

WEB ARCHÍVUMOK: A JÖVŐ(S)

30 Június 2011

Eric T. Meyer
Arthur Thomas
Ralph Schroeder
Oxfordi Internet Intézet
Oxfordi Egyetem



.....	Összefoglaló	3
.....	Bevezetés	4
.....	A jövőépítése	5
.....	Forgatókönyvek	5
A "Nirvána" forgatókönyv		5
Az "Apokalipszis" forgatókönyv		6
A "szingularitás" forgatókönyve		7
A "poros archívum" forgatókönyv		7
Navigálás a	jövő(k)	8
Tanulás az élő	webről	9
.....	Vizualizáció	9
Keresés mint Killer	App	9
Társadalmi	hálózelelemzés	10
.....	Alt-Metrics	12
Szociális	megjegyzés	12
Új	architektúrák	13
Szociális	gépek	13
.....	Hálózatokfeltérképezése	14
.....	Web Science	14
.....	tartalomhelyett az élmény megértése	15
Szemantikus webes és kapcsolt	adatgyűjteményekelemzése	16
A jelen és a jövő	kihívásai	17
A kumulatív háló:	WebarchívumEgy élő	
.....	webarchívum	17
A	változó világháló	18
Az archívumok és	weboldalak felhasználása	19
A szakosodott	web	20
A vizuális	web	21
A web, ahogy	volt	21
A	webszerkezete	22
Hogyan	burjánzanak az ötletek	22
Az	illegális háló	23
A digitális	lábnyom	23
Az	adatok hálója	23
A nemzeti	hálók	24
Következtetések: Az	előttünk álló út	25
.....	Hivatkozott hivatkozások	26
.....	Köszönetnyilvánítás	27

Ezt a jelentést az Oxford Internet Institute kutatói írták a Nemzetközi Internet Megőrzési Konzorcium (IIPC) számára. Célja, hogy további yitát ösztönözzön a webarchivátorok és kutatók között arról, hogy a jövőben milyen módon használhatják a kutatók a webarchivumokat.

A szakasz négy lehetséges jövőbeli forgatókönyvet vázol1 fel:

- **Nirvána:** ahol a webarchivumokat sok csoport széles körben használja, szabványosított, áttekinthető és hatékony hozzáférési felületekkel rendelkezik.
- **Apokalipszis:** az archívumok széttöredezettek, nem szabványosított, nehezen megtalálhatóak és hozzáférhetőek, ezért nem hasznosak és alig használják őket.
- **Szingularitás:** ebben a forgatókönyvben az archívumok szükségtelessé válnak, mivel egyetlen összekapcsolt intelligencia fejlődik ki, amely képes kapcsolatot teremteni a digitális tárgyak és az emberek között.
- **Poros archívumok:** ebben a forgatókönyvben a webarchiváló közösség soha nem válaszol a "na és?" kérdésre, és a webarchivumok nagyrészt kihasználatlanul, digitális porosodva maradnak.

Ezek a forgatókönyvek lehetővé teszik, hogy különböző módon gondolkodjunk az archívumok, a kutatók és a kutatók közötti kölcsönhatásokról.

A 2. szakasz a jelenleg az élő weben folyó kutatások különböző típusait ismerteti, amely jelenleg sokkal szélesebb körben alkalmazott technika, mint a webarchivumok használata. Az elképzelés az, hogy az élő web felhasználása inspirálhatja a webarchivumok lehetséges felhasználási módjairól való gondolkodást. Ezek a felhasználási módok a következők:

- **Vizualizáció:** nem csak a weboldalak, hanem a különböző típusú információk között is lehet linkeket létrehozni, hogy lehetővé váljon az archívumok rendszerezése és áttekintése.
- **Alt-metrika:** a scientometriát kutató tudósok az idézetelemzésen kívül új forrásokból is kezdenek adatokat szerezni - például a kutatók blogjaiból és a blogok közötti linkekből.
- Számos más technikát, például a felhasználók által generált tartalmak feltérképezését és a közösségi hálózatok elemzését is bemutatjuk.

A 3. szakasz a jelenlegi és jövőbeli kihívásokkal foglalkozik. A szakasz első része a web változásának néhány módját ismerteti. - és javasol néhány rövid, közép- és hosszú távú megoldást a levéltárak számára, hogy megbirkózzanak ezekkel a változásokkal. A szakasz az előttünk álló útra vonatkozó javaslatokkal zárul.

Ez a jelentés tervezet formájában 2011 májusában készült, és a 2011. május 9-10-én Hágában (Hollandia) tartott IIPC 2011. évi közgyűlésen került kiosztásra. A jelentést egy plenáris ülésen foglalták össze, és egy műhelytalálkozón vitatták meg. A jelentést e-maiban is eljuttatták az internetkutatók, valamint a könyvtár- és információtudományi közösséghez. A jelentéstervezet célja az volt, hogy provokáljon és ösztönözzön. Gondolkodásra ösztönözni. Vitára ösztönözni. A webarchivátorokat és kutatókat a tétlenségből való kipróbálásra ösztönözni. Ugyanezeket az embereket és másokat is cselekvésre ösztönözni. Provokálni és ösztönözni a változást. Már eddig is vitára serkentett; hogy a változás irányába tett lépésekre ösztönöz-e, az még kiderül.

Miért van szükség a változásra? Amikor az IIPC¹ megkeresett minket, hogy vállaljuk el ezt a projektet, az az érzésünk volt, hogy a webarchiválási közösségnek, és különösen az IIPC-nek, új módszereket kellene vizsgálnia a webarchivumok új felhasználóinak és felhasználásának, a webarchiválás új modelljeinek és a kutatókkal való együttműködés új módozatainak ösztönzésére.

Ezeket a kérdéseket korábban már két, a JISC² által finanszírozott jelentés is felvetette, amelyek a webarchivumok jelenlegi helyzetére (Dougherty, et al., 2010) és az új beruházások lehetőségeire (Thomas, et al., 2010) összpontosítottak. E két dokumentum néhány következtetést alább tárgyaljuk, de az egyik általános témá az volt, hogy "még mindig van egy szakadék a kutatók potenciális közössége között, akiknek jó okuk van arra, hogy részt vegyenek a webarchivumok létrehozásában, használatában, elemzésében és megosztásában, és a kutatók tényleges (általában még mindig kicsi) közössége között, akik jelenleg ezt teszik" (Dougherty, et al., 2010, 5. o.). A jelentésen való munka és az IIPC tagjaival és az internetes kutatóközösséggel folytatott beszélgetések tapasztalatai nem sokat változtattak a véleményünkön ebben a kérdésben; sőt, minden eddiginél jobban még vagyunk győződve arról, hogy a webarchivumok felhasználási esetei nincsenek jól megfogalmazva, és nem foglalkoztatják jelentős mértékben a kutatóközösséget. Ez a jelentés önmagában nem sokat fog változtatni ezen, de ha az érintett közösségek komolyan veszik a benne foglalt javaslatokat, lehetséges, hogy az internetes anyagok archivumai a jövőben fontosabbá válnak a kutatók számára.

Ez a jelentés először is úgy épül fel, hogy néhány spekulatív gondolatba bocsátkozik a web lehetséges jövőjéről, hogy elgondoljunk azon, mit kell tennünk *most annak érdekében, hogy a jövőben megbízhatóan és gyümölcsözően használhassuk a web archivumait*. Ezután rátérünk az élő web kutatására használt módszerek és eszközök vizsgálatára, hogy megmutassuk, milyen dolgokat lehet kifejleszteni az archivált web megértéséhez. Ezután rátérünk egy sor olyan témára és kérdésre, amelyeket a kutatók az archivált web segítségével akarnak vagy akarnának kezelni. Ebben az utolsó szakaszban meghatározunk néhány olyan kihívást, amelyet az egyének, szervezetek és nemzetközi testületek megelőzhatnak annak érdekében, hogy növeljük képességeinket e témák feltárára és e kérdések megválaszolására. A jelentést néhány következtetéssel zárjuk, amelyek az e gyakorlat során szerzett tapasztalatokon alapulnak.

¹ Az IIPC a Nemzetközi Internet Megőrzési Konzorcium (<http://www.netpreserve.org>), amely finanszírozta ezt a munkát, és biztosította a platformot a további vitákhoz, kezdve a 2011-es IIPC közgyűlésen tartott előadással és műhelytalálkozóval, amely a jelentés végleges változatát fogja kialakítani.

² A JISC a Közös Információs Rendszerek Bizottsága (<http://www.jisc.ac.uk/>), amely az Egyesült Királyság oktatási és kutatási ágazatában az IKT-kutatást és infrastruktúra-fejlesztést finanszírozza.

"A jövőt úgy lehet a legjobban megjósolni, ha kitaláljuk." (Kay, 1995)

Beszélgetésünk kezdetén egy kis futurologiával fogunk foglalkozni. Ennek nem az a lényege, hogy megjósoljuk a jövőt, mert az ostobaság lenne. Valójában meglehetősen szkeptikusan állunk a jövő előrejelzésére, forgatókönyvek készítésére és egyéb, a jövőre vonatkozó állítások megfogalmazására irányuló erőfeszítésekhez, általában abban a tudatban, hogy a legtöbb ilyen erőfeszítést soha nem fogják számon kérni.

Van azonban legalább egyfajta jövőkutatás, amely a mi szemünkben megfelelő. Ez az, amikor a feladat lényege nem az, hogy megjósoljuk a jövőt, hanem az, hogy ösztönözzük azokat az embereket, akik felelősek a jövő alapját képező rendszerek kiépítéséért, hogy gondolják át jelenlegi döntéseik következményeit a várható hosszú távú hatások szempontjából. Az IIPC számos ilyen emberből áll, akik jelenleg az internet tartalmának megőrzését szolgáló rendszerek, eszközök, szabványok és protokollok kifejlesztésén dolgoznak, szem előtt tartva, hogy az internet hasznos legyen a társadalom megértéséhez, amelyben élünk.

A számítógépes rendszerek fejlesztése során számos "építészeti választási pont" (Kling, McKim, & King, 2003; Meyer, 2006) van az út mentén - olyan pontok, ahol olyan döntések születnek, amelyek egy útelágazást választanak a többi lehetőséggel szemben. Bizonyíték van arra, hogy a webarchiválási közösség jelenleg és a közeljövőben jelentős választási pontokkal néz szembe. A választási lehetőségek egyike a webes archiválásban bekövetkező szeizmikus változást ígérő jelenséghez kapcsolódik: az egyes weboldalak és oldalak elérésétől a *gyűjtemény mint gyűjtemény* felépítése és használata felé való elmozdulás a gyűjtemény egyes részeinek egyszerű elérése helyett. Milyen döntéseket kell hozni annak érdekében, hogy a dobozos archivum hasznos, használható, fenntartható és hatásos legyen? Az "archivum a dobozban" gondolatára később még visszatérünk, más, döntéseket igénylő kihívásokkal együtt.

A feladat lényege tehát annak eldöntése, hogy a jelenlegi webarchiváló közösség milyen módon kívánja meghozni azokat a döntéseket, amelyek befolyásolni fogják a jövőt, és olyan lépéseket és döntéseket javasol, amelyek a jövőt egyik vagy másik irányba terelik.

SZENARIÓK

A webarchivumok és használatuk sokféle lehetséges jövőjét el tudjuk képzelni; a vita kedvéért négy lehetséges forgatókönyvet vázolunk fel, amelyek a következő egy-két évtizedben játszódhatnak le, megvizsgáljuk a következményeiket, és javaslatokat teszünk arra, hogy a webarchivum-közösség hogyan tudna megbirkózni velük. A dokUMENTUM későbbi részében részletesebben foglalkozunk néhány olyan elemmel, amelyek ezeket a forgatókönyveket alkotják, meghatározzuk a megvalósításuk útjában álló kihívásokat, és példákat mutatunk az "élő" webhez kifejlesztett eszközök széles skálájára, amelyek a történelmi adatokra alkalmazva különbséget tehetnek a legjobb és a legrosszabb forgatókönyvek között.

A "NIRVÁNA" FORGATÓKÖNYV

Minden lehetséges világ legjobbjában a webarchivumok egyszerre lennének robusztusak, szabványosítottak és biztonságosan megőrzöttek, ugyanakkor nyitlak, rugalmasak, széles körben használtak és a standard kutatási eszköztár részét képeznék az internettudomány, a politikatudomány, a közgazdaságtan, a szociológia, a kortárs történelem (és a jövőben a 20th. század végének és a 21.st század elejének története), az újságírás, a nyelvészet, a kommunikáció, az üzleti élet, a médiatudomány és más tudományágak számára. A tudományos életen túl a webarchivumok a nagyközönség, a kormányok, a politikai egységek és ágytrösztök, a vállalkozások és a nem kormányzati szervezetek számára is használhatóak és hasznosak lennének. Sajnos ez sok szempontból a legkevésbé valószínű forgatókönyv, mivel megvalósítása sokkal nagyobb erőfeszítést és nagyobb erőforrásokat igényelne, mint ami jelenleg a webarchivum közösségen belül megvalósíthatónak tűnik. Mindazonáltal hasznosnak találhatjuk, ha szem előtt tartjuk ezt az eszményképet, miközben megvizsgáljuk a lehetséges és a megvalósítható közötti kompromisszumokat.

Ahhoz, hogy ez a forgatókönyv akár csak vázlatosan is megvalósulhasson, számos dolognak meg kell történnie (egy későbbi szakaszban példákat mutatunk az élő webről, ahol már történtek ilyen dolgok). Ezek a következők:

- Sokkal **erősebb és hatékonyabb eszközök kifejlesztése a szöveges kereséshez, az információk kinyeréséhez és elemzéséhez**, a vizualizációhoz, a társadalmi megjegyzésekhez, a longitudinális elemzéshez és a hangulatelemzéshez.
- Sokkal jobb módszerek kidolgozása a felhasználók számára, hogy megértsék az egyes vagy több gyűjtemény "Gestát" -ját. Míg a szöveges tartalom kereshető, gazdag metaadatokra van szükség a tartalom széles körű áttekintéséhez, vagy a tartalom újfajta rendszerezésének támogatásához. Az emberek különösen jók a vizuális minták felismerésében, ami azt sugallja, hogy a grafikus eszközök valószínűleg az egyik legjobb módot jelentik ennek elérésére. Elképzelhető olyan virtuális környezetek létrehozása, amelyek lehetővé teszik a 3D-s "átrepülést" és egyéb

a tartalom szervezésének intuitív térbeli módjai. Szükséges esetben a teljesen magával ragadó CAVE-típusú (Schroeder, 2011) virtuális környezetek támogatják a térben szétszórta embercsoportok közös munkáját, és lehetővé teszik a hatékony megosztást és a társadalmi interakciót. Ily módon a webarchívumok összességét egy hatalmas "közösségi térként" lehetne elképzelni, amelyben az emberek egyedül vagy csoportosan barangolhatnak. A jelenlegi web (és így a webarchívumok is) kevésbé érzékelteti a térbeli szerveződést, és nincsenek jó térbeli jelzések vagy más, a navigációt és a felfedezést segítő "lehetőségek". A digitális dokumentumok egymástól viszonylag elszigetelten élnek, ellentétben a fizikai könyvtárakkal, ahol a dokumentumok egy 3D-s világban vannak elrendezve, amely lehetővé teszi a "bolyongással történő navigációt", és az olyan jelzések, mint a térbeli közelség, segítik a felfedezést.

