Műhely

Közép-európai webarchiválási körkép 2021

NÉMETH Márton

Bevezetés

A 2021. november 23–24-én megrendezett *„404 Not Found – Ki őrzi meg az internetet?”* című konferencia és workshop első alkalommal bővült kétnapos online rendezvénnyé, mivel meghívtuk a környező országok közgyűjteményi webarchiválással foglalkozó szakértőit is, alapot teremtve egy jövőbeni együttműködés számára.

Az első napon a cseh, horvát, osztrák, szlovák és szlovén webarchívumok munkatársai mutatták be gyűjteményeiket, hosszú távú munkafolyamataikat, tevékenységeiket, illetve a rendelkezésükre álló technológiai eszköztárat, a Lengyel Állami Levéltár pedig a webarchiválást érintő terveit ismertette. Ezt követően mintegy egyórás kerekasztal-beszélgetésre került sor a jövőbeni együttműködés lehetséges területeiről.

Az alábbiakban bemutatjuk az egyes országokban zajló webarchiválási tevékenységet az elhangzott előadások nyomán, majd áttekintjük, hogy a közös gondolkodás eredményeként milyen együttműködési irányokat vázoltak fel a résztvevők.

A hazai könyvtári szaksajtóban többször felmerült korábban az egyes közép-európai országok webarchiválási gyakorlata, ám ez a rendezvény kínál lehetőséget arra, hogy egységes tematikus keretek között, a legidőszerűbb áttekintést nyújthassuk olvasóinknak.

A magyar webarchiválási eredményeket, törekvéseket külön cikkekben mutatjuk be a lapban.

* Valamennyi előadás prezentációja elérhető a rendezvény weboldalán: https://webarchivum.oszk.hu/404-not-found-workshop-2021-november-23-24/. A videófelvételek a Videotóriumban érhetők el: https://videotorium.hu/hu/channels/5339/404-not-found-ki-orzi-meg-az-internetet-2021. A magyar szolgáltatás honlapján a cikkben előkerülő fogalmakról kislexikon tájékoztat: https://webarchivum.oszk.hu/mediawiki/index.php/MIA_WIKI.

A Cseh Nemzeti Könyvtár webarchívuma

Alapvető információk

Prágában a Cseh Nemzeti Könyvtárban létrehozott webarchívum elsődleges feladata a – terület, nyelv, szerzőség vagy téma alapján – cseh vonatkozású webhelyek megőrzése a következő generációk számára. A webarchívum honlapja a *https://webarchiv.cz* címen érhető el. Jelenleg több mint 400 TB-nyi anyagot tartalmaz a gyűjteményük. Ennek gyarapításában és kezelésében négy webkurátor működik közre, az ő tevékenységüket egészítik ki a technikai támogatást nyújtó munkatársak. A webarchívum szakmai oldaláról *Marie Haškovcová*, a Webarchiválási Osztály vezetője, a technikai háttérről pedig *Zdenko Vozár*, az informatikai részleg igazgatója adott áttekintést.
A webarchiválással kapcsolatos első kísérletek a Cseh Nemzeti Könyvtár, a Morva Tartományi Könyvtár, illetve a brnói Masaryk Egyetem együttműködésében 2000-ben kezdődtek, az első weboldalak mentése 2001-re datálható. A nemzeti könyvtár keretei között zajló üzemszerű webarchiválási tevékenység 2005-ben indult el. 2007-ben csatlakoztak az internet archiválásban érdekelt intézményeket és piaci szereplőket összefogó *International Internet Preservation Consortium* (IIPC) tagjai közé. 2020-ban ünnepelték tehát a webarchívum 15 éves fennállását.

Jogi háttér

Archiválási, hosszú távú megőrzési célra mindenféle megkötés nélkül gyarapítható a gyűjtemény. A szerzői jogi szabályok miatt azonban csak azok a webhelyek tehetők hozzáférhetővé korlátozás nélkül, melyek *Creative Commons* licenccel bírnak, vagy ahol egyedi szerződést kötöttek a jogtulajdonossal. Ennek következtében sajnos csak a gyűjtemény 0,4%-a érhető el nyilvánosan, a teljes archívumhoz így csupán a könyvtár épületében lehet hozzáférni. További kihívást jelent, hogy a cseh kötelespéldánytörvény nem tartalmaz a digitálisan született dokumentumokra vonatkozó rendelkezéseket. A legújabb európai uniós szerzői jogi irányelvet,[1] mely megkönnyítené a webarchívum kutatási célú felhasználását, még nem vezették be a cseh jogrendben. Ennek az új irányelvnek az érvényesülése a tudományos kutatási célú felhasználásnak tágabb teret adhat a jövőben.

