Pécsi Tudományegyetem
Felnőttképzési és Emberi Erőforrás Fejlesztési Kar

SZAKDOLGOZAT

Kornhoffer Mónika
informatikus könyvtáros szak

Pécs
2010

Pécsi Tudományegyetem
Felnőttképzési és Emberi Erőforrás Fejlesztési Kar

informatikus könyvtáros (BA) szak
„Az információs műveltség pedagógiája” szakirány

Kornhoffer Mónika

A világhálón található információk gyűjtésének és megőrzésének hazai és nemzetközi áttekintése

Dr. Sipos Anna Magdolna
egyetemi docens
konzulens

Tartalomjegyzék

Bevezetés

A mai kor embere azt gondolja, ha bármilyen információra van szüksége, még a könyvtárba sem kell elmennie, elég ha leül a számítógépe elé, „felmegy” a világhálóra és ott minden szükséges információt megtalál. Ez részben igaz is, hiszen ha egy színházi műsorra, vagy menetrendre vagyunk kíváncsiak, esetleg egy kutató legújabb kutatási eredményeire, az Internet segítségével és megfelelő keresési módszerekkel ezeket az információkat gyorsan meg tudjuk találni. De vajon ugyanilyen könnyű dolga lesz-e egy kutatónak is, aki ötven év múlva a mai weboldalakkal kapcsolatban szeretne tanulmányt írni?

A legtöbb ember, akinek ezt a kérdést feltesszük, gondolkodás nélkül rávágja, hogy persze, hiszen miért tűnne el bármi is a világhálóról? Gondoljuk mindezt annak ellenére, hogy az Interneten böngészve akár naponta előfordulhat, hogy a keresett oldal helyett csak egy hibaüzenetet találunk, mely arról tájékoztat, hogy a keresett oldal már nem található. Bár bosszankodunk ezen, de utána esetleg megváltoztatott paraméterekkel folytatjuk tovább a keresést, hátha valamilyen más módon el tudjuk érni a keresett információt.

Bár napjainkban mindennapossá vált, hogy a szükséges információkat a világhálón keressük meg, mégsem merül fel bennünk, hogy a már jelenleg is megtalálható hatalmas adatmennyiség napról napra bővül, és ennek a tárolása nem kis feladat. Ezen kívül az egyre több és több információ közül sokkal nehezebben tudjuk kiválogatni a számunkra fontosakat, hiszen csak félrevezeti a keresőt, ha mondjuk olyan cégek honlapjai is megtalálhatók az Interneten, amelyek már régen megszűntek. A folyamatosan működő intézmények, vállalatok honlapjainak is követnie kell a képviselt szervezett működésében, tevékenységében, elérhetőségében stb. bekövetkezett változásokat, amely a régebbi adatok felülírásával valósul meg. Ezekből a példákból

is látszik, hogy szükséges a régi, elavult információk törlése, módosítása.

De mi történik azokkal a weblapokkal, amelyeket már nem tudunk elérni? Gyakorló informatikusként azt gondoltam, hogy azokat a honlapokat, amelyeket nem tudok elérni, azok is biztosan valahol megtalálhatók elmentett, archivált formában, hiszen minden cég, minden adatáról (még a nagyon régiekről is) különböző biztonsági mentéseket tárol. Ehhez képest nagy meglepetést okozott számomra egy 2008. nyári HVG cikk, ahol pont az „eltűnő honlapok”[1] témáját járták körül. A cikk szerint a weblapok semmilyen magyarországi webarchívumban sincsenek eltárolva, hiszen igazából webarchívumunk sincsen. Ha régebbi vagy az Interneten már nem megtalálható weblapokat keresünk, akkor csak az Internet Archive[2] – amerikai nonprofit cég – webarchívumában érdemes keresgélni.

A nyomtatott dokumentumok nyilvántartására, megőrzésére különböző törvények, rendeletek, szabályozók, stratégiák léteznek, de mi a helyzet például azokkal a folyóiratokkal, melyek csak elektronikus formában léteznek? Ugyanúgy „elvesznek”, mint bármelyik másik weblap? Egyáltalán létezik-e bármilyen törvény, vagy szabályozás az elektronikus dokumentumok megőrzésére Magyarországon és a világ más országaiban? Létrehoztak-e a különböző országok Internet archívumokat, melyekben a saját nemzeti dokumentumaikat gyűjtik? Ha léteznek nemzeti Internet archívumok, akkor milyen elvek alapján válogatják ki és gyűjtik össze a bekerülő dokumentumokat?

A cikk elolvasása után többek között ezek a kérdések foglalkoztattak és ezekre szeretnék választ kapni a szakdolgozatomban úgy, hogy a weblapok feldolgozásával nem foglalkozom.

---

[1] Riba István: Eltűnő honlapok – Hibaüzenet. In: Heti Világgazdaság, 2008. (30. évf.), 32. sz., (augusztus 9.) 22-23. p.
[2] www.archive.org

1. Mi a web és miért kell archiválni a weben található anyagokat?

A web más néven Internet vagy világháló, egy nemzetközi információs hálózat, nyílt architektúrával, sokrétű szolgáltatással.[3] Az Internetet a hálózatok hálózataként is szokták emlegetni, mert sok kis hálózatot kapcsol össze egy nagy hálózattá a nyílt architektúra segítségével.

Az Internetet katonai szempontok figyelembevételével alkották meg, de mint sok egyéb katonai fejlesztés, ez is életünk nélkülözhetetlen részévé vált. Az 1960-as években az amerikai hadsereg célja az volt, hogy a katonai hálózatot ne egy központi számítógép köré szervezzék – hiszen így azt megsemmisítve az egész hálózat működésképtelenné válhat – hanem önálló és egyenrangú számítógépek (csomópontok) összekapcsolásából álljon a rendszer. Így egy-egy csomópont (számítógép) kiesésével is működőképes marad a hálózat megmaradt része. 1969-ben el is indult a megálmodott hálózat, de közben más hálózatokat is kifejlesztettek és felvetődött az igény, hogy a meglévő hálózatok kommunikálni tudjanak egymással.

Maga az Internet 1983-ban kezdte meg a működését az Amerikai Egyesült Államokban két különálló (katonai és polgári) hálózat összekapcsolását követően. A különálló hálózatok összekapcsolása miatt szükségessé vált az információ-csomagok formájának szabványosítása, hálózati címek kiosztása és átjáró (gateway) gépek használata. A szabványos formájú információ-csomagoknak köszönhetően már bármikor újabb hálózatokat lehetett hozzákapcsolni a meglévő rendszerhez úgy, hogy az eddigi felépítést nem kellett

[3] Pluhár Gábor: Informatikai értelmező szótár. Válogatás az informatikai szakirodalom tanulmányozásához. http://mek.oszk.hu/00000/00083/00083.pdf (2009. október 22.)

megváltoztatni.[4] Hamarosan az USA összes egyeteme rákapcsolódott a hálózatra, majd Kanadában és Ausztráliában is elkezdték kiépíteni és használni az Internetet. Európába ugyan kicsit később jutott el, de nagyon hamar kiszorította az európai hálózatnak szánt ISO/OSI megoldást (International Standard Organization/Open Systems Interconnect). Az 1990-es években az Internetet használók száma világszerte megsokszorozódott, és 1990-ben már Magyarországon is megkezdődött a világháló használata, bár az első kereskedelmi célú internetes szolgáltató csak néhány évvel később jelent meg. Elsőként nálunk is az egyetemek és kutatóintézetek kapcsolódtak az Internethez létrehozva ezzel a magyar felsőoktatási és kutatási hálózatot.[5] Napjainkban viszont már annyira elterjedtté vált, hogy akik használják ezt az eszközt, és a segítségével elérhető információs forrásokat, el sem tudják képzelni nélküle az életüket.

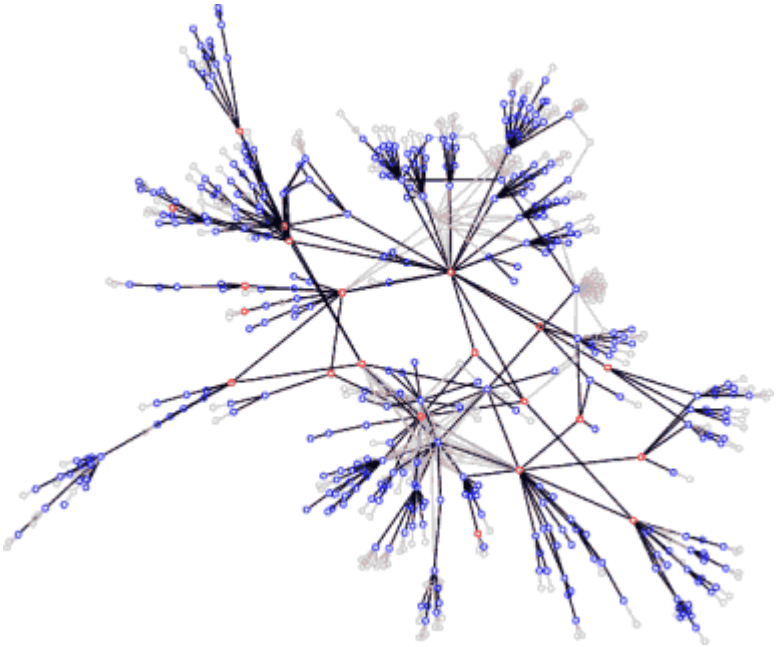
Az Internet működése sokkal egyszerűbb, mint ahogy gondolnánk. A világhálón különböző dokumentumok (weblapok) találhatók, melyeket hyperlinkek segítségével kötöttek össze, melyek segítségével újabb és újabb weblapokra tudunk eljutni. A dokumentumok által kialakult rendszert webböngésző programok használatával lehet elérni, melyeket a dokumentumok – más néven weblapok – megjelenítésére hoztak létre.

Ha megpróbáljuk elképzelni az előzőekben leírtakat, akkor láthatjuk, hogy az Internet nem csak fizikailag háló jellegű, hanem a dokumentumok – mint csomópontok – és a hyperlinkek – mint a háló szálai, melyek segítségével el lehet jutni egyik csomópontból a másikba – is ezt a jelleget erősítik.

---

[4] Debreceni Egyetem, Informatikai Szolgáltató Központ: Internet hálózat http://www.cic.klte.hu/iszkw3/kltenet/kltenet5.html (2009. október 22.)

[5] Máray Tamás: Hálózatok hálózata: az internet. http://www.mindentudas.hu/maray/20031201maray2.html?pIdx=2 (2009. október 22.)



1. ábra *Az Internet hálózat egy tipikus darabja. A körök a routereket jelölik. A hálózat szélén halványan jelölt csomópontokon találhatók az egymással kommunikáló számítógépek. (Forrás: http://www.cybergeography.org/atlas/topology.html)*

Bár az Internetet viszonylag rövid ideje használjuk, mégis a weben megtalálható dokumentumok száma és mérete hatalmas és nagyon gyorsan növekszik. Naponta több mint 7 millió oldal születik a világban, de ezzel egyidőben folyamatosan tűnnek is el olyan oldalak, melyeket betiltottak, vagy már nincs rájuk szükség (pl. befejeződött az a projekt, amire létrehozták). Egy cikk szerint a weboldalak átlagos élettartama 60 nap.[6] „Az összes weblap fele nem idősebb 100 napnál, s körülbelül a negyedük idősebb csak egyévesnél. A .com (üzleti szféra) területen a weblapok 40%-a naponta változik, míg az állami (.gov) és oktatási (.edu) szektorban az oldalak átlagos élettartama négy hónap. A világhálón megjelenő tartalom átlagos élettartama két év, az URL-ek átlagos élettartama négy év. A tudományos oktatásban használt URL-ek átlagosan ötvenöt hónapig élnek.”[7]

[6] Acobs, Neil - Chambers, Jenny - Morris, Anne: Dokumentumszolgáltatók weboldalai In: Tudományos és Műszaki Tájékoztatás 2000. (47. évf.), 11. sz. http://tmt.omikk.bme.hu/show_news.html?id=1505&issue_id=30 (2009. október 22.)

[7] Brooks, Terrence A.: Keresés a világhálón: hogyan változtatta meg az internet az információkeresést? In: Tudományos és Műszaki Tájékoztatás 2004. (51. évf.), 5. sz. http://tmt.omikk.bme.hu/show_news.html?id=3602&issue_id=450 (2009. október 22.)

Terrence Books pillanatfelvételhez hasonlította a világhálón meglévő forrásokat, amely megállapítást az előzőekben ismertetett adatok alapján el kell fogadnunk. Viszont ha ezt elfogadjuk, akkor mindenki számára világossá válik, hogy valamit nagyon gyorsan ki kell találni annak érdekében, hogy ezeket a „pillanatfelvételeket” valamilyen formában meg tudjuk őrizni az utókor számára. Ha ezt nem tesszük, úgy rövid időn belül, könnyen örökre elveszhetnek ezek a tartalmak, és mi is úgy járhatunk, mint az elődeink a festményekkel, könyvekkel, régi filmekkel és sok más egyéb kulturális értékkel.

2. Mi az, amit archiválni kell?

Ha megkérdezzük ismerőseinket, hogy szerintük mi az, amit érdemes lenne megtartani az utókor számára, valószínűleg legtöbben elsősorban a hírportálok vagy az internetes újságok anyagait említenék. Ám ha a jövő kutatóinak szemüvegén keresztül vizsgálnánk az Interneten fellelhető anyagokat, valószínűleg más témákat tartanánk fontosnak.

De pontosan mi is található az Interneten? Milyen anyagokat lehet archiválni, és azok milyen típusú weboldalakon találhatók? Az Internet mely szintjein helyezkednek el az adatok? Ezekre a kérdésre próbálok válaszokat adni ebben a fejezetben.

2.1. Mi a tartalom?

A tartalmat a különböző felfogások egymástól eltérően értelmezik. Az egyik meghatározás szerint minden, ami az Interneten található, az a tartalom kategóriájába tartozik. Egy másik definíció azonban szűkebben értelmezi a fogalom tartalmát: azokat a dokumentumokat sorolja ide, amelyek ellenőrzöttek, jól dokumentáltak és értékes információhordozó állományt tartalmaznak.[8]

Bármelyik definíció szerint is vizsgáljuk az Interneten található anyagokat, mindegyik besorolható a következő két csoport valamelyikébe:

- A valós élet melléktermékeiként értelmezhető információk: például az e-folyóiratok, az e-könyvek, a publikációk stb.

[8] Nagymélykúti Balázs: Tartalommegőrzés az interneten: webarchívumok: szakdolgozat. Szegedi Tudományegyetem, Juhász Gyula Tanárképző Főiskolai Kar, Könyvtártudományi Tanszék, 2007. www.szilleri.tvn.hu/nagymelykut.doc (2009. március 11.)

- A virtuális világ melléktermékének tekinthető információk: elektronikus levelek, web oldalak, chatszobák, blogok stb.

Jóllehet a jövő kutatói számára mindkét csoportba tartozó anyagok fontosak lesznek, ám az adott kor emberének köznapi életét, egymás közötti kommunikációját a virtuális világ melléktermékei mutatják majd be legjobban, ezért – véleményem szerint – az archívumoknak erre a csoportra is koncentrálniuk kellene.

2.2. A weboldalak típusai

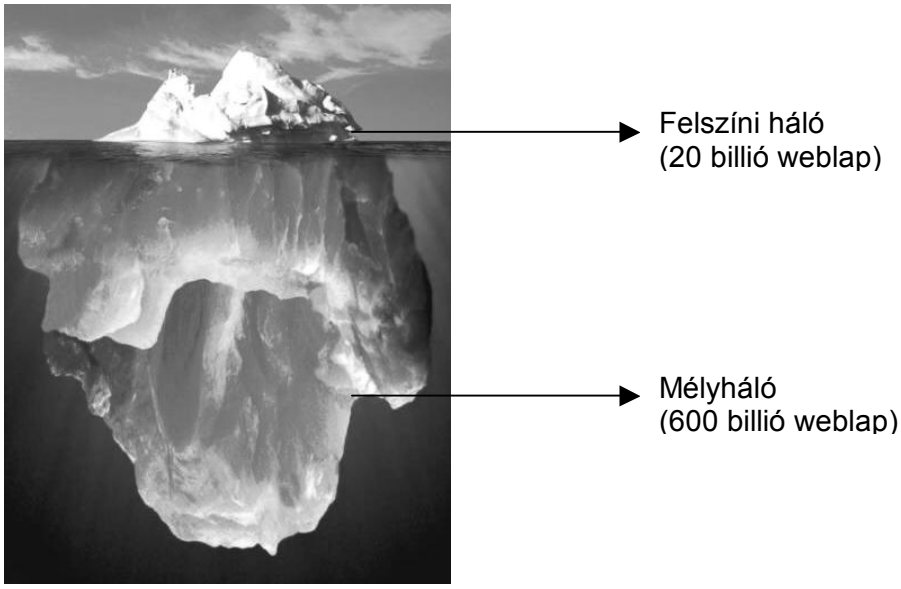
Az Interneten böngészve ahány weboldalt megnyitunk, annyi féle lehet, mégis az összes oldal két nagy csoportba sorolható: statikus vagy dinamikus oldalak.

