

WEBARCHIVÁLÁS

CÉLKITŰZÉS

A fejlesztés célja egy magyar internet archívum (webarchívum) koncepciójának, szervezeti kereteinek és informatikai infrastruktúrájának létrehozása, valamint egy bevezető tesztfázis után az üzemszerű működés feltételeinek megteremtése: a magyar és magyar vonatkozású internetes tartalmak folyamatos archiválásának elindítása. Egy olyan rendszert kívánunk létrehozni, amely a 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, webhelyek időbeli változásának elemzése és vizualizálása, stabil hivatkozhatóság, idődimenziót is tartalmazó szöveg- és adatbányászati alkalmazások futtatása, internettörténeti kutatások, hiteles másolatok szolgáltatása).

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 webarátásra, egyes webhelyek vagy weboldalak szelektív lementésére, valamint online tartalmak önkéntes beadására is. Az ehhez szükséges infrastruktúra megteremtése mellett kialakítjuk a rendszeres, nagy tömegű 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.

INDOKLÁS

Írott és egyéb módon rögzített kultúrájának megőrzése minden nemzet kötelessége. A tárgyiasult formában, vagyis valamilyen hordozón nyilvánossá tett dokumentumok szervezett gyűjtésének, rendszerezésének, őrzésének és szolgáltatásának feladatát nagyrészt a közgyűjtemények: könyvtárak, levéltárak és múzeumok látják el, évszázadok alatt kialakított módszertannal. Az 1990-es évek elejétől robbanásszerűen növekvő interneten megjelenő tartalomnak viszont nincs hazánkban őrzőhelye és nincs kialakult megőrzési metódusa, pedig ennek egyre nagyobb hányada már semmilyen más formában, semmilyen hordozón nem jelenik meg, így ha valamilyen okból elérhetetlenné válik az interneten, akkor lehet, hogy örökre elveszett a jövő számára. A weboldalak átlag élettartama néhány hét, és persze vannak olyanok is, amelyeken percenként változik a tartalom. Még a stabilnak gondolt site-ok is eltűnnek vagy teljesen átalakulnak néhány évenként, a mögöttük álló intézmények átszervezése vagy megszűnése, illetve a technológiaváltás miatt. Így veszítettük el a digitális kultúránk már több mint 20 évét, köztük olyan webhelyeket is, amelyekre készítők annak idején jelentős mennyiségű (köz)pénzt és szellemi munkát fordítottak. De nemcsak az elvesztegetett források és eltűnt értékek miatt problémás az online világ jelenidejűsége, hanem például azért is, mert a gyorsan változó digitális dokumentumok használata nem tud rendesen beépülni az oktatásba és a tudományos tevékenységbe, mivel a rájuk hivatkozó linkek előbb-utóbb „eltörnek”, az oldalak elérhetetlenné válnak vagy egészen más tartalom kerül rájuk.

A csak digitális formában létező (born digital) kultúra kiemelt veszélyeztetettségére az UNESCO Közgyűlése is felhívta a figyelmet 2003-ban a „Charta a digitális örökség védelméről” című dokumentumban. Ez a dokumentum meghatározza a felelősöket is: „A digitális örökség megőrzése a kormányok, alkotók, kiadók, releváns iparágak és az örökségvédelmi intézmények kitartó erőfeszítéseit igényli.” A nemzeti könyvtáraknak természetesen kiemelt felelősségük van ebben, és a világ számos országában üzemszerűen vagy legalább kísérleti jelleggel, határozott idejű projektek formájában működnek már ilyen webarchívumok – vagy a nemzeti könyvtár vezetésével, vagy legalább részvételével egy nagyobb konzorcium keretén belül. Mivel a nemzeti könyvtár évszázadokra létrehozott közintézmény és az ott dolgozó könyvtárosoknak rengeteg ismeretük és nagy gyakorlatuk van a megőrzendő kulturális értékek kiválasztásában, leírásában, gyűjteménnyé szervezésében és a rájuk épülő szolgáltatások kialakításában, ezért – a technikai feltételek megteremtése és folyamatos fejlesztése esetén – az OSZK biztosítani tudja a begyűjtött digitális tartalom hosszú távú megőrzését és annak közcélú hasznosulását. A kötelespéldány törvény értelmezésével és kiegészítésével megteremthetők a webarchiválás jogi feltételei; az egyéb digitális anyagok (pl. saját digitalizált állomány, kiadók által beküldött e-könyvek) könyvtári szolgáltatására kialakított rendszeren keresztül (pl. védett privát hálózat és olvasótermi gépek, a NAVA-pontokhoz hasonló OSZK-pontok) pedig biztosítható a webarchívum nem nyilvános részéhez való hozzáférhetőség is – ellenőrzött körülmények között – az arra jogosultak számára.