Gyűjteményszervezési irányelvek, együttműködés, kutatástámogatás

A cseh webdomén (*.cz*) általános aratásához szükséges az ebbe a tartományba bejegyzett doméncímek listája, melyet a Cseh Internetszolgáltatók Tanácsa bocsát a könyvtár rendelkezésére. Az általános aratásra évi egy vagy két alkalommal kerül sor, mintegy 1,4 millió domén anyagát rögzítve minimális mélységben, pillanatfelvétel-szerűen.
A *szelektív és eseményalapú gyűjtemények* témáit a webkurátorok választják ki, illetve ők válogatják össze a nagy mélységben, tartalmilag minél teljesebb mértékben aratandó webhelyek listáját is. A válogatás alapja a kiemelt kulturális, tudományos, történeti jelentőségű webhelyek különböző témakörökbe szervezett feltárása. Bárki javasolhat a webarchívum honlapján elhelyezett űrlapon archiválásra méltónak tartott webhelycímeket. Kialakítottak egy olyan gyűjteményt is, ahol vezető közéleti személyiségek ajánlottak személyenként tíz, általuk fontosnak ítélt webhelyet archiválás céljából. A Cseh Tudományos Akadémia Nyelvtudományi Intézete a profiljába tartozó weboldalakról külön tematikus gyűjteményt épít a nemzeti könyvtár szakmai támogatásával. A szükséges jogi feltételek megteremtésével mintegy 5000, szelektíven mentett weboldal anyaga érhető el nyilvánosan.
Az *eseményalapú aratások* témáit egyrészt előre lehet tervezni a különféle kiemelt események, évfordulók miatt, másrészt rugalmasan alkalmazkodni kell az aktuális történésekhez. Természetesen, amennyiben egy kiemelt témában az IIPC nemzetközi gyűjtési tevékenységet is szervez (pl. nagy sportesemények vagy a Covid19-járvány kapcsán), akkor a címlisták bekerülnek a közös gyűjteménybe is. A Pozsonyi Egyetemi Könyvtárral is építenek partnerségben tematikus, illetve eseményalapú gyűjteményeket. A prágai Károly Egyetem emellett napi szintű archiválást is végez a kiemelt jelentőségű hírportálok anyagáról.
A Cseh Nemzeti Könyvtár módszertani támogatást nyújt azon közintézmények, állami szervek számára, amelyek saját webarchívumot kívánnak kialakítani.
A szelektíven mentett webhelyekről készülő rekordok bekerülnek a Cseh Nemzeti Bibliográfiába, ahol az abban meghatározott tematikus rend szerint rögzítik őket.
Külön kutatási projekt foglalkozik egy *alkalmazáscsomag és szolgáltatói felület* fejlesztésével, mely nagy mennyiségű adatok kinyerését teszi lehetővé a webarchívum gyűjteményéből, kutatási célú felhasz-

nálásra. Ebben a Cseh Nemzeti Könyvtár mellett a Nyugat-morvaországi Egyetem Alkalmazott Tudományok Kara, illetve a Cseh Tudományos Akadémia Szociológiai Intézetének Kibernetikai Osztálya vesz részt.

Katalogizálás és bibliográfiai metaadatok

A nemzeti könyvtár az *Aleph* integrált könyvtári rendszert használja, melybe a bibliográfiai adatokat *MARC21* formátumban viszik be. 2015-től azonban a leíró jellegű metaadatokat már *RDA*-ban rögzítik, amihez külön útmutatót is készítettek a katalogizálást végző munkatársak számára. A weboldalak katalogizálását egy saját fejlesztésű, *WA-KAT* névre hallgató program segíti (*https://kat.webarchiv.cz/*), a leírásukhoz pedig külön ajánláscsomagot is készítettek (*https://webarchivcz.github.io/katalogizacni-manual*).

Szakmai kihívások

A nemzeti könyvtár egyik fontos feladata, hogy legalább az archivált webhelyekre vonatkozó leíró jellegű metaadatokat hozzáférhetővé tegye a nyilvánosság számára, olyan teljességben, amennyire csak szakmailag és jogilag lehetséges. Kiemelt kihívást jelent a közösségimédia-oldalak, illetve a dinamikus weboldalak mentése. Erre a célra a nemzetközi webarchiválási közösség által széleskörűen alkalmazott szoftvereket használják (*Webrecorder, ArchiveWeb.Page, Browsertrix*).

Fontos feladat a kutatói közösséggel való folyamatos kommunikáció biztosítása, az együttműködési lehetőségek elmélyítése. A webarchiváláshoz kötődő munkafolyamatok során a minőségbiztosítás kezelése, az automatizálás különféle újabb lehetőségeinek keresése, illetve a webarchívumokban tárolt személyes adatok megfelelő védelme szintén folyamatos kihívást támaszt.

Technikai háttér és fejlesztési tervek

Zdenko Vozár nagyon érzékletesen festette le azt a kihívásegyüttest, amelyre a webarchiválás szakmai tervezésénél, illetve a kapcsolódó technikai háttér felállításánál tekintettel kell lenni. A szoftver- és hardverhátteret úgy kell kialakítani, hogy az archiválás megbízhatósága garantálható legyen. Figyelni kell a deduplikációs szabályok megfelelő finomhangolására az egyes archiválási feladatok beállításánál, és ügyelni kell arra is, hogy az aratórobotok tevékenysége ne jelentsen aránytalan terhelést a webhelyeknél, különben ki lesz tiltva a robot. Optimalizálni kell a rendelkezésre álló személyi, technikai és anyagi erőforrásokat. Úgy kell megalkotni a feladatok modelljét, hogy minden tevékenységhez rendelkezésre álljon megfelelő személyi, technikai és anyagi háttér, még ha ez kényszerű szakmai és üzemeltetési kompromisszumok megkötéséhez is vezet. A Cseh Nemzeti Könyvtár webarchívumának éves gyarapodása 25–50 TB körül alakul, 2021. november elejéig a 23 TB-ot érték el. Ezt persze tovább lehet bontani a napi és havi tematikus, illetve eseményalapú aratásokra (melyek mindig duplumszűréssel zajlanak), illetve az évi 1–2 általános aratás adataira (itt deduplikáció nélküli, illetve deduplikáció futtatása utáni adatok is elérhetőek). A begyűjtött anyagokat tároló szervereket hierarchikus struktúrába szervezik. Az aratási környezet kialakításához szükséges virtualizáció *VmWare* segítségével történik. A rendelkezésre álló kapacitások függvényében alakítják különféle házi rendszabályokkal a hálózat általános működését, illetve határozzák meg az adott feladaton párhuzamosan dolgozó aratórobotok számát és paramétereit is. A szoftveres háttér áttekintésekor megemlítendő a saját fejlesztésű, *Seeder* nevű nyílt forráskódú alkalmazás (*https://github.com/WebarchivCZ/Seeder*), mely a webkurátorok munkáját segíti a gyűjteményszervezés alakításában és az aratási szabályok megalkotásában. A *Seeder*ből évente háromszor jelenik meg új verzió. A szelektíven gyűjtött anyagot a már említett *WA-KAT* segítségével katalogizálják. Egy külön szoftver szolgál az archivált anyag ellenőrzésére, illetve a hosszú távú megőrzési rendszerbe történő beszolgáltatáshoz szükséges metaadatok megadására. Jelenleg kísérletek zajlanak a webarchívumban tárolt nagy mennyiségű adat kutatási célú feldolgozására. Adatbányászati próbák is történnek szöveg, audio- és videoanyagok, linkek kinyerésére, a webhelyek kapcsolati hálójának vizsgálatára. A cél a különféle kutatási igényekhez illeszkedő adatkészletek előállítása. Ezzel összefüggő fejlesztési irány az adatkészletek vizsgálatára épülő tematikus tartalomelemzési modellek megalkotása, mégpedig a katalogizálási metaadatok felhasználásával, mesterséges intelligencia segítségével. Létre kívánnak hozni egy sokoldalú visszakereső felületet is, ahol facettaalapú szűrésre (aratások, aratási dátumok, tartalomtípusok, formátumok stb. szerint) nyílna lehetőség, és ahol saját szempontok szerint lehet majd gyűjteményeket építeni, illetve egyéni igények alapján adatszűrőket be-

