Az OKR fejlesztés keretében 2018 végéig megvalósítandó OSZK-s webarchiválási pilot projekt koncepciója

2017. április 17.

Készítette: Drótos László (OSZK E-könyvtári Szolgáltatások Osztály)

Felhasznált munkaanyagok és e-mailben kapott tájékoztatások:
Répászky Lipót (NAVA): Online Press Harvest 0.1-es PILOT projektterv (2017. április)
Holl András (MTA-KIK): Tudományos tartalmak webarchiválása (2017. április)
Máray Tamás (KIFÜ-NIIFP): Műszaki specifikáció (2017. április)
Kovács Péter (OSZK): Beszámoló az eddigi tesztekről (2017. április)
Drótos László (OSZK): Webarchiválás munkaterv a 3L projekthez (2016. október)

TARTALOM

Célkitűzés **2**
Feladatok **2**
Ütemezés és munkaerőszükséglet **3**
0. Előkészítő fázis (2017. március - 2017. április) 4
1. Tesztarchívum (2017. március - 2018. december) 4
2. Archiválandó webhelyek adatbázisa (2017. május - 2018. december) 4
3. Tájékozódás, tanulás, kapcsolatépítés (2017. március - 2018. december) 5
4. IIPC csatlakozás (legkésőbb 2017. októberig) 5
5. Gyűjtőkör (2017. május) 5
6. Magyar webtér (legkésőbb 2017. októberig) 5
7. Metaadatok (2017. május - 2018. december) 6
8. Engedélyek (2017. június - 2018. december) 6
9. Jogi keret (2017. május - 2018. december) 6
10. Informatikai infrastruktúra (2017. június - 2018. december) 7
11. Oktatás (2017. június - 2018. december) 7
12. Tájékoztatás (2017. április - 2018. december) 7
13. Fenntarthatóság (2018. szeptember - 2018. december) 8

Célkitűzés

A fejlesztés célja egy leendő magyar internet archívum (webarchívum) koncepciójának, szervezeti kereteinek és informatikai infrastruktúrájának megteremtése, valamint egy teszt-archívum felállítása és ennek segítségével tapasztalatok gyűjtése egy majdani nagy kapacitású, üzemszerűen működő rendszerhez. Végső célként egy olyan rendszert kívánunk majd létrehozni, amely az interneten nyilvánosan megjelenő magyar és magyar vonatkozású kulturális örökség hosszú távú megőrzésének feladata mellett képes kiszolgálni az oktatás, a tudományos kutatás, az állami szervek, az üzleti szféra és az egyes internet használók igényeit is. Az archívum megvalósulásával a most csak jelen időben létező magyar internetnek „múltja" is lesz, és olyan lehetőségek nyílnak meg a mai és a jövőbeli felhasználói számára, amelyek jelenleg nem, vagy csak nehézkesen valósíthatók meg (pl. megszűnt weboldalak megtalálása, régi vagy jelenkori, de feltört ill. véletlenül törölt webhelyek helyreállítása, weboldalak időbeli változásának elemzése és vizualizálása, stabil hivatkozhatóság, idődimenziót is tartalmazó szöveg- és adatbányászati alkalmazások futtatása, internettörténeti kutatások, hiteles másolatok szolgáltatása, igény szerinti webarchiválás megrendelésre).

A rendszert a külföldi legjobb gyakorlatok tapasztalatainak felhasználásával alakítjuk ki, együttműködésben a hazai közgyűjteményi és informatikai szféra egyes szereplőivel, és egyaránt alkalmas lesz nagy tömegű webaratásra, egyes webhelyek vagy weboldalak szelektív lementésére, valamint online tartalmak önkéntes beadására is. Az ehhez szükséges infrastruktúra megteremtése mellett kialakítjuk a webarchiválás és az archivált anyag metaadatolásának módszertanát; megfogalmazzuk az archívum és az eredeti tartalomgazdák, valamint az archívum és annak felhasználói közötti jogviszonyt szabályozó szerződésmintákat; továbbá ajánlások, tanulmányok, előadások és tanfolyamok segítségével ismertetjük meg az érintett szereplőkkel a webarchiválással és a webarchívum-használattal kapcsolatos ismereteket.

