Hegedűs István

HOGYAN SEGÍT ÚJRAGONDOLNI A LEVÉLTÁRAK SZEREPÉT A MESTERSÉGES INTELLIGENCIA ÉS A GÉPI TANULÁS?

Bevezetés

Az archívumokban található hatalmas információtömegek kiváló lehetőségeket biztosítanak a mesterséges intelligencia mint rendelkezésre álló technikai újítás kiaknázásához. Az adatok tömege nagy segítséget jelent nemcsak a kutatás, hanem a politikai döntések előkészítésében és a közigazgatás egyes területein is a nem túl távoli jövőben. Bár a mesterséges intelligencia egykor a sci-fi irodalom terméke volt, kutatása és alkalmazása jelenleg a számítástechnika egyik jelentős ágát képviseli, amely intelligens viselkedéssel, gépi tanulással és gépi adaptációval is foglalkozik. Olyan tudományággá vált, amely a való élet problémáit próbálja megválaszolni. A mesterséges intelligencia által szabályozott rendszereket ma már széles körben használják a közgazdaságtan és az orvostudomány vagy akár a design területén, akárcsak katonai célokra.

Globálisan fennáll a veszélye a helyi és a központi archívumok finanszírozásának csökkenésére. Emiatt is különösen képesnek kell lennie a levéltári közösségnek arra, hogy az archívumok fontosságát bizonyítsa, mind közgazdaságtani szempontból, mind pedig tágabb kulturális értelemben.[1]

Rendkívüli értékek és felfoghatatlan mennyiségű információk rejlenek a körülöttünk levő adatokban. Úgy gondolom, hogy ezek lehető legteljesebb felhasználása kulcsfontosságú mindenki számára. A levéltári adatok valójában a civilizációnk által felhalmozott legfontosabb értékek, amelyek a „múlt bölcsessége" révén az emberi tudás és tapasztalat jelentős részét képezik.[2] Az adatfeldolgozás és -felhasználás lehetősége exponenciálisan növekszik. Elég olyan területekre gondolni, mint az egészségügyi adatok vagy a marketingadatok felhasználásának különböző aspektusai. Nem kétséges, hogy az adatok önmagukban nem érnek sokat, valódi értéküket a feldolgozás számtalan változata határozza meg. A nagy mennyiségű adat feldolgozása a levéltárak esetében is új lökést adhat a fejlődéshez és az előrelépéshez, ezáltal növelve az archívumok befolyását. Mindannyian látjuk, hogy a világban végbemenő folyamatok végső soron az egész levéltárszakmát érintik. A levéltárosok

[1] *Do Archives Have Value?* Szerk. Michael Moss – David Thomas. Facet, 2019. 8.

[2] Bőgel György: *A Big Data ökoszisztémája*. Budapest, 2015, 23.

„alanyai" a „nyomtatott ipari társadalomról" a „tech-alapú internetes társadalomra" való nagy áttérésnek.[3]

Mint sok más szakma esetében, a levéltárosoknak is a lehető leghamarabb fel kell ismerniük, hogy a gépi tanulás eszközei és alkalmazásai sok feladatot elvesznek tőlük. Ugyanakkor egyre több más, kifinomultabb és magasabb minőségű, értékesebb feladat előtt nyílnak meg a lehetőségek. Kevesebb idő alatt is el lehet végezni olyan munkákat, amelyekről korábban álmodni sem mertünk volna. A jegyzékek és segédletek készítése, vagy a levéltári anyagok egymáshoz rendelése minden bizonnyal könnyen elvégezhető lesz egy mesterséges intelligencia algoritmusokkal támogatott gépi tanulási rendszerrel. Hasonlóképpen a gyűjtemények szervezése és rendszerezése, valamint a releváns tartalom rangsorolása is felgyorsul majd. Az olyan feladatok azonban, mint az archiválási adatok elemzése és az elemzéshez vezető út tervezése, új feladatokat ígérnek. Összefoglalva, ami ma nem túl egyszerű vagy rutinszerű feladat, a jövőben azzá válik. A munkatársak legtöbb idejét felemésztő feladatok robotokra és a gépi tanulás alkalmazásait is használó automatizált folyamatokra cserélhetőek a közeljövőben.