állítani, tiltószavak listáját meghatározni. A visszakereső felület mellett *REST API* alkalmazásfejlesztési felületen keresztül is hozzáférést terveznek nyújtani a webarchívumhoz, ami lehetővé teszi például az IIPC által a webarchiválás kutatási célú használatának bemutatásához tervezett, *Jupyter Notebook*-alapú interaktív tananyagokhoz történő kapcsolódást is. Természetesen az *on-demand* adatexport szolgáltatást is biztosítani kívánják, *JSON* és *CSV* formátumokban, illetve *fulltext* tartalmak, kollokációk elérésével, hálózati elemzési adatok rendelkezésre bocsátásával. Kísérletek folynak továbbá egy új tárolási technológiai háttér fejlesztésére is, ahol az elsődleges cél a nagy mennyiségű adatok objektumalapú biztonságos tárolása, a jelenleginél gazdaságosabban üzemeltethető rendszerben.

Cseh konklúzió

Összefoglalásképpen elmondhatjuk, hogy a Cseh Nemzeti Könyvtárban a webarchiválási tevékenység immáron jelentős múltra tekint vissza, megfelelő személyi, anyagi és infrastrukturális háttérrel, magas színvonalon. Kimondottan értékesnek bizonyult a technológiai vonatkozásokat bemutató előadás és a fejlesztési tervek felvázolása. Az esetleges eredmények itthoni implementálása nagy reményekkel kecsegtethet a kutatási célú felhasználási lehetőségek új, magasabb szintre emelése érdekében.

A szlovák webarchívum hatéves útja

Alapvető információk

Szlovákiában a webarchiválás a Pozsonyi Egyetemi Könyvtár keretei között történik. Az első kísérletek 2005–2006-ban zajlottak egy webes kulturális örökséggel foglalkozó projekt során, melyben a könyvtár is szerepet vállalt. Akkoriban nem voltak meg a jogi, technikai és szervezeti keretei sem a webarchiválásnak, sem az online született digitális anyagok gyűjtésének. 2015-ben indult el – egy társadalmi-informatikai operatív program részeként, a *Digitális Erőforrások* nemzeti projekten belül – a webaratásra és a digitálisan született anyagok begyűjtésére irányuló munka. Az alapvető technikai feltételeket 2015 végére teremtették meg, ekkor történtek az első kísérleti aratások is. A rendszerszerű webarchiválás mint önálló könyvtári tevékenység 2016-ban kezdődött. A *Digitális Erőforrások* projekt fenntartási periódusa 2021 novemberében zárult le, így a webarchiválás most már teljesen a könyvtári munka részeként zajlik. A Digitális Erőforrások Osztályon egy osztályvezető és három webkurátor dolgozik, a digitálisan születő dokumentumokkal kapcsolatos feladatokat pedig egy félállású munkatárs látja el. A technikai üzemeltetést és támogatást külső cég biztosítja, szolgáltatásmegrendelési szerződés alapján. Az előadás, amelyet *Jana Matúšková* osztályvezető és *Peter Hausleitner* webkurátor állított össze és utóbbi kolléga tartott meg, alapvetően a szakmai munkafolyamatokra és a gyűjteményszervezésre koncentrált.

Jogi háttér

Az 1997-es kötelespéldány-törvény nem tartalmaz rendelkezéseket a digitálisan született anyagokról. 2021 októberében jelent meg az új, digitális publikációkkal foglalkozó törvény tervezete. Ebben megkísérlik definiálni a webportál fogalmát mint naponta frissülő digitális hírforrást, amely ily módon periodikának minősül. A tervezet szerint a portál tulajdonosának engedélyeznie kell majd a Pozsonyi Egyetemi Könyvtár számára, hogy mentéseket készítsen, és az archivált anyagot elhelyezze a hosszú távú digitális megőrzést szolgáló gyűjteményében. Módosítani kellene azonban a szerzői jogi és a kötelespéldánytörvényt is, hogy a könyvtárnak jogi felhatalmazása legyen a Szlovákiában született, illetve szlovák vonatkozású webes anyagok gyűjtésére. A kötelespéldánykötelezettség kiterjesztése megkönnyítené a könyvtár dolgát, hiszen nem kellene magához a gyűjtéshez is szerződést kötnie a honlapok tulajdonosaival. A cseh fejezetben is megemlített uniós irányelvet még a szlovák jogrendbe sem ültették át, ami pedig megkönnyítené a webarchívum kutatási célú felhasználását.