Az OSZK stratégiai tervében évek óta szerepel ez a feladat, és már évtizedes gyakorlatunk van egyes dokumentumtípusok (digitális könyvek, elektronikus periodikák, digitális képek) egyedi archiválásában. 2006-ban egy koncepció és egy kis tesztállomány is készült Magyar Internet Archivum (MIA) néven, ugyanez év végén pedig az amerikai Internet Archive (IA), a világ legnagyobb webarchiválással foglalkozó projektje számára összeállítottunk egy 300 tételes listát fontos magyar kulturális honlapokról, melyeket azután az IA robotja be is gyűjtött. Azóta több próbálkozás is volt egy saját webarchiváló projekt indítására, pályázatok készültek és tesztaratók is történtek, de források hiányában az érdemi munka nem indulhatott el.

KONCEPCIÓ

A digitális kulturális örökség hosszú távú megőrzése és a ráépülő funkciók csak saját menedzselésű rendszerrel valósíthatók meg, így az olyan felhőalapú, külső szolgáltatások, mint például az IA által kínált Archive-It (archive-it.org), egy nemzeti webarchívum esetében nem jöhetnek szóba. De ez természetesen nem jelenti azt, hogy az archívum minden része csak és kizárólag az OSZK-ban legyen fizikailag, sőt kívánatos az elosztott működés és tárolás a partnerintézmények között.

Mivel belátható időn belül nem lesz más jelentősebb internet archívum Magyarországon, ezért egy olyan rendszert és gyűjteményt kell létrehozni, amely a könyvtár gyűjtési és feldolgozási szempontjai mellett valamilyen szinten képes más területek (pl. a levéltári, a kormányzati és az üzleti szféra) igényeit is kielégíteni. (Ez azért is hasznos, mert így megnő a fenntarthatóság esélye.) Vagyis a nemzeti bibliográfia gyűjtőkörét a webarchívum esetében ki kellene terjeszteni például a személyes és kormányzati dokumentumokra, vagy a különféle iparági információforrásokra is. És egy olyan rendszert kell létrehozni, amely külső intézmények (pl. más közgyűjtemények, állami szervek, tartalomszolgáltatók) által kezdeményezett vagy megrendelt webarchiválási munkákat is el tud végezni.

A fent említett okból a világban jelenleg már létező különféle internetes archívum-típusokból egy olyan kombinált rendszert kell kialakítani, amely egyaránt képes nagy tömegű aratásra

automatikus technológiákkal, de szelektív, félautomata módszerrel is lehet webhelyeket vagy azok egyes részeit lementeni vele, továbbá be tud fogadni önkéntes beadással (pl. valamilyen „push” technológiával, vagy csomagban való feltöltéssel) érkező webanyagokat is. (Utóbbi például a digitális hagyatékok beküldése, vagy az élő weben már nem létező, de még valahol megőrzött régi honlapok eljuttatása miatt érdekes, de elképzelhető az is, hogy egyes nagy hírportálok és egyéb gyorsan változó webhelyek így küldjék be a friss tartalmakat az archívumba.)

Az archívum tehát különböző minőségű, rendezettségű és feldolgozottságú részekből fog állni: egy – a megcélzott gyűjtőkörön belül – viszonylagos teljességre törekvő, automatikusan gyűjtött és metaadatolt, csak nagyobb időszakonként (pl. félévente) újraaratott halmazból, és egy válogatott, optimális időszakonként újramentett, emberi munkaerővel (is) feldolgozott halmazból, amely önmagában is több részgyűjteményből áll majd (pl. szektor, téma, esemény, műfaj, fájltypus stb. szerint szegmentált részekből, illetve egyes projektek vagy személyi hagyatékok anyagából).