Feladatok

A 2017 márciusától 2018 végéig terjedő pilot fázisban a webarchiválás, mint szakterület megtanulása, a majdani nagy kapacitású, üzemszerűen működő archívum létrehozásához szükséges kutatás-fejlesztési munka elvégzése tűzhető ki reális célként.

1. Az elméleti ismeretek megszerzése mellett gyakorlati tapasztalatokra is szükség van, ezért felállítunk egy kezdetben néhány száz (főként kulturális és tudományos) webhely, valamint az EPA-ban csak távoli forrásként nyilvántartott kb. 1700 elektronikus periodika néhányszori mentéséből álló **tesztarchívum**ot, továbbá megkísérreljük a .hu domain alatti webszerverek kellően nagy halmazának learatását egyszer vagy kétszer (egy ekkora aratás több hónapig fut a külföldi tapasztalatok szerint). Komolyabb nyilvános szolgáltatás még nem lesz ezekből a tesztekből, egyrészt a még tisztázandó szerzői jogi és személyes adatvédelmi kérdések miatt, másrészt mert még sok hiba, hiány lesz bennük, amelyeket a későbbiekben az aratógép beállításainak finomhangolásával, és a szelektív mentések utólagos automatikus és/vagy emberi ellenőrzésével, javításával lehet majd csökkenteni. Mivel ezeknek a mentéseknek 2018 végéig még alig lesz időbeli mélységük, különösebb értéket sem képviselnek egyelőre. De egy kisebb demót mielőbb kiteszünk az internetre az érdeklődők számára (pl. könyvtári honlapok mentéseit, melyekhez könnyű engedélyt kapni).

2. A tesztarchívum létrehozása mellett elkezdjük a szelektíven (is) **archiválandó webhelyek** vagy webhelyrészek kellően reprezentatív méretű halmazát feltérképezni. Ez majd végül egy tízezres nagyságrendű URL-lista lesz, kiegészítve néhány további metaadattal (pl. a webhely neve, fenntartója, műfaja, témaköre, mélysége, változékonysága, veszélyeztetettsége, arathatósága, jogi előírásai, a tartalomgazda elérhetősége). Ez az adatbázis lesz a majdani üzemszerűen működő archívum gyarapítási forrása, ez alapján indulnak majd - optimális gyakorisággal és konfigurációs beállításokkal - az egyes webhelyek mentései. A címlista bővítésébe és karbantartásába a projekt iránt érdeklődő partnerintézmények munkatársai is bekapcsolódhatnak, sőt a magyar internetes társadalom is (felhasználók és szolgáltatók egyaránt), akik szintén javasolhatnak majd megőrzésre érdemes tartalmakat.

Ezen gyakorlati feladatok mellett el kell végezni vagy legalább el kell kezdeni 2018 végéig különböző háttérmunkákat is, mint például:

3. jelenlegi külföldi **jó gyakorlatok** megismerése, és a téma utóbbi évekbeli szakirodalmának feldolgozása;
4. csatlakozás a webarchívumok **nemzetközi szervezet**éhez;
5. a magyar archívum **gyűjtőkör**ének meghatározása;
6. a **magyar webtér** lehatárolása;
7. a **metaadat**struktúra kialakítása a szelektív ill. a tömeges mentések esetében, és a szelektív mentések folyamatos metaadatolása;
8. a tartalomszolgáltatókkal kötendő **szerződés**nek és az engedélyt kérő e-mail szövegének megfogalmazása;
9. az archiválási tevékenység és az archívumhoz való hozzáférés **jogi keret**ének megteremtése;
10. a pilot fázis tapasztalatai alapján az üzemszerű működésre alkalmas, jól skálázható **informatikai infrastruktúra** megtervezése;
11. a webarchiválással, a webarchívumok működtetésével, használatával kapcsolatos ismeretek **oktatás**a az OSZK és más könyvtárak, közgyűjtemények érintett/érdeklődő munkatársai számára;
12. a magyar online tartalomszolgáltatók és a sajtó folyamatos **tájékoztatás**a a tervekről és a projekt előrehaladásáról;
13. a pilot fázis utáni **fenntarthatóság** megtervezése.

