A bölcsészettudományok kutatásának forradalmasítása

Beszélgetés Biszak Sándorral

Biszak Sándor 1989-ben alapította az Arcanumot*, aminek főigazgatójaként egy nemzetközi szinten is sikereket elérő, a hazai tudományos életben pedig pótolhatatlan adatbázist hozott létre. Az* Arcanum *alá ma az* Arcanum Digitális Tudománytár *(ADT), a* Szaktárs*, a* Mapire*, a* Hungaricana*, az* Arcanum Kézikönyvtár *és a* Matricula *adatbázisok tartoznak.*

Hogyan alakult meg a ma ismert Arcanum*? Milyen mérföldkövek voltak az adatbázis „pályafutásában"?*

Az Arcanum 1989-ben alakult. Érdekesség talán, hogy a Szabadalmi Hivatallal együtt alapítottuk meg, amivel már 1989-et megelőzően is sokat dolgoztunk együtt. A Szabadalmi Hivatalnak akkor a CD-ROM elterjesztése volt a célja a magyar piacon; mind úgy gondoltunk, hogy a CD-é a jövő. A Hivatallal egyébként a mai napig együttműködünk, és mostanra a teljes szabadalmi állomány felkerült az internetre. Tehát CD-ket adtunk ki kezdetben: a tizenhárom *Biblia*-fordítás és az ötven magyar költő összes versét tartalmazó *VersTár* voltak ekkor a legnagyobb sikerek. Ugyanígy CD-n adtuk ki például Marczali Henrik munkáit is. De már tulajdonképpen ekkor a digitalizálás volt a legfőbb célunk.

A *Századok* volt az első anyag, amit elkezdtünk legépelni. Az első három-négy évfolyamot sikerült, mert az utolsó három-négyet megkaptuk digitálisan. Mint kiderült, ez a módszer nagyon sok korlátot hordoz magában. Emlékszem, három évfolyam *Századok*at és öt évfolyam *Vasárnapi Újság*ot próbáltunk meg legépelni, de nyilvánvalóvá vált, hogy ez így reménytelen.

Akkoriban, a 2000-es évek elején jelent meg a *Google Books* szolgáltatás és a *Google Patents* is, ami a szabadalmakat dolgozta fel, és ezek új technológiát jelentettek: elterjedtek a kétrétegű PDF fájlok. Az ezzel a technológiával elért automatikus szövegfelismerés lehetővé teszi a szöveg kereshetőségét, kiemelhetőségét, átemelhetőségét. Ez gyakorlatilag forradalmasította az *Arcanum* világát is, hiszen így vált lehetővé a tömeges digitalizálás. A gépeléssel nem igazán lehetett haladni, korrektúrázásra is szükség volt, tehát borzasztóan lassan lehetett a munkát elvégezni, most azonban egy teljesen új világ nyílt ki előttünk. Érdekes, hogy szintén a *Századok* volt az az anyag, amit ezzel a technikával elsőként feldolgoztunk, a második pedig a *Vasárnapi Újság* – ezeknek a kiadását CD-n nem sikerült korábban megvalósítani.

Ez a forradalmi változás elsősorban a szakfolyóiratokat érintette: a *Századok* mellett például az *Etnográfiá*t és a *Filológiai Szemlé*t. Ezek rendkívül fontosak voltak a kereshetőség szempontjából a kutatóknak, vagyis az indulás nagyon ígéretes volt, az adatbázis pedig fokozatosan fejlődött a következő egy-két évben. A napilapok digitalizálása viszont ekkor még megvalósíthatatlan, óriási projektnek tűnt. Emlékszem, egyszer egy *Népszava* sorozatot akartak kidobni a Politikatörténeti Intézetben, és ekkor gondoltuk úgy, hogy mégis meg kellene próbálni egy ilyen végtelenül nagynak tűnő, több mint félmillió oldalt tartalmazó, ötven

polcfolyóméteren végignyúló lapnak a digitalizálását. Mivel ez kifejezetten jól sikerült, a digitalizálás még nagyobb lendületet vett.

A szakfolyóiratok digitalizálása a kutatók felől érkező igényekre, a napilapoké pedig a mindennapok iránt érdeklődők igényére adott válasz volt. A városi, megyei és országos napilapok digitalizálásával és elérhetővé tételével tehát már nemcsak a tudományos köröket, hanem a nagyközönséget is elértük.