A változatos és gyorsan változó online tartalomszolgáltató technológiák és formátumok miatt a rendszer archiváló komponensének rugalmasnak és folyamatosan továbbfejleszthetőnek kell lennie. Az alapvető cél a digitális tartalom megőrzése, de természetesen az ésszerűség határáig törekedni kell a megjelenés (pl. külalak, kívülről beágyazott elemek) és a használhatóság (pl. interaktív funkciók, belső linkek követhetősége) megtartására is. Ha másként nem megoldható, a külalak megőrzésére elfogadható megoldás lehet képernyőfotók automatikus készítése és elhelyezése a metaadatok közt. Ha a tartalom kinyerése is technikai nehézséget jelent (pl. CMS rendszerekből dinamikusan generált webhelyek esetében), akkor meg kell állapodni a szolgáltatóval, hogy vagy generáljon egy „robot-barát” verziót is (az akadálymentes felületekhez hasonlóan), vagy „push” technológiával ő maga küldje be a tartalmat az archívumba.

Az internet archívum elsődleges célja a szövegek és a képanyagok megőrzése, de – bizonyos formátum és terjedelmi korlátok mellett – begyűjti és tárolja a weboldalakba beágyazott animált, videó és hangzó dokumentumokat is. (A nagyméretű, illetve a sugárzott digitális hang- és videóanyagok archiválása érdekében munkamegosztást kell majd kialakítani a VIDEOTORIUM-mal és a NAVA-val.)

Az egyes webhelyek vagy weboldalak ismételt lementésének optimális gyakoriságát azok fontossága és tartalmuk változásának üteme határozza meg. Az archívum természetesen minden korábbi állapotot megőriz, de a tárterülettel való takarékoság érdekében mindig csak az előző állapothoz képest történt változásokat menti el, időbélyeggel és automatikus metaadatokkal ellátott tömörített, szabványos WARC (Web ARChive File Format) csomagok formájában.

A hosszú távú megőrzéshez, illetve az archívumra épülő egyes szolgáltatásokhoz szükség lesz bizonyos fájlformátumok nyílt formátumokra konvertálására (vagy legalább a nyers szöveg kimentésére); egyedi azonosítók (URI-k) generálására nemcsak az egyes weboldalakhoz, hanem akár azok egyes elemeihez (pl. fejezetek, ábrák, táblázatok) is, hogy ezek önállóan is hivatkozhatók, linkelhetők legyenek (pl. publikációkból vagy e-learning rendszerekből); valamint a lementett anyag hitelesítésére (digitális aláírás, tanúsítvány, ujjlenyomat). Ezek az „értéknövelő” megoldások természetesen inkább csak a válogatott és egyedileg mentett, illetve az archívumba beküldött anyagoknál jöhetnek szóba, mint ahogy a leíró metaadatok manuális előállítására is.

Az archívumba bekerülő digitális objektumok mindegyikéről készülnek automatikusan előállítható metaadatok (pl. eredeti URL, nyelv, formátum, dátum, kódolás, méret), melyek egy adatbázisba kerülnek, de az objektum mellett is letárolódnak. Mivel bizonyos

fájlformátumok (pl. HTML, PDF, DOCX, JPEG, MP3) önmagukban is tartalmazhatnak metaadatokat, célszerű ezeket is kinyerni (majd esetleg egy nyelvi elemzőn átfuttatni, hogy értelmes szövegek-e?) és szintén eltárolni őket. Szöveges állományok esetében az automatikus tárgyszavazás és kivonatkészítés is szóba jöhet, ilyen célra fejlesztett nyelvi elemző szoftver segítségével. Továbbá amiről lehet (pl. weboldalak, képfájlok, PDF dokumentumok első oldala) egy képernyőfotót vagy nézőképet is kell készíteni, amelyen látszik, hogy a mentés idején használatos megjelenítő programmal (pl. böngészővel) hogyan nézett ki és amely az archívumban való kereséskor a találati listában is megjelenik majd.