Ütemezés és munkaerőszükséglet

Az alábbiakban a fenti „Koncepció” fejezetben levő pontok sorrendjében próbáltam megbecsülni az egyes részfeladatok idő- és emberszükségletét. A becslés úgy készült, hogy nem áll még rendelkezésemre a webarchiválásra 2017-2018-ban fordítható költségkeret nagysága, sem pedig az, hogy ebből majd milyen jellegű kifizetésekre lesz lehetőség (pl. szakértői díjak, külső céggel való programfejlesztés, konferencia-részvétel, tanulmányi utak)? Nem tisztázottak továbbá még az egyes szereplők feladatai a webarchiválási projekt irányításában és a döntési jogköreik.

0. Előkészítő fázis (2017. március - 2017. április)

Feladatok: A koncepció elkészítése, egyeztetése és elfogadása, lehetséges partnerek megkeresése, munkaerő felvétele, munkacsoport megalakítása, a kommunikáció és a projektdokumentálás módjának egyeztetése, a munkakörnyezet megteremtése.

Érintett személyek: EMMI munkatársak, KIFÜ-NIIF projektfelelős, OSZK-s projektirányítók és témafelelősök, osztályvezetők, OSZK-s belső munkatársak, külső partnerintézmények képviselői.

1. Tesztarchívum (2017. március - 2018. december)

Feladatok:

1.1 (2017. március - 2017. május): Az aratást végző szerver terhelhetőségének tesztelése, az aratószoftver beállításainak próbálgatása, a begyűjtött tartalom méretének és összetételének elemzése, a naplófájlok és a hibaüzenetek átnézése. Az ebben a fázisban begyűjtött (sokszor töredékes) anyag nagy része még nem kerül megőrzésre.

1.2 (2017. június - 2018. december): Elektronikus periodikák és kulturális, tudományos, oktatási intézményi webhelyek megőrzési célú szelektív mentése, 1-3 hónapos ismétlődéssel.

1.3 (2017. május - 2018. december): A tesztarchívumhoz szükséges néhány száz honlap összeválogatása és engedélykérés a tartalomgazdáktól.

1.4 (2017. augusztus - 2018. december): A szelektíven archivált webhelyek metaadatolása.

1.5 (2017. október): A tesztarchívum bemutatása a szakmai köröknek.

1.6 (2017. november): Egy kisebb nyilvános demóállomány megjelentetése az OSZK honlapján, egyszerű keresési funkciókkal és magyarázattal. (Ez folyamatosan bővülhet a projekt során olyan webhelyekkel, amelyeknél a tartalomgazda erre engedélyt adott.)

1.7 (2017. november és 2018. április): Egy-egy .hu domain-szintű webaratás elindítása.

1.8 (2018. szeptember) A tesztarchívum metadat- és teljes szövegű keresőjének nyilvános szolgáltatása. Maguk az archivált weboldalak viszont csak az OSZK-n belül, illetve OSZK- és NAVA-pontokon, dedikált gépeken nézhetők meg.

Érintett személyek: informatikus, rendszergazda, programozó, témafelelős, webkönyvtáros, EPA-könyvtáros, adminisztrátor.

2. Archiválandó webhelyek adatbázisa (2017. május - 2018. december)

Feladatok:

2.1 (2017. május - 2017. június): A megőrzésre érdemes magyar webhelyek/webhelyrészek adatbázisának felállítása.

2.2 (2017. augusztus - 2017. szeptember): Munkamegosztás és munkafolyamatok egyeztetése a partnerintézményekkel.

