosítás céljaira. Óvatosnak kell lenni azonban az „ujjlenyomat" metaforával, mert azt a téves képzetet keltheti, hogy a szövegekből objektív módon volnának kiolvashatók a szerzőre jellemző, számszerűsíthető minták. A szerzői „ujjlenyomat" megalkotása egy kreatív digitális bölcsészeti feladat.

A megkeresés – melynek célja néhány olyan mű szerzőségének azonosítása volt, amelyekkel kapcsolatban fölmerült Jókai szerzősége – a Jókai kritikai kiadást előkészítő munkacsoporttól érkezett; a vizsgálatokat a kutatók közösen végezték. Egy ilyen kutatáshoz elengedhetetlen a magyar nyelvre, illetve Jókai prózájára vonatkozó távolságmérések és beállítások ismerete. A munka során a DH-Lab kutatói széles körű stilometriai elemzést készítettek, feltérképezték a Jókaira jellemző nyelvstatisztikai alapú mintázatokat, majd ezek alapján kísérletet tettek az írófejedelem szerzői „ujjlenyomatának" megalkotására.

A vizsgálatot a szövegelemzési célokra kiválóan alkalmazható Python programnyelvben végezték. Az Openscience elveknek megfelelően a futtatott kódokat, valamint a korpuszokat közzéteszik. Az előadás érdekes színfoltja volt egy néhány hete fölvetődött, a sajtóban futótűzként terjedő vélemény – miszerint érdemes volna átgondolni, Jókainak Az aranyember című regényében a mai kornak megfelelő szerepet szán-e Tímeának – alapján Jókai prózájában a női nemhez kapcsolódó kifejezések kvantitatív vizsgálata. A kutatók Jókai 66 regényét, továbbá naplórészleteit és egyéb írásait vetették össze 55 szerző 132 regényével.[10]

Az ELTEdata szolgáltatás

Sebestyén Ádám „Prozopográfiai adatbázis-fejlesztés" című előadása az ELTEdata szolgáltatást mutatta be, amely prozopográfiai, bibliográfiai és más történeti témájú kutatások információinak szemantikus adathálózatba rendezésével és közzétételével foglalkozik. Az ELTEdata mind a szemantikus állítások, mind pedig az entitások szintjén össze van kapcsolva a *Wikidata* megfelelő állításaival, melynek révén az ELTEdata a Wikidata részeként, de attól függetlenül, önálló hálózatként is szemlélhető és kereshető. Az adatbázis egyedi azonosítóval rendelkező elemekből épül fel, valamennyi szemantikus kijelentés a tulajdonság (property) és az érték (value) kettőséből áll.

Az egyetemen három ELTEdata-projekt kezdődött meg: a *Bölcsészettudományi Karon a Humanizmus Kelet-Közép-Európában Kutatócsoport* (HECE) gondozza a HECEdata szolgáltatást, amely az 1420 és 1620 között Magyarországon élt humanista szerzők életpályáját és szövegeit vizsgálja. A biográfiai rész a komplex életrajzi adatokkal elkészült, jelenleg a bibliográfiai adatok bevitele zajlik. Már hozzáférhető a HECEdata adatbázis, amelyben lekérdezhetők a közel 500 szócikkből manuális úton bevitt biográfiai adatok. Az előadó kitért a bibliográfiai rekordok bevitelének automatizálási lehetőségeire is. Az adatbázisban komplex lekérdezések hajthatók végre, az eredmények megjeleníthetők térképen vagy idővonalon – így például vizualizálható, hogy egy adott időpontban kik tanultak a heidelbergi egyetemen.

A Tudásáramlás a bölcsészkari Kora Újkori Történeti Tanszék projektje, melynek célja hét tudományterület önálló diszciplínává formálódásában szerepet játszó tudásáramlási folyamatok rekonstruálása az 1770 és 1830 közötti időszakban. Az ELITEdata a *Társadalomtudományi Kar Prozopográfiai és Családtörténeti Kutatócsoportjának* projektje, amely

a XIX. század végétől a második világháborúig terjedő időszakban a hazai egyetemi tanárok életének és munkásságának kutatása, az életrajzi adatbázis karbantartása, frissítése és elemzése. A kutatócsoport eddig négy kötetet adott ki a Történeti elitkutatások sorozatban.[11]

Digitális bölcsészeti kurzus

Smrcz Ádám az ELTE BTK több ezer hallgatóját érintő kurzusról és annak tanulságairól számolt be „A digitális bölcsészet oktatása digitális platformokon" című prezentációjában. Az ELTE BTK első ízben a 2019/20. tanévben indította azt a valamennyi hallgató számára kötelező új kurzust, melynek célja, hogy a bölcsészhallgatók megismerkedjenek a digitális bölcsészet alapjaival. A Canvas felületen megtervezett kurzus a 2019/20. tanévben hibrid környezetben, kis létszámmal indult, a 2020/21. évben azonban már kizárólag online környezetben lehetett a kurzust megtartani, lényegesen nagyobb hallgatói létszámmal.