Big Data, MI (mesterséges intelligencia), GT (gépi tanulás) és a levéltárak

Kétségtelen, hogy az állami intézmények és a közgyűjtemények egyre több információt és nyomtatott, valamint elektronikus adatot generálnak. Egyértelmű, hogy ezeken a területeken egyre több eszköz van arra, hogy az adatokat komolyabb adatbázisokba szervezzék, egyre több kapcsolatot hozzanak létre az adatbázisok között, egyre különfélébb lekérdezéseket alkalmazzanak, és összetettebb, de mégis könnyebben használható rendszereket biztosítsanak a társadalom, az adófizetők számára. Ahogy a közgyűjtemények és a levéltárak egyre növekvő mennyiségű adatot kezelnek, úgy egyre rugalmasabb és robusztusabb informatikai infrastruktúrákat szükséges kiépíteni. Ezáltal a „Big Data" használatának legnagyobb kihívása az, hogy miként mentsük el a nagy mennyiségű adatot az elérhető legjobb és legteljesebb formában egy adott időpillanatban.

Statisztikák szerint az intézmények csak adataik 20%-át használják fel a gyakorlatban, a fennmaradó 80% pedig mind az operatív, mind a döntéshozatali eljárásokból egész egyszerűen kimarad. A Big Data és a gépi tanulás közötti szorosabb együttműködés a jövőben áttörést hozhat ezen a területen. Azonban mielőtt a levéltárak és a gépi tanulás kapcsolatát vizsgálnánk, fontos tisztázni azt, hogy mi is az a Big Data. A *Forbes* cikke szerint ez egy olyan terület, ahol arra keressük a választ, hogy miként kezeljük, elemezzük és alakítsuk át a nagy mennyiségű adatot

[3] Kate Theimer: It's the end of the archival profession as we know it, and I feel fine. In: Caroline Brown: *Archival Futures*. Facet, 2018, 4.

hasznos információkká.[4] Az eredményként megkapott információk relevanciája szerint azután a rendszer képes visszajelzést is adni, ami növelheti a folyamat jövőbeni hasznosságát és pontosságát. A fő kérdés a mi esetünkben, a levéltárak kapcsán, hogy a mesterséges intelligencia, közismertebb nevén a gépi tanulás, miként támogatja ezt a folyamatot, és hogyan változtatja hasznosabbá magánál az adatelemzésnél. Létrehozza-e a közvetlen összefüggéseket, és esetleg még javaslatokat is tesz azok kapcsán a főbb adatok jobb rendszerezése érdekében? A MI és GT fontossága abban rejlik, hogy képes létrehozni a releváns korrelációkat, a kapcsolódó szálakat, kimutatja a két különböző adatcsoport közötti apró különbségeket, és azokat további elemzésre bocsátja.

Bruce, Malcolm és O'Neill szerint a kreatív ipar jelenleg több mint 84 milliárd fontra becsüli a brit gazdaság kulturális szektorát. Az iparág folyamatos sikerének hajtóereje a digitális tartalomfogyasztás növekedése.[5] Egy brit példát említve, a Nesta[6] az Arts & Humanities Research Councillal (AHRC) és az Angliai Művészeti Tanáccsal együttműködve, a Digitális K+F Művészeti Alap finanszírozásával szervezte meg az ArtsAPI nevű projektet, amely a digitális kultúra egy páratlan innovációs laboratóriumát hozta össze. Ez az interdiszciplináris K+F projekt azokat a kapcsolatokat vizsgálta, amelyek alátámasztják a művészeti szervezetek által létrehozott „relációs értéket" és az általuk fenntartott „hatáshálózatot". Módszertanként szociális hálózati elemzést használtak. Ez lehetővé tette egy informatikai eszköz, az úgynevezett ArtsAPI létrehozását, kifejlesztését. Ennek használatával a kulturális szervezetek kiaknázhatják a saját ökoszisztémájukon belüli kreatív online tevékenységet, felismerve a hálózataikat létrehozó és fenntartó tényezőket, lehetővé téve számukra, hogy megalapozottabb döntéseket hozzanak. De ez csak egy példa a sok közül.