A statikus oldalak csoportjába sorolhatók a weboldalak legegyszerűbb típusai, amelyeket úgy kell elképzelni, mint például egy Word dokumentumot, azzal a különbséggel, hogy ezek HTML kódokat tartalmaznak. A kódok segítségével képes bármelyik webböngésző program összefüggő dokumentumként elolvasni a statikus oldalakat. Az oldalak általában szövegeket és képeket tartalmaznak, és az oldalak között, továbbá azokon belül egyszerű linkek segítségével lehet navigálni. A statikus weboldalak esetében minden olvasó ugyanazt látja, az oldalak megjelenését nem befolyásolhatják. Mivel az egész oldal dokumentumszerű, ezért annak módosítása nehézkesebb. A legkisebb módosítás esetén is mindent ugyanúgy át kell írni, át kell szerkeszteni, mint bármelyik Word dokumentumnál. Ennek ellenére az Interneten található weboldalak körülbelül 80%-a statikus oldal, mert hatalmas előnyük, hogy a HTML kódokat könnyű elsajátítani, és minimális munkával, kis fantáziával szép weboldalakat lehet létre hozni.

A dinamikus oldalak kategóriájába tartozó weblapok egyre inkább divatba jönnek, hiszen sokkal fejlettebbek, mint a statikus oldalak. Ezek az oldalak már nem csupán szövegeket és képeket tartalmaznak, hanem interaktív elemeket is. Ezen túl a dinamikus oldalak „emlékezettel" is rendelkeznek, mivel a felhasználók adatait adatbázisokban tárolják. Az „emlékezetüknek" köszönhető, hogy ha például fellépünk egy online könyvesbolt weboldalára, az oldalon megjelennek azok a könyvek, amelyeket már egyszer megnéztünk, vagy azok a könyvek, melyeket az általunk kiválasztottak mellé már más felhasználók megrendeltek. A statikus weboldalakkal szemben ezeknek az oldalaknak a tartalmát – amely ugyancsak adatbázisban tárolódik – könnyen meg lehet változtatni egy egyszerű adminisztrációs űrlapon keresztül. A dinamikus weblapok csoportjába tartoznak például a különböző online boltok weblapjai, a hírportálok, fórumok, felhasználók regisztrációi stb. Ám bármilyen szépek is ezek az oldalak, bármennyire is könnyű tartalmuk módosítása, az avatatlan felhasználók számára mégis van egy hatalmas hátrányuk a statikus oldalakkal szemben: ezeket az oldalakat sokkal nehezebb elkészíteni, mert a kreativitáson kívül már programozói tudást is igényelnek.

2.3. A web szintjei

Amennyiben a keresők szemszögéből vizsgáljuk az Interneten található tartalmakat, elég hamar rá fogunk jönni, hogy azok nagyban hasonlítanak a jéghegyekhez. Mint a jéghegyeknél, itt is van látható (felszíni háló) és láthatatlan (mélyháló) rész. Sajnos, a jéghegyhez hasonlóan, a láthatatlan rész itt is sokkal nagyobb, mint a látható.



2. ábra *Felszíni és mélyháló*
(Forrás: http://googlediscovery.com/2009/03/26/deepdyve-onde-eles-estao-mergulhando/)

A felszíni hálót lényegében a statikus oldalak összessége alkotja, amelyekhez az általános keresőmotorok is könnyedén hozzáférnek és indexeket készítenek a tartalmukról. „Ennek nagyságát a teljes web méretének 16%-ára becsülik."[9]

A mélyháló az Internet „láthatatlan" része, melynek az angol neve deep web. „Az elnevezés a tartalom nehezebb elérhetőségére utal"[10], hiszen az adatok és információk a dinamikus weblapok „mögött" találhatók különböző adatbázisokban, strukturált adatokként. Az adatbázisokban az általános keresőmotorok általában nem tudnak keresni, ezért fordulhat elő, hogy ha például a T-Com weboldalán (www.t-com.hu) keresünk egy telefonszámot, akkor könnyen megtaláljuk a hozzá tartozó előfizetőt, viszont ha ugyanazt a telefonszámot beírjuk egy általános keresőbe, akkor nem biztos, hogy kapunk találatot.

[9] Rutkovszky Edéné - Rutkovszky Ádám: A láthatatlan web keresése https://nws.niif.hu/ncd2003/docs/ehu/EHU-61.htm (2009. december 8.)
[10] Kardkovács Zsolt - Magyar Gábor - Tikk Domonkos: A szavak hálójában: szabadszavas mélyháló kereső program. Budapesti Műszaki és Gazdaságtudományi Egyetem, Távközlési és Médiainformatikai Tanszék
http://categorizer.tmit.bme.hu/~domi/Publications/Htechnika.pdf (2009. december 8.)

A felszíni és a mélyháló csoportja között helyezkedik el egy sajátos, viselkedését és kereshetőségét tekintve, a kettő között meghúzódó kategória, az úgynevezett szürkeállomány. A jéghegy hasonlatnál maradva: míg a jéghegyeknél a víz alatti és feletti rész között éles határvonalként ott van a víz, addig a mélyháló és felszíni háló között nincs ilyen éles határvonal. A szürke zóna azokat a felszíni webhelyeket veszi körül, melyek elérhetőek a mély webhelyeken keresztül is.

Általánosságban a felszíni háló tartalma könnyedén bejárható a keresőmotorok által, viszont a mélyháló tartalmát csak a dinamikus weblapokon indított közvetlen lekérdezések segítségével lehet megjeleníteni. A dinamikus weblapok kezelését lényegesen megkönnyíti az a lehetőség, hogy ha egy dinamikus weblapról egyszer már lekérdeztünk valamit, úgy a lekérdezés létrehoz egy URL-t, amelyben megjelenik a lekérdezés szövege és általában az adatbázisrekord számát is tartalmazza. Ennek segítségével újra megtalálható a már egyszer használt dokumentum. Ezeknek az URL-eknek a segítségével a mély web tartalma a felszínre hozható, hiszen „bármely mélyweb tartalmat, amelyet egy statikus weboldal listáz, felfedezhetik a *„pókok”* („pókok”=keresőmotorok kereső funkciói), és ezáltal indexelhetik a keresőmotorok”[11]. Mindennek ellenére a mélyháló teljes állományát ezzel a módszerrel sem lehet felszínre hozni, de legalább egy része kereshetővé válik.

[11] Rabb Ágnes: Szöveggyűjtemény a mély web tanulmányozásához: Cikkek és tanulmányok, külföldi és magyar források alapján: szakdolgozat. Szegedi Tudományegyetem, Juhász Gyula Tanárképző Főiskolai Kar, Könyvtártudományi Tanszék http://szilleri.tvn.hu/rabb.doc (2009. december 1.)

3. Az internetes tartalom archiválásával kapcsolatos kérdések

Az előző fejezetben szó volt a weboldalak típusairól és a weboldalakon található adatokról, valamint az azokhoz való hozzáférésről. Ha csak ezekre gondolunk, máris rengeteg probléma merül fel, amelyekkel szembesülnek, szembesülni fognak az internetes archívumot létesítő szervezetek. De természetesen, ezen kívül még számos kérdés van, amelyeket az archívumok létrehozóinak és gondozóinak előre át kell gondolni.

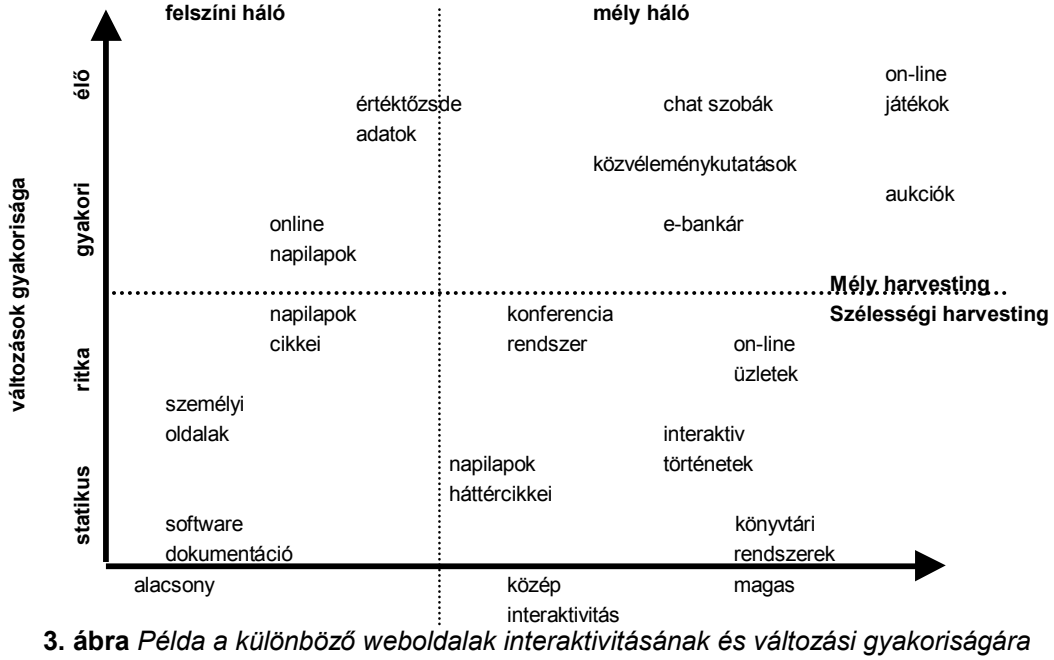
3.1. Mit archiváljunk, és mennyi ideig tartsuk meg? A dokumentum elavulása.

Ha ezt a kérdést feltennénk az „utca emberének”, biztosan értetlenül nézne ránk és azt válaszolná, hogy természetesen mindent, és minden anyagot őrizzünk meg az idők végezetéig. De biztos, hogy akkor cselekszünk jól, ha megfogadjuk a tanácsát?

Az előző fejezetben érintőlegesen szó volt arról, hogy milyen hatalmas mennyiségű és milyen sokféle anyag található az Interneten, melynek tükrében könnyen beláthatjuk, ha mindezt archiválni szeretnénk, nagyon hamar belefulladnánk a rengeteg elmentett tartalomba. Ennek köszönhető, hogy a már működő Internet archívumok általában a saját országaik tartalmának a megőrzésére törekednek, de így is kénytelenek differenciálni, mert még az így szűrt anyag mennyisége is egy idő után kezelhetetlenné válna. A további szelektáláshoz használható egyik szempont a dokumentumok életciklusa. Ennek segítségével viszonylag meggyőző bizonyossággal dönthető el, hogy meddig érdemes megőrizni egy-egy anyagot.

Bármennyire is furcsa, az Interneten található dokumentumok sokkal hamarabb elavulnak, mint nyomtatott társaik. Ennek egyik oka, hogy a publikációk sokkal gyorsabban és szélesebb körben jelennek meg a neten, mint nyomtatott formában. Ebből következik a következő kérdés: hogyan tudnánk megőrizni például az összes olyan tudományos cikket és a rájuk adott reakciókat, melyek különböző nyílt hozzáférésű folyóiratokban jelennek meg? Egyáltalán fontos-e az összes cikk és reakció megőrzése, hiszen jól tudjuk, bizonyos tudományágakban (pl. orvostudomány) szinte naponta születnek újabb és újabb eredmények, melyek az előző időszak eredményeit, felfedezéseit felülírják. Arról nem is beszélve, hogy a világhálón vannak jó és rosszminőségű, pontos és pontatlan, mások által megőrzött és illegális (kalóz szoftverek, kalóz filmek, gyermekpornó stb.) anyagok is. Mi az, amit ezekből érdemes az utókornak megmenteni? Érdemes-e azokat az anyagokat is archiválni, melyeket már működő archívumokban megtalálhatók (pl. MedLine, különböző hírügynökségek stb.)?

Ráadásul a weblapok tartalma – mint a következő ábrából is látható – különböző gyakorisággal kerül frissítésre.



3. ábra *Példa a különböző weboldalak interaktivitásának és változási gyakoriságára (Forrás: www.szilleri.tvn.hu/nagymelykut.doc)*

Ezért fontos meghatározni azt is, hogy milyen gyakorisággal célszerű archiválni az oldalakat. Ez a gyűjtőköri politika része.

Az itt felsorolt néhány kérdésből és felvetésből is látható: nem biztos, hogy jó ötlet mindig és mindenben a teljességre törekedni.

3.2. Ki végezze el az archiválást? Kinek a felelőssége legyen?

Nyomtatott dokumentumoknál egyértelmű, hogy az archiválás, megőrzés a nemzeti könyvtárak feladata. Az anyagok begyűjtésében segítségükre van a kötelespéldány-szolgáltatás törvényi háttere is. De kinek a feladata az Interneten található anyagok megőrzése?

Az archiválás tetemes anyagi ráfordítást igényel, viszont ezzel szemben a lehetséges bevétel minimális. Ezek után elég valószínűtlen, hogy bármely profitorientált cég szárnyai alá venné az Internet archívumok megvalósítását. Így csak a különböző nonprofit alapon működő cégek vagy a nemzeti könyvtárak foglalkozhatnának ezzel a kihívással, de megfelelő törvényi háttér és anyagi bázis nélkül egyik sem tudna erre a feladatra vállalkozni.

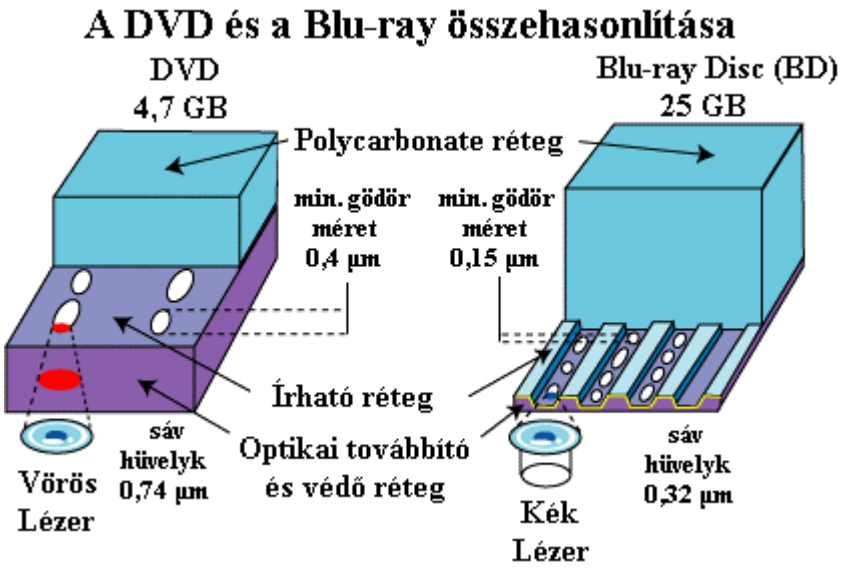
A nehézségek és pénzügyi gondok ellenére a világon több országban már megvalósult az internetes tartalmak feldolgozása, archiválása. Vannak olyan országok, ahol ezek a programok a nemzeti könyvtárhoz tartoznak, például a Svédi Királyi Könyvtár Kulturarw3 programja, vagy az Osztrák Nemzeti Könyvtár Austrian On-Line Archive (AOLA) programja. Más országokban pedig konzorciumokat alakítottak ezekre a feladatokra. Az utóbbi megoldásra példa az Egyesült Királyságban létrehozott UK Web Archiving Consortium, mely a British Library, a Joint Information Systems Commitee (JISC), a Walesi Nemzeti Könyvtár és a The Wellcome Library együttműködésével jött létre, vagy az Egyesült Államokban működő Minerva program, amely ugyan a Kongresszusi Könyvtár programja, de

együttműködnek az International Internet Preservation Consortium-al (IIP) is. A konzorcium a világ több mint harminc könyvtárát foglalja magába. Többek között a Katalán Nemzeti Könyvtárat, az Osztrák Nemzeti Könyvtárat és a British Libraryt is. A már működő Internet archívumok közül később néhányat részletesen is ismertetni fogok, csakúgy mint a magyar helyzetet.