Az archiválásra célzottan kiválogatott webanyagok esetében az automatikus metaadatok kiegészíthetők még emberi tudással előállított vagy javított adatokkal (pl. alkotó, tulajdonos, cím, téma, műfaj, jogi státusz). Az archiválandó tartalom hatalmas mennyiségére és változatosságára való tekintettel az anyag jellege és a várható felhasználási célok határozzák meg hogy mit, milyen mélységig és milyen részletességgel érdemes leírni (pl. elégséges lehet csak magát a webhelyet, vagy annak főbb alrészeit feldolgozni, és a legtöbb esetben elegendő egy minősített Dublin Core-szintű leírás). A metaadatolásba a teljes könyvtári hálózathoz bevonhatók erre kiképzett könyvtárosok (praktikusan a saját gyűjtőkörük – pl. helyismeret vagy adott szakterület – alá eső anyagok leírásába), valamint az internet felhasználók széles köre is (crowdsourcing). A kellő részletességgel leírt archiválási egységek rekordjai bekerülhetnek a Magyar Nemzeti Bibliográfiába, az OSZK és más könyvtárak katalógusaiba, valamint a közös MOKKA-ba; a teljes archívum, vagy annak valamely részhalmaza pedig külön keresőkkel kutatható, melyekkel az automatikusan kinyert metaadatokra is lehet keresni vagy szűkíteni.

Az archívum a metaadatkereső(k) mellett egyéb szolgáltatásokat is biztosít majd: többféle szempont (pl. fájl típus, nyelv, időszak, URL-részlet) szerint szűrhető teljes szövegű keresőt, képkeresőt, adott weboldal eddigi mentéseit és időbeli változását mutató megjelenítőt (wayback machine), képernyőfotókból összeállított látványos vizuális böngészőfelületet, statisztikai elemzőket, valamint API-kat külső fejlesztésű alkalmazások számára. Továbbá helyreállító szolgáltatást is tud nyújtani feltört vagy véletlen adatvesztés miatt tönkrement webhelyek számára.

Az archívum egyes részeihez különféle hozzáférési jogosultságok állíthatók be, amelyekkel szabályozható, hogy ki, honnan, mikortól és milyen szintig fér hozzá az adott archivált anyaghoz és milyen műveleteket (pl. megtekintés, nyomtatás, saját hordozóra mentés) végezhet rajta. Azt, hogy az eredeti tartalomgazdák egyedi engedélye nélkül az OSZK milyen szintű hozzáférést biztosíthat házon ill. a magyar könyvtári hálózaton belül, és milyen azon kívül, a hazai jogszabályok és a külföldi gyakorlatok tanulmányozása után külön szabályzatban vagy jogszabályban kell rögzíteni. Ezen általános előírásokat felülírhatják a tartalomgazdákkal kötött szerződések, amelyek szűkíthetik vagy tágíthatják a hozzáférési jogokat, illetve lobbizni kell azért, hogy a jövőben olyan előírások és pályázatkiírások szülessenek, amelyek a közpénzek felhasználásával működő webhelyek számára kötelezően előírják az archívumba való bekerülést és esetleg egy türelmi idő (pl. fél év) után az archivált anyaghoz való nyilvános hozzáférést.

PROJEKTTERV

Mivel a web és az egyéb internetes tartalomterjesztő technológiák komplexitása, változékonysága és mérete rendkívüli kihívást jelent gyarapítási, gyűjteményszervezési és informatikai szempontból, ezért egy nagy méretű internet archívum létrehozása és fenntartása meghaladja egyetlen intézmény kereteit, elkerülhetetlenül szükséges más szereplők: közgyűjtemények, internet- és tartalomszolgáltatók, informatikai fejlesztésekkel foglalkozó

intézetek és vállalatok, valamint az érintett kormányzati szervek bevonása is, hosszabb-rövidebb időre szóló együttműködések formájában. A feladat nagysága miatt pedig az egyetlen járható út egy pragmatikus, lépcsőzetes építkezés, az eredmények és tapasztalatok folyamatos visszacsatolása a következő fázis célkitűzéseinek meghatározásához.