Ahogy Anthea Seles jelezte egy közelmúltbeli ICA (Levéltárak Nemzetiközi Tanácsa) webináriumban, a kérdés az, hogy a levéltárosok mennyi hozzáférést szeretnének biztosítani a kutatóknak a nyilvántartásokhoz és az adataikhoz.[7] Véleménye szerint nem szabad túlságosan belemerülni a MI adta lehetőségekbe, fontos, hogy figyelmet fordítsunk a csatlakoztatott adatkészletek fejlesztésére és a szemantikai webre is. A vállalatok, amelyek gyakran teljes gyűjtemények digitalizálásáért cserébe egy ingyenes mesterpéldányt kaptak, rájöttek, hogy hatalmas

---

4 Jim Sinur: *AI & Big Data; Better Together*. 2019. https://www.forbes.com/sites/cognitiveworld/2019/09/30/ai-big-data-better-together/#6bdb564e60b3 (Utolsó letöltés ideje: 2020. szeptember 23.)

5 Fraser Bruce – Jackie Malcolm – Shaleph O'Neill: Big Data. Understanding how Creative Organisations Create and Sustain their Networks. *The Design Journal*, 2017. 1. sz. 435–443. DOI: 10.1080/14606925.2017.1352961 (Utolsó letöltés ideje: 2020. szeptember 23.)

6 https://www.nesta.org.uk/ (Utolsó letöltés ideje: 2020. szeptember 23.)

7 Athena Seles: Artificial Intelligence and Archives. Presentation at Emerging Technologies, Big Data and Archives. Webinar 9 June 2020. https://www.youtube.com/watch?v=noxwKS-cPh0 (Utolsó letöltés ideje: 2020. szeptember 23.)

információ és érték rejlik ezekben. Ezeknek az értékeknek a kiaknázása egy profitorientált vállalat számára mindig elsőbbséget fog élvezni a gyűjtemények érdekeivel szemben, ezért itt az ideje a levéltárosoknak is felvenni a kesztyűt gyűjteményeik védelmében.

A gépi tanulás lehetőségei a levéltárakban

A levéltárak az emberi erőfeszítés legrégebbi példái az információk és adatok összegyűjtésének. Az adatok és/vagy információk gyűjtésének közös célja ellenére a levéltárak és a gépi tanulásos rendszerek adatkészletei eltérőek. Ennek felismerése arra sarkalja a GT-kutatókat, hogy olyan programnyelveket és algoritmusokat építsenek, amelyek képesek kommunikálni a levéltári rendszerekkel is.[8]

A digitalizálás és a számítógépesítés folyamata természetesen nemcsak a levéltárakra, hanem a könyvtárakra is igen nagy hatással volt az elmúlt évtizedekben. Király Péter tanulmányában az adatkezelést, az adatközzététel független alterületként való kezelését, az adatelemzéssel való integrációt, valamint a decentralizált web és a szemantikus web hatását vizsgálta. Eun Seo Jo és Timnit Gebru már idézett cikke szintén fontos abban a tekintetben, hogy az archívumok, könyvtárak és más intézmények hogyan működhetnek együtt a gépi tanulás kutatóival. Amely együttműködés azért is különösen fontos, mert az említett területeken belül már tanulmányoztak és szabályoztak különböző etikai, reprezentációs és átláthatósági kérdéseket.