3.3. Hardver és szoftver gyors fejlődése

Egy weboldal megjelenéséhez elengedhetetlen bizonyos hardver és szoftver eszközök megléte, hiszen a weboldalak számítógépen jelennek meg és ahhoz, hogy látható legyen a rajta lévő tartalom legalább egy web böngésző programra szükség van. Ám sok esetben a böngésző program már kevés, hiszen sok tartalom lejátszásához, megjelenítéséhez szükség van például Acrobat Readerre, vagy valamilyen médialejátszó programra is. Általában ezeket a programokat csak egyféle, specifikus környezethez alakítják és csak adott operációs rendszeren, adott hardverigény mellett tud működni.

Ezeknek az anyagoknak a tárolása számítógépen történik, így az archiválásuk, megőrzésük is csak elektronikus adathordozón lehetséges. Ezeknél az adathordozóknál jelenleg meghatározhatatlan, hogy mennyi ideig marad fenn használható formában. Jelenleg a legtöbb helyen az archív anyagokat mágnesszalagon, CD vagy DVD lemezen tárolják. Ezek közül a CD és DVD lemezek elég tartósak, de folyamatosan jönnek az újabb és újabb technológiák, amik már nem feltétlenül lesznek kompatibilisek az előzőekkel. A legújabb technológia a blu-ray, mely már egészen más formátumú és sokkal nagyobb mennyiségű anyag tárolására alkalmas, mint akár egy DVD lemez.



4. ábra *A DVD és a Blu-ray összehasonlítása*
(Forrás: http://hu.wikipedia.org/wiki/Blu-ray_Disc)

Az egyetlen bíztató dolog, hogy bár egy CD lejátszóval nem lehet DVD vagy Blu-ray lemezt lejátszani, de ha van egy Blu-ray lejátszónk, abban le tudjuk játszani mind a CD, mind a DVD lemezeket, mert általában az új eszközök lefelé kompatibilisek.

Sajnos az sem megoldás, hogy az anyagokat számítógépre archiváljuk, mert egy hardvereszköz élettartama körülbelül 3-10 év, viszont minél régebbi egy eszköz, annál nehezebb és drágább a pótlása.

Jelenleg a legbiztonságosabb megoldás, ha az archivált anyagokat időről időre új tárolóra mentik át.

3.4. Az eltérő weblap-típusok kérdésköre

A statikus weblapok archiválása nem különösebben nehéz feladat, hiszen az információk szöveges formában tárolódnak, így bármikor ugyanolyan könnyen készíthető róluk pillanatfelvétel, mint bármelyik word dokumentumról és ugyanolyan könnyű a változások követése is. Ennek ellenére mégsem teljesen problémamentes a HTML-ben írt lapokat archiválni, mert az egyszerű leírónyelv nem ad lehetőséget

arra, hogy megfelelően kódolni lehessen az anyagokat egyedi tulajdonságaiknak megfelelően. Erre fejlesztették ki az XML programnyelvet, mely igazából nem is egy nyelv, hanem egy metanyelv. „Az XML egyik legnagyobb értéke az, hogy a dokumentumokat használhatóvá teszi különböző környezetben és elrendezésben."[12]

A statikus oldalakkal szemben a dinamikus weboldalak archiválása igen bonyolult, mivel ezek az oldalak rugalmasan igazodnak a felhasználókhoz és azok igényeihez. Ebből adódóan a felhasználók sokszor nem ugyanazt látják, ha megnyitják ezeket az oldalakat. Archiválási szempontból a legkomolyabb akadályokat az jelenti, hogy lehetetlenség egy weblap összes változatát letárolni. Ráadásul az információk nagy része csak az adatbázisokban található meg, amelyeknek a tartalma folyamatosan változik és az adatbázisokhoz történő hozzáférés sem egyszerű. Ezeket a weboldalakat a keresőmotorok is csak akkor tudják feldolgozni, ha erre a tartalomszolgáltató lehetőséget biztosít.

3.5. Szerzői és személyiségi jogi szabályozások

A legtöbb ember azt gondolja, hogy az Interneten található anyagok szabadon letölthetők, felhasználhatók, pedig ezekre a tartalmakra is ugyanúgy vonatkoznak a szerzői jogok, mint bármely nyomtatott dokumentumra. Magyarországon a szerzői jogokra a többször módosított 1999. évi LXXVI. törvény a szerzői jogról (továbbiakban Szjt.) vonatkozik, mely többek között meghatározza, hogy mely műfajok tartoznak a szerzői jog védelme alá (pl. számítógépi programok, reklám célra megrendelt művek, adatbázisok stb.). „Az Szjt. alapján általánosságban szerzői jognak a szerző azon

[12] Nagymélykúti Balázs id. m.

jogát nevezzük, hogy a szerző művel kapcsolatos szerzői minőségét mindenki elismerje, és szabadon rendelkezzen műve felett."[13]

A szerzői jogi szabályozás – mint a legtöbb jog – alapvetően területi hatályú, azaz a mű első megjelenésének vagy megírásának a helyétől függ, hogy mely ország törvénye vonatkozik rá. Sajnos teljes jogharmonizáció még az EU-n belül sincs, de ma már „az egyes országokban érvényes szerzői jogi törvényeket a nemzetközi egyezményekben és szerződésekben foglalt elvek alapján fogalmazzák meg. A jelenleg érvényben lévő nemzetközi szerzői jogi törvények alapja a Berni Egyezmény néven ismert szerződés."[14] Ennek ellenére mégis vannak különbségek, hiszen míg az európai kontinens országaiban a szerzői jog a magánjogba tartozik, addig az angolszász országokban a közjogba. Ez azt jelenti, hogy az európai kontinensen „a szerzőknek elidegeníthetetlen jogaik vannak a szellemi alkotásukhoz. Más szóval ez állampolgári joguk. A közjog szerint azonban a szerzői jog nem alanyi joga az egyénnek: azt törvényben kell garantálni."[15]

A legtöbb szabályozás szerint a jogok a szerző halála (ismeretlen szerző esetén a nyilvánosságra hozatal) után 70 évig (adatbázisok, adattárak esetében az utolsó jelentős módosítástól számított 15 év) érvényben vannak, azaz addig nem lehet semmilyen formában sem nyilvánosságra hozni a műveket a szerző, vagy örököseinek engedélye nélkül. Ez a szabályozás nem könnyíti meg az internetes archívumok működtetését, mivel a szerzői jogok megsértéséért minden esetben a tartalomszolgáltatók – internetes archívumok esetében az archívum üzemeltetői – felelnek. Ezt a helyzetet a különböző archívumok különböző módon próbálják áthidalni. Az Internet Archive például sok esetben (weboldalaknál nem) egyfajta nyilatkozatot kér a szerzőtől

[13] Farkas Szabó András: A szerzői jog és az Internet In: Internet és Politika 2003. (3. évf.), 4. sz. http://iroga.hu/internet&politika/farkas.htm (2009. december 4.)
[14] A Pulman Digital Guidelines magyar változata: digitális útmutató kiemelt fejezetei http://www.ki.oszk.hu/old/pulman/dg/szerzoijog.html (2010. január 3.)
[15] A Pulman Digital Guidelines magyar változata: digitális útmutató kiemelt fejezetei id. m.

(Creative Commons License[16]) a mű feltöltése előtt, amelyben engedélyezi a felhasználás különböző formáit. Ugyanakkor az Ausztrál Nemzeti Könyvtár a PANDORA-val archivált tartalom szolgáltatására a szolgáltatóktól kér engedélyt. Viszont azok a tartalmak, amelyeket az Ausztrál Nemzeti Könyvtár megbízásából az Internet Archive gyűjt be, nem publikusak.

Néhány begyűjtött weblap tartalmazhat különleges (érzékeny személyes) adatokat, melyek felhasználhatóságáról, szolgáltathatóságáról a különböző országok személyiségi jogi törvényei rendelkeznek. Magyarországon ez a 1992. évi LXIII. törvény a személyes adatok védelméről és a közérdekű adatok nyilvánosságáról.

A felhasználás ellenőrzését az adatvédelmi hivatalok és az adatvédelmi biztosok ellenőrzik. Mivel ilyen adatokat csak a begyűjthető weblapok nagyon kis hányada tartalmaz, ezért ezzel részletesebben nem fogok foglalkozni.

[16] http://creativecommons.org/about/licenses/

4. A weboldalak begyűjtésének módszerei

A „webgyűjtés egy olyan rendszer, amely weboldalakat tölt le, új URL-eket szűr ki a meglévő HTML kódokból, és egy olyan listába teszi őket, amelyek további letöltésre várnak, ha még nincsenek meg”[17]. A weboldalak begyűjtésekor úgynevezett pillanatfelvételek készülnek az oldalakról. Ezek a pillanatfelvételek manuális és automatikus módon is begyűjthetők. Bármely módszert is használja az archívum, bele kell törődnie abba, hogy a pillanatfelvételek nem lehetnek folyamatosak, hiszen egy-egy begyűjtési periódus a hatalmas adatmennyiség miatt hosszú hónapokig is eltarthat.

A manuális módszer legnagyobb előnye, hogy célirányosan végezhető és lehetőséget teremt arra, hogy minél több releváns oldal kerüljön begyűjtésre. Ennél a módszernél viszont elengedhetetlen az emberi erőforrás. Ugyanakkor bármekkora embertömeg is áll rendelkezésre, – az Internet mérete miatt – képtelenség átnézni és begyűjteni az összes szükséges weboldalt.

Az automatikus begyűjtést általában websiklók, vagy más néven robotok végzik, melyek végignéznek minden elérhető linket és oldalt. A letöltött weboldalakat automatikusan indexelik, így biztosítva a minél gyorsabb kereshetőséget. Ezzel a módszerrel viszonylag rövid idő alatt nagy mennyiségű weboldalt lehet begyűjteni, de a websiklók – kis kivételtől eltekintve – csak a felszíni hálót tudják végigpásztázni. A nyílt forráskódú websiklók közül a legismertebb a HTTrack[18], mely egy ingyen letölthető weboldalmásoló és kapcsolat nélküli böngésző program.

[17] Jodelis, Remigijus: Elektronikus források begyűjtése és archiválása Litvániában: úton egy virtuális könyvtár felé. In: Tudományos és Műszaki Tájékoztatás 2004. (51. évf.), 6. sz. http://tmt.omikk.bme.hu/show_news.html?id=3640&issue_id=451 (2009. április 11.)
[18] http://www.httrack.com/

A webarchívumok leggyakrabban az automatikus begyűjtést alkalmazzák, amelyet a manuális módszerrel egészítenek ki. A websiklók eredményes működéséhez meg kell adni néhány információt, többek között azt, hogy mely oldalakat gyűjtse be (például minden .hu domain végződésű oldalt), mikor kell visszatérniük a változások ellenőrzése érdekében, hogyan kerüljék ki azokat az oldalakat, amelyek túlterhelnék a működésüket stb. Mindezek mellett koordinálniuk kell a párhuzamosan működő websiklók munkáit is.

A websiklók figyelembe tudják venni az oldal változásának a gyakoriságát is, amely hatékonyabb begyűjtési terv felállítását teszi lehetővé. Ehhez az első alkalommal csak az oldalfrissítéseket ellenőrzi – melyek általában megtalálhatók a szervereken –, de nem tölti le az oldalakat. A robot a következő alkalmakkor már sokkal gyorsabban begyűjti a szükséges információkat, melyekből folyamatos begyűjtés esetén az archiválók akár órára pontosan is meg tudják határozni a várható változtatások idejét. Ha a frissítési információkat nem lehet automatikusan beszerezni, akkor kénytelenek az archiválók folyamatosan letölteni a weblapokat és így felállítani egy becslést a frissítések gyakoriságáról.

A robotok a legszűkebb peremfeltételek beállítása ellenére is sokszor kezelhetetlen mennyiségű anyagot gyűjtenek össze. Ennek a problémának a kiküszöbölésére hozták létre a fókuszált websiklókat, más néven témasiklókat. Ezek olyan robotok, melyek csak a megadott témákba tartozó anyagokat gyűjtik be. Működésükkor értékelik az oldalakat, és ezzel lehetővé teszik, hogy a begyűjtés a web egy specifikus részére koncentráljon. Teljesítményük attól függ, milyen gazdag az adott téma linkgyűjteménye, amellyel a siklók dolgoznak. A témasiklók segítségével a webarchívumok mély és naprakész gyűjteményt tudnak létrehozni.

Az archívumok gyűjteményei a manuális és automatikus módszereken kívül egyéb módon is bővíthetők. Ilyen például a letéti

forma, amikor a weboldalt, vagy az arról készült pillanatfelvételt a weboldalakat működtető webmester az archiváló szervezet rendelkezésére bocsájtja. Így az archívumba kerülhetnek olyan oldalak is, melyekhez a websiklók nem férhetnek hozzá.

A begyűjtést végezheti egy csoport, egy helyen (centralizáltan), vagy több intézmény elsősorban tevékenységi profiljának megfelelően (decentralizáltan). Jelenleg a legnépszerűbbek a különböző együttműködések az adattárak, a nemzeti könyvtárak és az információk előállítói között, amellyel segíthetik egymás munkáját. Ez az együttműködés költséghatékony, megkönnyíti a szabványosítást és csökkenti annak a valószínűségét, hogy ugyanazt az adatot több szervezet is elmentse és tárolja.

Ugyancsak hasznos módszer, ha a begyűjtést végző szervezet az adattár tulajdonosától kér engedélyt az adattárban található információk begyűjtésére, az adatok szolgáltatására. Ily módon hozzáférhetnek az adatbázisok tartalmához is, azaz lehetőségük lesz a mély hálóból is adatokat begyűjteni. Mivel az adatbázisok tartalma első alkalommal lementésre kerül, később elég csak a frissítéseket nyomon követni, így kevesebb adatcserével is könnyen követhető a dinamikus tartalom változása.

5. Néhány külföldi példa a webarchívumok működésére

A későbbiekben bemutatandó archívumokat a következő szempontok alapján választottam ki: Internet Archive – a világ legnagyobb archívuma; Ausztrália – webarchívumuk megvalósítása, működése példa értékű; Egyesült Királyság – nagyon jó példa működő könyvtári konzorciumra; Norvégia – hazánkhoz hasonlóan kis ország, Litvánia és Szlovákia – Magyarországhoz hasonlóan kis területű, volt szocialista ország és végül Katalónia, amelynek külön érdekessége, hogy nem önálló ország, hanem csak egy tartomány.

5.1. Internet Archive

1996-ban indult San Franciscóban „egy keskeny, szűkös, alig kétszáz négyzetméternyi egykori katonai barakk épületében”[19]. Jelenleg non-profit szervezetként működik. Az Internet Archive megálmodója és létrehozója Brewster Kahle, aki egy új „Alexandriai Könyvtárat” szeretne létrehozni, hogy mindenki számára hozzáférhetővé váljon az univerzális emberi tudás. Az Internet Archive szorosan együttműködött az Alexa Internettel[20], melynek egyik alapítója ugyancsak Brewster Kahle volt, jelenleg viszont az Amazon.com tulajdonában van.

Az archívum gyűjteményében nem csak weboldalak találhatók, hanem digitalizált könyvek, zeneművek, filmalkotások, fényképek, televízió-műsorok és számítógépes programok is, melyet a dokumentum tulajdonosa saját maga is feltölthet. Az anyagok feltöltésekor a feltöltést végzőnek ki kell tölteni a Creative Commons

[19] Rév István: Alexandriai könyvtár a pincében. In: Budapesti Könyvszemle 2004. (16. évf.), 4. sz. http://epa.oszk.hu/00000/00015/00036/pdf/06prrev.pdf (2009. november 5.)
[20] Legfőképp az internetes oldalak forgalmának megbecsülését és rangsorolását tartalmazó weblapjáról ismert cég.

License-t, amelyben a szerzők műveik különböző felhasználási szintjeit engedélyezik (például kereskedelmi forgalomba hozható-e, vagy sem; változtatható-e vagy sem). Természetesen a felhasználás során a szerző jogai megmaradnak, így a nevét is fel kell tüntetni a mű felhasználásakor, de így megvan a lehetőség, hogy szabadon másolhassák, terjeszthessék a művet. Ezzel ki lehet küszöbölni a szerzői jogi problémákat.