Tervezési fázis:

A projektnek ebben a fázisában össze kell gyűjteni és fel kell dolgozni az internetarchiválással kapcsolatos szakirodalmat, megismerni a legjobb külföldi gyakorlatokat, megvizsgálni a más szervezetek által már archivált magyar webanyagok átvételének lehetőségét, intézményileg csatlakozni az International Internet Preservation Consortiumhoz (netpreserve.org), kialakítani a hazai érintett szervezetek közötti együttműködés keretét és munkacsoportokba delegált szakemberek részvételével kidolgozni az archívum részletes terveit (műszaki terv, pénzügyi terv, munkafolyamatok, szabványok stb.) A következő 2-5 fős munkacsoportokra van szükség:

1. A válogatással és lehatárolással foglalkozó munkacsoport

Feladata: Megvizsgálni az archiválandó anyag kiválasztásának illetve lehatárolásának szempontjait, mind az egyedi, mind a generális begyűjtés céljára. Az egyedi site-ok mentéséhez meg kell határozni a válogatás kiindulópontjait (pl. link-gyűjtemények), a válogatás műfaji, tartalmi, minőségi szempontjait (pl. magyar intézmények vagy személyek által fenntartott, vagy magyar tartalommal rendelkező, kulturális, tudományos, közéleti stb. web-szolgáltatások). A generális aratáshoz meg kell határozni a magyar web-tér kiterjedését (a .hu domain és a rajta kívül eső, magyar tartalmat is szolgáltató szerverek), és létre kell hozni az együttműködést a domain-szolgáltatókkal a magyar web-térbe tartozó szerverek naprakész nyilvántartása céljából. Meg kell határozni továbbá, hogy a begyűjtés milyen mélységben és milyen típusú objektumokra terjedjen ki.

2. A begyűjtés és tárolás technikai kérdéseivel foglalkozó munkacsoport

Feladata: Áttekinteni az egyedi site-ok mentésének, valamint a robotokkal való aratás technológiájának jelenlegi állását, tesztelni és véleményezni az ezen a téren rendelkezésre álló szoftvereket, kezdeményezni ezek honosítását, illetve a hiányzó vagy túl drága komponensek hazai kifejlesztését. Kidolgozni a begyűjtött digitális objektumok tárolásának technikáját, az egyre inkább szabványosodó nemzetközi gyakorlatnak megfelelően. Felbecsülni a szükséges tárolási kapacitást és annak növekedési ütemét. Ajánlást kidolgozni a jól begyűjthető és jól archiválható web-site-ok kialakítására (illetve a problémás site-ok tartalmának ilyen célra alkalmas kiexportálására) és ennek figyelembevételére ösztönözni a nagyobb hazai tartalomszolgáltatókat.

3. A metaadatok kérdéseivel foglalkozó munkacsoport

Feladata: Áttekinteni az internet archívumok metaadatolásának nemzetközi gyakorlatát és javaslatot tenni az egyedi site-ok mentéseinek metaadataira. Mind az egyedi site-ok mentésénél, mind pedig a nagytömegű automatikus aratásnál ki kell dolgozni a dokumentumból automatikusan kinyerhető ill. generálható metaadatok előállításának és tárolásának technológiáját. Döntést kell hozni arról, hogy ezekből mi és milyen módon kerüljön át a nemzeti bibliográfiába, illetve a könyvtár(ak) katalógusaiba.

4. A hasznosítás/szolgáltatás kérdéseivel foglalkozó munkacsoport

Feladata: Az archívumba kerülő anyag lehetséges felhasználási formáinak áttekintése. Javaslatokat tesz a nyilvános és nem nyilvános, a non-profit és az üzleti célú hasznosításra.

Felméri a felhasználói igényeket. Demó szinten beüzemel egy vagy több keresőfelületet a teszt (pilot) fázisban begyűjtött anyagban való kereséshez és böngészéshez, valamint további szolgáltatások, az archivált tartalomra ráépülő alkalmazások terveit készíti el.