A kutatás és fejlesztés másik izgalmas területe az úgynevezett számítógépes levéltártudomány mint tudományág, amely ténylegesen egyesíti a Big Data és MI/GT irányokat és módszereket a legteljesebb és még relevánsabb adatbázisok létrehozása érdekében. A közelmúltból két nagyszerű amerikai program is említhető ennek alátámasztására, részletes bemutatásuktól azonban el kell tekintenem. Az egyik a Morgenthau Holokauszt Gyűjtemények Projekt, míg a másik a Rabszolgaság Örökségének Számítástudományi Kezelése munkacímet viselte.[9] Lilley és Moore szerint *„Vannak alapvető akadályok a művészeti és kulturális intézményekben a nagy adathalmazok használata kapcsán. Az első a finanszírozási környezethez kapcsolódik. Az ágazat jelenleg nagyrészt túl korlátozott megközelítést*

[8] Eun Seo Jo – Timnit Gebru: Lessons from Archives: Strategies for Collecting Sociocultural Data in Machine Learning. In: *Conference on Fairness, Accountability, and Transparency*. January 27–30, 2020, Barcelona, Spain. https://doi.org/10.1145/3351095.3372829 (Utolsó letöltés ideje: 2020. szeptember 23.)

[9] Richard Marciano – Jane Greenberg: *Computational Archival Science (CAS): a Paradigm Shift Across the Data*. July 6, 2020, CLIR & AERI, 2020. http://aeri.website/wp-content/uploads/2020/06/CLIR-AERI2020_Marciano-Greenberg-Students.pdf (Utolsó letöltés ideje: 2020. szeptember 23.)

alkalmaz az adatokkal kapcsolatban. Az adatok gyűjtését és a jelentések készítését túl gyakran tekintik a finanszírozás vagy a fenntartó terheként és követelményének, nem pedig a művészeti vagy kulturális intézmény javára felhasználható eszköznek. Ez azt a veszélyt hordozza magában, hogy visszatartja az ágazatot. Ez részben a függőség, a támogatás és a piaci kudarc filozófiájából fakad, ami a kulturális ágazat nagy részét, köztük a művészeti és közszolgálati műsorszolgáltatást is jellemzi. [...] *A második legnagyobb akadály az adatok felhasználásának korlátozott stratégiai megértése, vagy az iránta való érdeklődés a kulturális ágazat felsőbb szintjein."*[10]

Király Péter informatikus és könyvtáros kutatónak van egy érdekes ötlete a középkori dokumentumok feldolgozásáról. Kéziratában arról ír, hogy olyan többszintű eszközt kíván kidolgozni, amely képes importálni a már publikált adatokat, feldolgozza a természetes nyelvet, szemantikailag elemzi, dokumentumokban keresi és statisztikai elemzés alá vonja. Hipotézise az, hogy a dokumentumtípusok korrelálnak a szöveg szemantikai karakterével és relevanciájával. Ha be tudná sorolni a dokumentumokat, a kutatók kiválaszthatnák a szemantikai elemzésekhez megfelelő eszközöket és azok információs kontextusát.[11] Ehhez kapcsolódóan az elmúlt két év kutatási eredményei nyomán más kutatók is érdekes eredményeket értek el a biotechnológia területén középkori levéltári szövegek elemzésével.[12]

Mi szükséges manapság az archívumok megbízhatóságának és nyitottságának biztosításához? Goudarouli, Sexton és Sheridan szerint erre a kérdésre rendkívül magas színvonalú kutatás, a kutatás támogatása, a magán- és állami partnerekkel való kísérletezés és együttműködés lehet a válasz. A folyamatos innovációk kutatásával a módszerek és etikai követelmények feldolgozhatók, értékelhetők és alkalmazhatók. Ezek azután felhasználhatók olyan digitális fejlesztések megvalósítására, amelyek átláthatóbbá, elfogadhatóbbá teszik a dokumentumkezelés és az archiválás új formáit, és amelyek a jelenlegi és jövőbeli felhasználókra (a kormányra és az állampolgárokra) is alkalmazhatók.[13]

---

10 Anthony Lilley – Paul Moore: *Counting What Counts: What Big Data can do for the Cultural Sector*. 2013. http://www.cross-innovation.eu/wp-content/uploads/2013/04/CountingWhat Counts-Lilley-Moore.org_.pdf (Utolsó letöltés ideje: 2020. szeptember 23.)

11 Király Péter: *Medieval Data Mining*. 2015. https://www.academia.edu/16657649/Medieval_ Data_Mining (Utolsó letöltés ideje: 2020. szeptember 23.)