Bár a weboldalakról készített pillanatfelvételek egyfajta korlenyomatok, az archívum mindenféle személyiségi és egyéb jogokat igyekszik tiszteletben tartani, ezért a chateket, üzenő falakat, e-maileket és más online üzenő felületeket nem archiválják.

Mivel San Francisco a földrengések szempontjából a világ egyik leginkább veszélyeztetett pontja, ezért létrehoztak egy tükör-szervert, a 2003-ban, norvég építészek által tervezett és az UNESCO támogatásával újraépült Alexandriai Könyvtárban (Egyiptomban), ahol az Internet Archive teljes állománya elérhető.

Az archívum gyűjteményét a Wayback Machine-n keresztül lehet elérni.

5.1.1. Wayback Machine[21]

A Wayback Machine az Internetről készült pillanatfelvételek archívuma, melyet az Alexa Internet mérnökei alkottak meg, és maga az Internet Archive is az Alexa Internet cég által összegyűjtött elektronikus anyagokra épül.

Az automata rendszer néhány hónap alatt végez a világ weboldalainak átfésülésével. Tapasztalatok szerint az oldalak mintegy fele megváltozik a Wayback Machine előző „látogatása” óta. A begyűjtés időpontjától számítva átlagosan 6-14 hónap közötti idő telik el addig, ameddig az archivált anyagokat a felhasználók is látják az archívumban. Az archívumot bárki ingyenesen elérheti és

[21] Szabad fordításban a szolgáltatás neve: időgép.

használhatja, de a felhasználási feltételekben rögzítetteknek megfelelően, másolatot nem készíthet az archivált anyagokról. Természetesen, ha valaki a saját weboldaláról szeretne másolatot kérni, és megkeresi az Internet Archive-t, akkor kivételt tesznek, megkapja a másolatot.

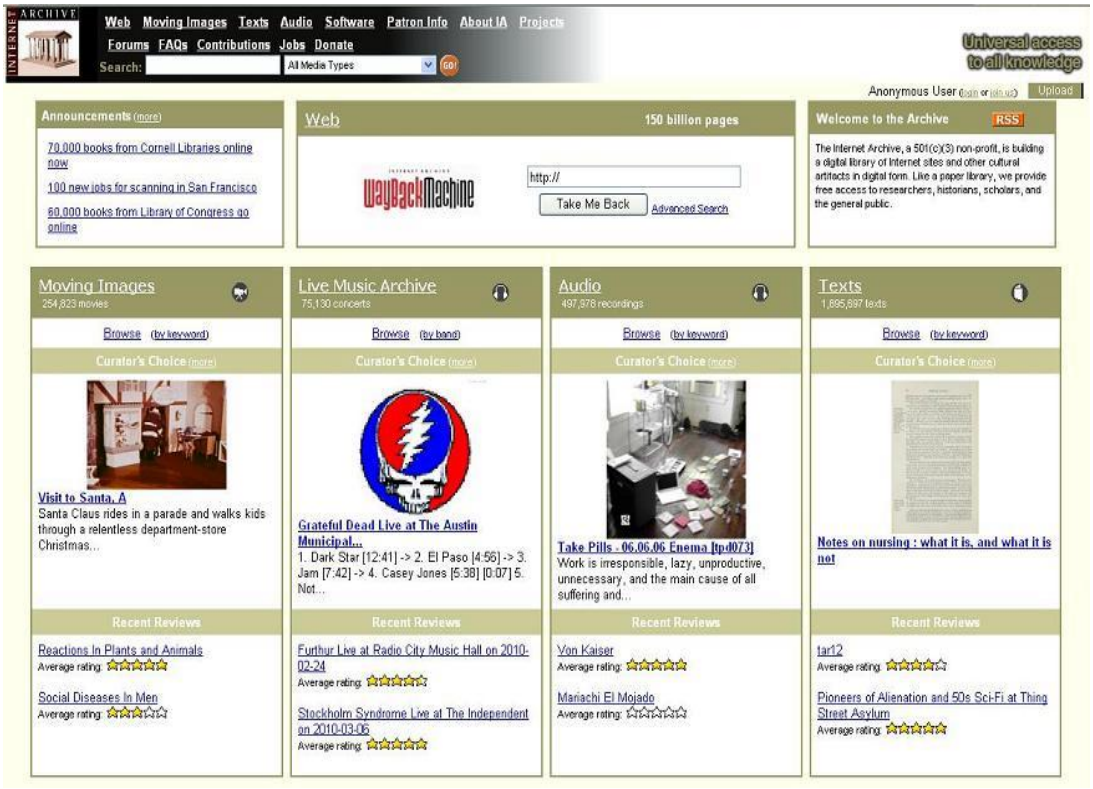
5.1.2. Archiválási problémák

A Wayback Machine sem tud minden oldalt archiválni. Ennek több oka is lehet, például az, hogy az oldal JavaScript-et, vagy egyéb olyan elemet tartalmaz, melyekhez szükség van a származtató szerverre, vagy az archiválandó oldal tulajdonosa elhelyezte az oldalán a robots.txt[22] fájlt, így ezek az oldalak nem kerülnek automatikus begyűjtésre. A Wayback Machine azokat az oldalakat sem archiválja, melyek mérete meghaladja a 10 MB-ot. Előfordulhat az is, hogy egy oldalt ugyan begyűjtöttek, de valaki úgy érzi, hogy az adott oldal sérti a személyiségi, vagy egyéb jogait. Ebben az esetben az érintett megkeresheti az Internet Archivot, és kérésre eltávolítják az adott oldalt/oldalakat.

5.1.3. Hozzáférés a gyűjteményhez

A gyűjteményhez a szervezet weblapján keresztül lehet hozzáférni.

[22] Ezt a fájlt más rendszerek is használják arra, hogy az adott oldal másolását, archiválását letiltsák.



5. ábra *Az Internet Archive gyűjteményének kereső felülete*
(Forrás: http://www.archive.org /)

A webarchívumban a keresett oldalakat az eredeti URL címe alapján lehet keresni. A többi archívumban viszont tartalom szerinti keresés működik.

5.1.4. A jövő

Komoly problémát okozhat az Internet gyors növekedése, már most havonta körülbelül 20 terabájttal növekszik az archívum tartalma, továbbá az egyre újabb és újabb technológiák megjelenése. Ezek későbbiekben arra kényszeríthetik az Internet Archive-t, hogy szelekciós begyűjtési módszert alkalmazzanak az eddigi tömeges begyűjtés helyett.

5.2. Ausztrália

1996-ban az Ausztrál Nemzeti Könyvtár kezdeményezésére, de több könyvtár együttműködésével jött létre a PANDORA program.

A program gyakorlati megvalósítása előtt rögzítették az on-line anyagok válogatási, súlyozási és begyűjtési szempontjait. A begyűjtendő anyagok között szerepelnek például olyan webhelyek – vagy azok részei –, amelyek lényeges vagy egyedi információt szolgáltatnak egy témáról, szervezetről, országos jelentőségű személyről, projektről vagy eseményről. Nem gyűjtik viszont a belső, intézményi célra létrehozott dokumentumokat, e-mail üzeneteket, zárt levelezőlisták anyagát, továbbá azokat az online dokumentumokat, amelyek nyomtatott formában is léteznek (többek között a nyomtatott napilapok online változatai). Hasonlóan nem kerülnek be a gyűjteménybe a blogok, a játékok, az egyéni szerzők cikkei, a hirdetések és az egyetemi szakdolgozatok.

A PANDORA tehát a szelekciós begyűjtést alkalmazza, így minden dokumentuma minősített, teljes mértékben katalogizálható, és ezért a nemzeti bibliográfia részévé válhat. Valamennyi dokumentum részben, vagy egészében hozzáférhetővé tehető Ausztráliában, és az archívumban lévő anyagok forrásainak tulajdonságai elemezhetők, meghatározhatók. A weboldalak archivált változatát rendszeres minőségvizsgálatnak vetik alá, hogy ellenőrizzék, minden fájlt megfelelően mentettek-e el.

„A webarchívum a Commonwealth hivatalos online kiadványait teljességgel gyűjti, de az állami kormányzatok dokumentumait válogatva. Az ausztrál szerzők műveit és az ausztrál vonatkozású online forrásokat válogatva az információtartalom, a kutatási érték, aktualitás és érdekesség alapján. Mintavétellel a következőket gyűjtik: egyéni honlapok, reklám-webhelyek, hirdetések. (Tervezeteket és még folyamatban lévő dokumentumokat nem regisztrálnak.) Az ausztrál Internet domainről pillanatfelvételeket készítenek.”[23]

[23] Dippold Péter: A nemzeti bibliográfiák gyűjtőköre, avagy elérhető-e a teljesség? In: Könyvtári Figyelő 2006. (52. évf.), 2. sz. http://www.ki.oszk.hu/kf/kfarchiv/2006/2/dippold.html (2010. február 28.)

5.2.1. Begyűjtés

A weboldalakat a PANDAS (PANDORA Digital Archiving System) nevű szoftverrel gyűjtik be, de a munkálatokat partnerintézmények is segítik. Mielőtt bármit is archiválnának tárgyalnak a szerzői jog tulajdonosaival és engedélyt kérnek a kiadóktól és a tulajdonosoktól.

Az aratás mélysége a weboldalaktól függ. Ha egy oldal túl nagy, akkor a PANDAS meg tudja oldani, hogy csak az oldal egy részét gyűjti be. Az Ausztrál Nemzeti Könyvtár a legteljesebb és legpontosabb adatmegőrzésre törekszik, ezért a PANDAS-al begyűjtött anyagoknál ellenőrzik, hogy minden szükséges fájl megvan-e, amelyhez egy Linkbot nevű program segítségét veszik igénybe. Ezen kívül manuális ellenőrzést is végeznek, hogy megnézzék, az oldal tökéletesen működik-e (például a JavaScriptek jól működnek-e).

Ezzel a módszerrel nem tudnak túl nagy mennyiségű anyagot összegyűjteni, ezért az Internet Archive és annak egy előfizetéses szolgáltatásának (Archive-It[24]) segítségével próbálják learatni az összes ausztrál tartalmat. Ám az így begyűjtött anyagok mennyisége olyan nagy, hogy azokat nem lehet olyan szisztematikusa ellenőrizni, mint a PANDAS szoftverrel összegyűjtött oldalakat.

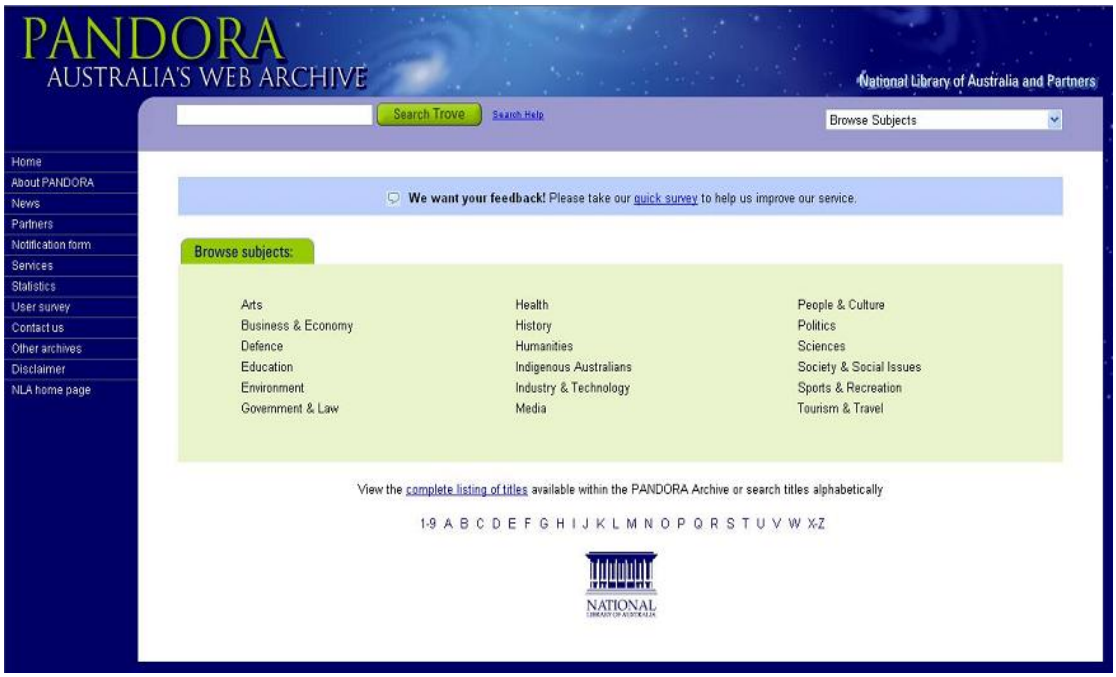
A mélyháló begyűjtésére kifejlesztettek egy nyílt forráskódú programot Xinq[25] néven, mely egy háló alapú hozzáférési eszköz. Feladata, hogy a strukturált adatbázisokat kutassa át és a másolatokat helyezze el az archívumban. Természetesen ez a szoftver sem tudja teljes egészében feldolgozni a mélyhálót, de legalább bizonyos információkat be tud gyűjteni.

5.2.2. Hozzáférés a gyűjteményhez

A PANDAS szoftverrel begyűjtött anyagok jogi helyzete tisztázott, hiszen a tulajdonosoktól, kiadóktól a könyvtár megkapta az engedélyeket, ezért ezeknek az anyagoknak nagy része on-line módon

[24] http://www.archive-it.org/
[25] http://www.nla.gov.au/xinq/

a PANDORA weboldalán[26] keresztül betűrendben, téma szerint, vagy kereső szerint hozzáférhető.



6. ábra *A PANDORA keresőfelülete*
(Forrás: http://pandora.nla.gov.au/)

A nem publikus anyagokhoz csak a felhasználók egy szűk rétege (kutatók) férhet hozzá a könyvtár épületén belül. A korlátozások okai különbözőek lehetnek, például személyiségi jogok, különböző kereskedelmi vagy kulturális okok.

A más forrásokból (például Archive-It) beszerzett anyagok esetében lehetetlen az összes anyagra engedélyt kérni a szerzőktől, ezért az összes ezzel a módszerrel begyűjtött oldal csak a kutatók számára a könyvtárban érhető el.

5.3. Egyesült Királyság

2004-ben konzorcium alakult az Egyesült Királyságban UK Web Archive[27] néven azzal a céllal, hogy megőrizzék az utókornak azokat a tudományos és kulturális értékeket, melyek csak weboldalakon

[26] http://pandora.nla.gov.au/
[27] http://www.webarchive.org.uk/ukwa/

léteznek. A konzorcium tagja 6 intézmény, többek között a Brit Nemzeti Levéltár, a Brit Nemzeti Könyvtár, továbbá a skót és a walesi nemzeti könyvtárak.

Első lépésként kidolgozták a készülő archívum céljait és irányelveit. Ilyenek voltak – többek között – a honlapok kiválasztási elve, a teljesen kereshető és böngészhető on-line webarchívum elkészítése és katalogizálása.

Maga a konzorcium és annak egyes partnerei is igyekeznek más webarchívumokkal, vagy különböző archiváló kezdeményezésekkel szoros kapcsolatos kiépíteni, hogy az ő tapasztalataikat is fel tudják használni. A konzorcium magát az infrastruktúrát és a PANDAS szoftvert is az Ausztrál Nemzeti Könyvtártól vette át, ennek ellenére a program a technológiától és a PANDAS szoftvertől független. Ha menet közben más megoldást választanak, minimális problémával át tudnak állni az új rendszerre.

5.3.1. Begyűjtés

Minden intézmény a saját szakterületéhez kapcsolódó oldalakat menti le, de a kiválasztást, gyűjtést és archiválást azonos módon, a PANDAS szoftver átalakított változatával végzik. Több szoftvert is kipróbáltak, de a PANDAS volt az egyetlen, amely „irányított környezetben a teljes munkafolyamatot felölelő szolgáltatást nyújtott”[28]. Bár az archiválandó oldalt minden tagintézmény saját maga választja ki, de az archiválás előtt az archiválónak ellenőriznie kell, hogy szerepel-e már az adatbázisban.

