Vezetői összefoglaló a 2017-2018 évi webarchiválási pilot projektről

Utoljára módosítva: 2018-11-12

Előzmények:

Az Országos Széchényi Könyvtárban (OSZK) az Országos Könyvtári Rendszert fejlesztő program keretében 2017-2018-ban lezajlott az a pilot-jellegű projekt, ami egy üzemszerűen működtethető magyar webarchívum technikai, szakmai és jogi feltételeinek megteremtését tűzte ki célul. A nyilvános weben található, nagyrészt már eleve digitálisan születő kulturális örökség megőrzésének fontosságára az UNESCO egy 2003-ban közreadott Charta útján hívta fel a memóriaintézmények figyelmét és a világ számos országában 10-20 éve működnek is webarchívumok, többnyire a nemzeti könyvtár keretében vagy annak irányításával konzorciumi keretek közt. Az OSZK-ban az elektronikus publikációk és egyéb könyvjellegű digitális dokumentumok gyűjtése 1999 óta folyik az 1994-ben indult Magyar Elektronikus Könyvtárban, és később ez a tevékenység az Elektronikus Periodika Archívum és Adatbázis, valamint a Digitális Képarchívum elindulásával az időszaki kiadványokra és a képi dokumentumokra is kiterjedt. 2006-ban pedig bekerült a tervek közé egy Magyar Internet Archívum létrehozása is, de a nyilvános magyar webtérben található kulturális, tudományos, oktatási és közéleti jellegű tartalmak tömeges archiválásnak előkészítése csak 2017 tavaszán kezdődhetett el. A nagyobb tesztaratásokhoz a Kormányzati Informatikai Fejlesztési Ügynökség (KIFÜ) biztosított egy szervert, a szoftvertesztekhez és a demó szolgáltatáshoz pedig egy kisebb teljesítményű OSZK-s szerver áll rendelkezésre. A rendszergazdai és fejlesztői feladatokat két informatikus végzi részmunkaidőben, a könyvtárosi munkákat pedig két főállású, erre a projektre felvett munkatárs látja el.

Eredmények:

Az internetes tartalmak archiválásával foglalkozó külföldi publikációk és jó gyakorlatok megismerése, a pilot projekt két éves időtartamára szóló koncepció és feladatlista összeállítása, a belső és külső kommunikációs csatornák létrehozása után 2017 nyarán elindultak az első tesztek a legtöbb nagy webarchívumnál használt ingyenes Heritrix szoftverrel. Ez a Google robotjához hasonlóan letölti a kiinduláskénti megadott weboldalakat, majd követi a bennük talált linkeket és lementi azokat a weblapokat és egyéb fájlokat is, amelyekre ezek hivatkoznak... Másfél év alatt 8 témában végeztünk 2-3 alkalommal megismételt tesztaratásokat: könyvtár, levéltár, múzeum és galéria, egyetem, kutatóintézet, kormányzat és önkormányzat, irodalom, időszaki kiadvány. Ezek mellett archiváltuk az OSZK saját online tartalmait, továbbá két eseménnyel kapcsolatban is mentettünk weboldalakat meghatározott ideig: a 2018-as téli olimpia és az országgyűlési választások. 2018 szeptemberében pedig 291 ezer, a .hu domén alá bejegyzett címre kiterjedően végeztünk egy nagy tesztet.

Összességében több mint 200 millió fájlt mentettünk le, melyek - duplumszűrés és tömörítés után - közel 10 terabájt (10^{12} bájt) tárhelyet foglalnak el. A webhelyek egy részénél képeket is csináltunk a kezdőlapról (a külalak pontos megőrzése érdekében), leindexeltük és kereshetővé tettük azok teljes szövegét, elvégeztünk valamilyen szintű minőségellenőrzést,

valamint metaadatokat és statisztikákat is készítettünk. Mindezen feladatokhoz a Heritrix mellett kb. egy tucat további, szintén ingyenes szoftvert használtunk. Egyes esetekben engedélyt kértünk a webhelyek tulajdonosaitól az archivált példányok nyilvános szolgáltatására és ezekből összeállítottunk egy kb. 120 címet tartalmazó demó gyűjteményt, amelyet nyilvánosan elérhetővé tettünk a projekt ideiglenes honlapján (mekosztaly.oszk.hu/mia). Itt egy űrlapon keresztül bárki javasolhat is megőrzésre érdemes magyar webhelyeket, melyeknek a címeit felvesszük az archiválandók listájára.