12 Erin Connelly– Charo del Genio – Freya Harrison: Data mining a medieval medical text reveals patterns in ingredient choice that reflect biological activity against infectious agents. *mBio*. 11:e03136-19. https://doi.org/10.1128/mBio.03136-19. (Utolsó letöltés ideje: 2020. szeptember 23.)

13 Eirini Goudarouli – Anna Sexton – John Sheridan: The Challenge of the Digital and the Future Archive: Through the Lens of The National Archives UK. *Philosophy and Technology*, 2018. 1. sz. 173–183.

Gépi tanulás a gyakorlatban

2019-ben az Egyesült Királyságbeli Művészeti és Bölcsészettudományi Kutatási Tanács 21 Big Data kutatási programot finanszírozott összesen 4,6 millió font értékben.[14] A támogatott programok alapvető célja annak biztosítása volt, hogy a művészeti és bölcsészettudományi kutatások élen járjanak a digitális kor olyan kulcsfontosságú kérdéseinek előremutató kezelésében, mint a szellemi tulajdon, a kulturális emlékezet és nemzeti identitás, valamint a kommunikáció és a kreativitás. Mérnökök, számítógépes tudósok és fejlesztők képesek ugyan infrastruktúrát és informatikai támogatást nyújtani mindehhez, de a művészeti és a humán tudományokon belüli innováció alapvető fontosságú lesz az új tudásban rejlő potenciál kiaknázásához, valamint megszervezésének, értelmezésének és felhasználásának kialakításában.

Az Egyesült Királyság Nemzeti Levéltárában számos párhuzamos projekt zajlott a közelmúltban, vagy zajlik jelenleg is, amelyek szorosan kapcsolódnak a Big Data és az MI / GT tudományokhoz. A gépi tanulási technológiában rejlő lehetőségek feltárására indított számos kezdeményezések között olyanokat is találhatunk,[15] mint például az ún. eDiscovery[16] eszközök az értékeléshez és a kiválasztáshoz; az Innsbrucki Egyetem által fejlesztett Transkribus kézírásfelismerő[17] alkalmazásának tesztelése; PhD-projektek a webes adatok nagy sebességű megértésére és felülvizsgálatára, vagy az ún. „crowdsourced"[18] adatok tisztításának nehézségei. 2017-ben a Nemzeti Levéltár még egy hackatlont[19] is szervezett,

---

14 *The Challenges of Big Data*. Arts & Humanities Research Council. https://ahrc.ukri.org/research/fundedthemesandprogrammes/themes/digitaltransformations/ (Utolsó letöltés ideje: 2020. szeptember 23.)

15 *Digital projects at The National Archives*.https://www.nationalarchives.gov.uk/documents/digital-projects-at-the-national-archives.pdf (Utolsó letöltés ideje: 2020. szeptember 23.)

16 Az Egyesült Királyság Nemzeti Levéltárának katalógusa. Bővebb információ: https://www.nationalarchives.gov.uk/help-with-your-research/discovery-help/what-is-discovery/

17 Az OCR (Optical Character Recognition, magyarul optikai karakterfelismerés) mellett napjaink legizgalmasabb területe a HTR, tehát a kézírást is felismerni tudó alkalmazások fejlesztése. (Lásd: transkribus.net.) Világos, hogy az OCR újradefiniálta és megváltoztatta a szöveges adatokról való gondolkodásunkat. Forradalmi hatással van a történelmi, néprajzi kutatásokra. A következő lépés azonban a kézzel írt szövegek felismerése (lásd HTR). A kézzel írott szövegek felismerésének kétségtelenül nagy jövője van, de itt talán még nagyobb a felelőssége az algoritmusok tanítását végző emberi csapatnak, amely megtaníthatja a gépeket bizonyos típusú írott dokumentumok felismerésére. Lásd Richard Dunley: *Machines reading the archive: handwritten text recognition software*. 2018. https://blog.nationalarchives.gov.uk/machines-reading-the-archive-handwritten-text-recognition-software/ (Utolsó letöltés ideje: 2020. szeptember 23.)