- Míg a közösségi megjegyzésekkel kapcsolatos eszközök kezdenek megjelenni, a webarchívumokból hiányoznak az egyéb együttműködési eszközök, például az ajánlómotorok (az Amazon online áruház kiváló példa arra, hogy mi lehetséges).
- Egyre inkább szükségünk van a felhasználók által generált ("Web 2.0") tartalmak archiválására nagyon nagy (Facebook méretben). Az ilyen heterogén és önmagában is rendezetlen tartalmakra azonban nehéz struktúrát erőltetni. A szemantikai gazdagság miatt a gépi feldolgozás technológiai szempontból nagy kihívást jelent, ezért egy alternatív megoldás a "tömegesen" megközelítés támogatása, azaz az archívum felhasználóinak lehetővé tenni a tartalom rendszerezését. Ez a közösségi annotáció egy szükséges formája, ahol a felhasználók nemcsak adatokat, hanem metaadatokat is létrehozhatnak (Gazan, 2008; van den Heuvel, 2009).

Ebben a Nirvánában a ma meghozott döntéseket a jövő kutatói dicsérni fogják, akik az emberi erőfeszítéseknek az interneten megtestesülő, megerősített és fűvábbfejlesztett információira és bizonyítékaira támaszkodnak, hogy mindenféle hatékony kutatási technikát lehetővé tegyenek.

AZ "APOKALIPSZIS" FORGATÓKÖNYVE

A lehető legraszababb esetben a folyamatosan változó internet továbbra is szédítő ütemben fejlődik és új technológiákat fejleszt (HTML5, végrehajtható tartalom, beágyazott videó és interaktív objektumok, adatbázis-alapú weboldalak, nem HTTP/HTML alapú mobiltelefonos alkalmazások stb.), a webarchiválási eszközök pedig nem tudnak lépést tartani, és egyre inkább lemaradnak. Még ha a webarchiválási technológiák képesek is lennének lépést tartani, a folyamatosan változó formátumok leküzdhetetlen kihívást jelentenek. Ebben a forgatókönyvben a tényleges tartalomnak csak egy kis része rögzíthető hűen, és még ha rögzítik is, a megtekintéshez szükséges speciális bővítmények nem karbantartottak, karbantarthatók, és a tartalom megtekintése lehetetlenné válik. Korunk online életének legtöbb feljegyzése végül olyan olvashatatlanná válik, mint az 1960-as évek lyukkartyái vagy tekerceses mágnesszalagjai. A formátum problémája mellett egyre nagyobb problémát jelent a méretarány. Ahogy az internet az IPv6 teljes körű használata felé halad, a "címezhető" objektumok száma (bőleértve egyre inkább a "tárgyak világában" a fizikai objektumokat is) valóban gigantikussá válik (10³⁸), és nagyságrendekkel meghaladja a címek tárolására szolgáló kapacitásunkat, nem is beszélve az általuk generált tartalomról. Következésképpen már keresni sem tudunk a dolgok után, mivel az indexelési és keresési technológia reménytelenül csődöt mond.

A szemantikus web fejlődésével a "tartalom" egész fogalma megváltozik. A tartalom már nem csak szöveg és kép, hanem tetszőleges adatelemek és a közöttük lévő kapcsolatokat is magában foglalja. A 2011. nyilvános "Linked Data" univerzumban⁴ (amely olyan gyűjteményeket foglal magában, mint a data.gov és a data.gov.uk) már most is több tízmilliárd adatelemet tartalmaz (egyre inkább RDF formátumban), amelyeket több százmillió, nem hivatkozható link köt össze. Az ilyen adathalmazok archiválásának kihívásával már kezdenek foglalkozni (pl. az Egyesült Nemzeti Levéltár Laboratóriumának PRONOM projektje⁵), de nagy a valószínűsége annak, hogy a Linked Data univerzum gyorsabban fog növekedni, mint ahogyan az kezelhető lenne, akár a gyűjtés, akár az elemzés szempontjából.

Ebben a forgatókönyvben még egy olyan vállalat, mint a Google hatalmas erőforrásai is eltörpülnek a probléma mellett, így a webes archiválásra adott elcsépejt válasz ("hagyjuk, hogy a Google csinálja") már nem megoldás.

Ha a mai döntések erre az útra vezetnek minket, akkor a holnap kutatókat megtanítják majd arra, hogy a web múltját hozzáférhetetlennek, megbizhatatlannak és olyannak tekintsek, amire csak anekdoták és másodlagos bizonyítékok révén emlékeznek az időből. A ma globálisan létrehozott hatalmas mennyiségű információ ugyanúgy íródhatott volna papírdarabokra, amelyeket milliárd cipősdobozban tároltak, bármennyire is jól tesz a világ fejleményeinek megértésében, ahogyan azt az interneten található tartalom tükrözi.

³ http://en.wikipedia.org/wiki/Web_of_Things

⁴ Kapcsolódó adatok - elosztott adatok összekapcsolása a világhálón, a www.linkeddata.org oldalon.

⁵ <http://www.nationalarchives.gov.uk/PRONOM/Default.aspx>

A "SZINGULARITÁS" SZKENARIA

Egy teljesen alternatív világban a legradikálisabb forgatókönyv szerint az internet, ahogyan mi ismerjük, valami teljesen új, esetleg saját intelligenciával rendelkező dologgá fejlődik (Kurzweil, 2005). Ahogy eléri a szingularitást, olyan komplex virtuális organizmussá fejlődik, amelyet talán alig értünk, és amelyet nem tudunk jobban archiválni, mint ahogy jelenleg az emberi agyban lévő tudatot sem tudjuk archiválni. Már most, 2011-ben is kezdjük látni, hogy a mesterséges és az emberi feldolgozás közötti különbségtétel kezd felbomlani. Az olyan szolgáltatások, mint a reCAPTCHA (Von Ahn, Maurer, McMillen, Abraham, & Blum, 2008) és az Amazon Mechanical Turk, amelyek embereket használnak "a hurokban" olyan problémák megoldására, amelyeket a gépek nehezen oldanak meg, megmutatják az utat egy olyan világ felé, amelyben az emberi és a gépi intelligencia elválaszthatatlanul összefonódik, és a köztük lévő határvonal elmosódik. Egy ilyen világban még az sem világos, hogy mit jelenthet az "archiválás", így az idő előrehaladtával a múlt elkerülhetetlenül és visszavonhatatlanul elveszik.

Ez a forgatókönyv science fictionnek tűnhet, de a mai technológiák közül sokan néhány évtizeddel ezelőtt még science fictionnek tűntek volna. Erdemes azonban nem elfelejteni, hogy a jövő kiszámíthatatlan, még akkor is, ha sikerül befolyásolnunk a választási pontokat az út mentén, hogy így vagy úgy nyomuljunk. Az általunk meghozott döntések elegendően bizonyulhatnak ahhoz, hogy valami teljesen új dologgal, például egy intelligens internettel foglalkozzunk. Ez nem jelenti azt, hogy ettől függetlenül ne próbáljuk meg azt tenni, amit helyesnek gondolunk.

A "POROS ARCHÍVUM" FORGATÓKÖNYV

Sajnos ez az a forgatókönyv, amely *jelenleg* meglehetősen valószínűnek tűnik: a webarchívumok a poros archívum digitális megfelelői lesznek, gyakran jól gondozottak és karbantartottak, de alig használják őket. Bár a webarchiváló közösség folyamatosan fejleszti az internet egyes részeinek megőrzésére vonatkozó szabványokat és gyakorlatokat, a kutatói közösségből kevés igazán lenyűgöző felhasználási mód tűnik fel. Az oldalakat online eszközökkel egyénileg lehet megtekinteni, és egyes kutatók továbbra is kis archívumokat fognak létrehozni bizonyos kutatási témákhoz, de az internétes kutatás továbbra is elsősorban az élő webre fog összpontosítani, és a közeljövőben kevés érdeklődés fog kialakulni a múltbeli web komoly kutatásra való felhasználása iránt.

Ez más, mint az apokalipszis forgatókönyve. Abban a forgatókönyvben a webarchiválási technológia nem tudott lépést tartani az internet technológiai változásaival. Ebben a forgatókönyvben a webes archiválás lépést tart a webes terjesztési technológiával. A megőrzött adatok azonban továbbra is csak azok maradnak - egy bizonytalan jövőbeli felhasználásra megőrzött példány.

E jelentés megírása során nyilvánvalóvá vált, hogy a webarchívumokkal való konzultáció helyett a felhasználók és a kutatók egyre inkább magát az élő webet tekintik archívumnak. Az élő web folyamatosan növekszik, és a legtöbb esetben az eltűnő adatokat sokan egyszerű kellemtelenségként tűrik el, amit többnyire ellensúlyoz az egyébként hatalmas adatmennyiség, amely bármikor a weben marad.

Az archívumról alkotott képünk olyan fizikai tárgyakról szól, mint például a fizikai helyen tárolt papírok és dokumentumok. Ez azonban nem a web lényege. A kutatók kutatási céljaik elérése érdekében nagy mennyiségű anyagot rögzíthetnek különböző élő webes forrásokból. Az archívumokat az utókor számára elzártként fogjuk fel: Maga a web azonban a kutatók számára potenciálisan érdekes, különböző típusú anyagok folyamatosan növekvő, hatalmas és változatos forrása, amelyet nem hagyományos archívumként, hanem egyszerűen adatforrásként tekintenek.

Ez egy pesszimista forgatókönyv, de úgy tűnik, hogy a bizonyítékok súlya az ő oldalán áll. Számos vezető kutatóval folytatott konzultációnk során azt tapasztaltuk, hogy a múltbeli webre vonatkozó kérdések feltevése és az internet mint történelmi fejlődés megértése iránt tartósan nem mutatkozik érdeklődés. Természetesen vannak kivételek, amelyeket a jelentésben később részletezünk, de nem tudtunk felfedezni olyan lappangó vágyat a webarchívumokkal való munka iránt, amely csak arra vár, hogy a megfelelő technológia felébresztesse. Talán a sarkon túl várakozik egy változás, amely készen áll arra, hogy egy új felhasználási mód vagy egy új technológia bemutatásán alapuló változást idézzon elő a kutatók fantáziájában. Ha azonban nem lesz ilyen, attól tartunk, hogy a webarchívumok továbbra is digitális porosodni fognak.

Ha ezt a forgatókönyvet el akarjuk kerülni, új típusú archivátorokra van szükségünk - olyanokra, akik együttműködnek a kutatókkal és a nyilvánossággal a szükséges adatok kinyerésében az élő webből, és ha az adatok eltűntek az élő webből, képesek helyreállítani őket oly módon, hogy láthatóvá és használhatóvá tegyék az élő web eszközei számára. Ahogyan az elmúlt évtizedben a levéltárakból származó történelmi dokumentumok digitalizálása hatalmas mennyiségű történelmi anyagot tett elérhetővé az interneten (Meyer, 2011; Tanner, 2010; Tanner & Deegan, 2011), a webarchívumok

⁶ Amazon Corp., Mechanical Turk a www.mturk.com oldalon.

nem a webről kell dobozokba helyezni, hanem vissza kell helyezni a webre, ha a bennük lévő tartalom egyébként eltűnt.

NAVIGÁLÁS A JÖVŐBE(S)

Miközben a webarchívum jövője felé haladunk, számos kérdést tehetünk fel magunknak azzal kapcsolatban, hogyan szeretnénk, hogy a jövő kinézzen.

Például a jövő archívuma egy fallal körülvett kert lesz-e, amelyet megőriznek és megóvnak a károktól, de amelyhez korlátozottan lehet hozzáférni? Vagy egy tágas, nyitott tér lesz, amely mindenki számára elérhető? Szilókból áll majd, vagy egyetlen összekapcsolt arena lesz? Vagy egy nyitott és összekapcsolt, de nagyrészt lakatlan szellemváros lesz?

Az archívumok részben a kutatók vagy tudósok, részben pedig a nyilvánosság számára készültek.⁷ Lehet-e a kettőt kibernetikai módon összekapcsolni - úgy, hogy a tudósok folyamatosan, valós időben nyomon követhetik, hogy a közönség (beleértve a tudósokat is) mihez fér hozzá, mit használ és mit hoz létre (és ezáltal javítják a globális tudatosság vagy a "tudati kollektiva" megértését), ugyanakkor úgy alakítják ki ezt a teret, hogy az optimális legyen a bővítésre, a gyűjtésre és a hatékony, élvezetes és az egész világot gazdagító használatra?

Még ha a szingularitás nem is következik be, az internet egyre inkább lehetővé teszi a kapcsolatokat oly módon, hogy a globális agyat metaforaként használva legalábbis érthetővé válik (Schroeder & Meyer, 2009). Ebben a globális agyban - élő közvetítésekkel és kapcsolatokkal - az egyes ágak bemeneti és kimeneti eszközökön keresztül kapcsolódnak egymáshoz. Hogyan tükrözhetik a webarchívumok a globális ágy összekapcsolt jellegét, ahelyett, hogy dokumentumok egymástól elszakított halmazainak tűnjének?

Ezek és sok más kérdés áll előttünk a jövőre nézve. A dokumentum következő részeiben megvizsgáljuk, hogy az élő web megértésének technikai milyen módon inspirálhatják a webarchiváló közösséget, majd felvázolunk néhány olyan kihívást, amely a webarchívumokat potenciálisan értékesebbé teszi a kutatás számára.

⁷ Nem foglalkoztunk a webes dokumentumok archívumainak üzleti és kormányzati felhasználásával, amelyeket gyakran jogi követelményeknek való megfelelés céljából alakítottak ki, mivel ez meghaladta hatáskörünket és szakértelmünket.

TANULÁS AZ ÉLŐ WEBRŐL

Fel kell tennünk a kérdést, hogy az internetes kutatás világában miért tűnnek a webarchívumok másodosztályú állampolgároknak? Jelenleg sokkal kevesebb kutató használja a webarchívumokat, mint azok, akik az élő webet tanulmányozzák, és kevés nem akadémikus épít eszközöket az archivált webhez, különösen az élő web tanulmányozására épülő eszközök hatalmas számához képest.

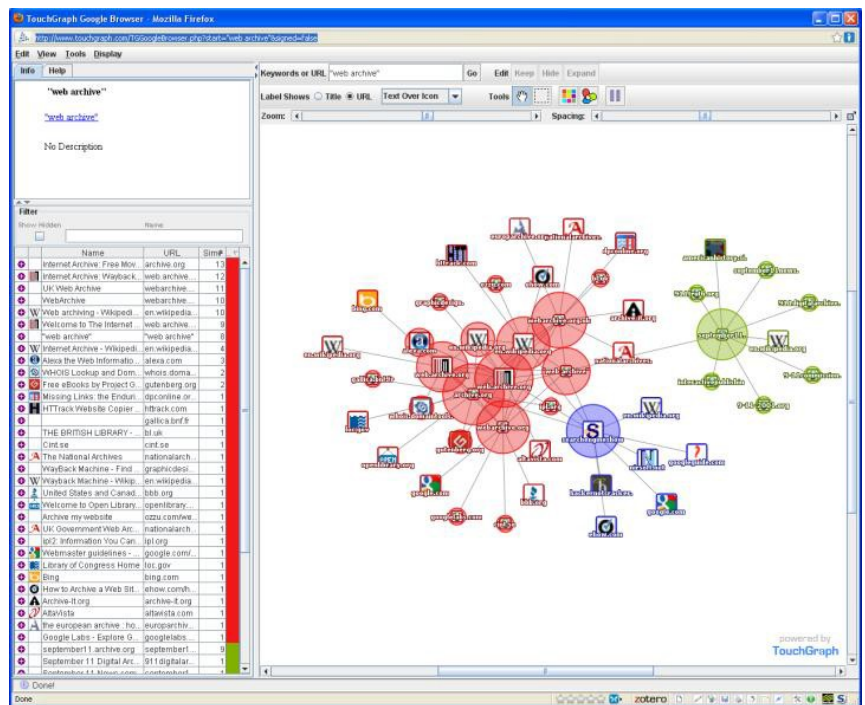
Az ebben a szakaszban kiemelt általános kihívás az, hogy a webarchiváló közösségnek össze kell kapcsolnia az általa épített erőforrásokat az informatikusok, kutatók, független fejlesztők és hackerek által az élő web tanulmányozására kifejlesztett élvonalbeli eszközökkel. Jelenleg az élő web tanulmányozására kifejlesztett eszközök nagyrészt nem alkalmazhatók könnyen a webarchívumokban fellelhetőre álló adatok tanulmányozására. Ez nagyban akadályozza, hogy a webet ne csak pillanatképként, hanem fejlődő ökoszisztémaként is megismerjük.