5. A jogi kérdésekkel foglalkozó munkacsoport

Feladata: Áttekinteni a webarchiválással kapcsolatban felmerülő jogi kérdéseket: a kötelezpéldány törvény kiterjeszthetősége, copyright és privacy kérdések, az archívum tartalmának tulajdonjoga és hasznosításának joga stb. Szerződéstervezetet dolgoz ki az egyedi webhelyek archiválásához, amely rögzíti az eredeti site gazdája és az archívum jogait és kötelességeit. Jogi nyilatkozattervezetet dolgoz ki az egyedi engedélyekkel, illetve az automatikusan (egyedi engedélyek nélkül) begyűjtött anyag státuszára és felhasználására vonatkozóan. Szabályokat alkot arra, hogy aki nem szeretne az általa szolgáltatott tartalommal bekerülni az archívumba vagy szeretne kikerülni onnan, az hogyan teheti ezt meg (pl. a robot.txt fájlban megadott paraméterekkel, illetve egy „opt-out” procedúrával). Szabályzatot alkot az archívum felhasználóira és felhasználási módjaira vonatkozó előírásokról.

6. A finanszírozás kérdéseivel foglalkozó munkacsoport

Feladata: Részletes pénzügyi terv készítése a projekt tervezési és fejlesztési fázisának időszakára, valamint a hosszú távú fenntarthatóság érdekében az üzemszerű működés várható költségeinek megbecslése és a lehetséges finanszírozási formák (pl. pályázatok, konzorciumi tagok hozzájárulása, állami források, szponzori támogatás, fizetős szolgáltatások) felmérése.

Teszt fázis:

Már a tervezési fázis félidejétől elindítható egy teszt archívum kialakítása, amely néhány száz webhely néhányszori mentését tartalmazza és valamilyen egyszerű visszakeresési mechanizmust is biztosít az archivált anyag megtekintéséhez. A cél az üzemszerű működéshez szükséges gyakorlati tapasztalatok megszerzése, a felmerülő problémákkal való szembesülés, a majdani nagy tömegű archiválás erőforrásigényeinek (pl. tárhely, gépidő, élőmunka, költségek) reális megbecsléséhez adatok gyűjtése. Ebben a „pilot” szakaszban kell összerakni a rendelkezésre álló kész szoftverekből egy együttműködő rendszert, illetve átírni vagy hozzáfejleszteni olyan komponenseket, amelyekre egy magyar archívumhoz szükség van. Ekkor kell elkezdeni egy olyan munkafolyamat-nyilvántartó és -vezérlő keretrendszer kialakítását is, amellyel az egész archiválási tevékenység végigkövethető és adminisztrálható.

Fejlesztési fázis:

A tervezési és a teszt fázis eredményei alapján ki kell alakítani a már üzemszerű működésre, nagy tömegű webarchiválásra alkalmas informatikai rendszert. Meg kell oldani a hosszú távú megőrzés biztonságos módját (pl. nyílt formátumokra konvertálás, ellenőrző kódok a fájl sérülések felismerésére). Fejlettebb visszakereső és statisztika-készítő funkciókat kell beépíteni. El kell készíteni a metaadatoláshoz és a munkafolyamat menedzseléshez szükséges szoftverek végleges változatait. El kell végezni a rendszer tesztelését és a felmerülő hibák és hiányok javítását/pótlását. Szabályzatokat kell írni az archiválási tevékenység egyes munkafázisaihoz és ki kell képezni a válogatással, az archiválással és a metaadatolással foglalkozó munkatársakat a rendszer egyes komponenseinek használatára.