18 A crowdsourcing során egy szervezet a hagyományosan belsőleg, saját dolgozók vagy alvállalkozók által elvégzett feladatokat a szervezettől független személyek nagy csoportjának szervezi ki, jellemzően online formában. Jellemzője, hogy a *crowd* (angolul: tömeg) minden tagja csak egy kis részlettel járul hozzá a teljes feladat elvégzéséhez. Számos altípusa alakult ki. https://www.merriam-webster.com/dictionary/crowdsourcing (Utolsó letöltés ideje: 2020. szeptember 23.)

amelyben 35 kollégájuk vett részt. A különböző csapatok gyakorlatba ültették át a korábbi tapasztalataikat, egyúttal sokféle adattal kísérleteztek, felvetették a digitális nyilvántartásokhoz való hozzáférés megőrzésének különböző problémáit is. Az esemény segítségével két főbb jövőbeni kutatási területet azonosítottak a további vizsgálatokhoz: a kódolt nyelvek automatizált felismerését, valamint katalógusleírások témamodellezését.[20]

2019. október 31-én fontos hír jelent meg a *Nature* folyóiratban. Az azt figyelemmel kísérő tudományos közösségben nagyot robbant a hír, hogy a velencei „Időgép" projektet felfüggesztették a nemzetközi partnerek egyet nem értése miatt.[21] A projekt a Lausanne-i Szövetségi Műszaki Egyetem, a Velencei Állami Levéltár és a velencei Ca' Foscari Egyetem együttműködésével indult, amikor 2014-ben mindhárom intézmény nem kötelező erejű egyetértési megállapodást írt alá a munka elvégzéséről. A projekt célja az volt, hogy digitalizálja az állami archívum több mint 80 kilométernyi dokumentumát. Azért függesztették fel, mivel a megállapodás nem határozta meg az eredménytermék felhasználásának pontos paramétereit, amelyek a digitalizált adatok kutatók általi felhasználását szabályozták volna az olasz törvényekkel összhangban. Hozzá kell tennem, hogy a konzorciumi megállapodások részletei döntő fontosságúak egy ilyen horderejű program esetében.

Ausztráliában, az Új-Dél-Wales-i Állami Levéltár Digitális Állami Levéltárában a gépi tanulást vizsgálták, különös tekintettel annak a nyilvántartás kezelésben és az elektronikus iratok selejtezésében való alkalmazására. 2017-ben és 2018-ban két pilot projekt, azaz próbavállalkozás indult azzal az elsődleges céllal, hogy a miniszterelnöki kabinet elektronikus iratait dolgozzák fel. Konkrét céljuk volt egy olyan MI eszköz kifejlesztése, amellyel hasznos terméket hozhatnak létre az igazságszolgáltatás számára.[22]

---

19 A verseny egy napindító (*bootphase*) előadás-sorozattal indul, amely kiválóan alkalmas az egyetemen megszokottól eltérő módszerű képzési környezet kialakítására. Az előadások után következő „Kérdezz, felelek"-blokkok a felvetődő problémák és kérdések megválaszolásával segítik a versenyzőket az elindulásban és ötleteik kidolgozásában. A második fázisban a résztvevők egy közös, plenáris ötletbörzén prezentálják az ötleteiket, amelyek közül a legjobbakat szavazással választják ki, majd a legjobb ötletek gazdái csapatokat kapnak oly módon, hogy a résztvevők megjelölik a nekik legjobban tetsző ötletet, és így csatlakoznak valamelyik csapathoz. A kialakuló *ad hoc* csapatok a rendelkezésükre álló időben kidolgozzák, élet- és versenyképessé alakítják az ötletet. Azaz a résztvevőket játékos módon, élményalapú, korszerű és hatékony eszközökkel – játék- és tapasztalati alapon – képezik, miközben jobban megismerik érdeklődési körüket és a szokásostól eltérő csoportban mutatott attitűdjüket, az ilyen szituációkban jelentkező kreativitásukat.