VISUALIZÁCIÓ

Bármilyen archívumot építenek vagy használnak, az hatalmas és áttekinthetetlen lesz, és nem lesz térkép vagy vizuálisan hozzáférhető és intuitív módja annak, hogy az archívumokat és azok összekapcsolódását lássuk. Az egyik legfontosabb megoldás itt a vizualizáció, de sokféle vizualizáció létezik.

Kihívások: Számos eszköz létezik a közösségi média felhasználói közötti kapcsolatok vizsgálatára, az idővonal megtekintésére, a térképek és a felhasználók lábnyomainak összehasonlására, az adatok összekapcsolódásának bemutatására szolgáló vizuális eszközök és hasonlók - de ezeket úgy kell integrálni, hogy a webarchívumokkal is működhessenek. Az információs vizualizáció a kutatás felejtett területe, de az, hogy hogyan lehet a legjobban látni egy archívumot (vagy keresni, vagy látni a bennük és közöttük lévő kapcsolatokat), beleértve az intuitív felületeket, a nézetek változásait, a 3D-s nézeteket és az időbeli dinamikát - még mindig nehezen megfogható. És - lehetséges-e például a gyűjteményeken belüli témák azonosítása vizuális szemlélettel, vagy a kapcsolatok halmazainak vizuális szervezése? Maska pp fogalmazva, a vizualizációs eszközöket a kutató szolgálatjává tenni?

Példák: Touchgraph⁸, Apple Time Machine⁹



Az itt látható TouchGraph ábra¹, a felhasználók számára lehetővé teszi a weboldalak közötti kapcsolatok felleldezését grafikus felület segítségével. Az adatok az élő webről származnak.

KERESÉS MINT GYILKOS ALKALMAZÁS

Mivel az interneten található információk egyre szaporodnak, mind mennyiségükben, mind tartalom típusaiban, sokkal összetettebb keresésekre lesz szükség ahhoz, hogy ebből a hatalmas gyűjteményből bármi értelmeset és használható ki tudjunk hozni. A keresés egyre összetettebb feladatok fele fordul, mint például a kép- és videokeresés.

Kihívás: Sokkal ambiciozusabb teljesítményszint létrehozása viszonylag szerény költségek mellett, különösen a gyűjteményekre alkalmazható keresőeszközök létrehozása. Ez megkövetelheti, hogy a fejlesztők agresszívan használják a felhalapú keresőmotorokat.

⁸ <http://www.touchgraph.com/>

⁹ <http://www.apple.com/macosx/what-is-macosx/time-machine.html>

```
Operations: LOAD, GROUP, COGROUP, FILTER, FOREACH, ORDER
Generate Query:
visits = LOAD visits.txt AS (user, url, time);
pages = LOAD 'pages.txt' AS (url, pagerank);
v_p = JOIN visits BY url, pages BY url;
users = GROUP v_p BY user;
useravg = FOREACH users GENERATE group, AVG(v_p.pagerank) AS avgr;
answer = FILTER useravg BY avgr > 0.5;
```

és jobb keresési nyelvek, amelyekkel rugalmasan lehet összetett kérdéseket feltenni a webarchívumokban tárolt adatokra.

Példák: (ma Apache) PIG Latin¹⁰ platformja, amely támogatja a nagyon nagy adathalmazok ad hoc elemzését (lásd az ábrát 2).

TÁRSADALMI HÁLÓZATELEMZÉS

A **társadalmi hálózatelemzés** (SNA) az internetkutatók, szociológusok, fizikusok és sokan mások jelentős kutatási érdeklődésének és tevékenységének területe. Az érdeklődés témái széles skálán mozognak, beleértve a barátok közötti kapcsolatok megértését az olyan közösségi oldalakon, mint a Facebook (Hogan, 2010), a politikai vitákhoz hozzájáruló politikai hovatartozásának vizsgálatát (Hindman, 2007), és az angol irodalomban szereplő cselekményhálózatok feltárása (Moretti, 2005, 2011). Az SNA-központú kutatások elvégzésére szolgáló eszközök egyre szaporodnak, többek között a NodeXL¹¹, a Voson¹², a Pajek¹³, a UCINET¹⁴ és sok más.¹⁵ Ezek közül az eszközök közül azonban kevés, vagy egyáltalán nem mindegyik lett engedélyezve vagy optimalizálva a webarchívumokkal való használatra.

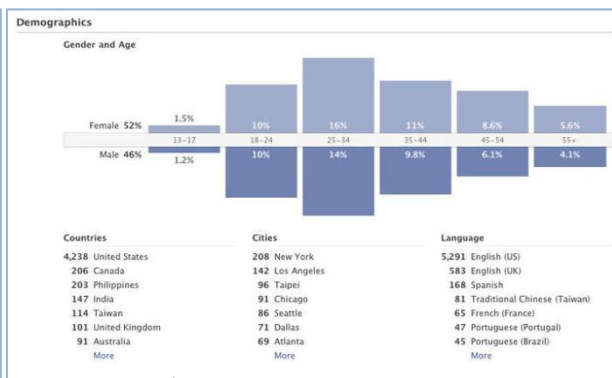
Kihívás: Először is, dolgozzunk együtt a főbb SNA-eszközök fejlesztőivel, hogy lehetővé tegyünk és optimalizáljuk őket a webarchívum-adatokkal való munkavégzésre. Emellett új, innovatív módszerek kifejlesztése, amelyek csak akkor lehetségesek, ha a hálózati adatokhoz hozzáadjuk az idő dimenzióját, hogy nyomon követhessük például a közösségi hálózatok időbeli fejlődését, nemcsak a közösségi oldalak állapotának archiválásával, hanem azzal is, hogy mikor hoznak létre linkeket, mikor tartanak fenn linkeket, mikor törölnek linkeket, mikor kommunikálnak egymással, mikor csatlakoznak csoportokhoz, mikor hagyják el a csoportokat és az oldalakat. Nem szabad elfelejtenünk, hogy a világháló a linkek hálózata, és a hálózatelemzés betekintést nyújt e hálózat természetébe.

Példák:

Facebook Analytics: számos eszköz áll rendelkezésre a Facebook-felhasználók közötti interakciók, a befolyás áramlásának és a közösségi grafnak az elemzésére. Ilyen például a Facebook Insight¹⁶:



A Facebook közösségi grafikonok megjelenítése:



Ábra Forrás3 : http://infosthetics.com/archives/2008/03/facebook_social_network_graph.html

Ábra Forrás4 : <http://www.facebook.com/notes/facebook-engineering/visualizing-friendships-769716598919/>

¹⁰ <http://pig.apache.org/>

¹¹ <http://mckel.codesplex.com/>

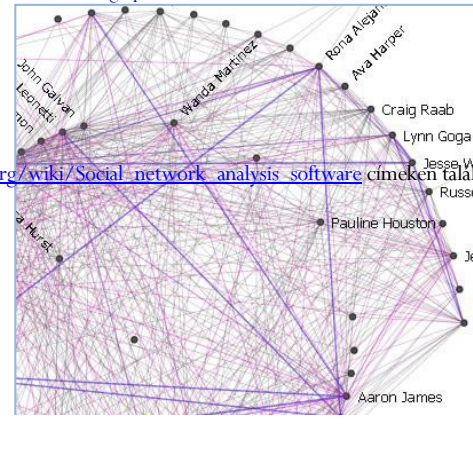
¹² <http://voson.uni.edu.au/>

¹³ <http://www.analyticshub.com/pajek/>

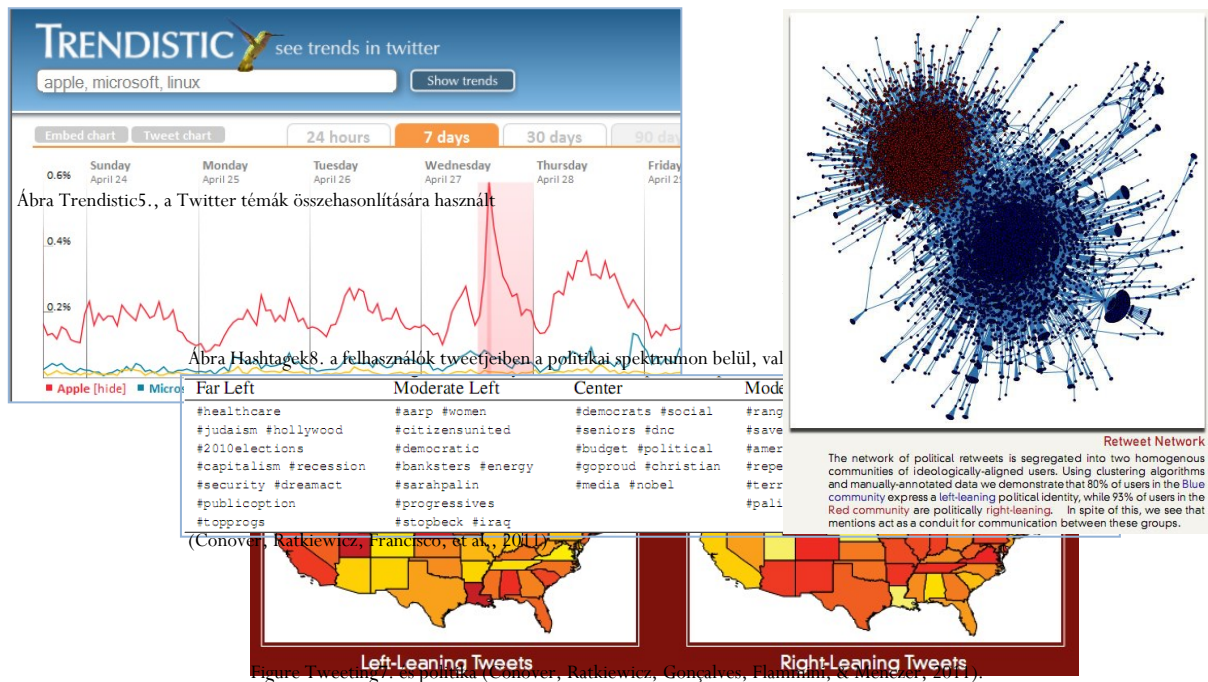
¹⁴ <http://www.analyticshub.com/ucinet/>

¹⁵ Lásd például a <http://www.ibm.com/cognitive/complexity/index.html> és a http://en.wikipedia.org/wiki/Social_network_analysis_software címeken található listákat.

¹⁶ <http://www.facebook.com/insights/>



Hasonlóképpen, a Twitter analitikák széles skálája létezik, mint például a Twitalyzer¹⁷, Trendistic¹⁸ és mások.



¹⁷ <http://www.twitalyzer.com/>

¹⁸ <http://trendistic.com/>

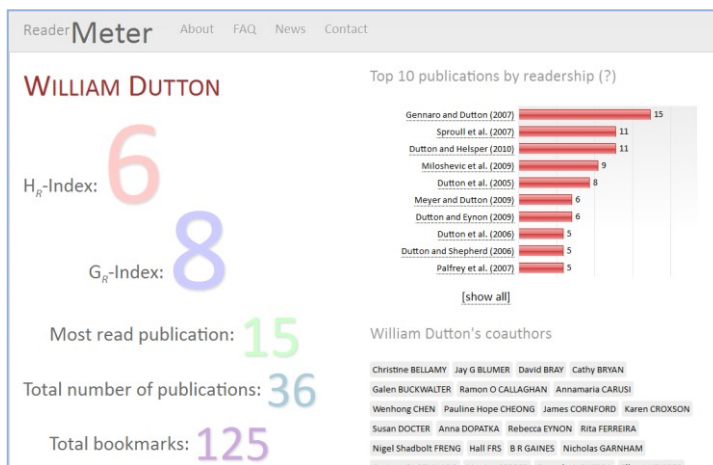
ALT-METRICS

Az alt-metrika egy olyan kifejezés, amely a tudományos hatás mérésének újszerű módjait jelöli a hagyományos bibliometriai, webometriai és szcintometriai méréseken túl.¹⁹ A tudósok közötti és a tudósok közötti, valamint a tudományos közösségek belüli kommunikáció egyre inkább a világhálón zajlik. A kutatást tanulmányozó kutatók újonnan kialakuló közössége a Twitter, a Mendeley, a blogok, a FriendFeed és sok más eszköz által hagyott nyomokat és linkeket használja fel arra, hogy megértse a kutatás gyorsan fejlődő hatásait, gyakran sokkal hamarabb, mint ahogy a hagyományos hatások kialakulhatnak.

Még ennél is tovább megyünk, ha úgy gondolkodunk, mint **alt-alt-metrika**: hogyan lehet nyomon követni a **nem akadémiai** hozzájárulásokat a tudáshoz. Alkalmazhatók-e a kutatókkal kapcsolatos kutatás eszközei más, nem akadémiai területekre? Mérhetjük-e például a hobbi csoportok egyéni közreműködőinek hatását az idő múlásával, hasonló mérőszámokkal, mint amilyeneket a kutatók befolyásának alakulásának megértésére fejlesztettek ki?

Kihívás: A digitális anyagok időintervallumának sokkal egyszerűbb meghatározási módjainak lehetővé tétele, hogy az alt-metrikus elemzés a bibliometriával közvetlenül analóg módon végezhető legyen: a formális publikációs modellekben minden publikációnak van szerzője és dátuma, és ezeket használják az egyes munkákra történő hivatkozások nyomon követésére. A formális publikációnak van egy kipróbált, megbízható és elfogadott archíválási rendszere: a tudományos folyóirat. A nem hivatalos internetes publikációnak azonban nincs hasonlóan jól kidolgozott módszere a tudáshoz való hozzájárulások archíválására oly módon, hogy azok idővel idézhetőek és áthelyezhetőek legyenek. Ez a hiányosság betöltésre vár, és a webarchívumok kézenfekvő kiindulópontot jelentenek. A nem tudományos hozzájárulások esetében a wikik és más kollaboratív kúrációk elemzésére használt eszközöket ki lehet terjeszteni a hosszú időn át tartó változások megértésére?