Digitálisan született folyóiratok, monográfiák és kapcsolódó weboldalak

A *Digitális Erőforrások* projekt alapozta meg a webarchiválási tevékenységeket, illetve a digitálisan született publikációk kezelését is. Ez magában foglalta a technológiai és szervezeti infrastruktúra kiépítését, a webhelyek és a digitálisan megjelenő periodikák rendszeres gyűjtéséhez, nyilvántartásához, hosszú távú megőrzéséhez kapcsolódó módszerek bevezetését, valamint a munkafolyamatok kialakítását. Szlovákiában tehát a Pozsonyi Egyetemi Könyvtáron belül a webarchiválás és az online periodikák gyűjtése azonos szervezeti keretek között zajlik. A digitálisan született periodikákra és monográfiákra

ISSN/ISBN számot kérnek, a szlovák nemzeti ISSN adatbázist pedig összekötötték a Pozsonyi Egyetemi Könyvtár *„born digital"* dokumentumokat nyilvántartó adatbázisával. Az önálló műveket PDF formátumban archiválják, ezeket vagy a kiadó küldi el, vagy a kurátor gyűjti be. A periodikákat, monográfiákat tartalmazó weboldalak gyűjtése webaratással történik a kurátorok jóvoltából. A kötelespéldánykörbe tartozó dokumentumok a *https://www.webdepozit.sk* oldalon teljes szövegű keresővel együtt érhetők el. 2021. október végéig 327 periodika és 67 monográfia került be a gyűjteménybe, 130 címet szolgáltatnak felhasználási engedéllyel, 167 cím nyílt hozzáférésű. 53 archivált folyóirat nem jelenik már meg, 25 pedig már teljesen elérhetetlen az élő weben.

A szlovák webarchívum gyűjteményszervezési, megőrzési és szolgáltatási keretei

A cseh jogi szabályozáshoz hasonlóan a begyűjtött weboldalak szolgáltatásához tartalomgazdai engedély, esetleg *Creative Commons* vagy egyéb nyílt hozzáférést biztosító licenc szükséges. Az archivált webes dokumentumokhoz való hozzáférésnek három típusa létezik: 1. nyilvános, 2. helyi könyvtári hozzáférés, 3. teljesen tiltott (különleges esetekben). A helyi hozzáférést az osztály 26 munkaállomással rendelkező kutatótermében biztosítják a felhasználók részére.

A Pozsonyi Egyetemi Könyvtár központi adatarchívuma tárolja a mentett anyagot, és gondoskodik annak hosszú távú megőrzéséről, a vonatkozó ISO szabványok előírásai szerint. 29 720 db *SIP* (beszolgáltatási) csomagban mintegy 4,5 TB mennyiségű tömörített webtartalom található az archívumban. Kétféle csomag van: a webről gyűjtött dokumentumokat tartalmazó, illetve a digitálisan született publikációkra kiterjedő típus, az előbbit szabványos *WARC* formátumban tárolják. Az archiváláshoz szükséges metaadatokat az összes beszolgáltatási csomag esetében *mets.xml* fájlok tartalmazzák. A digitális periodikáknál az anyagot PDF formátumban őrzik, *MARC XML*-alapú bibliográfiai leírások kíséretében. A webarchívum a *.sk* tartományban regisztrált domének listáját a szlovák Internetszolgáltatók Tanácsától kapja meg. A 2021-es adatok szerint 442 701 doménről van szó. Természetesen a gyűjtés kiterjed a szlovák vonatkozású, *.sk* doménen kívüli címekre is, különösen azokra, melyek digitálisan született periodikákat, monográfiákat is tartalmaznak. Összesen tehát 658 340 domén szerepel a webarchívum katalógusában, illetve 654 428 doménről kísérelnek meg tartalmat gyűjteni az évente legalább egyszer lefuttatott általános aratás során. 281 961 domén benne van az adatbázisban, de az élő weben már nem érhető el. 71 512 doménről gyűjtöttek be annyi tartalmat, hogy a *WARC* fájl mérete meghaladja az 5 MB-ot. A 2021-ben lezajlott általános aratás során 316 233 doménről sikerült tartalmat archiválni 16 TB összméretben, mintegy 255 millió webes objektumot gyűjtve be.

A szelektív archiválást tekintve az elmúlt hat évben 26 különféle tematikus kategóriában folyik, illetve folyt gyűjtés, amit kiegészítenek az eseményalapú gyűjtések. 229 URL-címre 223 felhasználási szerződést sikerült megkötni. Összesen 82 tematikus és 57 eseményalapú aratási művelet indult el 2021. november közepéig. Az eseményalapú gyűjtéseknél együttműködnek a cseh webarchívummal a mindkét országot érintő történések esetében, és címlisták cseréjére is sor kerül. 2019-ben a *Bársonyos Forradalom* húszéves évfordulóján a két webarchívum közös gyűjtést tartott, és jelenleg is folyik egy hasonló projekt a mindkét országban elismert művész, *Miroslav Žbirka* halálának apropóján. A szlovák webarchívum is tagja az IIPC-nek, és részt vesz a közös címgyűjtési tevékenységekben.

Az archiválás során természetesen itt is jelentkeznek a közösségi média, illetve a dinamikus webtartalmak esetében tapasztalható kihívások. Megfelelő szabályozási háttér hiányában nem kerülhetik meg az egyes webhelyeken található *robots.txt* fájlban beállított korlátozásokat, ami szintén hátráltatja a munkát.

Jövőbeni tervek

A könyvtárnak új stratégiát kell alkotnia a digitális dokumentumokkal kapcsolatos tevékenységeket, munkafolyamatokat illetően – mivel, mint említettük, az eddigi munkának keretet adó projekt kifutott –, és ebben rögzíteni kell a jövőbeni tevékenységekre vonatkozó szándékokat. Nagyobb hangsúlyt kellene helyezni a multimédia anyagok archiválására. Meg kell oldani a digitális dokumentumok katalógusának integrálását a Pozsonyi Egyetemi Könyvtár, illetve a túrócszentmártoni Szlovák Nemzeti Könyvtár központi katalógusaival. Szükséges lenne együttműködni a szlovák kormányzati szervekkel, tudományos könyvtárakkal, a Szlovák Tudományos Akadémia kutatóintézeteivel az őket érintő webes tematikus gyűjtőkörök formálása terén. Meg kell továbbá oldani a központi adatarchívumban tárolt állomány

szolgáltatását, és fejleszteni szükséges a hardveres és szoftveres hátteret az adatexport és adatbányászat céljára. Kifejezett szándékként merül fel a visegrádi országokkal történő együttműködés az információcsere, illetve a közös szolgáltatások fejlesztése érdekében (elég csak arra utalni, hogy a magyar nyelvű szlovákiai és a szlovák nyelvű magyarországi weboldalak esetén egymást metszi a magyar és a szlovák webarchívum gyűjtőköre). Végezetül bővíteni szeretnék nemzetközi tevékenységeiket az IIPC keretei között, valamint kérdőíves felmérést kívánnak folytatni a kutatói felhasználási igényekről is.