Tehát fogalmazhatunk úgy, hogy az Arcanum Digitális Tudománytár *feladata a nyomtatott sajtó archiválása és elérhetővé, illetve kutathatóvá tétele?*

Igen, az *Arcanum*hoz tartozó *Hungaricana, Szaktárs* és *Mapire* más-más feladatokat látnak el. A *Hungaricana* adatbázisába az állami és egyházi finanszírozású intézmények anyagai kerülnek be, ezért nyilvános. Ezzel szemben a *Szaktárs* esetében az egyes könyviadókkal működünk együtt, és digitalizáljuk a könyveiket, míg a *Mapire* teljesen saját vállalkozás az osztrák, horvát, szlovén és magyar levéltárakkal együttműködésben.

Hogyan zajlik a Szaktárs *és a* Hungaricana *adatbázisának a bővítése?*

A szakkönyvek, de általában a könyvek digitalizálása nagyon érdekes és nehéz téma, hiszen a könyvek a folyóiratokhoz, napilapokhoz képest megfoghatatlanok. A *Népszava* ötven polcfolyóméter, a *Századok* tíz polcfolyóméter, ezeket meg lehet fogni, le lehet emelni a polcról, de ilyen értelemben a könyveket már nem. A könyvek számbavétele egészen nagy kihívást jelent.

Mi havonta egy-másfél millió oldalt digitalizálunk, de egymillió oldalnyi könyvet képtelenség összeszedni. Ezért a könyvek esetében csak azt tudjuk csinálni, és csináljuk is, hogy az egyes kiadókkal való együttműködés mellett a nagy sorozatokat, amiket CD-n is kiadtunk, újra digitalizáljuk, illetve időnként tematikus gyűjteményeket digitalizálunk. Ilyenek például a 2014-ben megjelent első világháborús gyűjteményünk, a családtörténeti művek vagy olyan nagy sorozatok, mint a *Pallas Nagy Lexikona* vagy a *Révai Nagy Lexikona*. Ezek a sorozatok ugyanis megfoghatók, leemelhetők a polcról, akárcsak a nyomtatott sajtó. De a kiadókkal folytatott közös munkán és a sorozatokon kívül valóban nagyon nagy kihívás a könyvek digitalizálása, hiszen mi alapján válasszuk ki a digitalizálandó művet? Nincs meg az a fajta rendszer, ami a folyóiratok és napilapok esetében segíti a munkánkat.

Külön fontos kiemelni az Arcanum Mapire *adatbázist, ami térképek digitalizálását és elérhetővé tételét végzi el. Milyen térképek érhetők el ebben az adatbázisban, és ezek milyen kutatásokba illeszthetők be?*

Az osztrák hadsereg 1766–1785-ben, 1806–1869 között és 1872–1884-ben végezte el az első, második és harmadik katonai felmérést az Osztrák–Magyar Monarchiában, amelyeknek során rendkívül részletes térképek születtek. A térképek 1:28 800-as méretarányúak (az utolsó csak 1:25 000-es), tehát településszint alatt ábrázolják az országot, vagyis például minden falu minden háza látható rajtuk.

A Hadtörténeti Intézet és Múzeum birtokában van a harmadik katonai felmérés színes, fénymásolt példánya, amivel elsőként kezdtünk el foglalkozni. Ezt követően a felmérések Bécsben őrzött eredeti példányainak digitalizálásával bővítettük az adatbázist, amik a Magyarországon kívül eső területeket, így Csehországot, Ausztriát, Lombardiát és Galíciát is tartalmazzák. A kataszteri térképeink még részletesebbek, ott már a mezőgazdasági parcellák is láthatók.

Úgy gondolom, hogy a világhírt a térképekkel értük el, hiszen ilyen digitális adatbázisa a történeti térképeknek nincsen máshol a világon. Nagyon kevés ország rendelkezik ilyen régi térképekkel, amelyek ennyire részletesek, de teljesen egyedi az is, hogy a *Mapire* lehetővé teszi a georeferálást, amikor a korabeli térképet pontos koordináták alapján ráhelyezhetjük a mai térképre, így megfigyelhetjük az eltelt idő alatt bekövetkezett változásokat az adott településeken. A *Budapest időgép*, szintén georeferálással, a főváros esetében tesz lehetővé nagyon részletes történeti térképészeti kutatást. Így a kutató akár egyes épületek esetében nemcsak a korabeli állapotokkal vetheti össze azok jelenlegi helyzetét, de akár az épületekre vonatkozó terveket is megtekintheti, illetve már háromdimenziós megjelenítési formák is elérhetők. Ez a részletesség a helytörténeti, térképészeti kutatások mellett akár családtörténeti kutatásokban is segíthet.