Példák: DataCite²⁰



Abra ReaderMeter9, egy alt-metrikai eszköz, amely a Mendeley (<http://www.mendeley.com/>) statisztikáin alapuló, a felhasználók szerzői olvasási szokásainak megértésére szolgál. Forrás: <http://readermeter.org>

TÁRSADALMI MEGJEGYZÉSEK

A felhasználók szeretnék, ha közzétehetnék és megoszthatnák linkjeiket és könyvjelzőiket, valamint az ezekhez a forrásokhoz fűzött megjegyzéseiket és megjegyzéseiket. A kutatók számára fontos kérdés annak megértése, hogy ezek a közösségek hogyan alakulnak ki az idő múlásával és hogyan tartják fenn magukat. A Redditnek például több mint 8 millió egyedi olvasója és havi 1 milliárd oldalmegtekintése van (Jasra, 2011).

Kihívások: Az archívumok nem csak weboldalakat és weboldalkollekciókat, hanem az ezekre az oldalakra és kollekciókra mutató hivatkozásokat és megjegyzéseket is képesek tárolni. Legyen képes megválaszolni a kérdést: hogyan mutatnak egymásnak az emberek ezekre a forrásokra, és hogyan változik ez az idő múlásával? Vizsgálja meg a meglévő "mashup" technológiák használatát, és adaptálja a meglévő közösségi jegyzetelési eszközöket. Egy lépéssel tovább lépve: lehet-e az archivált gyűjteményekre mutató linkek és megjegyzések közösséget is ösztönözni hasonló eszközökkel, mint ahogyan az emberek egymásra mutatnak az élő webes elemekre?

Példák: Delicious²², Reddit²³, bookmarklet-alapú pl. MadCow²⁴

¹⁹ Lásd <http://altmetrics.org/manifesto/>

²⁰ <http://readermeter.org>

²¹ <http://datacite.org/>

²² <http://www.delicious.com/>

²³ <http://www.reddit.com/>

²⁴ <http://www.web-notes.com/>

ÚJ ARCHITEKTÚRÁK

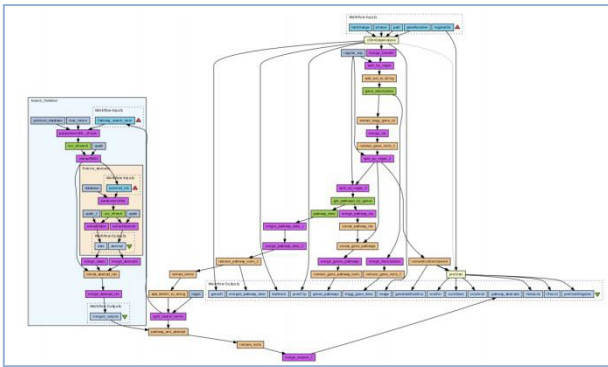
A világ megértéséhez az interneten található adatok felhasználására irányuló tevékenység egyre inkább az olyan erőforrásokra támaszkodik, mint az API-k és a kapcsolt adatok, amelyek az adatokat újrafelhasználásra, újrafelhasználásra és mashupra teszik elérhetővé. Az egyik fő oka annak, hogy az élő webet aktívabban kutatják, mint az archivált webet, az, hogy az eszközfejlesztők hozzáférhetnek az élő webes adatokhoz akár közvetlenül a weboldalak feltekerésével, akár a Google/Yahoo, Twitter, Facebook stb. API-kon keresztül. Az API-k némelyikének korlátai ellenére is ezek jelentik a fő okát annak, hogy az élő webes adatok felhasználásával virágzik a kutatási tevékenység.

Az API hatékony eszköz olyan új alkalmazások létrehozására, amelyek adatokra támaszkodnak, és több forrásból származó adatokat egyesítenek. Egy kutató elmondta nekünk: "Bizonyos hash-tagek használatát kutatom a Twitteren, és az API-használat korlátozását tartom a legzavaróbbnak, mivel az általam keresett tweetek még mindig online és elérhetőek, bár elég nehéz megtalálni őket, vagy más módon jelentéseket vagy aggregációkat futtatni róluk. Korlátoztak például a Twapper Keepert, az egyetlen általam ismert elérhető szolgáltatást, amely lehetővé tette a munkámhoz szükséges jelentések létrehozását a hash-ek körül" (Jeffrey Keeler, személyes közlés).

A kérdés tehát az, hogyan lehet az API-kat archiválni, és ha az API-kat korlátozzák vagy leállítják, az hogyan érinti az API-n keresztül már archivált anyagokat? Vannak-e olyan módszerek, amelyek a jogtulajdonosok és a licencfeltételek tiszteletben tartása mellett megőrzik a tartalmat?

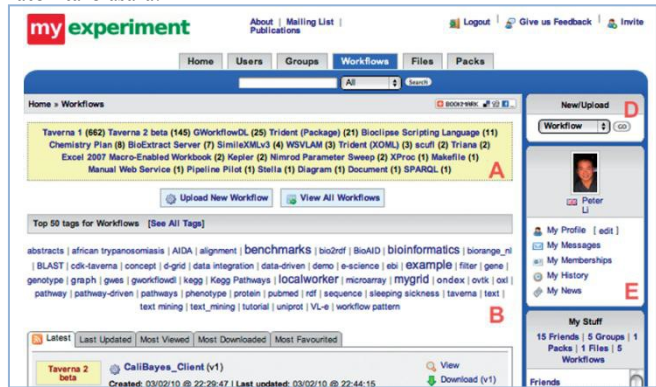
Kihívás: Hogyan lehet a múltbeli webre vonatkozó adatokat API-kon és összekapcsolt adatokon keresztül megnyitni, hogy az okos emberek új felhasználási módokat és a tudás létrehozásának új módjait alakíthassák ki belőlük? Azzal, hogy eszközöket biztosítunk az emberek számára, hogy saját, rugalmasabb munkafolyamatokat alakíthassanak ki, ahelyett, hogy monolitikus, egycelu eszközök használatára kényszerítenék őket. Az egyik megközelítés az lenne, ha az elemzési funkciókat webszolgáltatásként valósítanánk meg, amelyeket egy munkafolyamatok megvalósítására szolgáló motorral kombinálhatnánk. Ezt a megközelítést széles körben alkalmazzák a bioinformatika területén, ahol ugyanezzel a problémával kell szembenézni a több adattartóból származó adatok integrálásakor.

Példa: A Taverna²⁵ munkafolyamatok megvalósító motorját az elosztott számítástechnikai és tárolórendszerek által kínált webes szolgáltatások kombinálására használják; a munkafolyamatok ezután megoszthatók, újrafelhasználhatók és újrafelhasználhatók. myExperiment²⁶ egy példa a megosztható tudományos munkafolyamatok tárolására:



ábra Taverna11.

munkafolyamatok



ábra Forrás10.: A Bizottság a következő adatokat közölte: Carole Goble at al http://nar.oxfordjournals.org/content/38/suppl_2/W677.full

SZOCIÁLIS GÉPEK

Tim Berners-Lee és mások szerint a világháló "szociális géppé" fejlődik, azaz nem csupán egy információ-tár, hanem egy olyan infrastruktúra, amely a problémák közös megoldását szolgálja, ahol az emberek olyan feladatokat látnak el, amelyeket gépi úton nem lehet könnyen elvégezni. A technológia és a szociológia kölcsönhatásának megértése iránt érdeklődő társadalomtudósok tudni akarják, hogyan működnek együtt az emberek és a technológia olyan összetett feladatok megoldása érdekében, amelyeket egyikük sem tud egyedül megoldani.

²⁵ <http://www.taverna.org.uk/>

²⁶ <http://www.myexperiment.org/>



Figure Foldit12., <http://fold.it/portal/>

Kihívás: Hogyan lehet megragadni és megérteni a közösségi gépezetek felhasználói által a weben szerzett tapasztalatokat és interakciókat? Minden közösségi dolog az interakciókról szól. Ha nem értjük meg az interakciókat, soha nem érthetjük meg, hogy mi volt a gép szociális jellege.

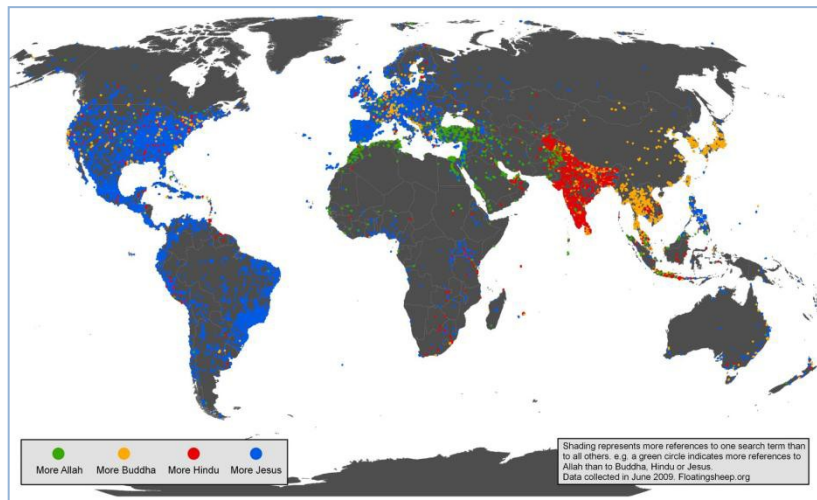
Példák: Az Amazon "Mechanical Turk"²⁷ egy olyan mechanizmus, amely problémákat oszt szét emberi szakértők között, és megoldásokat gyűjt. A tömeges forrásszerzés nehéz problémák megoldására is használható, pl. CAPTCHA az ember által támogatott optikai karakterfelismeréshez. Hogyan lépnek kapcsolatba a felhasználók a játékkal? Tanulságokat lehet levonni a kutatásból, amely arra irányul, hogy a játékosok hogyan interakcióznak online platformokon (pl. Williams, Yee, & Caplan, 2008).

HÁLÓZATOK FELTÉRKÉPEZÉSE

A geográfusok egyre gyakrabban használják az internetes adatokat az információk helyének, áramlásának és irányának, valamint a tartalom és a befolyás gazdagságának, szegénységének és változó alakjának megértéséhez időben és térben.

Kihívás: A földrajzi információk automatikus kinyerése a gyűjteményben található be- és kivezető linkekből, amelyek aztán feltérképezhetők. Ez jelenleg kihívást jelent az élő web esetében, és még inkább azzá válik, ha hozzáadjuk az időbeli változások összetettségét. A jelenleg kétdimenziós módszerekkel megjeleníthető információk nagy része három- vagy négydimenziós felületeket (például időbeli csúszkákat) igényel az idővel változó földrajzi információk értelmezéséhez. Például az egyetemen, kormányokon és vállalatokon belüli, illetve az egyetemek, kormányok és vállalatok közötti földrajzi befolyás időbeli megértése elméletileg lehetséges, de ehhez a földrajzi információknak az interneten található strukturálatlan adatokból való kinyerése szükséges.

Példa: FloatingSheep²⁸



Ábra FloatingSheep13. térkép a "Google's Geographies of Religion", <http://www.floatingsheep.org/2010/01/googles-geographies-of-religion.html>

WEB TUDOMÁNY

A webtudomány²⁹ a kutatók kísérlete arra, hogy a webet mint "információs műtárgyat" tanulmányozzák, és megértsék, hogyan növekszik és fejlődik, és hogyan alakulnak ki a "közösségek".

Kihívások: A "webgráf" mint matematikai objektum elemzéséhez hatékony eszközökre van szükség. Milyen a topológiája? Hogyan alakulnak ki a "klikkek"? Milyen skalázási törvények érvényesek (a webet valóban hatványtörvény szabályozza)? Hogyan terjed az információ a weben?

²⁷ <https://www.mturk.com/mturk/welcome>

²⁸ <http://www.floatingsheep.org>

²⁹ <http://webscience.org>

A világháló nem egyetlen információs tér, hanem alterek összetett és egymással összefüggő családja, amelyek információtartalmát néha egymástól elkülönülő közösségek határozzák meg. Hogyan történik az információ megosztása és terjesztése ezen részterületek között?

Ennek megválaszolásához olyan eszközöket kell kifejlesztenünk, amelyek képesek nyomon követni a fogalmak fejlődését és vándorlását az időben és a különböző részterületek között (pl. a blogoszféra és a "mainstream" média között).

Példák: MediaCloud³⁰, Recorded Future³¹



Ábra A MediaCloud15. földrajzi értelemben követi a hírek mozgását



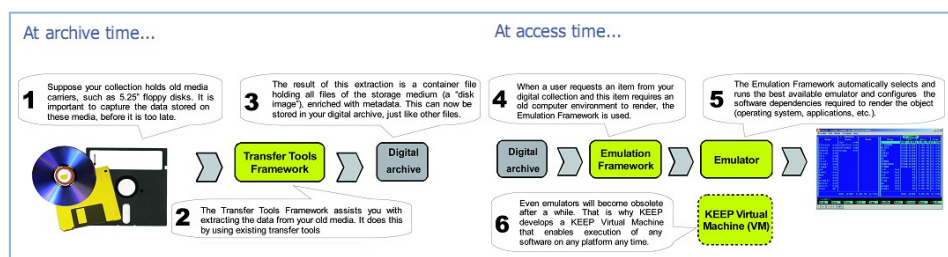
Ábra A Recorded Future a hírek mozgását követi nyomon az időben

A TARTALOM HELYETT AZ ÉLMÉNY MEGÉRTÉSE

A kutatók egyre inkább megértik, hogy fontos megérteni, hogyan használják az emberek a webes tartalmakat, nem csak magát a tartalmat. Ez figyelembe veszi a webes élmény és a végrehajtható tartalom állapotát: az élmény attól függ, hogy milyen platformot, milyen böngészőt/pluginokat/transzkódereket használnak, és egyre inkább a felhasználó tartózkodási helyétől is.

Kihívások: Az élmény megértéséhez képesnek kell lennünk arra, hogy újra létrehozzuk az élményt. A platformok, operációs rendszerek, böngészők stb. mind-mind megváltoztatják a webes élményt.

Példák: Browsershots³², KEEP (Keeping Emulation Environments Portable - Emulációs környezetek hordozható állapotban tartása)³³



KEEP ábra16. (Keeping Emulation Environments Portable - Emulációs környezetek hordozható állapotban tartása)

³⁰ <http://cyber.law.harvard.edu/research/mediacloud>

³¹ <https://www.recordedfuture.com/>

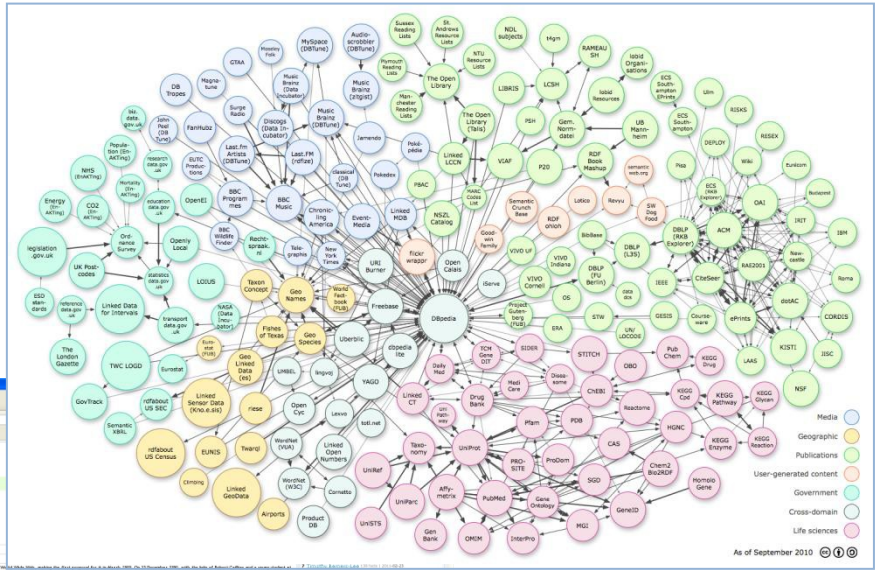
³² <http://browsershots.org>

³³ <http://www.keep-project.eu>

SZEMANTIKUS WEB ÉS ÖSSZEKAPCSOLT ADATGYŰJTEMÉNYEK ELEMZÉSE

A kapcsolt adatgyűjtemények száma gyorsan növekszik, az ismert gyűjteményekben legalább 28,5 milliárd tripla található. A szemantikus web/kötött adatok közössége által kifejlesztett eszközök radikálisan leegyszerűsítik az archívumok metaadat-kézelési problémáját, és lehetővé teszik, hogy a metaadatokban nagy léptékben lehessen keresni, valamint sokkal kifinomultabb módon használhatók a gyűjtemények közötti adatintegrációra.