Szlovák konklúzió

A pozsonyi könyvtár webarchívuma ugyan a cseh webarchívumnál jóval fiatalabb, de sokat tudtak meríteni szomszédjuk tapasztalataiból. Dinamikusan fejlődnek, és remélhetőleg ez az ív a webarchiválás eddigi kereteit adó fejlesztési projekt lezárultával sem törik majd meg. Sok területen körvonalazódnak potenciális együttműködési lehetőségek, melyekről a zárófejezetben lesz bővebben szó.

Az osztrák webarchiválás első évtizedének tapasztalatai és kihívásai

Bevezetés

Michaela Mayr, az Osztrák Nemzeti Könyvtár Digitális Könyvtárának osztályvezetője előadásában elsősorban azokra a tényezőkre fókuszált, amelyek a sikeres nemzeti könyvtári webarchiválási gyakorlat kialakításához szükségesek, és rámutatott, hogy ehhez képest milyen kihívásokkal szembesülnek a gyakorlati munka során. Az Osztrák Nemzeti Könyvtárnak mintegy 300 munkatársa van, a járvány előtt a nemzeti könyvtár által üzemeltetett nyilvános fizikai tereket (olvasótermek, múzeumok) csaknem 900 ezren látogatták évente. Az 1386-ban alapított intézményben kb. 11 millió fizikai objektumot őriznek. A 2035-ig szóló könyvtári stratégia az eddiginél is nyitottabb intézményt irányoz elő, a fizikai és a virtuális térben egyaránt.

Jogi háttér

Az osztrák médiatörvény 2009-től tartalmazza a doménszintű és szelektív aratások szabályozását, a nemzeti könyvtári gyűjtőkörbe tartozó, jelszóval védett oldalakat is beleértve. Az archivált anyaghoz mintegy 20 osztrák könyvtár épületeiben lehet hozzáférni. Az archív tartalmat digitális formában nem lehet újra feldolgozni, sem pedig másolatot készíteni róla, a nyomtatás viszont lehetséges. Nagyon nehézkes a technológiai változásokat követni a jogalkotással (új médiaformátumok megjelenése, technikai követelmények módosulása stb.). A webhelyek, weboldalak fogalmát nem is olyan egyszerű pontosan meghatározni, illetve elkülöníteni a webhelyeken lévő önálló digitális objektumok kezelését (pl. e-könyvek) az általános webes környezettől. Mindazonáltal az e-könyvek közvetlen hozzáférését sikerült megoldani az Osztrák Tudományos Akadémia számára.

Szervezeti keretek, munkafolyamatok

A webarchiváláshoz szükséges hardveres háttér, illetve tárhely a nemzeti könyvtárban áll rendelkezésre. A webarchívumnak csupán egy főállású munkatársa van, aki a Digitális Könyvtár alá tartozik. A webkurátorok munkáját, az archiválás szervezését segítő *Netarchive Suite* szoftvercsomagot közösen használják és fejlesztik a dán, a francia, a spanyol és a svéd nemzeti könyvtárakkal. A gyűjtemény mérete mintegy 188 TB, és 2 459 583 doménről 4 371 324 936 objektumot tartalmaz.

A létszámhiány nagyon megnehezíti a webarchiválás különféle vonatkozásainak időszerű követését. Az újabb és újabb *„paywall"* (fizetőfal) mögötti, térítés ellenében elérhető tartalomszolgáltatások megjelenése egyre több könyvtárosi munkát igényel. A hosszú távú digitális megőrzés kereteit most alakítják ki a webarchívum és a többi digitális gyűjtemény összefüggésrendszerében is. Weboldaluk a *https://webarchiv.onb.ac.at,* itt kapott helyet a keresőfelület, a webhelyajánló űrlap, illetve itt lehet hozzáférni az alkalmazásfejlesztési felülethez (*API*-hoz) is.

Tapasztalataik arra utalnak, hogy alapvető fontosságú a kommunikáció szerepe. Miután a gyűjtemény zárt hozzáférésű, a könyvtárak fizikai tereiben nagyon kevés a felhasználójuk. Kulcskérdés tehát, hogyan érjék el az embereket, hogyan lehet a gyűjtemény társadalmi jelentőségét bemutatni. Nélkülözhetetlen a nemzeti könyvtár vezetőségének pozitív hozzáállása is. Belső és külső szövetségeseket kell találni a webarchiválás ügyének előmozdítására, és ez nem egyszerű feladat, mert a webarchiválás mintegy szigetszerűen létezik a nemzeti könyvtáron belül. Munkafolyamatai és infrastruktúrája élesen elkülönülnek a könyvtár többi részéről, sokan nem is értesülnek a webarchívum létezéséről az intézményen belül sem. Fontos feladat tehát, hogy a webarchiválást be

lehessen kapcsolni a hagyományos könyvtári munkafolyamatok közé, az eredményesebb munkához pedig szakreferenseket és webkurátori feladatokkal megbízható kollégákat kell találni házon belül. Az együttműködés kényszere abból is adódik, hogy végesek az anyagi és humán erőforrások, meg kell osztani a tapasztalatokat, a szakmai tudást, az eszközöket. Ez azonban nem mindig megy könnyen. Időt és energiát kell rá szánni, számolni kell a házon belüli eltérő munkakörülményekkel, nem is említve, hogy az intézményen belül eltérőek a prioritások és a célkitűzések is, amelyekhez illeszkedni kellene. Külön kihívást támaszt a kutatástámogatás. A kapcsolódó szoftvereszközök használatához sokszor személyre szabott segítség szükséges a kutatók számára. Nehezíti a munkát, hogy az adatok elkülönült silókban találhatók (weboldalak, digitalizált könyvek, e-könyvek, digitalizált történeti fényképek, digitálisan született fotóanyag stb.) – ezeket kiterjedt elemzésnek kell alávetni, és meg kell találni annak a módját, hogy az eddig elkülönülő silókat egységes környezetben lehessen kezelni. A kutatók munkáját a jogi korlátozások is nehezítik (pl. a különféle anyagok eltérő szerzői jogi státusza), és kihívásként jelenik meg a könyvtáron belüli megfelelő munkakörülmények biztosítása is.