Üzemszerű működés:

Az üzemszerű archiválást csak fokozatosan célszerű felfuttatni, nehogy túlterhelődjének az erőforrások. A lementett webanyagot folyamatosan monitorozni kell, elemezni az összetételét, a belső linkek működését, a hibákat és hiányokat, az egyes mentések közötti változások mértékét, és mindezeket a tapasztalatokat visszacsatolva finomítani az archiváló szoftver beállításain, és optimalizálni az egyes webhelyek ismételt mentésének gyakoriságát. A szoftverrobotokkal nem jól archiválható, de értékes website-ok tartalomgazdái meg kell keresni és megállapodni velük valamilyen más mentési módban (pl. a tartalom kiexportálása egy, a robot számára bejárható statikus formába, vagy a tartalom automatikus feltöltése az archívumba egy alkalmazásprogramozási felületen (API-n) keresztül vagy valamilyen más szabványos protokollal). Az archivált anyag mielőbbi és minél sokirányúbb hasznosítása érdekében különböző szolgáltatásokat kell kifejleszteni, valamint API-kat kell biztosítani külső alkalmazások számára (pl. hogy ha egy felhasználó egy webhelyen 404-es hibába fut, akkor a webszerver automatikusan az archívumban található mentés(ek)hez irányítsa át).

Mivel az internetes tartalomszolgáltató technikák folyamatosan változnak, ezért ezeket az archívumnak is követnie kell. Egy üzemszerűen működő rendszernek nemcsak az exponenciálisan növekvő digitális tartalommal kell lépést tartania, hanem magát a rendszert is állandóan tovább kell fejleszteni, és emellett migrációval vagy emulációval azt is meg kell oldani, hogy a lementett anyag a távolabbi jövőben is megjeleníthető maradjon.

ÜTEMEZÉS:

A Tervezési fázis teljes végigvitele kb. 1 évet igényel, de annak félidejétől már elindulhat a Teszt fázis, ami szintén legalább 1 év. Ezek után lehet nekiállni a tervek és a tapasztalatok alapján a Fejlesztési fázisnak, amelyre minimum két év kell. A rendszer első működőképes változatának elkészülte után indulhat be az Üzemszerű működés a legalapvetőbb funkciókkal és fokozatosan növekvő archiválási kapacitással.

FORRÁSIGÉNY:

A fejlesztési időszakban szükséges informatikai infrastruktúra (gyors hálózat, nagy teljesítményű szerverek és munkaállomások, nagyméretű és biztonságos tárolókapacitás) addigra már adottnak tekinthető. Az üzemszerű működés során az évi plusz tárhelyigény néhányszor 10 terabyte-ra becsülhető, de ez természetesen idővel növekedhet és akár egy nagyságrenddel több vagy kevesebb is lehet annak függvényében, hogy milyen döntések születnek a gyűjtőkörre, az ismételt mentések gyakoriságára és a növekedési ütemre vonatkozóan.

A projektmenedzsment költségei az ilyen nagyságrendű projektek esetében szokásosnál magasabbak lehetnek, ha egy konzorcium keretében történik majd a fejlesztés és megvalósítás. A teljes időszakra megbízott projektvezetőn kívül a jogi vonatkozások tisztázásához, a szükséges szerződésminták megfogalmazásához egy, a média- és internet-jogban járatos jogász megbízására is szükség lesz néhány hónapig.

Mivel hazánkban még nincs webarchiválási tapasztalat és tudás, ezért a tervezés és fejlesztés idejére 3-5 fő magyar és (esetleg) külföldi külső szakértő időszakos alkalmazása mellett a vonatkozó szakirodalom beszerzésére, feldolgozására, a külföldi archívumok üzemeltetőivel való kapcsolatfelvételre, a külföldi szakmai konferenciákon, tanfolyamokon való részvételre is tervezni kell költségeket.

A tapasztalatcserét és a már kifejlesztett, de nem open source szoftverekhez való hozzájutást nagyban segítené az International Internet Preservation Consortium-hoz való csatlakozás. (Ennek 12 hónapra érvényes tagdíja 10-50 millió euró közötti éves költségvetésű intézményeknek 4000 euró, ez alatt pedig 2000 euró.)

Az archívumhoz szükséges szoftverek egy része open source vagy legalább freeware, vagy kedvezményesen megkapható IIPC tagként, így ezeket csak megtanulni és esetleg a magyar szükségletekhez adaptálni kell. A teljes archiválási folyamat menedzseléséhez és a metaadatoláshoz szükséges keretrendszert helyben kell kifejleszteni, erre nincsenek kulcsrakész megoldások. Ez az ELDORADO-hoz hasonló komplexitású rendszerré nőhet az üzemszerű működés elindulásának idejére.