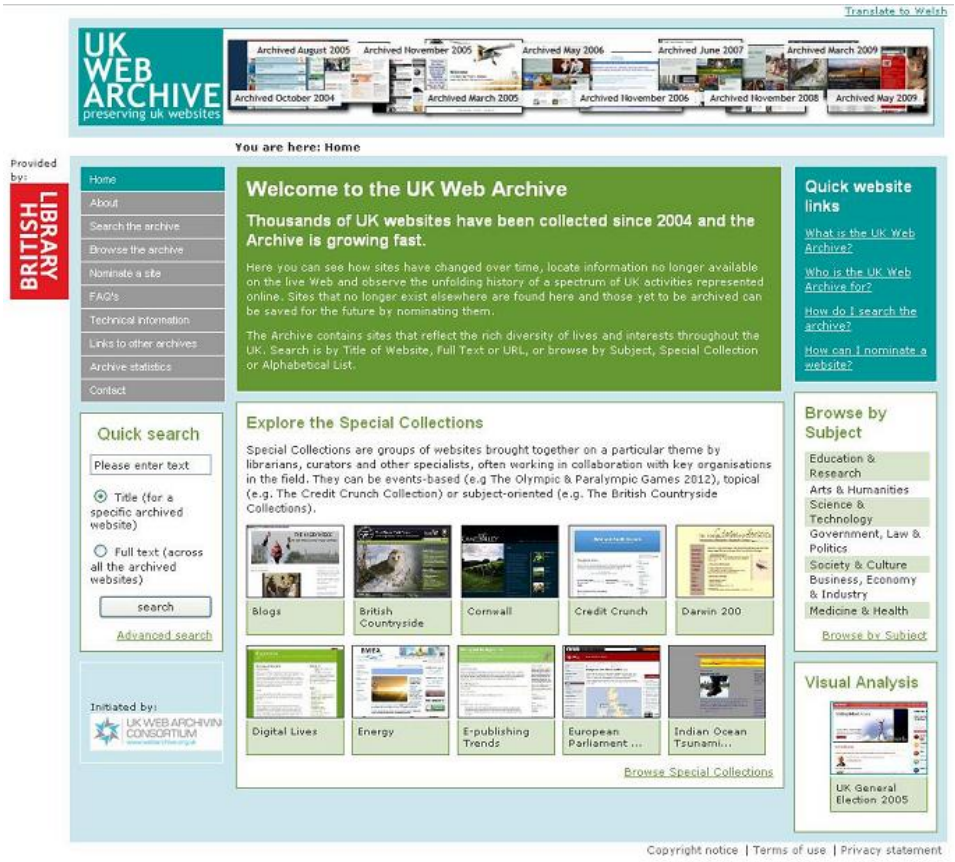
„Archiválás előtt a partnerek írásos engedélyt kérnek a honlapok tulajdonosaitól. Az engedélykéréshez azonos űrlapot használnak,

[28] Bailey, Steve – Thompson, Dave: E-tmt – Az első nyilvános webarchívum az Egyesült Királyságban. In: Tudományos és Műszaki Tájékoztatás 2006. (53. évf.), 10. sz. http://tmt.omikk.bme.hu/show_news.html?id=4555&issue_id=476 (2009. március 11.)

amelyet levéllel és a „Gyakori Kérdések Fájljával” látnak el.”[29] Így mindenki ugyanazt az információt kapja.

5.3.2. Hozzáférés a gyűjteményhez

Az archívum a konzorcium weblapján keresztül mindenki számára elérhető és bárki kérheti saját anyagainak ingyenes feltöltését.



7. ábra *UK Web Archive kereső felülete*
(Forrás: http://www.webarchive.org.uk/ukwa/)

A felhasználók téma, tartalom és a weblap címe alapján tudnak keresni a gyűjteményben. Az archívum a tartalom szerinti kereséshez a Lucene nevű keresőmotort használja.

5.3.3. Jövőbeli tervek

A konzorcium szeretné a begyűjtendő tartalom mennyiségét növelni, ezért folyamatosan gyűjtik a felhasználók tapasztalatait és ezeknek megfelelően alakítják a meglévő szoftvereket. A projekt

[29] Bailey, Steve – Thompson, Dave id. m.

weboldalát is fejlesztik, hogy a felhasználóknak minél jobb keresőképességet eredményezzen. Szeretnék, ha a felhasználók több különféle archívumban is keresni tudnának, ám ehhez szükséges az archívumok valamiféle átjárhatósága, összekapcsolása.

5.4. Norvégia

Norvégia a világ első olyan országa volt 1990-ben, ahol a frissen elfogadott kötelespéldány törvény kiterjedt az elektronikus publikációkra is. A törvény alapelve az, hogy minden általánosan elérhető információt, amit Norvég kiadó készített vagy kifejezetten a norvég közönség számára készült – tekintet nélkül a formátumára és a médiumra – archiválni kell, és a forrásnak elérhetőnek kell lennie kutatási és dokumentációs célokra. A törvény hatálya kiterjed az elektronikus publikációkra is – bármilyen adathordozón is vannak –, így abban a Norvég Nemzeti Könyvtár felhatalmazást kapott arra, hogy begyűjtse az e-dokumentumokat, beleértve a learatott weboldalakat is.

Ennek a törvénynek köszönhetően a Norvég Nemzeti Könyvtár 2001. augusztusban elindította a Paradigma-projektet azzal a céllal, hogy a norvég és számi (lappok) digitális dokumentumokat kötelespéldányként megőrizze, és hozzáférést biztosítson a felhasználóknak ezekhez az anyagokhoz.

5.4.1. Gyűjtemény építése

Kezdetben minden digitálisan elérhető dokumentumot begyűjtöttek a .no domainről és minden norvég vonatkozású információt a .com, a .org és a .net domainekről, mert nem lehet tudni, hogy később mire lesz szükségük a kutatóknak. Természetesen, ez bármikor szűkíthető, hiszen a könyvtárosoknak lehetőségük van automatikusan rangsorolt listák készítésére.

Jelenleg a Norvég Nemzeti Könyvtár több forrásból jut digitális dokumentumokhoz: Internetről automatikus gyűjtés, előfizetett

folyóiratok, levelezőlisták, kötegelt formában érkező adatbázis frissítések, fizikai adathordozók (például CD-ROM). A projekt kezdetétől a köteles példány osztály félautomata módon a HTTrack rendszer segítségével gyűjti a különböző eseményekhez kapcsolódó weboldalakat (például politikai pártok weboldalai választások előtt és után, királyi esküvő stb.). Ezen kívül elkezdték a különböző internetes folyóiratok gyűjtését, és tervbe vették néhány folyóirat teljes adatbázisának letöltését is.

5.4.2. Archívum kialakításának fő kérdései

Az egyik fontos kérdés az volt, hogy mely felhasználói csoportoknak készítsék az archívumot, és ezek a felhasználók mit fognak majd keresni benne. Mivel a kutatók lehetséges kérdéseit elég nehéz előre meghatározni, ezért két lehetséges csoportba sorolták azokat. Így már könnyebben meg tudták határozni: mit és hogyan gyűjtsenek.

Míg az első csoport tagjai a dokumentumot, mint médiumot tanulmányozzák és a nyelvhasználatra, technológiai fejlődési trendek, valamint a tartalom közötti összefüggésre, továbbá a webdizájnra lehetnek kíváncsiak. A másik csoport a különböző tudományterületek képviselőit tömörítheti, akik a dokumentumokat forrásanyagként vizsgálnák. Számukra nem a külső, hanem a kereshetőség és az információ, a tartalom lesz a lényeges.

5.4.3. A program lezárása és eredményei

A Paradigma program 2004. decemberében zárult le. Az elkészült archívum a Norvég Nemzeti Könyvtár weblapján keresztül elérhető és kereshető.

5.4.4. A webarchívum jelene

A Norvég Nemzeti Könyvtárnak írott e-mailemre kapott válasz alapján megtudtam, hogy az előzőekben leírt program eredeti formájában sosem valósult meg.

Jelenleg a webaratásokhoz a Heritrix[30] domain aratót használják, a weblapok szelektív begyűjtéséhez pedig a Web Curator Tool-t[31]. Az aratást általában évente 1-2 alkalommal végzik. Mivel a weblapok különleges adatokat is tartalmazhatnak, ezért a begyűjtéshez és a szolgáltatáshoz is az adatvédelmi biztos felhatalmazása szükséges. A Norvég Nemzeti Könyvtár a begyűjtéshez időszakos engedéllyel rendelkezik, melyet rendszeresen meg kell újítani. Jelenleg éppen tárgyalást folytatnak az adatvédelmi hivatallal, hogy meghosszabbítsák engedélyüket, így a webaratás szünetel. A learatott anyagokat viszont a könyvtár nem tudja szolgáltatni, mert egyelőre senkinek sincs engedélye, hogy megtekintse azokat.

Az általános gyűjtésen túl megpróbálják növelni a weblapok szelektív gyűjtését és az egyedi dokumentumok (például .pdf) gyűjtését is, melyeket eddig csak kis mértékben gyűjtötték. Egyedi dokumentumok lehetnek többek között tudományos jelentések, sorozatok, e-újságok és e-könyvek. 2008-tól a gyűjtés magába foglalja azoknak a norvég cégeknek a weblapjait is, akik nem tartoznak a .no domain alá. A norvég társadalmi és kulturális élet dokumentumai közül csak 8 on-line hírlapot és kismennyiségű Interneten publikált zenét gyűjtöttek be. A téma alapú begyűjtés eddig nem élvezett prioritást, de ennek főleg erőforráshiány volt az oka. Ugyancsak részben erőforráshiány, másrészt pedig technikai nehézségek miatt a blogokat sem gyűjtik. Az országos jelentőségű eseményekhez kapcsolódó oldalak begyűjtését továbbra is fontosnak tartják, így az ezekkel

[30] http://crawler.archive.org/
[31] http://webcurator.sourceforge.net/

kapcsolatos munkálatok (például gazdasági válság, Knut Hamsun norvég író születésének 150. évfordulója) továbbra is folynak.

Sajnos az angol nyelvű weblapjukon a webarchívum rész éppen átdolgozás alatt van, így a begyűjtött anyagok megjelenítési módjáról nem tudok írni.

5.5. Litvánia

Az Elektronikus Források Archívumának (webarchívum) működtetése a Litván Nemzeti Könyvtár feladata. Az archívum jogi alapját „a kiadványok kötelespéldány-másolatainak és egyéb dokumentumok elosztásának rendjéről" szóló 1996. évi kormányrendelet teremtette meg, amely nem fedte le az elektronikus dokumentumokat, forrásokat. Ennek a hiányosságnak a pótlására külön meg kellett határozni, hogy mely tartalmakat gyűjtheti a webarchívum (például az .lt domain névvel azonosított oldalakat, kiadók által jóváhagyott és hivatalosan bejegyzett összes elektronikus kiadványt) és melyeket nem (például magánszemélyek nem hivatalos dokumentumai, e-mailok, levelezőlisták).

5.5.1. Gyűjtemény építése

A program kezdetétől 2007-ig az Elektronikus Kiadványok Letéti Rendszerének modelljére és a NEDLIB (Letéti Könyvtárak Európai Hálózata) program dokumentumaira épült. Begyűjtéshez a NEDLIB gyűjtő programot használták, de annak folyamata nem volt hatékony. A meta-adatokat és a forrás fájlokat a program külön tárolta, ezért az adatbázisban a visszakeresés hosszú időt vett igénybe, továbbá a szoftverekhez sem volt támogatás. A szoftvereket 2008-ban lecserélték, azóta a begyűjtéshez a NetarchiveSuite[32], míg az adatok indexeléséhez, publikálásához az Internet Archive-nál már megismert Wayback Machine-t használják. Az új rendszer az 'ARC'

[32] http://netarchive.dk/suite

fájlformátumot[33] használja, ami nem csak a fájlt, hanem a hozzá tartozó adatokat (például katalógus adat) is tartalmazza.

Az eredeti tervek szerint a gyűjtés az összes .lt domainre és a kapcsolódó weboldalakra vonatkozott, mely kb. 300 .com és .net végződésű litván oldalt is tartalmazott, és ahonnan csak a statikus dokumentumokat akarták begyűjteni.

Az első ciklus 2002. októberben indult, mellyel az elérhető dokumentumok nagy részét összegyűjtötték. A második ciklust (2002. novemberben) már azzal a céllal indították, hogy szelektíven összegyűjtse az időszakos webkiadványokat, de hamar szembesültek azzal, hogy a legtöbb weboldal dinamikus formában készült és ezeket nem könnyű begyűjteni. Ezért úgy döntöttek, hogy elsőbbséget élveznek a közvetlenül a kiadóktól megszerezhető elektronikus kiadványok.

2002. október és 2003. május között négy gyűjtési ciklus volt, melyek eredménye olyan nagy mennyiségű anyag lett, hogy a későbbiekben már csak évi két gyűjtési ciklust terveztek, és ezt tartják a mai napig is.

Az így összegyűjtött információk közül kiválogatták azokat, amelyek a válogatás kritériumainak megfeleltek, majd indexelték és katalogizálták őket. Az ismérveknek megfelelő rekordok bekerültek a Nemzeti Bibliográfiai Adatbázisba is. Az oldalakról a bibliográfiai rekordok UNIMARC formátumban készülnek a Dublin Core szabvány alkalmazásával.

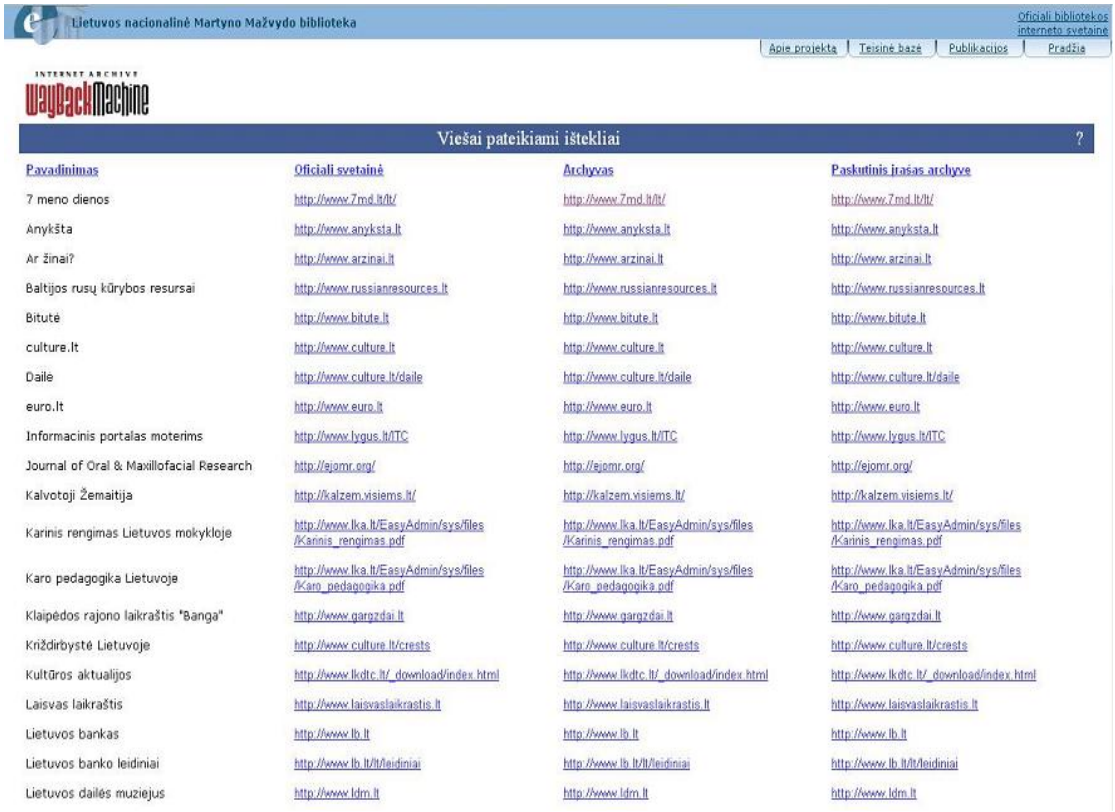
5.5.2. Hozzáférés a gyűjteményhez

A felhasználók egy weblap[34] segítségével férhetnek hozzá az archívum tartalmához, melyek lehetnek online és offline források. A weblapot még jelenleg is fejlesztik, változtatják.

[33] http://www.archive.org/web/researcher/ArcFileFormat.php
[34] http://eia.libis.lt/

Az archivált dokumentumok nagyobb része offline forrás, melyeket különböző kategóriákba rendeztek, például e-könyvek, magazinok, hírlapok, napilapok, értekezések és cikkek.

Az online források négy oszlopban jelennek meg.



8. ábra *Elektronikus Források Archívumának online források keresőfelülete (Forrás: http://eia.libis.lt:8080/archyvas/viesas/istekliai.jsp)*

Az első oszlopban a forrás neve, a másodikban a hivatalos weboldal linkje, a harmadikban az archivált oldal linkje és a negyedik oszlopban az archívumban található utolsó forrás linkje látható.