2.3 (2017. október): Crowdsourcing elindítása, az internetes közösség bevonása a megőrzésre érdemes magyar tartalom kiválogatásába, egy űrlapos felületen keresztül.

2.4 (2017. augusztus - 2018. december): Az adatbázis feltöltése és karbantartása.

Érintett személyek: informatikus, programozó, webkönyvtáros, elektronikus forrásokat gyarapító munkatársak, tájékoztató könyvtárosok, adminisztrátor, külső partnerek munkatársai.

3. Tájékozódás, tanulás, kapcsolatépítés (2017. március - 2018. december)

Feladatok:

3.1 (2017. március - 2018. december): Egy wiki építése az internetes tartalmak megőrzésével kapcsolatos fogalmakról, formátumokról, szoftverekről, szolgáltatásokról, a webarchiváló projektekről, szervezetekről, rendezvényekről és fórumokról. Ez a tudásbázis a szakterület magyarországi megismertetését és a könyvtár-informatikai képzésbe való beépülését is segíti majd.

3.2 (2017. május - 2017. július): A webarchiválás aktuális helyzetének áttekintésére alkalmas szakirodalmi szemle összeállítása az elmúlt években megjelent külföldi publikációkból.

3.3 (2017. június - 2018. december): Kapcsolatok felvétele néhány európai projekt szakembereivel (pl. a cseh, az osztrák, az angol, a német, a holland és a finn nemzeti webarchívumokkal), valamint az amerikai Internet Archive-val, ahol 1996-ig visszamenőleg igen jelentős magyar webtartalom van és aminek a megvételéről tárgyalni kellene velük. (Ahogy pl. a britek tették 2010-ben és 2013-ban.)

3.4 (2017. június - 2018. december): Részvétel a témában rendezett nemzetközi konferenciákon és workshopokon.

Érintett személyek: EMMI munkatársak, projektirányítók, témafelelős, webkönyvtáros, külső szakértő.

4. IIPC csatlakozás (legkésőbb 2017. októberig)

Feladatok: Az OSZK tagfelvételi kérelmének beadása a webarchívumokat összefogó, az együttműködéseket és a szoftverfejlesztéseket koordináló International Internet Preservation Consortium-hoz, a http://www.netpreserve.org/join-iipc oldalon szereplő feltételek és az oldal alján letölthető konzorciumi szerződés elfogadásával. A szerződés 2020-ig szól, ezért jó lenne az éves tagdíj fizetését legalább addig biztosítani.

Érintett személyek: EMMI munkatársak, OSZK főigazgató, OSZK-s jogász.

5. Gyűjtőkör (2017. május)

Feladatok: Az archívumban szelektíven gyűjtendő webhelyek, webhelyrészek, illetve azokon belüli digitális objektumok halmazának lehatárolása és prioritások meghatározása. Szempont lehet azok tartalma, műfaja, létrehozója, célja, formátuma, értéke, változékonysága, veszélyeztetettsége, lementhetősége, mérete, mélysége, jogi helyzete stb.

Érintett személyek: gyűjteményszervezési igazgató, kötelespéldány osztály vezetője, témafelelős, webkönyvtáros.

6. Magyar webtér (legkésőbb 2017. októberig)

Feladatok:

6.1 (legkésőbb 2017. szeptemberig): Dönteni kell, hogy a teljes körű aratás esetén meddig terjedjen ki az aratógép által bejárt webtér (pl. csak a .hu alatti szerverek; vagy a magyar állampolgárok ill. a magyar tulajdonú és/vagy Magyarországon működő intézmények/szervezetek/cégek által bejegyzett egyéb felső szintű domain-ek (pl. .com, .ro,) alattiak is; vagy minden olyan webhely, ahol magyar nyelvű tartalom van domaintól függetlenül; vagy minden, ahol magyar vonatkozású tartalom van domaintól és nyelvtől függetlenül).