Az archívum üzemeltetéséhez többféle szakemberre van szükség, akik közül egyeseknek már a fejlesztési fázisban is részt kell venniük. Szükség van egy rendszergazdára és egy-két programozóra (ha a teljes rendszer fejlesztése házon belül történik, akkor többre is), valamint egy operátorra, aki napi karbantartási munkákat elvégzi. Kell egy rendszerkönyvtáros, aki könyvtári szempontból felügyeli és menedzseli az archívumot, meghozza a szükséges gyűjteményfejlesztési döntéseket. A szelektíven archiválendő források kiválasztásához és a mentési paraméterek (pl. bejárasi mélység, mentendő fájltypusok, mentési gyakoriság) beállításához legalább két gyarapító munkatárs kell, de idővel ezek számát növelni célszerű (esetleg más könyvtárak bevonásával). Az archiválás sikerességének ellenőrzéséhez, a mentés során keletkezett hibák és hiányok felderítéséhez, dokumentálásához és lehetőség szerinti javításához szintén legalább két, informatikailag jól képzett könyvtáros kell. A szelektíven archivált webhelyek leíró metaadatokkal való ellátása 3-5 fő katalogizáló szakembert igényel (de ezek számát is növelni kell az archívum növekedésével, részben más könyvtárak munkatársainak bevonásával is).

További költségeket jelent, hogy már a fejlesztési fázisban el kell kezdeni az archívum leendő üzemeltetőinek és a partnerintézményekből bedolgozóknak a képzését szabályzatok, elektronikus tananyagok és tanfolyamok formájában. Az eredeti tartalomgazdák, valamint az archívum potenciális felhasználói számára pedig tájékoztató anyagokat kell készíteni, előadásokat tartani és létrehozni egy honlapot, ahol az érdeklődők naprakészen minden információt megtalálnak a készülő internet archívummal kapcsolatban.

IRODALOM:

Charta a digitális örökség védelméről

UNESCO, 2003

<http://www.unesco.hu/archivum-2009-vegeig/kommunikacio-informacio/charta-digitalis-orokseg>

Crook, Edgar (ref. Drótos László): Webarchiválás a webkettes világban

In: Tudományos és Műszaki Tájékoztatás, 2010. (57. évf.), 2. sz.

http://tmt.omikk.bme.hu/show_news.html?id=5276&issue_id=512

Dancs Szabolcs: Webarchiválási politikák

In: Könyv, könyvtár, könyvtáros, 2011. (20. évf.), 10. sz.

http://epa.oszk.hu/01300/01367/00248/pdf/EPA01367_3K_2011_10_14-20.pdf

Dippold Péter: A hagyományos nemzeti bibliográfia és az Internet : Válaszlehetőségek az új kihívásokra

Budapest : ELTE BTK, 2005

<http://mek.oszk.hu/03500/03557>

Drótos László: Mi a MIA? : Javaslat egy Magyar Internet Archívum létrehozására
In: Tudományos és Műszaki Tájékoztatás, 2006. (53. évf.), 6. sz.
http://tmt.omikk.bme.hu/show_news.html?id=4431&issue_id=473

Hegyközi Ilona: Hol tart ma a webarchiválás?
In: Könyvtári Figyelő, 2014. 4. sz.
<http://ki.oszk.hu/kf/2015/01/hol-tart-ma-a-webarchivalas/>

Kornhoffer Mónika: A világhálón található információk gyűjtésének és megőrzésének hazai és nemzetközi áttekintése
Pécs : PTE FEEK, 2010
<http://mek.oszk.hu/10400/10424/>

Nagymélykúti Balázs: Tartalommegőrzés az interneten: webarchívumok
Szeged : SZTE JGYPK Könyvtártudományi Tanszék, 2007

PoWR: The Preservation of Web Resources Handbook
London : JISC, 2008
<https://jiscpowr.jiscinvolve.org/wp/handbook/>