Példák: Sindice/Sig.ma³⁴RDF-alapú keresés



Sindice/Sig18..ma RDF-alapú keresés. Forrás: <http://sig.ma/search?q=Tim%20Berners%20Lee>

³⁴ <http://sig.ma/>

De milyen lépéseket tehet a webarchiváló közösség?

Néhány dolog, amire a kutatóknak szükségük van, nyilvánvalónak tűnhet, de ez nem jelenti azt, hogy jelenleg rendelkezésre állnak. A JISC korábbi jelentéseiben, amelyeket kollégáinkkal közösen írtunk, és amelyeket fentebb már tárgyaltunk (Dougherty, et al., 2010; Thomas, et al., 2010), számos ajánlást fogalmaztunk meg, amelyek három fő téma köré csoportosulnak: **közösségépítés, eszköz- és forrásépítés, valamint gyakorlatépítés** (Dougherty, et al., 27-29, 2010, .). Nem szeretnénk itt az említett jelentések ajánlásainak teljes listáját reprodukálni - a Dougherty, et al. 22 konkrét ajánlást tartalmaz, a Thomas, et al. pedig további 20-at -, ezért arra biztatjuk az olvasót, hogy tekintse meg ezeket a jelentéseket és ezt a jelentést is. A mi céljaink érdekében azonban kiemelhetünk néhány kulcsfontosságú témát, amelyek iránt a kutatók érdeklődnek, és azonosíthatjuk azokat a kihívásokat, amelyekkel az archívumok részt vehetnek annak érdekében, hogy a webarchívumok a különböző tudományágak kutatóinak standard eszköztárába kerüljenek.

A jelentés ezen szakasza néhány olyan témát és kérdést vázol fel, amelyeket a kutatók a webarchívummal - a dobozos archívummal - kapcsolatban fel akarnak vagy fel akarnak majd tenni, és amelyekkel kapcsolatban azonosítottunk néhány kihívást és lehetséges megoldást. E megoldások némelyike, különösen a rövid távú megoldások, az intézmények szintjén is megvalósítható. A hosszabb távú megközelítések közül sok esetben szélesebb körű megközelítésre lenne szükség, nemzeti, regionális vagy nemzetközi szinten, olyan szférvetéseken keresztül, mint az IIPC.

Azért említjük az archívum-a-dobozban-t, mert egyesek úgy érzik, hogy a webes archiválás túlmutat a korai időközön, amikor az *oldalakat* későbbi elemzés céljából elérhetővé tették (a WayBack Machine klasszikus felületének értelmében, ³⁵amely lehetővé tette a felhasználó számára, hogy elsősorban az archívum egyes oldalaihoz férjen hozzá és nézze meg azokat), és a *gyűjtemények* kutatási eszközként való hozzáférhetővé tétele felé halad. Például, ha egy kutató "az Egyesült Királyság kormányzati .gov.uk domainjét 2011-2020 között" látja, mit képzelhet el, mit tud vele kezdeni? Milyen kérdéseket lehet feltenni egy olyan gyűjteménynek, amely például a Wall Street-i nagybankok teljes webes tartalmát és a hozzájuk közvetlenül kapcsolódó oldalakat tartalmazza, ha az archívum egy olyan időszakot ölel fel, amely során bankválság alakult ki? Más szóval, ahelyett, hogy mikroszinten csak egyetlen webhelyet elemeznénk, vagy makroszinten az egész webet elemeznénk, mit tehetünk a web mézoszintű célzott részhalmozásával? A társadalomtudományi kutatások nagy része az offline térben vizsgálja a mézoszintű interakciókat; használhatjuk-e az internetes archívumokat ugyanerre a **h** hogy megértsük, hogyan tükrözi, erősíti és módosítja a társadalmi valóságot a változó internet?

Néhány dolgot lehetetlen lesz támogatni a meglévő webarchívumokban - az ehhez szükséges adatokat vagy tartalmakat nem biztos, hogy összegyűjtötték, és már elvesztek. A jövőre nézve azonban **milyen változtatásokat tudunk végrehajtani a webarchívumokban ma és az elkövetkező években**, hogy a kutatók vagy 2015, 2020, képesek 2050 legyenek a most gyűjtött forrásokra támaszkodni, hogy megválaszolják ezeket a kérdéseket? Mit fognak a jövő kutatói elvárni tőlünk 2011-ben és a jövőben, amit most nem teszünk meg? Mit tehetnek az egyes intézmények? Mi történhet jobban vagy hatékonyabban, ha az IIPC közösen dolgozik, kihasználva a több archívum erejét?

A KUMULATÍV HÁLÓ: EGY ÉLŐ WEBARCHÍVUM

Kérdés: Miért kell a webarchívumoknak archívumoknak lenniük? Miért nem lehet integrálni őket az élő webbe, átláthatóan elérhetővé téve őket a nyilvánosság és a kutatók számára? Elképzelhető egy olyan réteges web, amelynek felszínén a jelenlegi élő web az adatok és információk alapértelmezett forrásaként áll rendelkezésre. Ez a felszín azonban a múltbeli web mögöttes rétegeire épülne, és bárki számára könnyen elérhető lenne, aki érdeklődik, egyszerűen egy vagy több réteggel lejjebb érve. Ha az élő web kutatásához rendelkezésre álló számtalan eszköz egyszerű mechanizmusok segítségével alkalmazható ezekre az alsóbb rétegekre, akkor nő annak a valószínűsége, hogy a kutatók felhasználást találnak a múltbeli webet alkotó adatokra és információkra.

Valószínűleg ez a jelentés legnagyobb és legambiciózusabb kihívása, mivel a web infrastruktúrájának megváltoztatását igényli. Bár ez azt jelenti, hogy ennek megvalósulásának valószínűsége alacsony, a potenciális előnyök nagyok. A kutatási érteken túlmenően, hogy a múltbeli web a jelenlegi web alatt rétegesen elérhetővé válik, ez a felhasználók webes szemléletében is tektonikus változást eredményezhet. A jelenlegi webet sokan megbízhatatlannak tartják a halott linkek, a hiányzó információk, az eltűnő oldalak, a megváltozott URL-ek és a változó információk miatt, amelyek felülírják a régebbi verziókat, anélkül, hogy a korábbi verziókat meg lehetne tekinteni vagy vissza lehetne térni hozzájuk. Ha az internet struktúrája többretegűvé változna az időben visszafelé haladva úgy, hogy a legfelső rétegben keletkező lyukak nem lyukat ütneek a weben, hanem egy alsóbb réteget tennének szabaddá, lehetővé, hogy a webet stabil, megbízható információforrásnak tekintenek, amely ellenáll az információvesztésnek.

³⁵ <http://classic-web.archive.org/>

A linkek eltűnése, más néven linkrot, állandó probléma az élő web felhasználóinak. A probléma tovább súlyosbodik, ha az archivált webet vesszük figyelembe, amely nagyrészt nem rendelkezik a weboldalak archivált változatainak tartós azonosítóival. Történt néhány erőfeszítés. A WebCite³⁶ például lehetővé teszi a szerzők számára, hogy archiválják egy weboldal egy példányát, és létrehozzanak egy megőrzött linket vagy DOI-feloldót. A DeadURL³⁷ más megközelítést alkalmaz, és többek között az Internet Archive és a Google gyorsítótárára támaszkodva próbálja megtalálni a halott linkek mentett példányait. Az ehhez hasonló erőfeszítéseket azonban a kutatók túlnyomó többsége nem használja ki, és többnyire nem is tudnak a létezésükről. A kényelmetlenségi tényezőn túl van egy másik, alattomosabb, nem szándékolt következménye is: a kutatók azon szokása, hogy minél erősebb URL-címet adjanak meg tudományos munkájukban. Ennek több hatása is van. Először is, amikor az online források megpróbálják felmérni hatásukat olyan technikákkal, mint a webometrika, a linkek hiánya miatt a forrásuk kisebb hatásának tűnik. Másodsorban, az információ forrásának felkutatására törekvő olvasóknak sokkal nehezebb dolguk lesz, mivel nem csak az idézet helyes forrását próbálják megtalálni, hanem a forrás idézett változatát is, mivel az oldalak jelentősen megváltozhattak. Ha a webarchívumok megbízható forrassá válnak az online információk idézésére, az javítani fogja a tudományosságot, és általánosságban is növeli a webarchívumok ismertségét.

Ebben a kumulatív hálóban a hálóban megtestesülő tudás növekszik és fejlődik, de nem vetődik el ugyanúgy, mint a jelenlegi hálóban. A kumulatív web kereshető lenne a Google-hoz hasonló keresőmotorok segítségével, feltérképezhető, törölhető, összekapcsolható és elemezhető lenne. A linkek nem halnának meg, hanem a múltból származó, a web jelenlegi aktív rétegében már nem létező anyagokra irányulnának.

Hosszú távú kihívás: A kérdés két kihívást foglal magában, amelyek mindegyike számos érdekelt felet és szereplőt érint. Először is, újra kellene gondolnunk, hogyan látjuk és tervezzük az internetet, egy egyrétegű, sok oldalirányú háló egységből egy többretegű egységgé valva, amely a jelenlegi oldalirányú linkekkel, de a régi anyagokra mutató jelenlegi linkekkel és a régi vagy jelenlegi anyagokra mutató régi linkekkel is rendelkezik. Ez nem triviális kihívás, és nehéz lenne meggyőzni a jelenlegi és jövőbeli internet építésében érdekelt számos szereplőt, hogy fogadják el. Ez azonban olyan infrastruktúrát eredményezne, amely a múltbeli webet sokkal jobban hozzáférhetővé tenné a referenciák és a kutatás számára. A második nagy kihívás az lenne, hogy a webarchivátoroknak újra kellene definiálniuk szerepüket, valójában nem is a hagyományos értelemben vett archivátoroknak kellene lenniük, hanem olyan szakembereknek, akik segítenek a kutatóknak értelmet adni az internetes trendeknek és forrásoknak az idő múlásával, és akik szakértői a többretegű internet rétegeihez való hozzáféréshöz és azok manipulálásához szükséges eszközöknek, hogy a kutatókat és a nyilvánosságot irányítsák, miközben a növekvő web új, előre nem látott kérdéseket tesz fel.

A VÁLTOZÓ VILÁGHÁLÓ

Kérdés: Hogyan tudnak a kutatók reagálni a világ változó eseményeire, vagy nyomon követni a folyamatban lévő eseményeket? A helyi és a világméretű események egyre inkább a világhálón játszódnak le. Ezek lehetnek nemzetközi jelentőségű és érdeklődésre számot tartó események, mint például a közelmúltbeli észak-afrikai és közel-keleti politikai események vagy a haiti, japán és más földrengések; lehetnek helyi regionális jelentőségű események; lehetnek kisebb, de fejlődő vagy folyamatban lévő események is, amelyek elsősorban egy kis csoportot vagy akár egyetlen kutató érdeklődésére tarthatnak számot. Ezáltal különféle betekintést nyerhetünk abba, hogy milyen jellegű információkat oszthatnak meg az emberek, milyen témák és események kerülnek előtérbe, hogyan reagálnak az egyének, kormányok és szervezetek a válságokra, és idővel hogyan gyarapodnak és hanyatlanak az események a közbeszédben.

Bizonyos szempontból ez a legegyszerűbb kihívás, mert ez a legnyilvánvalóbb. Ráadásul számos kutató már dolgozik ezen a területen. A 2011-es IPCC-ülésen az eseménygyűjtés több előadás témája volt, az előadók az eseménygyűjtés számos példáját tárgyalták a 2011-es arab világbeli forradalmak, a Deepwater Horizon olajkatasztrófa és a londoni 2012 olimpia kapcsán.

A webarchívumokkal való munka egyik központi kérdése, hogy a webet nem keresztmetszeti adatként kell értelmeznünk, hanem egy változó, fejlődő hálózatnak kell tekinteniünk, amely idősorokat és más longitudinális megközelítéseket igényel. Vannak erőfeszítések ezen a fronton. Például az európai *Longitudinal Analytics of Web Archives* projekt³⁸ egy *webes megfigyelőközpontot* hoz létre, amely lehetővé teszi a longitudinális elemzést. Más erőfeszítésekre is szükség van.

Közvetlen kihívás: A kutatók számára olyan mechanizmusok létrehozása, amelyek segítségével gyorsan javasolhatják a nagyobb részletességű és megfelelő terjedelmű archiválást a változásban lévő oldalak és témák archiválásához. Jelenleg egy képzett kutató beállíthat olyan eszközöket, amelyekkel ismételt feltérképezéseket végezhet, de a technikailag kevésbé képzett, erős intézményi támogatással nem rendelkező kutatóknak meredekebb a tanulási görbe. Ha egy gyorsan változó esemény

³⁶ <http://www.webcitation.org/>

³⁷ <http://deadurl.com/>

³⁸ <http://www.lawa-project.eu/>

fejlesztésével a szakkutatóknak, aki érdeklődik az adott esemény iránt, de nincs tapasztalata a webes archíválásban, módot kell találnia arra, hogy összegyűjtse az adatokat elemzésre, mielőtt azok elvesznének. Azok a szervezetek, amelyek rendelkeznek az ilyen kialakuló helyzetekre vonatkozó adatok gyűjtésére szolgáló eszközök megfelelő szintű granularitású beállításához, szükséges készségekkel, módot biztosíthatnak a kutatók vagy mások számára weboldalak, témák, kulcsszavak stb. jelölésére, hogy gyorsan reagálhassanak a változó webes eseményekre.

Fejlődési kihívás: Használjon olyan eszközöket, mint például az RSS-feedek, hogy jelezzék, hogy a weboldalakon bekövetkezett változásokat archiválni kell. A következőkkel kapcsolatban

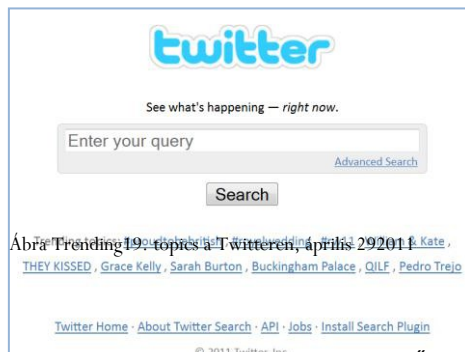
eseményeket, lehetőség van olyan dolgok nyomon követésére, mint az RSS

feedek vagy az újonnan

olyan alkalmazások fejlesztése, amelyek rendszerei a weboldalak rögzítésének gyakoriságának növelésével reagálnak a növekvő aktivitásra, vagy a

a humán kurátorok értesítése a potenciális érdeklődésre számot tartó területek kialakulásáról.