A tesztek mellett nagy hangsúlyt fektettünk a kapcsolatépítésre és az ismeretterjesztésre, mert annyira új és nagy feladatról van szó, hogy a külföldi tapasztalatok megismerése és hazai partnerek bevonása nélkül nem lehet üzemszerűen és hosszú távon is fenntartható módon működtetni egy nemzeti webarchívumot. Ezért az OSZK 2018 januárjában csatlakozott az internet archiválásával foglalkozó intézményeket tömörítő szervezethez, az International Internet Preservation Consortium-hoz, melynek az oktatási munkacsoportjában már aktívan részt veszünk. Személyesen vagy levelezés útján felvettük a kapcsolatot holland, osztrák, szlovák és dán kollégákkal, sőt a dán fejlesztésű SolrWayback kereső és megjelenítő szoftver tesztelésébe is bekapcsolódtuk.

Elindítottunk egy levelezőcsoportot a téma iránt érdeklődő hazai szakemberek részére, melynek már több mint 30 tagja van. „404 Not Found - Ki őrzi meg az internetet?” címmel 2017 és 2018 őszén két workshopot szerveztünk, ahol a mindkétszer kb. 70-80 résztvevőt tájékoztattuk a projekt előrehaladásáról és ahová vendégelőadókat is meghívtunk. Tucatnyi további hazai és külföldi rendezvényen, konferencián tartottunk előadásokat, bemutatókat, és körülbelül ugyanennyi cikket és tanulmányt publikáltunk különböző szaklapokban és egyéb kiadványokban.

Az internetes tartalmak archiválásával kapcsolatos ismeretekről 2017 eleje óta folyamatosan szerkesztünk egy magyar nyelvű wikit, ami már közel 600 szócikket tartalmaz és összeállítottunk egy 450 tételes bibliográfiát a külföldi szakirodalomból. Elkészítettük egy 30 órás továbbképző tanfolyam tervét, melyet 2019 tavaszától fog meghirdetni a Könyvtári Intézet, valamint létrehoztunk egy multimédiás e-learning tananyagot is “Az internet archiválása mint közgyűjteményi feladat” címmel.

Az EMMI Könyvtári és Levéltári Főosztályának kezdeményezésére megszületett egy, a webarchiválást szabályozó törvénytervezet, melynek véleményezésében mi is részt vettünk. Végezetül megírtuk a pilot projekt záródokumentumát, amely a részletes projektbeszámolón, a munkafolyamatokat szabályozó útmutatókon és az eredményeket bemutató mellékleteken kívül tartalmazza egy üzemszerűen működtethető, nemzeti szintű webarchívum megvalósíthatósági tanulmányát is.

Tervek:

Várhatóan még 2018 végén beszerzésre kerül az a nagy teljesítményű hardver infrastruktúra (szerverek és tárolók), amelynek teljes üzembeállítása 2019 nyarára várható. Onnantól kezdve ezen fognak futni a nagy tömegű aratások: a néhány száz vagy néhány ezer webhelyből álló tematikus gyűjtések (kb. 15-20-féle témában, ill. műfajban), jellemzően negyedéves mentési gyakorisággal; a néhány napig/hétig tartó eseményalapú archiválások; valamint a .hu domén alatti magyar webtér “pillanatfelvételszerű”, kis mélységű aratása félévente egyszer. A jelenleg az adminisztrációs, tesztelési és szolgáltatási funkciókat ellátó OSZK-s szerver helyett két virtuális gépet állítunk majd be, hogy szétválaszthassuk ezeket a funkciókat. Az egyiken alakítjuk ki az eddig kipróbált és bevált szoftverekből azt a

keretrendszert, amivel a válogatástól kezdve a mentések ütemezésén és az engedélykéréseken át a minőségellenőrzésig és a metaadatolásig minden munkafolyamatot menedzselni tudunk. Ezen a gépen végezzük majd az egyedi beállításokat vagy a Heritrix helyett más aratószoftvert igénylő speciális webhelyek mentését is. A szolgáltató gép pedig a metaadatokban és a teljes szövegben való keresést, illetve a különféle szempontok szerinti böngészést teszi lehetővé, és a hozzáférési engedélyektől függően jeleníti meg a találatokat vagy a nyilvános interneten, vagy a könyvtári zárt hálózaton, vagy az erre a célra dedikált munkaállomásokon. Ezek a kereső és böngésző funkciók be lesznek építve a webarchívum tervezés alatt levő új honlapjába, ahol az egyszerű felhasználók, a tudományos kutatók, a webhelytulajdonosok, a webarchiválás iránt érdeklődő szakemberek és a média munkatársai egyaránt megtalálnak minden információt, nagyrészt angolul is.