Osztrák konklúzió

Michaela Mayr alapvetően *menedzsmentszempontú* előadása kitűnő betekintést engedett abba, hogy egy webarchívum kialakítása és eredményes működtetése során nemzeti könyvtári szinten milyen követelményeknek, kihívásoknak – például amikor nincs elegendő személyi erőforrás a munka minél teljesebb körű elvégzéséhez – kell és lehet megfelelni. Külső és belső szövetségeseket kell találni, rá kell mutatni a webarchiválás fontosságára, és megfelelő érdekérvényesítő erőt kell kifejteni az intézményen belül, máskülönben nem lehetséges az előrehaladás a jövőben.

Horvát webarchívum

Bevezetés

A horvátországi webarchiválás, melyről *Karolina Holub* és *Inge Rudomino* könyvtári tanácsosok nyújtottak áttekintést, a zágrábi Egyetemi és Nemzeti Könyvtár keretei között zajlik. Előadásuk fő vezérfonalát fokozatosan táguló tevékenységük főbb szakmai állomásainak bemutatása adta.

Horvátországban az internetszolgáltatás 1991-ben indult meg, a *.hr* domén pedig 1993-tól él. A kötelespéldány-törvényt 1997-ben módosították, az online publikációkra is kiterjesztve azt. A zágrábi Egyetemi és Nemzeti Könyvtár 1998-tól kezdett online anyagokat katalogizálni. 2003-ban indult el a kötelespéldány-törvény körébe tartozó webhelyek gyűjtését és katalogizálását vizsgáló kísérleti projekt, és 2004 szeptemberében kezdődtek az első mentések. A könyvtár 2008-tól vált az IIPC tagjává. A webarchiválásra vonatkozó kötelespéldány-szabályokat 2019–2020-tól fogalmazták újra.

A horvát webarchiválás keretei

A webarchiválás szakmai gondozását 2004-től az Egyetemi és Nemzeti Könyvtár látja el, a technikai hátteret a Zágrábi Egyetem Számítástechnikai Központja biztosítja. 2011-ben kezdődött a *.hr* domén általános mentése, illetve a tematikus és eseményalapú gyűjtemények kialakítása. A webarchívum építésében két főállású és egy félállású munkatárs vesz részt. A szelektív és eseményalapú archiváláshoz kötődő gyűjtőköri alapelvek angolul is elérhetők a *https://haw.nsk.hr/en/selection-criteria/* címen. Minden szelektív, illetve eseményalapú archiválás során kialakított gyűjteményt katalogizálnak, és az adatokat áttöltik a webarchívum rendszeréből a könyvtár integrált rendszerébe. A szerzői jogi státusztól függően a könyvtár épületén belül vagy nyilvánosan érhetők el az archívum tételei. Teljes szövegű kereső is rendelkezésre áll, részletes keresési opciókkal és tárgyszóalapú böngészéssel kiegészítve. *URN-NBN* azonosítót is hozzárendelnek minden begyűjtött webhelycímhez, illetve az archivált példányokhoz is. A metaadatok aratását *OAI-PMH* alapú felület segíti. Mintegy 8700 archivált tételhez 87 415 példány tartozik, ezek 22 tematikus és 4 eseményalapú gyűjteménybe szerveződnek.

Jövőbeni terveik között szerepel egy, a szelektív archiválást segítő új szoftvereszköz fejlesztése és üzembe állítása, valamint a közösségi média archiválásának megkezdése (elsőként a *Twitter*rel). Bővíteni szeretnék továbbá kapcsolataikat a tudományos kutatókkal is.

A *.hr* domén általános aratása 2011-től évente egy alkalommal történik, eddig tíz mentés készült el a *Heritrix* segítségével. Az archiválási műveletek számának növelése napirenden van, a hardverképességek függvényében. A mintegy egymilliárd webes objektumot tartalmazó archivált anyag az *OpenWayback*

megjelenítő program segítségével, URL-címek alapján érhető el a könyvtáron belül.
2021 tavaszától 4 közkönyvtár segítségével 4 új helytörténeti jellegű tematikus gyűjteményt alakítottak ki. Az egyes intézmények illetékes munkatársainak megtanították a webhelyek válogatásának módszereit, és hogy miként lehet tematikus gyűjteményt szervezni a begyűjtött webhelyekből, illetve hogyan lehet népszerűsíteni azt. A jövőben ezt az együttműködést több intézményre is ki kívánják terjeszteni. Ehhez testreszabható ajánló űrlappal segítenék a partnerkönyvtárakat a címek gyűjtésében, és évente közösen összegeznék az elvégzett munkát.

Horvát konklúzió

A zágrábi Egyetemi és Nemzeti Könyvtárban jelentős hagyományokkal bíró, átfogó és folyamatosan fejlődő webarchiválási munka zajlik, megfelelő technikai támogatással a háttérben. Nagyon jó a nemzetközi beágyazottságuk is: a járvány előtti utolsó személyes részvételen alapuló IIPC-konferencia helyszíne Zágráb volt.