Ezen az innovatív projekten az ELTE Geofizikai és Űrtudományi Tanszékével együtt dolgoztunk, és annyira sikeres lett, hogy ma már jelentős nemzetközi érdeklődéssel is számolnunk kell.

Hogyan változott a tartalmakhoz való hozzáférés az elmúlt évtizedekben, és elsősorban kik keresik fel az Arcanum*ot?*

Eleinte az Elektronikus Információszolgáltatáshoz (EISZ) csatlakozó intézmények, elsősorban egyetemek és akadémiai intézetek kaptak hozzáférést, ekkor még magánelőfizetőink nem voltak. Amikor azonban a napilapok digitalizálásába is belevágtunk, akkor már magánelőfizetőket is fogadtunk, ami azt jelentette, hogy egyre szélesebb körben váltak hozzáférhetővé az *Arcanum* adatbázisai.

A *Hungaricana* és a *Mapire* esetében gyakorlatilag korlátlanul hozzáférhet bárki a dokumentumokhoz a világ bármely tájáról, de a *Szaktárs* és az *ADT* esetében is vannak nemcsak magyar, hanem külföldi előfizetőink is, például amerikai egyesült államokbeli egyetemek, de szingapúri ügyfelek is.

Főleg a bölcsészettudományokkal foglalkozó szakemberek keresnek minket, hiszen akár fogalmazhatnánk úgy is, hogy az *Arcanum* forradalmasította a bölcsészettudományok esetében nélkülözhetetlen dokumentumokhoz való hozzáférést. Különösen jelentősek a történészek és a néprajzkutatók, de muzeológusok és levéltárosok is rendszeresen keresik fel az adatbázisainkat. Ráadásul ma már a tudományos kutatást végzők mellett nagy számban vannak „amatőr" kutatók is.

Hogyan érintette a koronavírus járvány az Arcanum*ot? A könyvtárak korlátozott elérésével megnőtt az online adatbázisok szerepköre. Jelentkeztek új ügyfelek? Nőtt az* Arcanum *látogatottsága?*

Némileg furcsa a helyzetünk, hiszen az *Arcanum*ot egyértelműen pozitívan érintette a pandémia. A könyvtárak, levéltárak bezártak, viszont mi mindig „nyitva" voltunk az online térben. Így nagyon népszerűvé váltak az adatbázisaink, jelentősen megnőtt az érdeklődők száma. 2020-ban bizonyos időszakokban nyilvánosan hozzáférhetővé tettük az adatbázisokat, ekkor nagyon megugrott a látogatottságunk, illetve ennek révén nagyon sokan megismertek minket, így később egyértelműen nőtt az előfizetőink száma is.

Ugyanakkor természetesen számunkra is kihívást jelentett a járvány, hiszen a könyvtárak, levéltárak bezárásával mi sem tudtunk úgy anyagot gyűjteni, ahogy korábban, de szerencsére a problémák kisebb mértékben érintettek minket, könnyen áthidalhatók voltak.

Jelenleg milyen új projekteken dolgoznak?

Nemrég elkezdtünk Romániában található dokumentumokkal foglalkozni. 2021 áprilisában vágtunk bele Marosvásárhelyen magyar nyelvű anyagok digitalizálásába. Már régóta terveztük, hogy külföldi, idegennyelvű anyagokat is megpróbálunk gyűjteni, eddig azonban ez nem sikerült. Ahogy azonban Marosvásárhelyen végeztünk a részben 19. századi, de elsősorban 20. századi magyar nyelvű sajtótermékekkel, szinte adva volt, hogy belevágjunk a román nyelvű lapokkal való munkába. Így megkezdtük a *Scînteia,* a Román Kommunista Párt napilapjának, valamint ennek kapcsán a magyar nyelvű *Romániai Magyar Szó* és ennek utódja, az *Előre,* illetve a szintén bukaresti, de német nyelvű *Neuer Weg* című lapok digitalizálását is. Ebben a projektben most tartunk körülbelül kétmillió oldalnál, és egyre több könyvtár keres fel bennünket, például Nagyszebenből és Aradról, de már könyvkiadók is felvették velünk a kapcsolatot.