20 Mark Bell: *Machine learning in the archives*. 2018. https://blog.nationalarchives.gov.uk/machine-learning-archives/ (Utolsó letöltés ideje: 2020. szeptember 23.)

21 Davide Castelcchi: Venice 'time machine' project suspended amid data row. *Nature*, 2019. október 31.

22 Glen Humphries: *Case Study – External Pilot – Machine Learning and Records Management*. 2018. https://futureproof.records.nsw.gov.au/case-study-external-pilot-machine-learning-and-records-management/ (Utolsó letöltés ideje: 2020. szeptember 23.)

A Smithonian Intézet tudósai a növénycsaládok közötti különbözőségeket és hasonlóságokat vizsgálták olyan, ún. „mély tanulásos" algoritmusok segítségével, amelyek során 90% fölötti azonosítási pontosságra törekedtek. Az első kísérletben nyolcezer mintát használtak, és azt szerették volna tudni, hogyan katalogizálhatja a rendszer a növénypéldányokat. A kísérlet végén a pontosság több mint 90 százalék volt. A gépi tanulás az addigi időpazarló folyamatot egy néhány napos gyors automatizált elemzésre változtatta.[23]

2018-ban egy remek kísérletet folytatott a BBC a brit Nemzeti Levéltárral együttműködve. A BBC kutatási és fejlesztési részlege egy olyan MI-t használó alkalmazás fejlesztését tesztelte, amely a műsorkészítők és szerkesztők munkáját szerette volna megkönnyíteni. A cél a BBC 4-es csatornája esti programkínálatának összeállítása volt, teljes egészében a csatorna archívumában a gép által válogatott programokból.[24] Osztályozták a BBC műsorait és speciális, BBC 4-es csatornajellemzőket is gyűjtöttek, majd azt kérték az MI alkalmazástól, hogy válassza ki a legjobb 150-et a több mint 270 ezer archív műsorból. A kiválasztott programokat ezután egy másik algoritmus segítségével osztották jelenetekre, amelyeket kísérleti adásukban le is játszottak. A projekt a kutatás szempontjából jelentős siker volt. Az általuk használt technikák és technológiák szélesebb körben is alkalmazhatók olyan termékek, szolgáltatások és eszközök fejlesztése során, amelyek lehetővé teszik az archivált adatokhoz, ebben az esetben a műsorokhoz való könnyebb hozzáférést, valamint a levéltár információinak új és kreatív módon történő újrahasznosítását.

A Magyar Nemzeti Levéltár hadifogoly-projektje

2019. április 8-án a Magyar Nemzeti Levéltár és az Orosz Állami Hadilevéltár munkamegállapodást kötött a Magyar–Orosz Levéltári Vegyesbizottság munkatervével összhangban.[25] Az egyezmény értelmében az Orosz Hadilevéltár digitalizálja a gyűjteményében levő úgynevezett nyilvántartó kartonokat, amelyek azoknak a magyar nemzetiségű személyeknek az adatait tartalmazzák, akiket a Vörös Hadsereg egységei foglyuk ejtettek a második világháború során, majd a Szovjetunióban hadifogolyként őriztek. Az együttműködési egyezmény értelmében 2019. december 1-jéig öt szakaszban összesen 682 131 nyilvántartó karton másolata kerül a Magyar Nemzeti Levéltár állományába, ami összesen 1 364 262 digitális fel-

23 Ryan P. Smith: *How Artificial Intelligence Could Revolutionize Archival Museum Research*. 2017. https://www.smithsonianmag.com/smithsonian-institution/how-artificial-intelligence-could-revolutionize-museum-research-180967065/ (Utolsó letöltés ideje: 2020. szeptember 23.)

24 Tim Cowlishaw: *Using Artificial Intelligence to Search the Archive*. 2018. https://www.bbc.co.uk/rd/blog/2018-10-artificial-intelligence-archive-television-bbc4 (Utolsó letöltés ideje: 2020. szeptember 23.)