Webarchiválás Szlovéniában

Általános keretek

A szlovén webarchiválás történetéről, jelenlegi helyzetéről és jövőbeni terveiről *Janko Klasinc*, a ljubljanai Nemzeti és Egyetemi Könyvtár webarchívumának munkatársa adott áttekintést.
Szlovéniában a jelenleg hatályos kötelespéldánytörvény 2006-ban született meg, és 2007-ben bővítették ki a digitális dokumentumokkal. Szelektív webarchiválás 2008 óta zajlik, 12 különféle kategóriában: több mint 1600 webhelyet gyűjtenek változó gyakorisággal, a hetitől az éviig. Az eseményalapú archiválás elsősorban a kiemelt politikai, közéleti és sportesemények webes lenyomataihoz kötődik. A tematikus és eseményalapú gyűjtemények kezelése a *Web Curator Tool* (eredetileg angol–holland, mostanság holland–új-zélandi fejlesztésű keretrendszer) segítségével történik.
A *.si* doménről kétévenként készül általános aratás a *Heritrix* programmal, mintegy 100 ezer doménre kiterjedően. Szlovéniában egy olyan egyedi jogértelmezés nyert teret, mely szerint a tematikus és eseményalapú gyűjteményekben tárolt tartalmak szabadon hozzáférhetők. A keresőfelület számos szűrési és keresési funkcióval a *http://arhiv.nuk.uni-lj.si/* webhelyen érhető el. Az Egyetemi és Nemzeti Könyvtár 2007 óta tagja az IIPC-nek, 2013-ban az éves közgyűlésnek és konferenciának is ez az intézmény volt a házigazdája. 2021-ben az afganisztáni események nemzetközi visszhangját rögzítő tematikus gyűjtemény bővítéséhez járulnak hozzá a konzorcium keretei között.
A jövőt tekintve szélesebb közönséget kívánnak bevonni az archiválásra javasolt címek ajánlásába. Fontosnak tartják az archiválás szükségességéről való webkurátori döntések alaposabb dokumentálását. Terveik között szerepel továbbá a böngészőalapú új archiválószoftverek használata a dinamikus weboldalak, a jelszóval védett tartalmak és a közösségimédia-oldalak archiválásának elősegítésére. A szelektíven és eseményalapúan begyűjtött anyagok vonatkozásában a már rendelkezésre álló metaadatkészletet és a tartalmat szeretnék gazdagítani, új felhasználási módokat és lehetőségeket keresnének, és az archivált tartalom visszakeresését is hatékonyabbá kívánják tenni.

Szlovén konklúzió

Az osztrák gyakorlathoz hasonlóan a ljubljanai Nemzeti és Egyetemi Könyvtárban is csupán egy főállású munkatárs foglalkozik a digitális könyvtári részlegen belül a webarchiválással. Ennek ellenére a jelentős múlttal bíró archívum fejlesztése folyamatos, és immár egy komoly gyűjteménnyel büszkélkedhetnek. A tervezett közép-európai együttműködés az intézményi erőforrások megosztásával új lehetőségeket kínálhat számukra a webarchiválási tevékenység további fejlesztésére.

A Lengyel Állami Levéltár webarchiválási tervei

A lengyel webarchiválási helyzet hátteréhez tartozik, hogy törvényi felhatalmazás hiányában a Lengyel Nemzeti Könyvtár nem foglalkozhat a webarchiválással. A Lengyel Állami Levéltár viszont, úgy tűnik, biztosítani tudja a szükséges feltételeket ehhez a tevékenységhez. Az intézmény terveiről *Nicola Herbert* rendszergazda adott rövid áttekintést. A szükséges jogi háttér megteremtése után a jövő évtől tervezik a webarchiválási munka elkezdését, és 2022 elejétől IIPC-tagok is lettek. Levéltárként elsősorban a *.gov.pl* kormányzati domén anyagainak archiválására szeretnének összpontosítani. 2021. november 16–17-én nemzetközi workshopot szerveztek, melyen az első napi előadások, illetve online felvételek segítségével különféle könyvtári és levéltári webarchiválási modellek kerültek terítékre. Előre elküldték a résztvevőknek, hogy milyen főbb elemekre,

összefüggésekre lennének kíváncsiak, ami nagyban megkönnyítette az OSZK webarchívumát bemutató előadásunk[2] összeállítását. A második napon a lengyel tudományos-egyetemi és közgyűjteményi szféra képviselőinek részvételével műhelybeszélgetésekre került sor a webarchiválás szakmai, jogi, technológiai összetevőiről és arról, hogy miképpen lehetne egy saját modellt felépíteni. Remélhetőleg a jövő évben lengyel kollégáink már konkrét kezdeti eredményekről is számot tudnak majd adni.

Kerekasztal-beszélgetés a közép-európai webarchiválási együttműködésről

A workshop délutáni programjában mintegy egyórás kerekasztal-beszélgetésre került sor e tanulmány szerzőjének moderálásával, cseh, szlovák, horvát és szlovén előadóink, kollégáink részvételével, melynek témája a közép-európai webarchiválást érintő közös tevékenységek körvonalazása volt. Az ötletbörze megalapozásához összeállítottunk egy vitaindító dokumentumot.

A résztvevők közül többen megerősítették azt a kezdeti állításunkat, hogy a digitálisan született anyagok archiválása és szolgáltatása, hosszú távú megőrzése olyan komplex kihívást jelent, melyre nemzeti keretek között csak korlátozottan lehet választ adni. A nemzeti webteret nem lehet az országhatárokat követve lehatárolni, mindenképpen lesznek olyan tartalmak, amelyek több webarchívum gyűjtőkörébe is beleillenek. Felvetődik a lehetőség közös gyűjtemények, tartalomszolgáltatások létrehozására is, melyek leképezik kulturális örökségünk egymással összefüggő szeleteinek webes reprezentációját. Erre a felvetésre is pozitívan reagáltak a résztvevők. Minél több helyen, minél több időpontban mentünk le egy adott webhelyet, annál nagyobb az esély arra, hogy az archivált tartalom ténylegesen megmaradjon. Létre lehetne hozni közös címlistákat, és közös tematikus, illetve eseményalapú gyűjtéseket is lehetne folytatni, melyek témája régióspecifikus, így nem illik az IIPC által kezdeményezett tevékenységek általános keretei közé.