Az előadó ismertette a kurzus tematikáját és az első két alkalom eredményeit, tanulságait. Az egyes modulok során a hallgatók megismerkedhettek az információs műveltség, az információmenedzsment, a digitálisan létrejött tartalmak, a forráskeresés, az információs túlterheltség, a Big Data és a Smart Data, a mesterséges intelligencia alapjai, a hálózatelmélet, a webarchiválás stb. témaköreivel. A digitális tudomány modul főleg a szerzőség és a plágium kérdéseit járta körül. Valamennyi téma tekintetében igyekeztek a bölcsészettudományi szempontokra építeni. Modulonként 6-7 szövegrészlet feldolgozása, TED-videók és oktatófilmek megtekintése volt a hallgatók számára előírva.

Az egyik legnagyobb nehézséget az okozta, hogy a BA-, MA- és PhD-hallgatók tudásszintje és a téma iránti érdeklődése nagyon heterogén; nehéz a különböző képzési szinten álló hallgatók eltérő ismereteinek megfelelő tananyagot összeállítani. Az oktatók számára a legnagyobb kihívást a számonkérés jelentette. A hallgatóknak kvízeket kellett kitölteniük, órai feladatokat kellett teljesíteniük, de házifeladatokat is kaptak. Az oktatók kérték a visszajelzést, melynek során arra voltak kíváncsiak, hogy mennyi új ismeretet nyújtott a kurzus, az anyag mennyire volt érdekes, illetve követhető.

A nehézségek ellenére az első tapasztalatok jónak mondhatók, a digitális bölcsészeti képzést mindenképpen érdemes folytatni; a jövőben azonban a különböző tudományos háttérrel rendelkező hallgatókat eltérő módokon kell megszólítani.[12]

Az ARCANUM mesterséges intelligencia alapú fejlesztései

A „Digitális bölcsészet a gyakorlatban, az ARCANUM mesterséges intelligencia fejlesztései" című előadást a cég két vezetője, *Biszak Sándor* és *Biszak Előd* jegyezte. Az ARCANUM Adatbázis Kft. havi mintegy egymillió oldal digitalizálását, feldolgozását végzi el és publikálja az ADT, a SZAKTÁRS, a MAPIRE és a HUNGARICANA oldalakon – közülük a legismertebb szolgáltatás a mintegy 32 millió oldalnyi szöveget tartalmazó ADT. Ezen adatbázisok egyre inkább nélkülözhetetlenek a hazai bölcsészettudományi kutatásokban, melyeket a közel 60 millió digitalizált oldal szinte forradalmasított.

A cégvezetés az ARCANUM keresőszolgáltatására a legbüszkébb: nemcsak a gyorsasága, de szofisztikált keresési lehetőségei is komoly elismerést váltanak ki. A cég alapítása óta, 30 éve folyamatosan fejlesztik az ARCANUM keresőt, teljes mértékben saját erőből. A keresőrendszer egyik legfontosabb eleme az Unicode támogatás, amelynek köszönhetően az európai írásrendszerektől eltérő szövegeket is hatékonyan tudják kezelni. Röviden a technológiai háttérről: a csonkolási műveleteket n-gram technológiával gyorsítják, szomszédossági keresésre is van mód. A terheléseloszlás tekintetében meghatározó jelentőségű a sharding, amely lehetővé teszi az adatbázisok szétosztását kisebb adatbázisokra, amelyeket akár több példányban is el tudnak indítani – ennek köszönhetően tetszőlegesen sok felhasználót tudnak egyidejűleg kiszolgálni, villámgyors válaszidőkkel. A keresés során a jól bevált BM25 relevanciafüggvényt használják.