6.2 (2017. októberig): Egy rövid kiinduló címlista (ún. *seed list*) összeállítása, amely URL címekről elindulva az aratógép be tudja járni a 6.1-ben meghatározott magyar webteret. Célszerű a fontosabb magyar linkgyűjteményekből kiindulni (pl. http://lap.hu/abc/).

6.3 (2017. október - 2018. december): A magyar webteret alkotó webhelyek URL cím-adatbázisának építése folyamatosan, részben az aratógép által felderített címekből, részben a magyar DNS szolgáltatóktól (domain-name regisztrátoroktól), kormányzati szervektől (pl. NMHH: http://webext.nmhh.hu/sajto_kozzetetel/), esetleg a Google Magyarországtól elkért nyilvántartásokból. Ez egy milliós nagyságrendű adatbázis lesz idővel, így csak automatikus módszerekkel bővíthető és tartható karban.

Érintett személyek: EMMI munkatársak, KIFÜ-NIIF projektfelelős, gyűjteményszervezési igazgató, OSZK-s jogász, témafelelős, webkönyvtáros, informatikus.

7. Metaadatok (2017. május - 2018. december)

Feladatok:

7.1 (2017. május - 2017. június): Dönteni kell arról, hogy milyen leíró, adminisztrációs és technikai metaadatokat akarunk hozzárendelni a szelektíven mentett és a teljes körű aratással begyűjtött webhelyekhez (előbbinél Dublin Core leíró adatokat célszerű használni, utóbbinál pedig csak automatikus megoldások jöhetnek szóba).

7.2 (2017. június - 2017. július): Egy adatbázis létrehozása a szelektíven mentett webhelyek metaadatolása számára, egyelőre egy egyszerű kezelőfelülettel.

7.3 (2017. augusztus - 2018. december): A tesztarchívumba szelektíven és már megőrzésre érdemes minőségben lementett webhelyek metaadatolása folyamatosan.

Érintett személyek: témafelelős, webkönyvtáros, informatikus, programozó, adminisztrátor, külső partnerek munkatársai.

8. Engedélyek (2017. június - 2018. december)

Feladatok:

8.1 (2017. május - 2017. június): Meg kell fogalmazni azt a szerződésmintát és kísérőlevelet, amelyet a szelektíven archiválandó webhelyek tartalomgazdáinak az OSZK kiküld, és amelyben engedélyt kér az archiválásra, valamint nyilatkozatot arról, hogy az archív verziókat szolgáltathatja-e nyilvánosan, vagy csak valamilyen késleltetéssel, vagy csak dedikált gépekről zárt hálózaton, vagy egyáltalán nem szolgáltathatja, vagy nem is archiválhatja, vagy csak a tartalomgazda által megadott webhelyrészt archiválhatja.

8.2 (2017. június - 2018. december): Az engedélykérő levelek és szerződésmelléklet elküldése az archiválásra kiválogatott webes tartalmak gazdáinak, a válaszok iktatása, a kérdések megválaszolása, a kapott engedélyben megadott feltételek rögzítése a metaadatok között.

Érintett személyek: OSZK jogász, internetes jogban jártas külső szakértő, webkönyvtáros, adminisztrátor.

9. Jogi keret (2017. május - 2018. december)

Feladatok: A kötelespéldány törvény és/vagy más jogszabályok kiegészítése, valamint OSZK-s belső szabályzatok elkészítése, amelyek meghatározzák a nemzeti könyvtár jogait és kötelességeit a webaratással kapcsolatban (pl. milyen webtérre terjedhet ki, milyen engedélyeket kell beszereznie, megkerülheti-e a robots.txt-ben megadott tiltásokat), valamint szabályozzák a webarchívumhoz való hozzáférést, a ráépülő szolgáltatások használatát (pl. a

metaadatok nyilvánosak, de a teljes körű aratással gyűjtött tartalom csak helyben kutatható és még ott sem másolható, a szelektíven mentett webhelyekhez való hozzáférés pedig a tartalomgazda által megadott feltételekkel történhet; az eredeti tartalomgazda vagy az erre általa feljogosított jogi személy bármikor kivetetheti az anyagát a nyilvános szolgáltatásból vagy akár az archívumból/metaadatbázisból is („opt-out" ill. „notice and take down" lehetőség); az erre feljogosított hatóságok bármilyen mentett tartalomhoz hozzáférést igényelhetnek vagy hiteles másolatot kérhetnek).