Egyéb webkurátori területeken is megvizsgálhatjuk a feladatmegosztás lehetőségeit. Sajnos arra a szerzői jogi korlátozások miatt kevéssé van mód, hogy konkrét archivált tartalmakat is megosszunk egymással, átemeljünk egymás archívumaiba. Ugyanakkor lehetőség nyílik közös publikációk írására, a szakmai megjelenések igény szerinti egyeztetésére. Egy olyan informális hálózat létrehozására is kísérletet tehetünk, melynek révén megosztjuk a jó gyakorlatokat, szakmai tapasztalatokat egymás munkájának segítése érdekében.

Partnereink azt a felvetést is támogatták, hogy szükség lenne egy *közös kereső- és megjelenítőfelület* megalkotására, ahol a jogi követelményekhez illeszkedve legalább az archivált webhelyek metaadatai visszakereshetőek lennének, és egy soknyelvű portálon jelennének meg egységes keretben az egyes országok webarchívumainak szolgáltatásai.

Többen is szóvá tették, hogy *technológiai területen is mód nyílhatna az együttműködésre*. Fel tudjuk mérni, hogy ki milyen szoftvereket, milyen beállításokkal, milyen hardverinfrastruktúrával használ. Segíthetünk egymásnak új megoldások implementálásában. Ennek főleg azon archívumok esetében van fokozott jelentősége, ahol nagyon alacsony az informatikai támogatás mértéke, illetve a webkurátori tevékenységeknél is korlátozott humán erőforrás áll rendelkezésre. Sok esetben az egyes archívumok munkatársai jól fel tudják mérni, hogy milyen lépésekre lenne szükség, ehhez támogatást azonban akár intézményen belülről, akár azon kívülről alig kapnak. Az erőforrások részleges, feladatközpontú egyesítésével a technológiai fejlesztések is új lendületet kaphatnának.

Az IIPC jelentős erőfeszítéseket tervez tenni a közösségi média, illetve a dinamikus weboldalak archiválási lehetőségeinek felmérésére. Ehhez kapcsolódva közép-európai keretek között is megoszthatjuk egymással a jó gyakorlatokat, azokat a specifikus ismereteket, melyek az egyes nemzeti archiválási modellekhez illeszkednek, de globálisan talán kevésbé relevánsak, miközben a környező országoknak azok lehetnek.

A kerekasztal-beszélgetés végén abban állapodtunk meg, hogy 2022 januárjában folytatjuk a felvázolt témákban az egyeztetéseket annak érdekében, hogy meg tudjunk alapozni konkrét cselekvési terveket az együttműködésről, melyet a *Central European Web Archives* (CEWA) hálózat keretei között szeretnénk megvalósítani.

Epilógus

Jelen összefoglalóban áttekintést igyekeztünk adni arról, hogy a közép-európai országokban milyen webarchiválási tevékenységek zajlanak. Tanulságos ez a körkép abból a szempontból is, hogy az egyes országok webarchívumainak munkatársai milyen tényezőkre fektették a hangsúlyt. Egyedül a csehek esetében volt példa arra, hogy külön előadás foglalkozott

a webarchiválás technikai és gyűjteményszervezési, fejlesztési és webkurátori oldalával. Igen érdekes volt az Osztrák Nemzeti Könyvtár képviselőjének a többiekétől némiképp eltérő, a szolgáltatásmenedzsmentet a középpontba helyező előadása is. Az általános tanulságokon túl, kimondva vagy kimondatlanul megjelentek azok a kihívások is, melyek az adott országbeli webarchiválás fejlődésének főbb akadályát képezik. A szlovén esetben szembetűnő volt a technikai támogatás korlátozott volta, horvát szomszédjuk ezen a téren ellenben jóval szerencsésebb helyzetben vannak, mert a csehekhez hasonlóan biztosított náluk a technikai és a szakmai háttér is. Szlovákiában az eddigi lendületes fejlődést elindító projekt kifutása után újra kell tervezni a webarchiválás keretrendszerét, Lengyelországban pedig a közelmúltban az állami levéltár vállalta magára a webarchiválás feladatát, amelyhez még ki kell dolgozni a jogi hátteret. Workshopunknak kiemelt célja volt, hogy a bemutatkozó áttekintések körképe mellett megtegyük az első lépéseket az együttműködés feltételrendszerének megalkotása felé. A délutáni kerekasztal-beszélgetésen körvonalazódott, melyek azok a területek, melyeken a közös cselekvés elkezdődhet. Reményeink szerint 2022-ben mindezt konkrétumokká is lehet formálni, és a regionális összefogásból mindegyik partnerintézmény profitálni tud majd.

Irodalmi hivatkozások

1. Szerzői és szomszédos jogok az információs társadalomban [online]. Utolsó frissítés: 2021.09.23. Hozzáférhető: https://eur-lex.europa.eu/legal-content/HU/TXT/HTML/?uri=LEGISSUM:l26053&from=HU [Megtekintve: 2022.03.01.]
2. Elérhetők a lengyel webarchiválási workshop videói [online]. == OSZK Webarchívum. Hírek aloldal. Feltöltve: 2022.02.26. Hozzáférhető: https://webarchivum.oszk.hu/blog/2022/02/16/elerhetok-a-lengyel-webarchivalasi-workshop-videoi/ [Megtekintve: 2022.03.01.]

(Beérkezett: 2022. február 26.)

NETWORKSHOP – kötetek

Elérhető a 2021-es NETWORKSHOP 23 cikket tartalmazó konferenciakötete az MTA KIK REAL-ban – csakúgy, mint az ezt megelőző három év kötetei. Letölthetőek egészben, és az egyes cikkek külön-külön is.

NWS2021: http://real.mtak.hu/132242/
NWS2020: http://real.mtak.hu/119181/
NWS2019: http://real.mtak.hu/104936/
NWS2018: http://real.mtak.hu/87065/

A konferenciák archívuma:
http://ocs.mtak.hu/index.php/nws/index/schedConfs/archive