Évekkel ezelőtt kezdtek mesterséges intelligencia alapú fejlesztésekkel foglalkozni. Az első állomást a sajtótermékek illusztrációinak kezelése jelentette – a tapasztalatok alapján ugyanis a legtöbb felhasználó a képekre keres. Kezdetben a hagyományos képfeldolgozási eszközökkel és a klasszikus gépi tanítással próbálkoztak, de később egy másik megoldásra, a neurális hálókra esett a választás, és a tapasztalatok szerint ez az eljárás lényegesen jobb eredményt produkál. A képkeresésen belül legtöbben személyekre keresnek, ezért az ARCANUM az arckereséssel is elkezdett foglalkozni. Az arckeresésre első lépésként a nyílt forráskódú Single Shot MultiBox Detectort alkalmazzák. A megtalált arcképeket elküldték a szolgáltatásaikat működtető Amazon Web Service-be (AWS), ahol azokat felindexelték. A sikeresnek látszó projektről az Amazon egy terjedelmes blogbejegyzést íratott két szakértő munkatársával.[15]

15 Arcanum makes Hungarian heritage accessible with Amazon Rekognition. https://aws.amazon.com/blogs/machine-learning/arcanum-makes-hungarian-heritage-accessible-with-amazon-rekognition/

Jelenleg zajlik a legnagyobb szabású fejlesztés, az oldalszegmentálás, amelyhez 100 ezer oldalnyi annotált tanuló adatot hoztak létre. A cél a sok-sok elemre tagozódó újságoldalak szerkezeti egységeinek és metaadatainak – cikk, kép, cím, szerzői név stb. – kezelése.

A szövegfeldolgozással is sokat foglalkoztak: egy 10 milliárd szavas adatbázisra építettek egy BERT modellt. A fejlesztésnek köszönhetően az ADT-ben elérhetővé vált a tulajdonnevek felismerése 9 féle entitásként, amely lehet például személy, földrajzi hely, intézmény stb. Ugyancsak a BERT-re épül az OCR-javítás, vagyis a hibásan felismert betűkből adódó hibák korrigálása. A tömeges digitalizálás során komoly nehézséget okoznak a régi újságok esetén a kopott, elmosódott szövegrészek, a ritkított betűk, az elválasztások stb. A megoldást a BERT alapú end-to-end neurális hálóban látják.

Az ARCANUM egyik új, szintén a mesterséges intelligenciára épülő szolgáltatása a Kérdés megválaszolása, amely a cég által digitalizált összes lexikont, illetve a magyar Wikipédia szócikkeit használja föl a válaszok megadására.[13]

Felhasznált források

1 Bánki Zsolt: Küszöbérték – tartalom- és szolgáltatásfejlesztések radar felett. https://kifu.videotorium.hu/hu/recordings/42207

2 Mihály Eszter – Cséve Anna: A digitális szövegkiadások nehézségei és lehetőségei a közgyűjteményekben. https://kifu.videotorium.hu/hu/recordings/42405

3 Szűcs Kata Ágnes – Mihály Eszter: Automatikus kézírás-felismertetés Kiss József levelezésén. https://kifu.videotorium.hu/hu/recordings/42399

4 Fellegi Zsófia: Digitális filológiai korpusz mint Big Data? Szövegstatisztikai és nyelvelemző vizsgálatok TEI XML fájlokon. https://kifu.videotorium.hu/hu/recordings/42408

5 Kalcsó Gyula: Born digital workflow tervezése a PIM Digitális Bölcsészeti Központjában. https://kifu.videotorium.hu/hu/recordings/42402

6 Sárközi-Lindner Zsófia és Indig Balázs: A Digitális Örökség Nemzeti Laboratórium internetes újságcikk-kereső és archiváló szolgáltatása. https://kifu.videotorium.hu/hu/recordings/42822

7 Horváth Péter – Kundráth Péter – Palkó Gábor: Magyar líra a »távoli olvasás« horizontján: az ELTE Verskorpusz fejlesztése. https://kifu.videotorium.hu/hu/recordings/42417/

8 Bajzát Tímea Borbála– Szemes Botond– Szlávich Eszter: A magyar regény korpusza és a »távoli olvasás«.pptx

9 Kiss Margit – Palkó Gábor – Szakács Béla Benedek: Szöveghasonlósági vizsgálatok automatizálása. https://kifu.videotorium.hu/hu/recordings/42369

10 Timári Mária – Bajzát Tímea Borbála – Nemeskey Dávid – Palkó Gábor: A szerzői „ujjlenyomat” stilometriai koncepciója Jókai Mór prózájának szövegterében. https://kifu.videotorium.hu/hu/recordings/42366

11 Sebestyén Ádám: Prozopográfiai adatbázis-fejlesztés.pptx
12 Smrcz Ádám: A digitális bölcsészet oktatása az ELTE BTK-n. https://kifu.videotorium.hu/hu/recordings/42414
13 Biszak Sándor – Biszak Előd: Digitális bölcsészet a gyakorlatban, az ARCANUM mesterséges intelligencia fejlesztései. https://kifu.videotorium.hu/hu/recordings/42393

Tószegi Zsuzsanna
c. egyetemi docens, ELTE BTK,
tudományos újságíró.