Hosszú távú kihívás: Olyan algoritmusok létrehozása, amelyek az online aktivitási trendeket (például a Google trendeket vagy a Twitter trendek számított témákat) használják fel az adott témához kapcsolódó weboldalak fokozott archiválásának kiváltására. Ez nagyobb kifinomultságot és készségeket igényel, hogy a létrehozott archívumok alkalmasak legyenek az újrafelhasználásra és megosztásra, szabványosításra, és fenntarthatóak legyenek. Az ilyen algoritmikusan gyűjtött archívumok használatában érdekelt kutatóknak ismerniük kell a felvétel vagy kizárás logikáját, és még kell tudni érteniük a gyűjtemény jellegét és tartalmát. Az egyik központi kérdés az lesz, hogy hogyan illeszkednek ezek a gyűjtemények a felhasználható és hozzáférhető kutatási források ökoszisztémájába?



Ábra: Trending 190 topics a Twitteren, április 29. 2011. Kate, THEY KISSED, Grace Kelly, Sarah Burton, Buckingham Palace, QILF, Pedro Trejo

AZ ARCHÍVUMOK ÉS WEBOLDALAK HASZNÁLATA

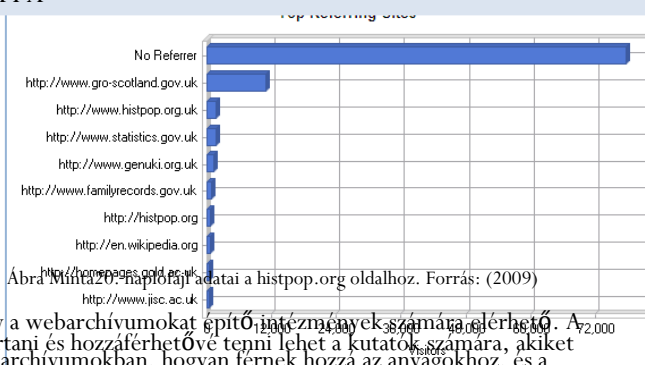
Kérdés: Hogyan használják az emberek a webarchívumokat, és ami még fontosabb, a weboldalakot? Jelenleg nagyon is lehetséges, hogy a meglévő webarchívumok és webarchiválási infrastruktúra segítségével megnézzük azokat a weboldalakot, amelyek egy adott időpontban jelen voltak a világhálón. A tudományos és ipari kutatók számára a szervernapló-elemzés és az analitika elemzése gyakori technika az élő web használatának, hatásának és forgalmi mintáinak értékelésére. Ezek a technikák azonban nem lehetségesek az archivált web esetében, így a múltbeli web és a webarchívumok használatának megértéséhez szükséges adatok nem állnak rendelkezésre.

Közvetlen kihívás: A webarchívum oldalak szervernaplóinak archiválása, hogy a kutatók tanulmányozhassák a webarchívumok működését.

használnak. Ez a legegyszerűbb és legegyszerűbb megoldás, amely a webarchívumokat építő intézmények számára elérhető. A webarchívumokkal kapcsolatos szervernaplókat tárolni, karbantartani és hozzáférhetővé tenni lehet a kutatók számára, akiket érdekel, hogy megértsék, hogyan navigálnak a felhasználók a webarchívumokban, hogyan férnek hozzá az anyagokhoz, és a webarchívum mely részeit használják a leggyakrabban. Ezek az információk elsősorban a webarchíváló közösséget érdekelnék, de ez az első lépés.

Hosszú távú kihívás: Ez egy ambiciózusabb erőfeszítés, de sokkal szélesebb körű potenciális érdeklődésre tarthat számot: olyan infrastruktúra létrehozása, amely lehetővé teszi az archivált weboldalakhoz kapcsolódó és azokhoz kapcsolódó szervernaplók és elemzések archiválását, hogy a kutatók ne csak azt lássák, mi volt a weben, hanem azt is, **hogyan használták azt**. Ez sokkal ambiciózusabb cél, mivel a szervernaplók és az analitikai fiókok csak belsőleg, védett üzemmódban láthatók a szerver- és fiókadminisztrátorok számára. A szerveradminisztrátorok általános gyakorlata nem feltétlenül a szervernaplók hosszú távú tárolása, mivel azokat helytakarékoság és a szerver zsúfoltságának elkerülése érdekében rutinszerűen törlik vagy felülírják. Ezek az adatok azonban potenciálisan értékesek lehetnek a kutatók számára, akik nem csak azt szeretnék tudni, hogy egy webhely egy adott állapotban létezett, hanem azt is, hogy hogyan használták, mennyire használták, milyen forgalmi forrásokból és egyéb, a naplóból és az analitikai adatokból kinyerhető tényekből. A lehetséges megoldások közé tartozik, hogy a szerveradminisztrátorok olyan mechanizmusokat hozzanak létre, amelyek segítségével a naplótárolt az archivált weboldalakhoz lehet társítani, és az elemzési szolgáltatók, például a Google, lehetőséget biztosítanak arra, hogy a webhelyek elemzési adatait a webarchívumokhoz lehessen csatolni, esetleg megszabva egy embargós időszakot az adatok kiadása előtt.

Ambiciózus kihívás: Olyan rendszerek tervezése, amelyek nemcsak a szervernaplókat, hanem **magát a webes forgalmat is** biztonságos, anonimizált módon archiválják. Ez a megoldás még ambiciózusabb, mivel a webes forgalom elemzésére szolgáló mechanizmusok nagyrészt nem nyilvánosak. Vannak



Ábra: Múltidő naplótárolt adatai a histpop.org oldalhoz. Forrás: (2009)

adatvédelmi aggályok az olyan dolgokkal kapcsolatban, mint a mélyreható csomagvizsgálat, amely az elemzők számára információt szolgáltat a világhálón keresztül áramló forgalomról. Az egyik központi kérdés az, hogy az emberek webes viselkedésének megértéséből származó előnyök mikor haladják meg az egyénekre jelentett kockázatokat. Ezért érdemes megfontolni, hogy vannak-e módok arra, hogy ezeket az adatokat archiválni és biztonságosan tárolni lehessen későbbi elemzés céljából, amikor az egyénekre vagy szervezetekre vonatkozó kockázatokat az idő múlásával és az adatok anonimizálásával már kellőképpen csökkentették.

A SZAKOSODOTT WEB

Kérdés: Lehetséges-e olyan gyűjtemények azonosítása, amelyek a történelmi háló méretét és alakját mérik a speciális érdeklődési területekre vonatkozóan? Ha feltételezzük, hogy az idők folyamán számos csoport vagy testület hoz létre egy sor weboldalt, akkor hogyan lehet őket meghatározni, azonosítani webes jelenlétük koherenciáját, és összegyűjteni a kutatáshoz szükséges oldalakat? Számos példa juthat eszünkbe: hobbicsoportok, speciális tudományos témák, örökség tárgyak, például hangok és képek, politikai csoportosulások honlapjai és így tovább. Ezek tanulmányozása magában foglalhatja meta-kollekciók létrehozását: gyűjtemények gyűjteményei, az archívum a fobozban. A különböző internetes források archívumgyűjteményeinek létrehozásakor milyen útmutatásra van szükség a gyűjtemények értékének növeléséhez? Mennyire lehet egy meta-gyűjteményt összehasonlítani egy másikkal, milyen mértékben lehet őket összekapcsolni és még nagyobb meta-gyűjtemények ~~szélesíteni~~?

A szakosodott web elismeri, hogy az eseményeken vagy témákon alapuló, kisebb léptékű, szelektív adathalmazok iránti igény továbbra is igen fontos, és összhangban van a meglévő kutatási gyakorlatokkal és elvárásokkal. Az ilyen típusú "korpusz" továbbra is egyetlen személy vagy csapat által megfigyelhető és kereshető marad, aki ezt az adathalmazt egy tudományág vagy egy kutatási téma saját szemszögéből hozza létre és elemzi, továbbra is fontos a tudományos világban, különösen azokon a területeken, ahol erős hagyománya van annak, hogy a kutatók saját korpuszuk alkotói, kezelői és elemzői. Azonban még ezek a speciális gyűjtemények is potenciálisan hozzáadott értékkel bírnak, ha olyan szabványos módon hozzák létre őket, amelyek később más kutatási kérdések megválaszolásához újra kombinálhatók és atkonfigurálhatók.

Ez felveti a skálázhatóság kérdését, ami számos kérdést vet fel a speciális webarchívumokkal kapcsolatban: Van-e kritikus méret a webarchívum gyűjteményének lefedettsége vagy terjedelme szempontjából, hogy a kutatók számára hosszú távon hasznos legyen? Hogyan felelnek meg a kutatók elvárásainak a különböző intézményi webes feltérképezési stratégiák és politikák (például tömeges/domain harvesting vagy szelektív/eseményes harvesting), és milyen felhasználási módok támogatására alkalmasak a legjobban különböző stratégiák?

Néhány domáinspecifikus példa a társadalomtudományi eszközökben található speciális gyűjteményekre, amelyeket ki lehetne fejleszteni a Világbank adatainak és a Wolfram Alpha egészségügyi vagy más népszerűvel kapcsolatos elemzéseknek a kombinálására. Ehhez olyan eszköz (például a Wolfram Alpha) használata szükséges, amely folyamatosan frissül új adatokkal, valamint algoritmusaival és vizualizációs eszközeivel, így egy élő adathalmaz elemző "készletet" biztosít. A természettudományok területén a kutatókat érdekelné, ha az éghajlatra (hőmérsékletre) vonatkozó adatokat vennék, és összekapcsolnák ezeket az amatőr természetjárók (pl. madármegfigyelők) által a világ különböző csoportjaiban (olyan eszközökkel, mint a Google+ vagy a Facebook) gyűjtött információkat, és egyesítenék erőiket annak elemzésére, hogy az időjárás hogyan változtatja meg a madarak vonulási szokásait, vagy hogy a madárvonulás mit árulhat el az éghajlatváltozásról. Ehhez ismerni kell a polgári tudományt, az online közösségeket és a környezetvédelmet. A bölcsészettudományok területén egy eszmétörténet összefasonlíthatja Alan MacFarlane interjúit az Oxford iTunes U előadásában szereplő jelentős kortárs társadalmi gondolkodókkal, hogy szembeállítsa a társadalomtudományok elképzeléseit a jelentős gondolkodókról és az iTunes U legnépszerűbb gondolkodóiról (más szóval, összehasonlíthatja a "kánont" a "népszerű gondolkodással").

Közvetlen kihívás: Útmutatás nyújtása a különböző típusú metakollekciókhoz, hogy a webarchívumban szabványos elemek szerepeljenek, valamint hogy a terjedelem és a koherencia meghatározott és biztosított legyen.

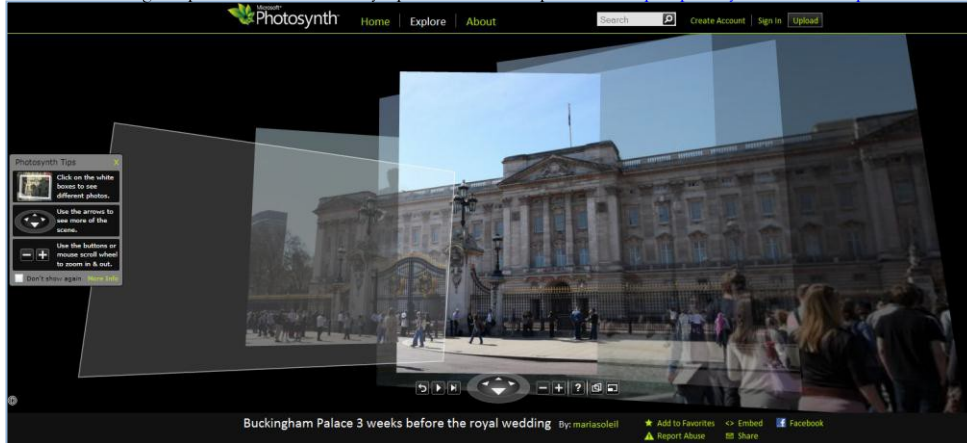
Fejlődő kihívás: Olyan eszközök és szervezeti infrastruktúrák biztosítása, amelyek biztosítják, hogy az egyes kutatók leküzdjék a bizonytalanságokat, hogy olyan meta-kollekciókat hozhassanak létre, amelyek legalábbis potenciálisan az egyes kutatókon túl is használhatóak. Szabványok meghatározása annak érdekében, hogy az archív meta-kollekciók egymással is használhatóak legyenek.

Hosszú távú kihívás: Olyan szervek létrejöttének ösztönzése, amelyek támogatják és ösztönzik a hasznos és széles körű metakollekciókat.

A VISUÁLIS HÁLÓ

Kérdés: Ha az interneten található képeket arra akarom használni, hogy megértssem, hogyan változik a világ, ki tudok-e nyerni képeket az internetes archívumokból, hogy vizuálisan is megértssem ezt a folyamatot? Lehetséges lesz-e például vizuális elemzést végezni az idő múlásával azáltal, hogy ugyanazokról a helyekről változó képeket vonok ki weboldalakról és olyan oldalakról, mint a Flickr? Az újrafényképezés az a gyakorlat, amikor egy korábban már lefényképezett helyszínt felkeresünk, és új fényképet készítünk, hogy dokumentáljuk a folytonosság és a változás pontjait az idő múlásával. Az egyik legkorábbi ilyen jellegű projekt 120, a kormányzat földmérői által 100 évvel korábban lefényképezett helyszínt vizsgált újra az amerikai Nyugaton (Klett, Manchester, & Verburg, 1984). Most képzeljük el, hogy évek 100 múlva nem csak egyetlen, ugyanazon a helyen készült fényképet tudunk összehasonlítani, hanem a webarchívumokból egy egész fotósorozatot tudunk kinyerni, amely a körülöttünk lévő változó és statikus világot dokumentálja az idők során.

21. ábra. A Buckingham-palota 220 különböző fényképből összeállított képe. Forrás: <http://photosynth.net/view.aspx?cid=34e49d3e-2d1e-4118-bbad-d2f5d74ce340>



[bbad-d2f5d74ce340](http://photosynth.net/view.aspx?cid=34e49d3e-2d1e-4118-bbad-d2f5d74ce340)

Közvetlen kihívás: Biztosítani, hogy a képek, amelyek túl gyakran hiányoznak az archivált weboldalakról, prioritást élvezzenek a megőrzés szempontjából.

Fejlődő kihívás: Az olyan technológia, mint a PhotoSynth³⁹, amely képes nagyszámú fényképet egy hely vagy tárgy panorámaképevé összeilleszteni, időbeli információkkal együttműködve hasonló nézeteket állíthat össze az idő múlásával.

Hosszú távú kihívás: A világ fényképeit tartalmazó archívum létrehozása, amely a lehető legtöbb idő- és helyinformációt tartalmazza az EXIF- és weboldalak adataiból, hogy a fényképek felhasználhatók legyenek a kutatásban. Eszközökre lenne szükség a képek kereséséhez, kinyeréséhez, kombinálásához és manipulálásához.