5.6. Szlovákia

Északi szomszédunk 2006. tavaszán kezdte meg a web minőségi és mennyiségi felmérését, melynek eredményeként 2006. májusában 92 ezer regisztrált szlovák nemzeti domain nevet (.sk) találtak[35].

[35] Androvič, Alojz: Web-archívum made in Slovakia: Kísérleti projekt az elektronikus információforrások gyűjtésére és archiválására. In: Tudományos és Műszaki Tájékoztatás

5.6.1. Kísérleti projekt

Az általános felmérés mellett a Pozsonyi Egyetemi Könyvtárban kísérleti projekt indult azzal a céllal, hogy összegyűjtsék az ISSN számmal rendelkező webforrásokat. 260 ilyen forrást találtak melyekből 164 csak elektronikus formában érhető el.

A tényleges archiválásra viszonylag szerény hardvereszköz-állomány állt rendelkezésre, melyeken nyílt forráskódú szoftvereket (például Debian GNU/Linus operációs rendszert) használtak.

Összesen 34,5 GB-nyi anyagot archiváltak, amelynek túlnyomó része HTML és JPEG formátumú volt, de ezen kívül GIF, PDF, TXT és DOC stb. formátumú anyagok is megtalálhatók voltak az archivált dokumentumok között.

5.6.2. Eredmények

A felmérés és a kísérlet eredményeként fogalmazták meg azokat a stratégiai célokat, amelyek az internetes források begyűjtéséhez és megőrzéséhez feltétlenül szükségesek:

- az elektronikus források kötelespéldányainak feldolgozásához ki kell alakítani egy önálló rendszert, melyhez szükséges a törvényi háttér megteremtése is: kötelespéldányok, szerzői jog stb.;
- ki kell dolgozni az elektronikus források hosszútávú megőrzésének módszereit és az időszakosan megjelenő elektronikus források azonosítását;
- ki kell dolgozni a nemzeti domaineken megjelenő elektronikus források begyűjtésének, archiválásának és a webarchívum létrehozásának szervezeti és a technikai feltételeit.

Szerettem volna megtudni, hogy a 2006-ban meghatározott stratégiai célokat sikerült-e megvalósítani, illetve elkezdték-e a

---

2007. (54. évf.), 10. sz. http://tmt.omikk.bme.hu/show_news.html?id=4788&issue_id=487 (2009. március 11.)

webarchívum építését, ezért e-mailt írtam a Szlovák Nemzeti Könyvtárnak és a Pozsonyi Egyetemi Könyvtárnak, amelyben az előbb leírt kérdéseket tettem fel. A kérdéseimre a dolgozat beadási idejéig nem érkezett válasz.

5.7. Katalónia

Bár Katalónia csak egy tartománya Spanyolországnak, mégis fontosnak tartják saját bibliográfiai örökségük gyűjtését, feldolgozását és terjesztését. Ennek köszönhetően a Katalán Könyvtár 2005. júniusában elindította a Katalónia Digitális Öröksége Programot (Patrimoni Digital de Catalunya = PADICAT) Programot, mely magába foglalja minden olyan digitális tartalom összegyűjtését, feldolgozását, állandó elérhetőségét, amelyet a katalán közösségnek szántak. Ez az összes olyan katalán vagy más nyelvű oldalt jelenti, mely a .cat domain alatt van, valamint a földrajzilag, vagy tematikusan Katalóniához köthető.

A projekt 2005-ben előkészítő, tervező szakasszal kezdődött, ahol többek között elemezték a részprojekteket, a rendelkezésre álló erőforrásokat, azokat a szervezeteket, akik megőrzendő katalán weboldalakat készítenek, valamint az egész projekt jogi vonatkozásait.

2006-2008-ban tesztelték a meghatározott módszereket és magát a rendszert. Ezen kívül teszt jelleggel elkezdték begyűjteni a katalán webet.

2009-re összeállt a rendszer és 100%-os teljesítménnyel üzemel. Ezzel a Katalán Könyvtár olyan rendszer hozott létre, amely nemcsak Spanyolországban, hanem egész Európában példaértékű.

A PADICAT által használt összes program ingyenes, nyílt kódú nonprofit cégek, szervezetek (az International Internet Preservation Consortium tagjai) készítették el. A Katalán Könyvtár is az IIPC tagja.

5.7.1. Gyűjtemény építése

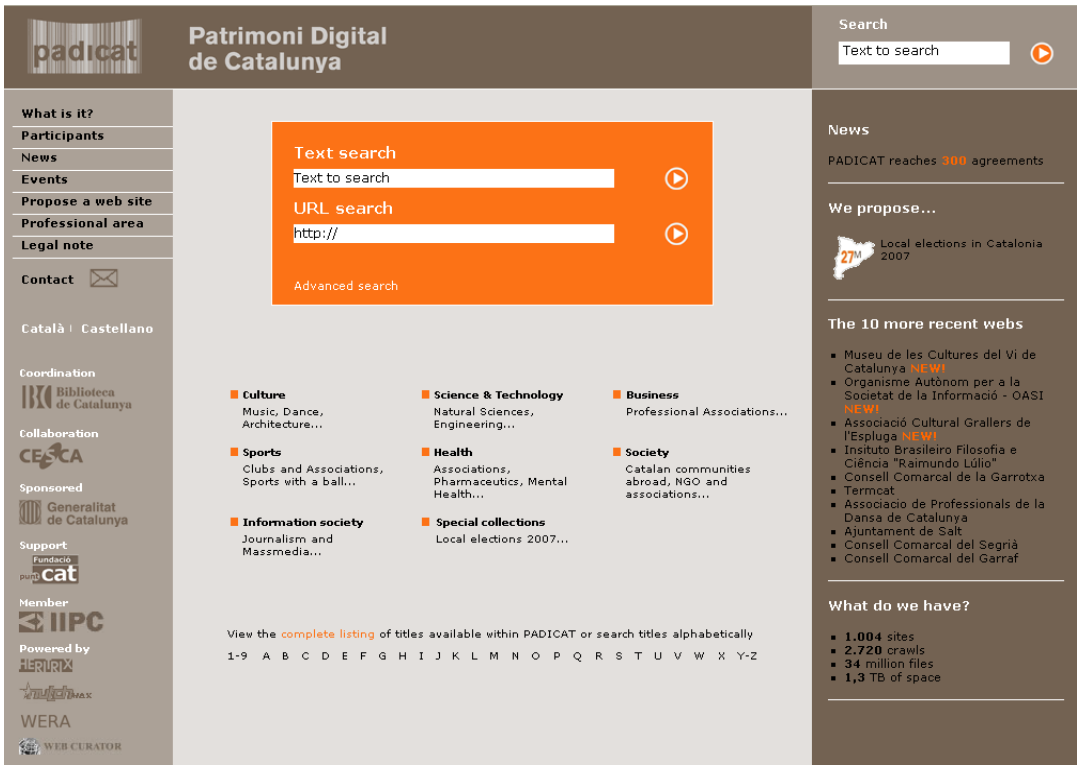
A Katalán Könyvtár, más nemzeti könyvtárak gyakorlatának megfelelően, hibrid begyűjtési módszert használ. Ez azt jelenti, hogy az Internetről tömegesen gyűjtik a nyílt, digitális anyagokat, de emellett leszerződnek különböző katalán intézményekkel is, hogy rendszeresen megkapják a saját készítésű webes anyagaikat.

A weboldalak begyűjtéséhez a Heritrex programot használják és a begyűjtött dokumentumokat 'ARC' formátumban tárolják. Az összegyűjtött anyagokat a megfelelő programmal indexelik, így később könnyebb a gyűjteményen belüli visszakeresés.

Az összegyűjtött forrásokat a Web Curator Tool program segítségével katalogizálják.

5.7.2. Hozzáférés a gyűjteményhez

Az archívumot a PADICAT[36] program weboldalán keresztül lehet elérni:



9. ábra *PADICAT keresőfelület*
(Forrás: http://www.recercat.cat/bitstream/2072/9258/2/A4.pdf)

[36] http://www.padicat.cat/

Több mit 300 szervezettel van különböző stádiumban lévő együttműködési szerződésük, így a gyűjtemény jelentős része on-line, nyílt hozzáférésű.

Az archívum tartalmához betűrendes, vagy téma szerint keresővel lehet hozzáférni. A tematikus integrációnak köszönhetően lehetővé vált a katalán közélet egyes eseményeinek kutatása. Ezen kívül lehetőség van a kulcsszavas szűrésre, továbbá a Wayback segítségével, az adott weblap URL-je segítségével közvetlen keresésre is van lehetőség.

5.7.3. További fejlesztések

Mivel a jelenleg használt hardver és szoftver nem teszi lehetővé a teljes katalán web archiválását, ezért a jövőben egyik legfőbb feladatuknak tartják a növekedési képesség javítását, a munkafolyamatok és a rendelkezésre álló erőforrások optimalizálását, amelyhez először fel kell mérni a jelenleg rendelkezésre álló infrastruktúrát. Ezen kívül rendszeresen át kívánják vizsgálni a katalán webet, és felmérik azokat a formátumokat, melyek már rövidtávon is olvashatósági problémákat okozhatnak. Ezeknek az ismeretében majd ki kell dolgozniuk valamilyen eljárást arra, hogy azokat hosszú távon is használni tudják. Ugyancsak fontos feladatnak tartják a sorozatkiadványok rendszeres begyűjtésének kidolgozása is.

A Katalán Könyvtár megpróbálta egységesíteni a PADICAT szoftvereivel készített bibliográfiai rekordokat a más rendszerek által készítettekkel. Tapasztalatuk azt mutatja, hogy egyelőre nem lehet hatékonyan kicserélni, és így integrálni a meglévő adatbázisokba az egyes rekordokat. A próbálkozás további fontos eredménye: az egységes nyelvek használata, valamint az átjárók kiépítése olyan feladat, melyet a többi archívummal közösen kell elvégezniük.

6. Magyar Internet Archívum

A Magyar Internet Archívum (MIA) terve – amely Drótos László kezdeményezéséhez fűződik – 2006-ban vetődött fel először a nyilvánosság előtt. Az archívum ötlete abból indult ki, hogy „a hagyományos dokumentumok digitalizálásának szükségességét már evidenciának tekintik a hazai közgyűjteményekben és – a lehetőségek függvényében – folyik is ez a tevékenység. De közben egyre nagyobbra nyúlik az a "sötét" korszak, amiről semmilyen lenyomat nem marad ezekben a gyűjteményekben, mert a fontos webhelyek archiválását nem sikerült megszervezni."[37]

A konkrét elképzelések ellenére a MIA a mai napig sem került megvalósításra, sőt – a releváns szakirodalom áttanulmányozása után állítható – hogy a könyvtáros szakmai közéletben is kevés szó esik róla.[38] Pedig Rácz Ágnes szerint az OSZK „technikailag már felkészült akár a webaratásra is"[39], tehát nem lehet a technikai feltételek hiányára hivatkozni amiatt, hogy MIA-ügyben egyelőre nem történt előrelépés. Mindemellett az Internet archívumok fontosságát már az UNESCO is megfogalmazta: „A digitális örökség megőrzése a kormányok, alkotók, kiadók, releváns iparágak és az örökségvédelmi intézmények kitartó erőfeszítéseit igényli."[40] Ennek ellenére ma Magyarországon a törvényhozók és az illetékes hatóságok nem érzik át egy webarchívum létrehozásának szükségességét. Ellenkezőleg. Szinte elzárkóznak attól, mert úgy gondolják, amennyiben létezne egy webarchívum, úgy

[37] Drótos László: Egy gondolat az internet archiválásról. In: Katalist 2009. december 21. https://listserv.niif.hu/pipermail/katalist/2009-December/019825.html (2010. január 10.)
[38] Ezt bizonyítja az is, hogy a könyvtáros szaksajtó szinte alig foglalkozik a kérdéssel, sőt a könyvtárügy stratégiai fejlesztési dokumentumaiban sem található erre vonatkozó utalás. Ez az oka annak is, hogy a MIA-val kapcsolatos diskurzus bemutatásánál nem az elsődleges, hanem a másodlagos nyilvánosságra támaszkodhattam.
[39] Rácz Ágnes: Egy gondolat az internet archiválásról. In: Katalist 2009. december 21. https://listserv.niif.hu/pipermail/katalist/2009-December/019827.html (2010. január 10.)
[40] Magyar Unesco Bizottság: Charta a digitális örökség védelméről http://www.unesco.hu/informacio-kommunikacio/digitalis-orokseg/charta-digitalis-orokseg (2010. március 14.)

„nagy pénz lehet benne: a vitákhoz, sajtóperekhez, cikkekhez, könyvekhez elő-elő kapni a múltat... szolgáltatni belőle”[41].

6.1. Drótos László tervei a Magyar Internet Archívum létrehozására

A külföldi tapasztalatok azt mutatják, hogy egy ilyen archívum létrehozása nagy és komplex feladat, hiszen a web tartalmát a begyűjtés után fel is kell dolgozni. Ezen kívül kezelni kell a felmerülő műszaki és jogi problémákat is. Megvalósítását, finanszírozását egyetlen intézmény sem tudja egymaga vállalni, ezért célszerű lenne ezekre a feladatokra konzorciumot létrehozni. A konzorcium tagjai lehetnének például informatikai intézmények, egyetemi tanszékek, cégek stb.

Az archívum feladatainak ellátásához a szükséges technikai feltételek jelenleg is adottak. A szoftverek egy része is rendelkezésre áll (például saját fejlesztésű kereső), a további szoftverek pedig részben szabadon hozzáférhetők, részben az IIPC (International Internet Preservation Consortium) rendelkezésre bocsájtaná, ha Magyarország csatlakozna hozzájuk. Ezek után már csak a kormányzati akaratra, a szükséges jogi környezetre és költségvetési támogatásra lenne szükség.

Az archívum létrehozásához célszerű különböző munkacsoportokat felállítani. Elsőként egy előkészítő csoportra lenne szükség, amely javaslatokat tenne a konzorcium tagjaira és elkészítené a stratégia tervet. A konzorcium megalakítása és a

[41] Kokas Károly: Egy gondolat az internet archiválásról. In: Katalist 2009. december 21. https://listserv.niif.hu/pipermail/katalist/2009-December/019823.html (2010. január 10.)

stratégiai célok elfogadása után már felállíthatók lennének az újabb munkacsoportok[42], melyek már a konkrét munkához kapcsolódnának:

A válogatással és lehatárolással foglalkozó munkacsoport: az ő feladatuk lenne annak meghatározása, hogy ki legyen a felelős a válogatásért. Ezen kívül meg kellene határozniuk a magyar web kiterjedését (például .hu domain és a magyar tartalmat szolgáltató szerverek), továbbá együtt kellene működniük a domain szolgáltatókkal, nyilvántartásba kellene venniük a magyar webtérbe tartozó szervereket. Meg kellene még határozniuk a begyűjtés mélységét és a begyűjtendő objektumok típusait.

A begyűjtés és tárolás technikai kérdéseivel foglalkozó munkacsoport: át kellene nézniük és tesztelniük kellene a rendelkezésre álló arató robotokat, valamint ki kellene választaniuk a feladatnak leginkább megfelelőt. A nemzetközi szabványoknak és gyakorlatoknak megfelelően ki kellene dolgozniuk a begyűjtött dokumentumok tárolásának technikáját is.

A metaadatok kérdéseivel foglalkozó munkacsoport: javaslatokat kellene kidolgozniuk a mentett honlapok metaadatainak mentésére, mint például arra, hogy ki lássa el a weblapot Dublin Core leírással.

A hasznosítás/szolgáltatás kérdéseivel foglalkozó munkacsoport: át kellene tekinteniük, hogy az archívumba kerülő anyagokat milyen formában lehet hasznosítani; fel kellene mérniük a felhasználói igényeket; piackutatást kellene végezniük; meg kellene becsülniük a várható forgalmat és a lehetséges bevételek nagyságát.

A jogi kérdésekkel foglalkozó munkacsoport: át kellene tekinteniük az Internet archiválásával kapcsolatos jogi problémákat (például kötelespéldány-törvény kiterjesztése az internetes dokumentumokra, szerzői és személyiségi jogi törvények stb.), és különböző

[42] A munkacsoportok neveit a következő cikkből vettem: Drótos László: Mi a MIA? – Javaslat egy Magyar Internet Archívum létrehozására. In: Tudományos és Műszaki Tájékoztatás 2006. (53. évf.), 6. sz. http://tmt.info.omikk.bme.hu/show_news.html?id=4431&issue_id=473 (2009. március 9.)

szerződésterveket kellene kidolgozniuk a weblapok archiválásához (weboldal tulajdonosának, archívumnak a jogai és kötelességei).