Jelenleg szinte már minden hazai magyar nyelvű folyóiratot és napilapot elérhetővé tettünk digitális formában; amit nem, azt vagy azért nem, mert mi sem tudunk hozzájutni, vagy pedig azért, mert nem kapunk engedélyt a digitalizálásukra. Most harminckétmillió digitálisan elérhető oldalnál tartunk, ehhez minden hónapban igyekszünk még félmillió oldalt hozzátenni, de a magyar nyelvű anyagok esetében már nem számítunk olyan jelentős mérföldkövekre, mint amiket a kezdetekkor elértünk. Vannak magyar nyelvű gyűjteményeink, amiket az Egyesült Államokból vagy Ausztráliából szereztünk be, és bár ezek komoly oldalszámokat jelenthetnének, sajnos alig őrizték meg őket teljességükben, esetenként az adott folyóirat 30–40%-a elpusztult, teljes évfolyamok hiányoznak, mint például az Egyesült Államokban nyomtatott *Amerikai Magyar Népszava* című lap esetében. Nagyobb mennyiségű magyar anyagra a romániai anyagok példáján keresztül tehát főleg a szomszédos országokból számíthatunk, és tervezzük is a szlovákiai vagy horvátországi magyar nyelvű sajtó begyűjtését és digitalizálását.

Milyen trendekre lehet számítani a digitalizálás és az adatbázisok fejlődésének, bővítésének terén? Hogyan látja az Arcanum *jövőjét?*

Az új trendek tekintetében a legígéretesebb jelenleg a mesterséges intelligencia alkalmazása. Eddig tulajdonképpen a gyűjtőmunkát végeztük el, most pedig a mesterséges intelligencia révén rátérhetünk ennek a hatalmas mennyiségű anyagnak az adatfeldolgozására.

A mesterséges intelligencia első elemei már használhatók nálunk. Ilyen például az arcfelismerés funkció, ami abban nyújt segítséget, hogy a feltöltött kép alapján a program az arc vizsgálata után kiválasztja és megmutatja a keresett személyről az adatbázisban található képeket. Ezt ma már nagyon sokan használják például az Országgyűlési Könyvtárban, a Fővárosi Szabó Ervin Könyvtárban vagy a Nemzeti Filmintézetben a képeken, filmkockákon szereplő ismeretlen személyek azonosítására. A következő nagy alkalmazásunk a tulajdonnévfelismerés, ami régi probléma a számítógépes nyelvészetben, hiszen a keresőprogramok jelentős része nem ismeri fel a tulajdonneveket, holott a keresések 99%-a helynévre vagy személynévre vonatkozik. A felismerésükön kívül továbbá a mi programunk azt is megkülönbözteti, hogy milyen tulajdonnévről van szó: hely-, személy- vagy intézménynévről. Ezt is a mesterséges intelligencia segítségével, az arcfelismerést ráadásul a világ egyik legnagyobb digitális szolgáltatójával, az Amazonnal együttműködésben dolgoztuk ki.[1]

[1] Az Arcanum és az Amazon együttműködéséről lásd https://aws.amazon.com/blogs/machine-learning/arcanum-makes-hungarian-heritage-accessible-with-amazon-rekognition/

A jelenlegi automatikus szövegfelismerést is mesterséges intelligenciával próbáljuk javítani. Magyar nyelven az ilyen új innovációkat mi alkalmazzuk először, tehát a technológiai változásokat naprakészen követjük, így igyekszünk a szövegek kereshetőségének a minőségén folyamatosan dolgozni. Most például a napilapok esetében az egy oldalon szereplő cikkeket próbáljuk egymástól elkülöníteni, hogy a keresés ne egy oldalon, hanem egy cikken belül történjen. De szeretnénk, hogy lehetőség legyen csak képaláírásra, szerzőre, címre is keresni. Ezek mind mesterséges intelligenciával való képfeldolgozás révén érhetők el. A *Hungaricaná*ban található hangadatbázis esetében is dallamra történő keresési funkciót szeretnénk beállítani, ami pedig a hangfelismerésen alapul. Úgy tűnik tehát, hogy most a mesterséges intelligenciáé a jövő.

A beszélgetést készítette: BESSENYEI VANDA