A WEB, AHOGY VOLT

Kérdés: Hogyan láthatom a világhálót úgy, ahogy volt? Ha úgy szeretnék navigálni a weben, ahogyan az mondjuk 2011, január 01-én volt, és úgy tudnék kattintani az oldalakra, képekre, linkekre és egyéb tartalmakra, ahogyan az akkor volt, hogyan tudtam ezt megtenni? A Wayback Machine újrajátszási verziójának jelenlegi béta verziója ilyen funkciót "úgy" ("Surf the Web as it was - BETA version!") az oldal jelenleg

Ábra A WayBack Machine visszajátszási változatának béta22. verziója. Forrás:



<http://replay.web.archive.org/2004100185532/http://netpreserve.org/about/index.php>

³⁹ <http://photosynth.net/Background.aspx>. A Microsoft korai PhotoSynth munkájáról szóló videós bemutatót lásd a http://www.ted.com/talks/blaise_aguera_y_arcas_demos_photosynth.html oldalon.

⁴⁰ <http://web.archive.org/>

buzdít), de milyen egyéb erőfeszítéseket tehet az IPC vagy az egyes archívumok a lejátszható web lehetőségeinek bővítése és javítása érdekében?

Kihívás: A jelenlegi web jövőbeni újrakészítésére irányuló erőfeszítések kiterjesztéséhez és fokozásához a jelenlegi web összegyűjtésére, tárolására és újbóli feltárással szemben szükség van a múltbeli webdé válásra. A központi kérdés, amit itt fel kell tenni, az, hogy hogyan lehet ezt a pusztán kuriozumon vagy alkalmi referenciaként túl felhasználni. Milyen kiaknázatlan igény vagy elképzelhetetlen kutatási kérdés támaszkodna a múltbeli web kézi keresésére és böngészésére? Vajon a jövő történészei, akiket a mai világ érdekel, ugyanúgy fogják-e olvasni a világhálót, mint ahogyan a múlt korok híreit, kiadványait és címerjeit olvasták? Alkalmazásokat akarnak majd használni, vagy csupán dokumentumokat akarnak majd tanulmányozni? Más szóval, a legnagyobb kérdés az, hogy milyen felhasználási eseteket kell kialakítani a lejátszható web számára, majd olyan interfacseket kell létrehozni, amelyek támogatják ezeket a felhasználási eseteket. Ehhez a webarchivalással foglalkozó szakembereknek konzultálniuk kell a szakterületen dolgozó szakemberekkel, köztük történészekkel és a múlt rekonstrukciójában érdekeltekkel.

A WEB SZERKEZETE

Kérdés: Mi alkotja a webet, és hogyan változik ez az idő múlásával? A web mint rendszer megértésére irányuló fokozódó erőfeszítésekhez olyan nagy léptékű elemzési képességekre lesz szükség, amelyek képesek lesznek az idő múlásával kialakuló minták és tendenciák feltárással. Ennek során el kell kezdenünk feltenni a kérdést, hogy milyen megközelítések állnak rendelkezésre érvényes elemzési módszerek kidolgozásához? Hogyan tudjuk validálni az archív webes adatokra mint adathalmazra vonatkozó feltételezéseket? Milyen statisztikai eszközök alkalmazhatók webarchívum-gyűjteményekre, és milyen új eszközöket kell kifejleszteni?

Jelenleg még az egyszerű statisztikákat sem triviális kinyerni a webről. Például mennyi volt a weboldalak éves száma (világszerte, egy adott országban, egy adott témában) az elmúlt X évben? A dobozos archívumgyűjteményekben mi az oldalak létrehozásának dátuma? Milyen nyelven vannak az oldalak? Vannak-e tendenciák az oldalak létrehozásának időpontjában? Van-e klaszterezés? Vagy ez egy egyenletesen növekvő folyamat? Bizonyos témák jobban kapcsolódnak egymáshoz, mint mások? Egyes típusú gyűjtemények nagyobb vagy kisebb valószínűséggel hivatkoznak külső forrásokra? Lehet-e a weboldalakat olyan kategóriákba sorolni, amelyeket klaszterelemzéssel feltehetünk?

Összehasonlíthatjuk-e az oldalakat olyan statisztikák alapján, mint a weboldal átlagos mérete a különböző kategóriákban, a linkek átlagos száma, a nem szöveges adatok mennyisége (képek, képek stb.), a tartalom kora a frissítések között, a frissítések gyakorisága, a felület típusa (például statikus vs. dinamikus).

Hogyan segíthetnek ezek a statisztikák a gyűjtemények és a web szerkezetének megértésében?

Kihívás: Eszközök és módszerek létrehozása a web mint hatalmas adathalmaz, nem pedig dokumentumok gyűjteménye használatára. Jelenleg, ha valaki meg akarja tudni, hogy milyen alapvető kérdéseket kell feltennie akár a jelenlegi web, akár a korábbi időpontokban létező web méretéről és felépítéséről, akkor ezek az adatok nem állnak a kutatók rendelkezésére. Ezért olyan eszközöket kell létrehozni, amelyek képesek a web vagy egy archív gyűjtemény oldalainak összeírására.

HOGYAN TERJEDNEK AZ ÖTLETEK

Kérdés: Hogyan nyerne teret és terjednek el az ötletek az interneten? Az internet egyik lenyűgöző aspektusa, hogy hihetetlenül képes támogatni a memék, azaz a kulturálisan növekvő és terjedő eszmék terjedését. Ha arra vagyunk kíváncsiak, hogyan terjed egy videó, hogyan terjed egy vicc, vagy hogyan jut el egy információ vagy felretájékoztató gyorsan a köztudatba, milyen eszközök segítenek ebben? Hogyan építhetünk be olyan képességeket az eszközökbe, amelyek lehetővé teszik, hogy egy archívumot ne a fizikai vagy virtuális földrajz, hanem egy eszme mozgása alapján építsünk fel? Elképzelhető, hogy képesek leszünk meghatározni egy eszmét, és aztán kimászni, hogy kövessük az eszmét, ahogyan az idővel fejlődik.

Továbbá, milyen tágabb kontextusba helyezi az archívumban látható tartalmat? Például mit kerestek az emberek a weben, amikor a felvételek készültek? A Google Zeitgeist⁴¹ és a Google Trends⁴² elárul valamit arról, hogy az emberek mire kerestek; mit tudunk még gyűjteni kontextus megértéséhez? A Twitter említések például segítenek megérteni a tartalom kontextusát, ha megnézzük, milyen dolgokat említenek együtt. Az IBM Watson például egy saját fejlesztésű rendszer, amely sok archív webes anyagot használ fel az archivált webes anyagokból.

⁴¹ <http://www.google.com/press/zeitgeist2010/>

⁴² <http://www.google.com/trends>

DeepQA⁴³ motor, amely segít nekünk a Jeopardy győzelemhez a Jeopardy-ban Hogyan2011. lehet az ilyen típusú eszközök szélesebb körben elérhetővé tenni a kutatás előmozdítása érdekében?

Kihívás: A web idődimenzióját, ahogyan az létrejön, meg kell őrizni, ki kell nyerni és elemezhetővé kell tenni. Finomabb részletességre van szükség ahhoz, hogy látni lehessen, hol kezdődtek az ötletek, hogyan terjedtek, és milyen tevékenységek növeltek vagy csökkentették a terjedés sebességét. Az ötletek organikusán jelennek meg a világban, és csak akkor lesz érdekelte az embereket, hogy visszavessék őket eredetükig. Az archiválás megfelelő szemcsézettisége és mélysége nélkül azonban az eredeti ötlet elveszhet, mire valakinek eszébe jut, hogy megkeresse.

A TILTOTT HÁLÓ

Kérdés: Hogyan használják a világháló a tiltott tevékenységek támogatására és lehetővé tételére, és hogyan változik ez az idő múlásával? Az egyik olyan tartalomtípus, amely elszaporodott az interneten, de meglehetősen kevés tudományos figyelmet kapott, a világhálón található tiltott anyagok közé tartozik. Ezek a széles körben elterjedt szexuális tartalmaktól kezdve a tiltott kábítószer-használatra, illegális szerencsejátékokra, gyűlöletcsoportok anyagaira, terrorizmussal kapcsolatos tartalmakra és egyéb illegális vagy társadalmilag problematikus anyagokra terjednek ki. A kérdés itt az, hogy kinek, ha egyáltalán kinek kellene archiválnia a világháló illegális és legális, de társadalmilag kevésbé elfogadott tartalmát? Hogyan lehet ezt megtenni anélkül, hogy törvényt sértene, és hogyan lehet ezt a kutatók számára hozzáférhetővé tenni anélkül, hogy veszélybe kerülne akár a kutató, akár a hozzáférést biztosító intézmény? Annak ismerete, hogy mely illegális tevékenységek mennyisége és népszerűsége növekszik, melyeké csökken az idő múlásával, és milyen újonnan megjelenő illegális tevékenységek jelennek meg, nemcsak a kutatók, hanem a közpolitikai döntéshozók, a veszélyes viselkedés eredményeit kezelő egészségügyi szakemberek, a közegészségügyi szakértők, a szociális támogató ügynökségek és szakemberek, valamint a kiszolgáltatott népességcsoportok jólétének védelméért felelős személyek számára is hasznos.

Kihívás: A legnagyobb kihívást az jelenti, hogy bár a tiltott anyagok gyakoriak az interneten, a bűnüldözésen kívül kevés szervezet hajlandó vállalni az ilyen anyagokra vonatkozó adatok összegyűjtésével járó kockázatot. Az illegális anyagokhoz való hozzáféréssel és tárolással kapcsolatos kulturális tabuk és jogi kockázatok - még az olyan pozitív célok, mint a modern társadalom e kevésbé vizsgált aspektusának kutatása esetében is - elriasztják a legtöbb webkutatót és webarchiválókat az ilyen anyagoktól. Ugy tűnik, hogy a legfontosabb mechanizmus, amelyet be kellene vezetni, egy olyan jogi védelmi rendszer lenne, amely lehetővé tenné, hogy jól képzett és esetleg tanúsított személyek és szervezetek archiválhassák és kutathassák az interneten elérhető tiltott adatokat anélkül, hogy attól kellene tartaniuk, hogy veszélybe sodorják a szervezeteiket vagy a gyűjtéseket használó kutatókat.

A DIGITÁLIS LÁBNYOM

Kérdés: Hogyan lehet (és kell) archiválni egy személy digitális lábnyomát? Egy személy online tettei és tevékenységei potenciális érdeklődésre tarthatnak számot, különösen akkor, ha az illető közismert (vagy ismertté válik). Azzal érveltek (Garfinkel & Cox, 2009), hogy egy személy életművének katalogizálása az archivátorok feladata lesz. Egy személy webarchívuma tartalmazhatja weboldalait, közösségi halozati profiljait és bejegyzéseit, kommunikációit, publikációit és egyéb, a digitális életére vonatkozó anyagokat.

Kihívás: A központi kihívás annak kitalálása, hogyan lehet olyan eszközöket létrehozni, amelyek lehetővé teszik az egyének számára, hogy manuálisan megadják, hogyan gyűjtsék össze automatikusan a digitális lábnyomukat. Lehetséges-e, hogy az eszközök automatikusan összeállítsák a digitális lábnyomot, esetleg opt-in alapon? Továbbá, hogyan tudjuk elérni, hogy ezek a rendszerek ne csak az emlékezés, hanem a felejtés lehetőségét is biztosítsák, lehetővé téve az emberek számára, hogy később törölhessék (és elfelejthessék) a lábnyom egyes részeit vagy egészét, ahogyan azt egyes tudósok alapvető jogként állítják (Mayer-Schönberger, 2009)? Az ezen a területen előrelépések különösen bonyolultak, mivel számos adatvédelmi és jogi kérdéssel kell majd foglalkozni.

AZ ADATOK HÁLÓJA

Kérdés: Hogyan lehet az adatokat újra kinyerni a webes archívumokból? Az elmúlt években jelentősen nőtt az adatvezérelt web (szemben a dokumentumok webjével). Számos kérdés merül fel, például hogyan lehet az adatokat a dokumentumok mellett archiválni? Milyen adattisztító eszközökre lesz szükség az adatokkal való munkához?

Gondoljunk például a tudás létrehozására szolgáló, tartományok közötti tudományos vagy ipari folyamatokra, mint például a repülőgép-tervezés, a gyógyszerkutatás és így tovább. Hogyan őrizhetnénk meg a mérnöki tervezés ellátási láncából származó adatok megértésének képességét évekig, amikor egy repülőgép lezuhan, és a nyomozók újraértékelni akarják az eredeti mérnöki számításokat? A tervezés digitális volt, a tudást 100 partner generálta, akik közül néhány időközben megszűnt, és akik mindannyian személyzetet foglalkoztatnak.

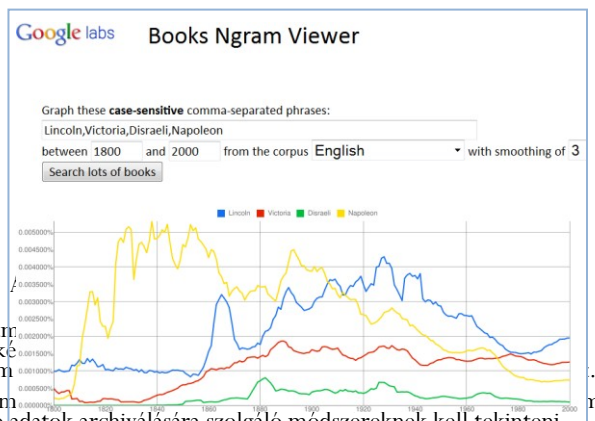
⁴³ <http://www.research.ibm.com/deepqa/deepqa.shtml>

egy teljesen más embercsoport által. Ebben az esetben a tudás életciklusa sokkal hosszabb, mint az üzleti életciklus, és az archívumok szerepet játszhatnak ezen információk megőrzésében.

Ehhez kapcsolódó kérdés, hogy hogyan archiválhatjuk és elemezhetjük a védett adatokat? Egyes adatok védettek, és a jogmegállapodások tiszteletben tartása érdekében meg kell őrizni őket. Sok elemzést azonban a nyers adatokhoz való hozzáférés nélkül is el lehet végezni, ha a megbízható szervek tartolják a nyers adatokat, és csak az elemző eszközök számára engedélyezik az adatokhoz való hozzáférést. Ezután az összesített és összevont eredmények a kutatók rendelkezésére bocsáthatók anélkül, hogy a védett adatokat felfednék. Léteznek olyan eszközök, amelyek modellként szolgálnak, és a nyers adatok darabjaihoz biztosítanak hozzáférést a nyers adatok letöltése helyett, mint például az Elixer⁴⁴, a csillagászati adatok elérésére szolgáló program. Egy másik példa a Google Books Ngram Viewer,⁴⁵ amely lehetővé teszi a felhasználók számára az adatok elemzését a nyers adatokhoz való hozzáférés nélkül.

Ha az adatok már elemezhető adattárakban vannak, ez további lehetőségeket teremt, például az adatok összekapcsolását eszközökkel, a olyan komponensek könyvtárai, amelyek lehetővé teszik a kutatók számára a módokon teremtik meg a keverés és illesztés lehetőségeit. Ha ezen adatkezelési föderáció sokkal könnyebbé válik, és növeli az adatok korábban fel nem

Kihívás: Az internet azon részei esetében, amelyek nem dokumentumok, a dokumentumok megőrzésére, hanem a dokumentumokban található adatok archiválására szolgáló módszereknek kell tekinteni. Ehhez az adatok tárolásának és kinyerésének új modelljeire lenne szükség, amelyek a strukturált adatok adatelemzésére alkalmas modelleket követik a strukturálatlan adatok dokumentumelemzése helyett.