25 2020-ban indul a szovjet fogolykartonok feldolgozása. https://mnl.gov.hu/mnl/ol/hirek/2020_ban_indul_a_szovjet_fogolykartonok_feldolgozasa (Utolsó letöltés ideje: 2021. február 4.)

vételt jelent. A kartotékrendszer nemcsak azoknak a személyeknek az adatait tartalmazza, akik hadifogságba estek, hanem azokét is, akiket civilként internáltak, majd deportáltak a Szovjetunióba. A kartonokról tudni kell, hogy magyarul nem értő szovjet katonák töltötték ki őket, mégpedig kézírással, cirill betűkkel, hallás alapján. Ez azt jelenti, hogy az adatok szinte biztosan torzultak, így korrekciójuk külön feladat lesz. A projekt sikeressége érdekében a későbbiekben sor kerülhet kézírás-felismerést használó, valamint gépi tanulásos programok fejlesztésére és alkalmazására is.

Következtetések

Kétségtelen, hogy a levéltárak szerepe világszerte változik. Ebben a grandiózus átalakulásban a levéltáraknak az élen kell járniuk a saját jövőjük érdekében, hogy képesek legyenek irányítani azt, és ne veszítsenek teret. Mivé válhat a levéltáros szakma a jövőben? A levéltári egységek leírását és azok védelmét főként gépek, betanított személyzet és papírrestaurátorok fogják végezni. A levéltártudomány inkább egyfajta módszertan lesz, semmint önálló tudományág, amellett, hogy az interdiszciplinaritás még komolyabb teret fog nyerni. A levéltárak elsődleges feladata (a közigazgatási funkció mellett) a kutatás, a kutatás támogatása és az oktatás lesz a megőrzendő és digitalizálandó gyűjtemények szaporítása mellett. Az intézményi modellek teljesen átalakulnak és összeolvadnak más közgyűjteményekével. A múzeumok, könyvtárak és levéltárak közötti intézményi különbségek megszűnnek. A megfelelő gyűjtemények digitálisan egy helyen lesznek elérhetőek, még akkor is, ha fizikailag máshol tárolják őket. A digitális tárgyak közötti kapcsolatokat nem csak a kurátorok, hanem maguk a felhasználók is létrehozhatják majd (ahogy erre már láthatunk is példákat), ami akár intézményi elhelyezésüket is befolyásolhatja majd.

A levéltárak hasznossága egyáltalán nem lebecsülendő, mivel az új technológiák új lehetőségeket jelentenek, és a munka fontosságának újra-elismerését (és elismertetését) is magukban hordozzák. A jövőben az eddigi tapasztalatokra és készségekre is szükség lesz, de alkalmazkodóképesnek kell lenni, valamint az új körülményeknek, kihívásoknak kell megfelelni.[26] A levéltárnak teljes mértékben a felhasználóra kell összpontosítania (legyen az a kutató vagy a közigazgatás egy intézménye), és a tartalom fejlesztését kell szolgálnia, a lehető legnagyobb mértékben fenntartható és automatizált folyamatokat kell bevezetnie, továbbá a technológiát még okosabban és racionálisabban kell használnia.

A jövő egyértelműen digitális. Ha a szakma gondolkodása nem változik, nem alkalmazkodik az állandó változásokhoz, akkor az egyébként elkerülhetetlen kockázatok mellett más veszélyekkel is szembe kell majd néznünk.

26 Caroline Brown: *Archival Futures*. Facet, 2018. V.

István Hegedűs

HOW DOES ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING HELP TO RETHINK THE ROLE OF ARCHIVES?

Although the artificial intelligence is the product of science-fiction literature, it currently represents a significant branch of computer science dealing with intelligent behavior, machine learning, and machine adaptation. It became a discipline that attempts to answer real-world problems. Artificial intelligence systems are nowadays widely used in economics, medicine, design or by the armed forces. The role of archives is changing worldwide. In this grandiose transformation, archives need to be at the forefront of their own future, so they need to steer, guide themselves, and try to not lose out. The vast masses of archival records provide an excellent platform for the exploitation of artificial intelligence. The plethora of data could be a great help not only for research but also for preparation of policies and in some areas of the public administration in the not too distant future.