A finanszírozás kérdéseivel foglalkozó munkacsoport: feladata a célzott támogatások és pályázati lehetőségek felkutatása, valamint az archívum működtetéséhez szükséges pénzösszegek előteremtése lenne.

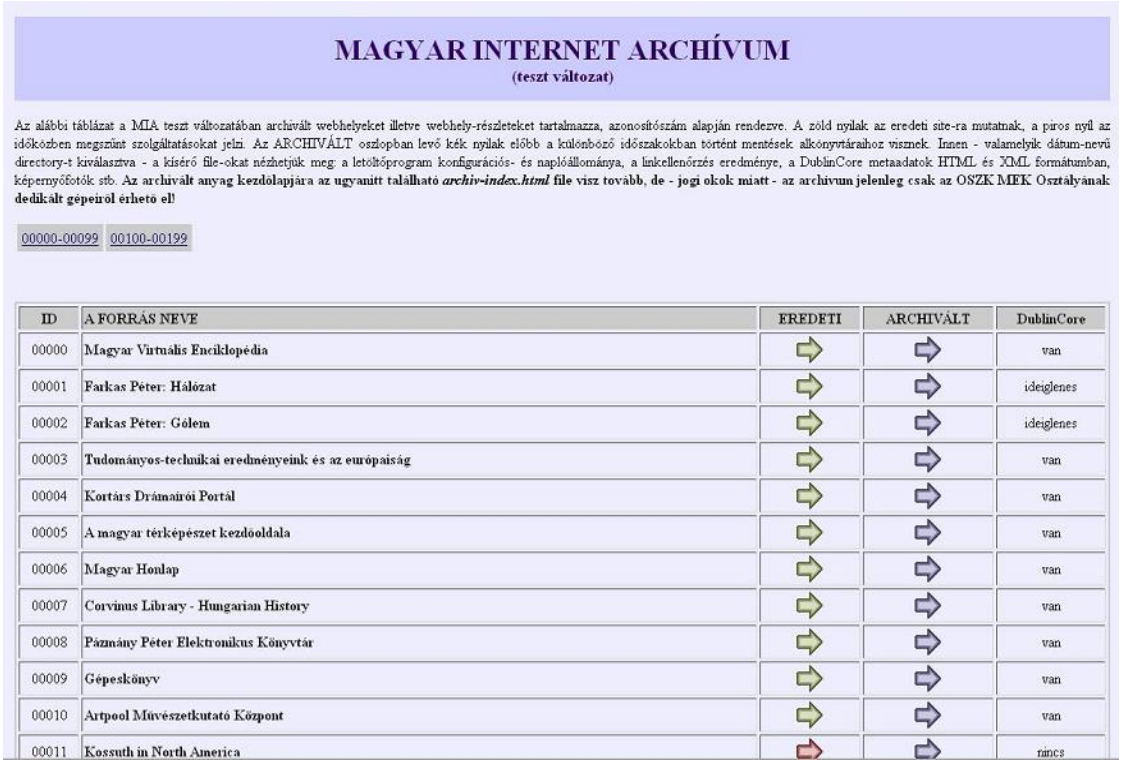
Egy jó webarchívum gyűjteményére közhasznú és üzleti célú szolgáltatások is építhetők (például tematikus összeállítások eseményekre, évfordulókra, vagy szöveg- és adatbányászati rendszerek, tématérképek felépítése), amelyekkel el lehet érni, hogy hosszú távon az archívum eltarthassa saját magát.

6.2. Kísérletek a MIA létrehozására

Drótos László néhány évvel ezelőtt megpróbált egy kis archívumot[43] (magángyűjteményt) létrehozni az OSZK egyik szerverén. Az erre a célra használható tárhely betelt és elfogyott a szabadidő is, így a kísérlet függőben maradt. Az „archívumban” jelenleg 189 weblap található, melynek tartalmát időnként frissíti.

A gyűjtemény megjelenítési felülete nagyon egyszerű és átlátható. Keresési lehetőség nincs benne, csak egy táblázatba szerkesztett lista az egész ahol a bekerülés sorrendjében láthatóak a weblapok címei.

[43] http://mekosztaly.oszk.hu/mia/



MAGYAR INTERNET ARCHÍVUM
(teszt változat)

Az alábbi táblázat a MIA teszt változatában archivált webhelyeket illetve webhely-részleteket tartalmazza, azonosítószám alapján rendezve. A zöld nyilak az eredeti site-ra mutatnak, a piros nyíl az időközben megszűnt szolgáltatásokat jelzi. Az ARCHIVÁLT oszlopban levő kék nyilak előbb a különböző időszakokban történt mentések alkönyvtáraihoz visznek. Innen - valamelyik dátum-nevű directory-t kiválasztva - a kíséró file-okat nézhetjük meg: a letöltőprogram konfigurációs- és naplóállománya, a linkellenőrzés eredménye, a DublinCore metaadatok HTML és XML formátumban, képernyőfotók stb. **Az archivált anyag kezdőlapjára az ugyanitt található *archiv-index.html* file visz tovább, de - jogi okok miatt - az archívum jelenleg csak az OSZK MEK Osztályának dedikált gépeiről érhető el!**

00000-00099 00100-00199

ID	A FORRÁS NEVE	EREDETI	ARCHIVÁLT	DublinCore
00000	Magyar Virtuális Enciklopédia			van
00001	Farkas Péter: Hálózat			ideiglenes
00002	Farkas Péter: Gólem			ideiglenes
00003	Tudományos-technikai eredményeink és az európaiság			van
00004	Kortárs Drámairói Portál			van
00005	A magyar térképészet kezdőoldala			van
00006	Magyar Honlap			van
00007	Corvinus Library - Hungarian History			van
00008	Páznány Péter Elektronikus Könyvtár			van
00009	Gépeskönyv			van
00010	Artpool Művészetkutató Központ			van
00011	Kossuth in North America			nincs

10. ábra *MIA teszt változatának kereső felülete*
(Forrás: http://mekosztaly.oszk.hu/mia/)

A táblázat oszlopaiban látható a weblapok azonosítója, neve, egy link az eredeti változathoz és egy másik, amely az archívumban található helyére mutat. Az utolsó oszlopban látható, hogy az adott weblap fel lett-e dolgozva, van-e már Dublin Core-adata. Az weboldalak archivált változatai a szerzői jogi gondok miatt csak egyes OSZK gépekről érhetők el.

A Szegedi Egyetemi Könyvtár ugyancsak kísérletezett a MIA létrehozásával, amikor megpróbáltak a TÁMOP-tól (Társadalmi Megújulás Operatív Program) pályázati támogatást nyerni, annak érdekében, hogy legalább elkezdődhessen a program. A pályázat mellett aláírtak egy keretmegállapodást a NIIF-el (Nemzeti Információs Infrastruktúra Fejlesztési Program) és az OSZK-val. A NIIF adta volna a hardvert az archívumhoz, míg az OSZK végezte volna a keresést, szelektálást és biztosította volna a szakmai hátteret. A szegediek vállalták volna, hogy hozzáfognak az alapok kiépítéséhez, és amikor a

projekt beindul, akkor azt visszaadták volna az OSZK-nak és a NIIF-nek.

Sajnos a pályázat elbukott, így – bár igény lenne rá – továbbra sem indulhat el a MIA-projekt.

Összegzés

Dolgozatomban megpróbáltam összefoglalni azt, hogy miért van szükség egyáltalán Internet archívumokra, és milyen nehézségekkel szembesül az, aki bármiféle Internet archívumot szeretne létrehozni.

Az előzőekben leírtak alapján látható, hogy napjainkban már a fejlett országok jelentős része –, sőt egyes országok tartományai is – a felmerülő problémák ellenére is fontosnak tartja, hogy saját internetes dokumentumaikat a jövő nemzedék számára valamilyen formában tárolják. Sajnos Magyarország a kivételek között szerepel, hiszen egy kisebb csoporton kívül senki sem foglalkozik érdemben ezzel a problémával. Ez azért is különösen érthetetlen, mert közvetlen környezetünk, a szomszédos országok nagy része már megtette kezdeti lépéseit saját Internet archívumuk kialakítására. Ausztriában már működő archívum van, Szlovákiában, Szlovéniában, Csehországban már elkezdték a webarchívumuk előkészítő projektjeit, Horvátországban pedig kötelespéldány rendeletben szabályozták az on-line publikációk beszolgáltatását.

Sajnos ma Magyarországon a törvényhozók és a hatóságok teljes elzárkózása miatt csupán alulról jövő kezdeményezéssel lehetne bármiféle Internet archívumot létrehozni. Feltehetően ilyen kezdeményezésre gondolt Ládi László is, amikor civil mozgalom szervezését javasolta. Véleménye szerint fel kellene hívni az országos gyűjtemények figyelmét, hogy „a nemzetnek fontos anyagokat mentsék, amennyire telik az erejükből. (5 is több mint a semmi!)” [44]. Ez persze sokkal több plusz időt és energiát követel az elkötelezett kevesektől, továbbá számos veszélyt is hordoz magában a koordinálatlan munka, mintha a projekt mögött kormányzati támogatás

[44] Ládi László: Egy gondolat az internet archiválásról. In: Katalist 2009. december 21. https://listserv.niif.hu/pipermail/katalist/2009-December/019826.html (2010. január 10.)

is és egységes koncepció-tervezet állna. Mindezek ellenére úgy gondolom, nem lehetetlen a jelen pillanatban is rendelkezésre álló lehetőségek kihasználásával elindítani a MIA-projektet. A szükséges hardveres hátteret például a szegedi próbálkozáshoz hasonlóan a NIIF-től meg lehetne igényelni. Ezen kívül fontos lenne, hogy Magyarország belépjen az IIPC tagjai közé, mert a szervezet nemcsak a webarchívum működtetéséhez szükséges szoftverekkel tudja segíteni az új tagjait, hanem a régi tagok tapasztalatai is nagyon hasznosak lehetnek.

A webarchiválást támogatja a Szjt. 2003. évi módosítása is, hiszen engedélyezi a közgyűjtemények számára, hogy „archiválási célra készíthetnek digitális másolatot a jogvédett művekről”[45]. Ezek után a begyűjtéshez első lépésként nem kellene többet tenni, csak betartatni a 60/1998. (III. 27.) Kormányrendeletet, mely szabályozza a sajtótermékek kötelespéldány szolgáltatását és hasznosítását. Ebben a rendeletben nem csak a papírlapú sajtótermékekről, hanem az elektronikus dokumentumokról is rendelkeznek. A rendelet a 19.§-ban a következő formában határozza meg az elektronikus dokumentumokat: „csak számítógéppel olvasható (mágneslemezen, CD-ROM-on vagy egyéb digitális formában megjelenő) dokumentum, beleértve azt a szoftvert is, amely az elektronikus dokumentum része, illetve annak használatához szükséges”[46]. Az előző meghatározás alapján egyértelmű, hogy az elektronikus dokumentum kategóriájába kell besorolni az összes weblapot, így a hírportálokat is. A rendelet alkotói mégsem bíztak semmit a véletlenre, ezért az előbb említett paragrafusban konkrétan azt is kimondták, hogy a különböző sajtótermékek elektronikus változatai is e rendelet hatálya alá

[45] Tószegi Zsuzsanna: A digitalizálás és a szerzői jogok. In: Tudományos és Műszaki Tájékoztatás 2004. (51. évf.), 2. sz.
http://tmt.omikk.bme.hu/show_news.html?id=3510&issue_id=448 (2009. március 9.)
[46] A Kormány 60/1998. (III. 27.) Korm. rendelete a sajtótermékek kötelespéldányainak szolgáltatásáról és hasznosításáról.
http://www.oszk.hu/hun/helyi/koteles/koteles_jogszab_hu.htm (2009. március 11.)

tartoznak. A rendelet alapján a sima „elektronikus dokumentum esetén az előállított példányszámtól függetlenül kell kötelespéldányokat szolgáltatni”[47], a sajtótermékeknél – így az elektronikus sajtótermékeknél is – „minden kiadási és előállítási változatából”[48]. Természetesen, azért egy egészséges határt meg kell húzni, hiszen ha csupán az Origó, vagy az Index minden változtatásából újabb és újabb példányokat kapna a könyvtár, akkor egy idő után belefulladnának már csak ennek a két hírportálnak a weblapjaiba is.

Első lépésként célszerű lenne egy kampányt indítani a különböző hírportálok tulajdonosainak, a kiadóknak, az oktatási intézményeknek, a színházaknak és a közgyűjteményeknek, melyben felhívnák a figyelmüket arra, hogy ha eleget tennének a beszolgáltatási kötelezettségeiknek, akkor azzal a magyar kulturális örökség megőrzésében komolya szerepet vállalnának.

Amennyiben a kampány sikertelen lenne, az OSZK következetesen alkalmazhatná a törvényt, és szigorúan élhetne a számára biztosított szankcionálási lehetőségekkel, melyeknek segítségével a kezdeti nehézségek és kemény viták után valószínűleg működőképes lenne a rendszer. Így nem fordulhatna elő olyan eset, hogy a különböző hírportálok tulajdonosai, a kiadók és a többi szervezet semmibe venné őket. „Van pl. olyan kiadó, amelyik minden, csak elektronikusan létező kiadványából az OSZK számára nyomtat egy példányt, és azt küldi be kötelespéldány gyanánt.”[49]

Ezen felül a törvény módosításával elérhető lenne az is, hogy az OSZK az így befolyó pénz egy részét a projekt további folytatására fordíthassa.

[47] A Kormány 60/1998. (III. 27.) Korm. rendelete a sajtótermékek kötelespéldányainak szolgáltatásáról és hasznosításáról. id. m.
[48] A Kormány 60/1998. (III. 27.) Korm. rendelete a sajtótermékek kötelespéldányainak szolgáltatásáról és hasznosításáról. id. m.
[49] Rácz Ágnes id. m.

Ha a MIA-projekt eddig eljut, akkor innentől már szinte mindegy, hogy melyik nemzet Internet archívumát tekintjük példának, hiszen a legtöbb archívum ugyanolyan elvek alapján működik: fontosnak tartják a szolgáltatókkal való együttműködést, amelyből származó adatokat szabadon – a különböző jogi feltételek betartásával – lehet szolgáltatni. Ezen kívül mindegyik archívum alkalmaz valamilyen automatikus arató-robotot. A robot által begyűjtött adatok, dokumentumok azonban csupán egy szűkebb kör, főleg kutatók számára érhetők el. Ugyancsak mindegyik archívum használ valamilyen szoftvert a gyűjteményhez való hozzáférés biztosításához. Vannak olyan archívumok, ahol téma és betűrend szerint lehet keresni (például PANDORA), és vannak olyanok is, ahol az eredeti weblap URL címe alapján (például Internet Archive). Véleményem szerint a katalánok PADICAP projektje által használt megjelenítés a legkomplexebb, mert az ötvözi a téma és betűrend szerinti keresést az URL cím alapján történő kereséssel.

Az Szjt. szerint a szabad felhasználás feltétele a következő: a dokumentumok a közgyűjtemények helyiségeiben az „ezzel a céllal üzembe állított számítógépes terminálok képernyőjén tudományos kutatás vagy egyéni tanulás céljára a nyilvánosság egyes tagjai számára szabadon megjeleníthetők, és ennek érdekében (…) szabadon közvetíthetők, ideértve a nyilvánosság számára történő hozzáférhetővé tételt is, feltéve, hogy az ilyen felhasználás jövedelemszerzés vagy jövedelemfokozás célját közvetve sem szolgálja”.[50] Eszerint ha a MIA elkészül, akkor a gyűjteménye a zárt könyvtári hálózaton belüli szabadon szolgáltathatók lennének. Ha a későbbiekben a gyűjteményt az Interneten is szolgáltatni szeretnék, ahhoz persze szükség lesz külön engedélyek beszerzésére (szolgáltatóktól, kiadóktól stb.), vagy törvénymódosításra.

[50] 1999. évi LXXVI. törvény a szerzői jogról. http://www.complex.hu/jr/gen/hjegy_doc.cgi?docid=99900076.TV (2009. december 5.)