A NEMZETI HÁLÓK

Kérdés: Mi értelme van nemzeti webarchívumok létrehozásának, amikor a web egy határokon átvéhető jelenség? A finanszírozás és a jogi korlátozások realitását tekintve az archívum-in-a-box erőszítések egy része biztosan nemzeti szinten fog megvalósulni. Az Egyesült Királyságban British Library jelenleg készül a várható szabályok hatálya lépésére, amelyek az Egyesült Királyság webes terének archiválására vonatkoznak, mint az Egyesült Királyságban megjelent összes publikáció letéti könyvtára. 2011 májusában a dán kutatási és innovációs minisztérium úgy döntött, hogy számos nemzeti kutatási infrastruktúrát hoz létre a különböző kutatási területeken (természettudományok, bölcsészettudományok stb.), és az egyik hangsúlyt az archivált webes anyagok elemző eszközeinek használatára helyezi.

Közvetlen kihívás: még mindig nem világos, hogy a kutatók hogyan fogják felhasználni ezeket a nemzeti archívumokat. Sokan még csak tervezési fázisban vannak. Azt állítjuk, hogy az egyik legfontosabb teendő az, hogy olyan kutatókat vonjunk be, akik nem csak az internetes kutatásban, hanem olyan területeken is jártasak, mint a szociológia, a politikatudomány, más társadalomtudományok, a fizika és más tudományok, a művészetek és a bölcsészettudományok, és más területek, mivel ezeket az infrastruktúrákat úgy tervezik meg, hogy a nemzeti kutatók igényei tükröződjön a létrehozott gyűjteményekben. Ez egy időigényes folyamat, és a szakterületi szakértők bevonása nehezségeket ütköztethet. Ennek elmulasztása azonban csökkenti annak valószínűségét, hogy az új infrastruktúrák széles körben elterjedjenek.

⁴⁴ <http://www.cfh.hawaii.edu/Instruments/Elixir/home.html>

⁴⁵ <http://ngrams.googlelabs.com/>

KÖVETKEZTETÉSEK: AZ ELŐRE VEZETŐ ÚT

Ez csak néhány azok közül a dolgok közül, amelyeket a mi kis csoportunk mások segítségével ki tudott találni - még sok más is létezik. Az általunk leírtak egy része általános, mivel egy webes gyűjtemény jellemzésére vagy a webes tartalom időbeli változásának tendenciáinak elemzésére szolgáló konkrét, lépésről lépésre kidolgozott technikákhoz egy erre a célra létrehozott kutatási projekt erőforrásaira, egy szakterület-szakértőkből álló csapatra és olyan releváns gyűjteményekre lenne szükség, amelyeken ezeket a módszereket tesztelni lehet. Nincs tehát csodafegyverünk. Ugyanakkor megmutattuk, hogy bár számos általános kihívás áll a webarchívumokkal dolgozni kívánó kutatók előtt, az egyik legfontosabb dolog, amely az interjúk és a megbeszélések során újra és újra felmerült, az a stabil, felhasználóbarát interfészek jelenlegi hiánya a webarchívumok létrehozásához, illetve a webarchívumok létrehozását követően a bennük található adatokhoz való hozzáféréshez és azok elemzéséhez. A tanulási görbe jelenleg túl meredek a nem műszaki felhasználók számára, és a legtöbb intézményben a rendelkezésre álló támogatás minimális, ha egyáltalán létezik. Ezen változtatni kell. Ha ez nem történik meg, akkor a webarchívumokat biztonságosan, pörös dobozokban fogják tarolni.

Hosszú távon reméljük, hogy ebből az erőfeszítésből más eredmények is szülehetnek. Elképzelhető például egy Hágát követő munkacsoport létrehozása, amely a műhelymunkák ötleteinek továbbfejlesztésére és a lehetséges jövőbeli felhasználási esetek megfogalmazására, valamint az eszközeik fejlesztésére összpontosít. Ennek a munkacsoportnak lennie egy webes komponense, amelyet nemcsak az IIPC tagjainak hirdetnénk meg, hanem azon kutatók kapcsolódó közösségeinek is, akik nem vesznek részt a webarchívumok közösségében, de nagy valószínűséggel használják a webarchívumokat. Ilyenek például az internetkutatók (például az AoIR tagjai), az informatikusok (például az IFIP és az ASIS&T tagjai), valamint a digitális humán tudományok iránt érdeklődő listák és egyesületek. Határozottan javasoljuk, hogy az IIPC küldjön képviselőket az ilyen szervezetek éves találkozóira, és szervezzen paneleket és workshopokat, hogy a kutatókat bevonja a webarchívumokban rejlő lehetőségekbe. Kiemeltünk néhány ötletet, de ezek a közösségek még sok más ötletet is generálhatnak. Ne várjuk meg, hogy eljőjenek az IIPC-hez. Az IIPC-nek kellene ellátni a feladatot.

Egy másik ötlet a jövőbeni tevékenységre egy hackathon, ahol a számítógépes programozók és hacktivisták 2-3 napra összejönnek a kutatókkal, és hozzáférést kapnak a webarchívum adataihoz. Csoportokba lehetne őket osztani, és azt a feladatot kaphatnák, hogy innovatív és kreatív megközelítéseket találjanak a meglévő adatokkal és eszközökkel való munkához, valamint gyorsan új eszközöket és interfészeket hozzanak létre. A kutatókat azért választanák ki, mert vannak olyan kérdéseik, amelyekre szeretnének választ kapni, a programozók pedig a saját képességeikkel segítenének nekik abban, hogy elérjék (vagy közelebb kerüljenek) kutatási céljaikat. Ismétlem, az élő webes adatokkal dolgozó számítógépes programozóknak megvannak a készségeik ahhoz, hogy sok kreatív dolgot csináljanak az eszközökkel; ha hozzájuk fordulunk, az nagyobb hasznot hoz, mintha arra várnánk, hogy webes archívumokat készítsenek.

Elérjük-e a Nirvánát, apokalipszist, leszünk-e kárhóztatva, a Szingularitás kiszorít-e minket, vagy felügyeljük a Poros Archívum létrehozását? Ezt nem tudhatjuk. Elerkeztünk azonban egy olyan ponthoz, ahol érdemes feltenni a kérdést: milyen lépéseket tehetünk ma annak érdekében, hogy az, ami a jövőben rendelkezésünkre áll, ne csak sok, alig megfontolt döntés felhalmozódása legyen, hanem része annak az erőfeszítésnek, hogy a webarchívumok robusztusak, fenntarthatóak, hozzáférhetőek, értékesek és - mindenekelőtt - a jövő kutatói által használhatóak legyenek?

- Conover, M. D., Ratkiewicz, J., Francisco, M., Gonçalves, B., Flammini, A., & Menczer, F. (2011). *Politikai polarizáció a Twitteren*. A barcelonai 2011, ICWSM: International Conference on Weblogs and Social Media konferencián bemutatott tanulmány.
- Conover, M. D., Ratkiewicz, J., Gonçalves, B., Flammini, A., & Menczer, F. (2011). *A visszhangkamra*. Előadás a Journal of Information Technology & Politics 2011 konferencián: The Future of Computational Social Science, Seattle.
- Dougherty, M., Meyer, E. T., Madsen, C., Van den Heuvel, C., Thomas, A., & Wyatt, S. (2010). *Researcher Engagement in Web 2.0*. *Archives: State of the Art*. Report. London: JISC. Letölthető a következő honlapokról: <http://ssrn.com/abstract=1714997> és <http://ie-repository.jisc.ac.uk/544/>.
- Garfinkel, S. & Cox, D. (2009. február 9-11.). *Az internetes lábnymegtalálása és archiválása*. Előadás a First Lives Research Conference konferencián: Személyes digitális archivumok a 21. században, London.
- Gazan, R. (2008). Társadalmi megjegyzések a digitális könyvtári gyűjteményekben. *D-Lib Magazine*, 14 (11/12).
- Hindman, M. (2007). A "nyílt forráskódú politika" újrarendelése: Emerging Patterns in Online Political Participation. In V. Mayer-Schönberger & D. Lazer (szerk.), *Kormányzás és információs technológia: From electronic government to information government* (pp. 183-207). Cambridge: The MIT Press.
- Hogan, B. (2010). A Facebook-hálózatok elemzése. In D. Hansen, M. Smith & B. Schneiderman (Eds.), *Analyzing Social Media with NodeXL*. New York, NY: Morgan Kaufman.
- Jasra, M. (2011. február 3.). Reddit Surpasses 1 Billion Monthly Page Views Retrieved 30 April, 2011, from <http://www.webanalyticsworld.net/2011/02/reddit-surpasses-1-billion-monthly-page.html>. (Archiválva <http://www.webcitation.org/5yKdMBKNC> címen)
- Kay, A. (1995). A jövő megjósolásának legjobb módja a jövő feltalálása. *Mathematical Social Sciences*, 30, 326-326.
- Klett, M., Manchester, E., & Verburg, J. (1984). *Second View: The Rephotographic Survey Project*. Albuquerque: University of New Mexico Press.
- Kling, R., McKim, G., & King, A. (2003). Egy kicsit több informatika: Tudományos kommunikációs fórumok mint társadalmi-technikai interakciós hálózatok. *Journal of the American Society for Information Science and Technology*, 54 (1), 46-67.
- Kurzweil, R. (2005). *A szingularitás közel van: Amikor az ember túllép a biológián*. New York: Viking.
- Mayer-Schönberger, V. (2009). *Törlés: a felejtés erenye a digitális korban*. Princeton, NJ: Princeton Univ Press.
- Meyer, E. T. (2006). Szociotechnikai interakciós hálózatok: Kling STIN-modelljének erősségeiről, gyengeségeiről és jövőjéről. In J. Berleur, M. I. Numinen & J. Impagliazzo (Eds.), *IFIP International Federation for Information Processing Volume 225, Social Informatics: Egy információs társadalom mindenkinek? Rob Kling emlékére* (pp. 37-48). Boston: Boston: Springer.
- Meyer, E. T. (2011). *Fröccsenések és hullámok: A digitális erőforrások hatására vonatkozó bizonyítékok összegzése*. Jelentés. JISC. Letölthető a <http://ssrn.com/abstract=1846535> honlapról.
- Meyer, E. T., Eccles, K., Thelwall, M., & Madsen, C. (2009). Zárójelentés a JISC számára a *Kultur* digitálizálási projektek 1. fázisának használati és hatásvizsgálatáról és a digitalizált tudományos források hatásvizsgálatának eszköztáráról (TIDSR). Letölthető a http://microsites.oii.ox.ac.uk/tidsr/system/files/TIDSR_FinalReport_20July2009.pdf honlapról.
- Moretti, F. (2005). *Gráfok, térképek, fák: Absztrakt modellek az irodalomtörténethez*. London: London: Books.
- Moretti, F. (2011). Hálózatelmélet, cselekményelemzés. *New Left Review*, 68., 80-102.
- Olston, C., Reed, B., Srivastava, U., Kumar, R., & Tomkins, A. (2008). *Pig Latin: Egy nem túl idegen nyelv az adatfeldolgozáshoz*. Előadás az ACM SIGMOD'08 konferencián, Vancouver, BC, Kanada.
- Schroeder, R. (2011). *Együtt ott lenni: Social Interaction in Shared Virtual Environments*. New York, NY: Oxford University Press USA.
- Schroeder, R., & Meyer, E. T. (2009). Egy fejlődő globális agy: Hogyan forradalmasítja az internet a tudományos kutatást. *Nagy-Britannia 2009 Economic & Social Research Council Annual Magazine*, 113.

- Tanner, S. (2010). *Inspiráló kutatás, inspiráló ösztöndíj*. Report. London: JISC. Letölthető a <http://www.jisc.ac.uk/media/documents/programmes/digitisation/12pagefinaldocumentbenefitssynthesis.pdf>.
- Tanner, S., & Deegan, M. (2011). *Inspiráló kutatás, inspiráló ösztöndíj: A digitalizált források értéke és előnyei tanulás, a tanítás, a kutatás és a szórakozás területén*. Jelentés. London: JISC. Letölthető: http://www.kdcs.kcl.ac.uk/fileadmin/user_upload/documents/Inspiring_Research_Inspiring_Scholarship_2011_SimonTanner.pdf.
- Thomas, A., Meyer, E. T., Dougherty, M., Van den Heuvel, C., Madsen, C., & Wyatt, S. (2010). *Researcher Engagement in Web Archives: Challenges and Opportunities for Investment*. Jelentés. London: JISC. Letölthető: <http://ssrn.com/abstract=1715000> és <http://ie-repository.jisc.ac.uk/543/>.
- van den Heuvel, C. (2009). MAPS: Manuscript Map Annotation and Presentation System: Formális ontológiák összekapcsolása társadalmi címkézésrel a kéziratlektépek és a kontextuális dokumentumok közötti kapcsolatok (újra)konstruálásához. *Digital Humanities (2009)*. Maryland, Maryland Institute for Technology in the Humanities (MITH) Abstracts, 138-141.
- Von Ahn, L., Maurer, B., McMillen, C., Abraham, D., & Blum, M. (2008). reCAPTCHA: Human-Based Character Recognition via Web Security Measures. *Science*, 321(5895), 1465-1468.
- Williams, D., Yee, N., & Caplan, S. E. (2008). Ki játszik, mennyit és miért? A sztereotipikus játékosprofil megcáfolása. *Computer-Mediated Communication*, 13 (4), 993-1018. doi: 10.1111/j.1083-6101.2008.00428.x

KÖSZÖNETNYILVÁNÍTÁS

A szerzők köszönetet mondanak az IIPC-nek a munka támogatásáért.

Ezenkívül szeretnénk köszönetet mondani az alábbi személyeknek, akik észrevételekkel, javaslatokkal és véleményekkel járultak hozzá a jelentés tartalmának alakításához:

A május 10-ban Hágában tartott IIPC workshop résztvevői 2011

Robert Ackland, Ausztrál Demográfiai és Társadalomkutató Intézet, Ausztrál Nemzeti Egyetem Michael

Boniface, IT Innovation, Southamptoni Egyetem

Niels Brügger, Informatikus és Média tudományi Tanszék, Aarhusi Egyetem, Dánia Cristobal

Cobo, Oxfordi Internet Intézet, Oxfordi Egyetem

Lewis Crawford, The British Library

Meghan Dougherty, Loyola Egyetem, Chicago

Alex Halavais, Quinpiac Egyetem

Helen Hockx-Yu, British Library

Gildas Illien, Département du Dépôt légal, Bibliothèque nationale de

France Jeffrey Keefer, University of Lancaster

Sean Martin, The British Library

John Postill, IN3, Katalán Nyílt Egyetem; Sheffield Hallam Egyetem Burkhard

Stiller, Informatikai Tanszék, Zürichi Egyetem

Charles M. J. M. van den Heuvel, Holland Királyi Művészeti és Tudományos Akadémia, Huygens ING

Intézet Sally Wyatt, e-Humanities Group, Holland Királyi Művészeti és Tudományos Akadémia (KNAW)