A webarchívum rendszerét össze fogjuk kapcsolni a fejlesztés alatt levő Országos Könyvtári Platform egyes moduljaival (pl. közös keresés az OSZK többi digitális gyűjteményével, metaadatkezelés, névterek, jogkezelés, felhasználói azonosítás, hozzáférés szabályozás, dokumentum hitelesítés, használati- és állomány-statisztika, digitális raktározás, hosszú távú megőrzés). Bizonyos fontos archivált webhelyek (pl. időszaki kiadványok) metaadatait meg kell majd jelentetni olyan adatbázisokban is, mint az OSZK központi katalógusa, az országos MOKKA rendszer, vagy akár a Magyar Nemzeti Bibliográfia. Ehhez a leíró metaadatok MARC21 mezőknek való megfeleltetésére lesz szükség.

Hosszabb távon meg fogjuk oldani, hogy a szoftveres robotokkal való aratások kívül egyéb módon is bekerülhessen tartalom a webarchívumba: csomagban való beadással (pl. technikai korlátok miatt nem aratható, vagy már megszűnt webhelyek esetében), illetve szabványos adatcsere csatornákon való “valós idejű” automatikus beküldéssel (pl. a folyamatosan változó hírportálok esetében). Szeretnénk az élő weben már nem elérhető, de valahol még létező régebbi magyar webtartalmat is integrálni valamilyen módon a magyar archívumba. Ezzel kapcsolatban már kaptunk is ajánlatot az amerikai Internet Archive-tól, ahol több mint 1 milliárd, a .hu domén alól lementett webes dokumentumot őriznek egészen 1996-ig visszamenőleg, tömörítve 44 terabájtnyi összméretben. A környező országok nemzeti könyvtárainak webarchívumaiban található magyar vonatkozású tartalom közös kereshetőségét az ún. memento protokoll segítségével kívánjuk megoldani és mi is biztosítunk majd olyan programozható csatolófelületet (API-t), amin keresztül más rendszerek le tudják kérdezni az OSZK archívumban levő webhelyek metaadatait.

Amennyiben a Nemzeti Audiovizuális Archívum (NAVA) elkezdi az online videó és audió tartalmak gyűjtését, akkor azzal is meg kell majd teremteni a kapcsolatot, mert az OSZK-ban megvalósítandó webarchívum elsősorban a szöveges és képi tartalmakra koncentrál, a nagy méretű médiafájlok nem tartoznak a gyűjtőkörébe azok jelentős tárhely-szükséglete és speciális feldolgozási ill. szolgáltatási igényei miatt. Kiemelten fontosnak tartjuk a hazai közgyűjteményekkel való munkamegosztást olyan formában, hogy ezek “hagyományos” gyűjtőkörüknek megfelelően bekapcsolódhassanak a válogatás, az engedélykérés, a minőségellenőrzés és a metaadatolás élőmunka-igényes feladataiba, s akár saját részgyűjteményeket is kialakíthassanak és szolgáltathassák ezeket a felhasználóiknak. Ehhez egy olyan felhő-alapú rendszert kell kifejlesztenünk, amellyel különösebb informatikai ismeretek nélkül is lehet webes tartalmakat archiválni és kezelni a lementett anyagot. A webarchívum rendszerét továbbá ki fogjuk majd egészíteni olyan funkciókkal is, amelyek segítik a tudományos kutatást, elvégzik az automatikus metaadatolást a lementett fájlok kielemzése alapján, megoldják a metaadatok gazdagítását külső forrásokból szemantikus

kapcsolatokon keresztül, valamint különböző vizualizációs technikákkal áttekinthetővé teszik a nagy tömegű adathalmazokat.

A webarchiválás jogi kereteit megteremtő, előkészítés alatt levő törvény és majd az azt követő, egyes részterületekre vonatkozó miniszteri rendelet fogja meghatározni a webarchívum üzemszerű működtetésének szabályait, az OSZK jogait és kötelességeit, valamint a webhelytulajdonosok számára a bejelentési kötelezettséget. Ahhoz, hogy ezeknek, a várhatóan 2020-tól érvénybe lépő előírásoknak meg tudjunk felelni, belső szabályzatokat kell alkotnunk, meg kell fogalmaznunk a tartalomgazdákkal kötendő szerződés szövegét, egy bejelentő felületet kell létrehoznunk és egy olyan ügyfélszolgálatot kell működtetnünk, amely a beérkező kérdések és kérések mellett abban is segítséget tud adni a webmestereknek, hogy hogyan tudják archívumbaráttá alakítani a webhelyüket, hogy az hosszú távon és jó minőségben megőrizhető legyen.

Mindezen tervek megvalósulása természetesen elsősorban azon múlik, hogy az elkövetkező években rendelkezésre álljon a szükséges emberi, műszaki és anyagi forrás, és hogy a webarchívumot gondozó munkacsoport megtalálja a helyét az OSZK megújuló szervezeti rendszerében.