Érintett személyek: EMMI munkatársak, OSZK jogász, internetes jogban jártas külső szakértő.

10. Informatikai infrastruktúra (2017. június - 2018. december)

Feladatok:

10.1 (2017. június): A jelenlegi tesztelést végző aratószerver mellé kell egy kisebb teljesítményű szerver a keretrendszer adatbázisának és webes felületének szolgáltatására. (Ez a rendszer a szelektíven mentendő webhelyek metaadatainak, a tartalomgazdáktól kapott engedélyek, valamint a munkafolyamatok nyilvántartására szolgál, és távolról is hozzáférést kell adni a partnerintézmények bedolgozói számára.)

10.2 (2017. szeptember): A kb. féléves próbaidőszak tapasztalatai után felül kell vizsgálni a tesztszerver és tárhely paramétereit, eldönteni, hogy elegendő lesz-e a 2017. novemberében és 2018. áprilisában indítandó teljes körű aratáshoz, és szükség esetén bővíteni. És valószínűleg kell majd egy külön gép a szelektíven archivált webhelyek anyagának indexeléséhez a teljes szövegű kereső számára.

10.3 (2017. október): Az internetre nyilvánosan kitehető demóarchívum szolgáltatásához kell egy kisebb teljesítményű webszerver.

10.4 (2018. november - 2018. december): A pilot projekt lezárásakor a tesztaratások és a tesztarchívum gyakorlati tapasztalatait felhasználva meg kell tervezni az üzemszerű működéshez, a nagyméretű archiváláshoz, a visszakereséséhez, az archívumra épülő szolgáltatásokhoz, valamint a lementett digitális tartalom hosszú távú és biztonságos megőrzéséhez szükséges informatikai infrastruktúrát.

Érintett személyek: KIFÜ-NIIF projektfelelős, informatikus, rendszergazda.

11. Oktatás (2017. június - 2018. december)

Feladatok: Ismertetők és oktatóanyagok készítése, szakmai napok szervezése, konferenciaelőadások tartása, belső és külső tanfolyamok indítása, javaslat tétel a könyvtár-informatikus képzés tantervének kibővítésére, hogy minél több kolléga tudjon hozzájárulni a nemzeti webarchívum építési munkálataihoz, illetve legyen képes internetes forrásokat archiválni a saját intézménye ill. annak használói számára.

Érintett személyek: témafelelős, webkönyvtáros, a Könyvtári Intézet képzéssel foglalkozó munkatársai.

12. Tájékoztatás (2017. április - 2018. december)

Feladatok: A projekt céljáról és előrehaladásáról folyamatosan sajtóanyagok és a nagyközönség számára is érthető tájékoztatók írása, ezek terjesztése az OSZK honlapon és egyéb kommunikációs csatornákon. A Magyarországi Tartalomszolgáltatók Egyesületének és az Internet Szolgáltatók Tanácsának közvetlen tájékoztatása a webarchiválás elindulásáról.

Érintett személyek: projektirányítók, témafelelős, webkönyvtáros, OSZK Sajtóiroda, Webtartalom Koordinációs Osztály

13. Fenntarthatóság (2018. szeptember - 2018. december)

Feladatok: A pilot fázis vége felé ki kell dolgozni a webarchiválás és a webarchívum hosszú távú fenntarthatósági tervét, és elkezdeni dolgozni a szükséges pénzügyi, technikai, élő-munka, jogi stb. feltételek megteremtésén.

Érintett személyek: EMMI munkatársak, KIFÜ-NIIF projektfelelős, OSZK főigazgató, gazdasági igazgató, projektirányítók, témafelelős, webkönyvtáros, jogász, külső szakértők.