Természetesen, a projekt megkezdése előtt nulladik lépésként célszerű lenne néhány olyan munkacsoportot létrehozni, amelyek eldöntenék, hogy például a weblapok leírását ki végezze el, mit írjanak le, hogyan gazdálkodjanak a meglévő pénzügyi kerettel, illetve honnan, hogyan szerezzenek újabb anyagi forrásokat. Az összes többi, Drótos László által javasolt munkacsoport létrehozása ráérne addig, ameddig webarchívum el nem jut arra a fejlettségi szintre, hogy arató robotokat állítson munkába és a gyűjteményt szolgáltassa.

Végül is mindegy, milyen módszerrel kezdik el építeni a MIA gyűjteményét, csak minél hamarabb kezdődjön el, hiszen azok az információk, amelyek eltűntek a tétlenség évei alatt már pótolhatatlanok. Talán az Internet Archive-tól meg lehet szerezni néhány általuk archivált weboldalt, esetleg a domain szerverek tulajdonosaitól is el lehet kérni az adatvesztés elkerülése miatt archivált anyagaikat, de ezek bonyolult és valószínűleg nem is túl olcsó lehetőségek. Ám amennyiben nem történik hamarosan valami előrelépés ebben az igen fontos kérdésben, úgy „a történelem, a Nagy Moloch fölfal mindent, az egész elektronikus kultúránk úgy tűnik el mint a rendszerváltáskori tv-archívumok sokasága: gróf Apponyi Albert cilinderes trianoni bevonulását több filmszalag is őrzi, de hogy "hogyan is foglaltuk el a tévét Cserhalmival", arról alig van valami, holott mindenki láthatta, 100 kamera vette. Digitális volt, letörölhető, eldobható. Amennyiben így állunk hozzá értékeink megőrzéséhez, akkor végül is Snydernek lesz igaza: a jövő gengszterei biztonságban vannak. S mi, akik állítólag a digitálissá váló múlt őrzői is vagyunk, mit szólunk ehhez?”[51]

[51] Kokas Károly id. m.

Bibliográfia

Folyóiratcikkek, tanulmányok

ACOBS, Neil – CHAMBERS, Jenny - MORRIS, Anne: *Dokumentumszolgáltatók weboldalai*. In: Tudományos és Műszaki Tájékoztatás 2000. (47. évf.), 11. sz. http://tmt.omikk.bme.hu/show_news.html?id=1505&issue_id=30 (2009. október 22.)

ANDROVIČ, Alojz: *E-tmt – Web-archívum made in Slovakia: Kísérleti projekt az elektronikus információforrások gyűjtésére és archiválására*. In: Tudományos és Műszaki Tájékoztatás 2007. (54. évf.), 10. sz. http://tmt.omikk.bme.hu/show_news.html?id=4788&issue_id=487 (2009. március 11.)

BAILEY, Steve – THOMPSON, Dave: *E-tmt – Az első nyilvános webarchívum az Egyesült Királyságban*. In: Tudományos és Műszaki Tájékoztatás 2006. (53. évf.), 10. sz. http://tmt.omikk.bme.hu/show_news.html?id=4555&issue_id=476 (2009. március 11.)

BROOKS, Terrence A.: *Keresés a világhálón: hogyan változtatta meg az internet az információkeresést?* In: Tudományos és Műszaki Tájékoztatás 2004. (51. évf.), 5. sz. http://tmt.omikk.bme.hu/show_news.html?id=3602&issue_id=450 (2009. október 22.)

CERBOVÁ, Ludmila: *A cseh web és a kötelespéldány-rendelet*. In: Könyvtári Figyelő, 2009. (55. évf.) 3. sz. p. 518-520.

CROOK, Edgar: *Web archiving in a Web 2.0 world*. In: The Electronic Library 2009. (27. köt.), 5. sz. http://www.emeraldinsight.com/10.1108/02640470910998542 (2010. március 2.)

DIPPOLD Péter: *A nemzeti bibliográfiák gyűjtőköre, avagy elérhető-e a teljesség?* In: Könyvtári Figyelő 2006. (52. évf.), 2. sz. http://www.ki.oszk.hu/kf/kfarchiv/2006/2/dippold.html (2010. február 28.)

DIPPOLD Péter: *Kötelespéldány-szolgáltatás és nemzeti bibliográfia*. In: Könyvtári Figyelő 2005. (51. évf.), 1. sz. http://epa.oszk.hu/00100/00143/00054/dippold.html (2010. március 2.)

DRÓTOS László: *Mi a MIA? – Javaslat egy Magyar Internet Archívum létrehozására*. In: Tudományos és Műszaki Tájékoztatás 2006. (53. évf.), 6. sz. http://tmt.info.omikk.bme.hu/show_news.html?id=4431&issue_id=473 (2009. március 9.)

FARKAS Szabó András: A szerzői jog és az Internet. In: Internet és Politika 2003. (3. évf.), 4. sz. http://iroga.hu/internet&politika/farkas.htm (2009. december 5.)
GOLDEN Dániel – TÓTH Tünde – TURI László: *Virtuális örökkévalóság: objektumok a digitális könyvtárban*. In: Tudományos és Műszaki Tájékoztatás 1998. (45. évf.), 8-9. sz. http://tmt.omikk.bme.hu/show_news.html?id=2017&issue_id=3 (2009. március 11.)
HAAVISTO, T.: *A szerzői jog és a könyvtárak Kelet- és Közép-Európában: helyzetkép*. In: Tudományos és Műszaki Tájékoztatás 1999. (46. évf.), 11-12.sz. http://tmt.omikk.bme.hu/show_news.html?id=1712&issue_id=21 (2009. március 9.)
ILLIEN, Gildas: *Webarchíválás a francia gyakorlatban*. In: Könyvtári Figyelő, 2009. (55. évf.) 3. sz. p. 553-554.
JÁKI Éva: *Az Internet Archívum ad helyet a szabad elérésű Szövegarchívumnak*. In: Tudományos és Műszaki Tájékoztatás 2005. (52. évf.), 5. sz. http://tmt.omikk.bme.hu/show_news.html?id=3951&issue_id=462 (2009. március 14.)
JODELIS, Remigijus: *Elektronikus források begyűjtése és archiválása Litvániában: úton egy virtuális könyvtár felé*. In: Tudományos és Műszaki Tájékoztatás 2004. (51. évf.), 6. sz. http://tmt.omikk.bme.hu/show_news.html?id=3640&issue_id=451 (2009. április 11.)
KENNEY, Anne R.: *Hogy áll ma az e-folyóiratok megőrzése?* In: Tudományos és Műszaki Tájékoztatás 2007. (54. évf.), 10. sz. http://tmt.omikk.bme.hu/show_news.html?id=4794&issue_id=487 (2010. február 28.)
KOLTAY Tibor – HORVÁTH Péter: *Digitális könyvtárak a világban*. In: Tudományos és Műszaki Tájékoztatás 1998. (45. évf.), 7. sz. http://tmt.omikk.bme.hu/show_news.html?id=2011&issue_id=1 (2009. március 9.)
NUYS, Carol van – ALBERTSEN, Ketil – PEDERSEN, Linda et al.: *A Paradigma projekt*. In: Tudományos és Műszaki Tájékoztatás 2005. (52. évf.), 11-12. sz. http://tmt.omikk.bme.hu/show_news.html?id=4227&issue_id=467 (2009. március 11.)
RÉV István: *Alexandriai könyvtár a pincében*. In: Budapesti Könyvszemle 2004. (16. évf.), 4. sz. http://epa.oszk.hu/00000/00015/00036/pdf/06prrev.pdf (2009. november 5.)

RIBA István: *Eltűnő honlapok – Hibaüzenet*. In: Heti Világgazdaság, 2008. (30. évf.), 32. sz., (augusztus 9.) p. 22-23.
SCHWENS, Ute – WIECHMANN, Brigitte: *Távoli elérésű kiadványok a Német Nemzeti Könyvtárban*. In: Könyvtári Figyelő, 2009. (55. évf.) 3. sz. p. 531-532.
TAPOLCAI Ágnes: *Nyílt könyvtári gyűjtemények az interneten. Szabványos metaadatok: átjárhatóság*. In: Tudományos és Műszaki Tájékoztatás 2003. (50. évf.), 1. sz. http://tmt.omikk.bme.hu/show_news.html?id=1628&issue_id=47 (2009. március 11.)
TÓSZEGI Zsuzsanna: *A digitalizálás és a szerzői jogok*. In: Tudományos és Műszaki Tájékoztatás 2004. (51. évf.), 2. sz. http://tmt.omikk.bme.hu/show_news.html?id=3510&issue_id=448 (2009. március 9.)
VIDA Andrea: *Könyvtári honlapok megújítása: kinek, miért és hogyan? CMS-rendszerek a könyvtárak szolgálatában*. In: Tudományos és Műszaki Tájékoztatás 2006. (53. évf.), 3. sz. http://tmt.omikk.bme.hu/show_news.html?id=4332&issue_id=470 (2009. március 11.)

Elektronikus dokumentumok

1992. évi LXIII. törvény a személyes adatok védelméről és a közérdekű adatok nyilvánosságáról
http://www.complex.hu/jr/gen/hjegy_doc.cgi?docid=99200063.TV (2010. március 31.)
1999. évi LXXVI. törvény a szerzői jogról
http://www.complex.hu/jr/gen/hjegy_doc.cgi?docid=99900076.TV (2009. december 5.)
60/1998. (III. 27.) Korm. rendelet a sajtótermékek kötelespéldányainak szolgáltatásáról és hasznosításáról
http://www.oszk.hu/hun/helyi/koteles/koteles_jogszab_hu.htm (2009. március 11.)
Alexa Internet
http://www.alexa.com
ARTISJUS – Magyar Szerzői Jogvédő Iroda Egyesület: *Jogszabályok*.
http://www.artisjus.hu/aszerzoijogrol/jogszabalyok.html (2009. december 5.)
BERKE Barnabásné: *Elektronikus dokumentumok tipizálása, kötelespéldány-szolgáltatása, gyűjtőköri szempontok*.
http://mekosztaly.oszk.hu/oszkdb/anyagok/Gyujtokor/Tip_Koteles_Gyujtokor_Berke.doc (2010. február 28.)
CALIMERA Útmutató
http://www.ki.oszk.hu/old/calimera/ (2010. február 22.)
COMMUNIA: *About*

http://communia-project.eu/about (2010. március 8.)
Creative Commons: *About Licenses* http://creativecommons.org/about/licenses/ (2010. március 3.)
DE LA VEGA, R. – TORRES, N. – CAMBRAS, J.: *Patrimoni Digital de Catalunya, a year and a half experience*. http://www.recercat.cat/bitstream/2072/9258/2/A4.pdf (2010. március 8.)
Debreceni Egyetem, Informatikai Szolgáltató Központ: *Internet hálózat*. http://www.cic.klte.hu/iszkw3/kltenet/kltenet5.html (2009. október 22.)
DIPPOLD Péter: *A hagyományos nemzeti bibliográfia és az Internet. Válaszlehetőségek az új kihívásokra*. http://mek.niif.hu/03500/03557/html/index.htm (2010. február 28.)
DRÓTOS László: *Egy gondolat az internet archiválásról*. In: Katalist 2009. december 21. https://listserv.niif.hu/pipermail/katalist/2009-December/019825.html (2010. január 10.)
DRÓTOS László: *Hálózati értelmező szótár* http://mek.niif.hu/01200/01280/html/1.02/index.htm (2009. október 22.)
HEGYKÖZI Ilona: *Az Ausztrál Nemzeti Könyvtár (ANK) kiválasztási irányelvei*. http://mekosztaly.oszk.hu/oszkdb/anyagok/Gyujtokor/hegykozi_pandora.doc (2010. február 28.)
Internet Archive http://www.archive.org (2009. október 22.)
KARDKOVÁCS Zsolt – MAGYAR Gábor – TIKK Domonkos: *A szavak hálójában: szabadszavas mélyháló – kereső program*. Budapesti Műszaki és Gazdaságtudományi Egyetem, Távközlési és Médiainformatikai Tanszék http://www.mft.hu/publications/tikk/Htechnika.pdf (2009. december 8.)
KIT Hírlevél: *Pusztuló internetes örökségünk. Internet archívum -- mi mindig, mindenhonnan...?* KIT Hírlevél 2009., 10. sz. http://www.kithirlevel.hu/index.php?kh=pusztulo_internetes_oroksegunk_internet_archivum_mi_mindig_mindenhonnan (2010. március 14.)
KOKAS Károly: *Egy gondolat az internet archiválásról*. In: Katalist 2009. december 21. https://listserv.niif.hu/pipermail/katalist/2009-December/019823.html (2010. január 10.)
KÖMLŐDI FERENC: *Ahova a Google sem jut el*. http://index.hu/tech/net/2009/03/08/ahova_a_google_sem_jut_elahova_a_google_sem_jut_el (2009. október 22.)

LÁDI László: *Egy gondolat az internet archiválásról*. In: Katalist 2009. december 21. https://listserv.niif.hu/pipermail/katalist/2009-December/019826.html (2010. január 10.)
LYMAN, Peter: *How Much Information*. http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/ (2009. március 9.)
Magyar UNESCO Bizottság: *Charta a digitális örökség védelméről* http://www.unesco.hu/informacio-kommunikacio/digitalis-orokseg/charta-digitalis-orokseg (2010. március 14.)
MÁRAY Tamás: *Hálózatok hálózata: az internet.* http://www.mindentudas.hu/maray/20031201maray2.html?pIdx=2 (2009. október 22.)
MURAKÖZI Gergely, Dr: *A szerzői jog és az Internet – Az Internet technikai megvalósítása a szerzői jog tükrében*. http://www.jogiforum.hu/files/publikaciok/drMurakozi-A_szerzoi_jog_es_az_internet%28jf%29.pdf (2009. december 4.)
NAGYMÉLYKÚTI Balázs: *Tartalommegőrzés az interneten: webarchívumok: szakdolgozat*. Szeged, Szegedi Tudományegyetem, Juhász Gyula Tanárképző Főiskolai Kar, Könyvtártudományi Tanszék 2007. www.szilleri.tvn.hu/nagymelykut.doc (2009. március 11.)
NEDLIB – Networked European Deposit Library http://nedlib.kb.nl/
PALKÓ Mária: *Tudományos kutatás az Akadémiai Láthatatlan Weben* http://www.korunk.org/?q=node/8&ev=2009&honap=1&cikk=9501 (2009. március 14.)
PANDORA – Australia's Web Archive http://pandora.nla.gov.au (2010. március 3.)
Patrimoni Digital de Catalunya: *Frequently Asked Questions* http://www.padi.cat/en/pmf.php (2010. március 8.)
PLUHÁR Gábor: *Informatikai értelmező szótár. Válogatás az informatikai szakirodalom tanulmányozásához*. http://mek.oszk.hu/00000/00083/00083.pdf (2009. október 22.)
A Pulman Digital Guidelines magyar változata: digitális útmutató kiemelt fejezetei http://www.ki.oszk.hu/old/pulman/dg/szerzoijog.html (2010. január 3.)
RABB Ágnes: *Szöveggyűjtemény a mély web tanulmányozásához: Cikkek és tanulmányok, külföldi és magyar források alapján: szakdolgozat*. Szegedi Tudományegyetem, Juhász Gyula Tanárképző Főiskolai Kar, Könyvtártudományi Tanszék. Szeged, 2006.

www.szilleri.tvn.hu/rabb.doc (2009. december 1.)
RÁCZ Ágnes: *Egy gondolat az internet archiválásról*. In: Katalist 2009. december 21. https://listserv.niif.hu/pipermail/katalist/2009-December/019827.html (2010. január 10.)
RUTKOVSZKY Edéné – RUTKOVSZKY Ádám: *A láthatatlan web keresése*. https://nws.niif.hu/ncd2003/docs/ehu/EHU-61.htm (2009. december 8.)
SZABÓ Gergely: *Új magyar keresőmotor turkál a mélyweben*. http://index.hu/tech/net/polymeta0604/ (2009. március 11.)
TREFIL Rita: *A német nyelvű mélyweb forrásai az Interneten: szakdolgozat*. Szeged, Szegedi Tudományegyetem, Juhász Gyula Tanárképző Főiskolai Kar, Könyvtártudományi Tanszék 2005. www.szilleri.tvn.hu/trefil.doc (2009. december 6.)
UK Web Archive http://www.webarchive.org.uk/ukwa/ (2010. március 3.)
XINQ http://www.nla.gov.au/xinq/ (2010. március 3.)
Webkeresők működése http://www.cs.elte.hu/~hexapoda/jegyzet2.doc (2010. február 15.)

Kulcsszavak

Internet archívum
Magyar Internet Archívum
világháló
weboldal