Az első nyilvános webarchívum az Egyesült Királyságban

Sokak számára a web az elsődleges információforrás, eddig mégis kevés figyelmet fordítottak a weboldalak hosszú távú megőrzésére, ami azzal a veszéllyel jár, hogy felbecsülhetetlen tudományos és kulturális értékek vesznek el a jövő generációi számára.

A probléma megoldására hat vezető brit intézmény dolgozik közösen egy tesztelési környezet kidolgozásán, amely alapján kiválaszthatók az archiválni kívánt weboldalak. A hat intézmény: Brit Nemzeti Levéltár, Brit Nemzeti Könyvtár, Közös Információs Rendszerek Bizottsága (JISC), a skót és a walesi nemzeti könyvtárak és a Wellcome Könyvtár, megalakította az *Egyesült Királyság Webarchiválási Konzorciumát (UK Web Archiving Consortium = UKWAC).* Az archiválásra az Ausztrál Nemzeti Könyvtár által kifejlesztett PANDAS (PANDORA Digital Archival System = Pandora Digitális Archiváló Rendszer) szoftvert használják. A partnerek az adott intézmény szakterületéhez kapcsolódó oldalakat mentik el.

A konzorciumi partnerek évente négyszer találkoznak, hogy nyomon kövessék a fejlődést, megvitassák a problémákat, és tervet készítsenek a jövőre vonatkozóan. A projekt kezdetén a következő célokat tűzték ki a résztvevők:

- a PANDAS szoftver használati jogának megszerzése;
- külső fél szerződtetése az infrastruktúra létrehozására;
- összefogás egy közös, kereshető honlapadatbázis kidolgozásáért, a következő problémák megoldásával: kiválasztás, jogkezelés és digitális megőrzés;
- a webarchívum infrastrukturális fejlődésének értékelése, illetve egy ilyen közös vállalkozás hosszú távú megvalósíthatóságának és fenntarthatóságának megállapítása.

A projekt *Célok és irányelvek* c. dokumentumában a következő feladatok szerepeltek:

- közös engedélykérő űrlap kidolgozása a honlapok archiválásához,
- a honlapok kiválasztási elveinek meghatározása,
- teljes mértékben kereshető és böngészhető online webarchívum elkészítése és katalogizálása,
- konzorciumi honlap és levelezőlista készítése a partnereknek,
- értékelő jelentés összeállítása a projekt folytatására vonatkozó ajánlásokkal.

Mindezeket 2005 májusára sikerült megvalósítani.

Módszertan

A konzorciumi partnerek a kiválasztást, gyűjtést és archiválást azonos módon végzik a PANDAS szoftverrel, betartva a közös szabványokat és irányelveket.

Az archiválásra szánt oldalakat ugyan önállóan választják ki a partnerek, a közös adatbázis építéséhez ellenőrizni kell, hogy a kiválasztott oldal nincs-e még elmentve. Ha a honlap még nem szerepel az adatbázisban, akkor az archiválást végző személy beviszi az alapvető metaadatokat, és a továbbiakban ő lesz felelős az oldal kezeléséért. A hatékonyság érdekében a honlapért felelős személy lép kapcsolatba a fenntartóval a konzorcium nevében.

A partnerek az érdeklődési köröknek megfelelően honlapokat cserélnek egymás között, hogy az oldalakat a legmegfelelőbb résztvevők archiválják. Így minden partner hozzájárul az archívum épüléséhez, és annak maguk is részesei lesznek.

Archiválás előtt a partnerek írásos engedélyt kérnek a honlapok tulajdonosaitól. Az engedélykéréshez azonos űrlapot használnak, amelyet levéllel és a „Gyakori Kérdések Fájljával" látnak el. Mindez arra szolgál, hogy valamennyi honlaptulajdonos ugyanazokat az információkat kapja meg.

A PANDAS-nak van ugyan központi metaadattára, a partnerek a katalogizálást saját katalógusukban kezdik, hogy használóik keresni tudjanak a helyi gyűjteményben. Így az archívum szélesebb közönségre talál, az archivált honlapok pedig hasonló tartalmú hagyományos dokumentumokkal együtt lesznek kereshetők. A partnerek nemcsak a felelősségen, hanem a költségeken, illetve a dokumentumok módosulásának kockázatán is osztoznak.

A digitális megőrzés példaértékű és újszerű fejlesztése

A projekt fő célja a digitális megőrzés. A PANDAS rendszer igazoltan hatékony a honlapok „begyűjtésében". A projekt résztvevői kihasználták ezt a funkciót, és tökéletesítették a kiválasztott weboldalak sikeres feldolgozása érdekében. A PANDAS átvételével a brit honlapokat a fejlesztésre szánt minimális idővel, erőfeszítéssel és költséggel lehetett archiválni. A PANDAS-ban módosították a begyűjtő funkciót, hogy csökkentsék a tárolószerverek alapját képező web crawl motor HTTrack kellemetlen hatását. Csökkentették az alkalmazás által létrehozott egyidejű nyílt kapcsolatokat és a maximális letöltési sebességet. Amint a partnerek a jogtulajdonosok engedélyével archiválnak, a mentési tevékenységet azonnal egy tulajdonlást igazoló bizonylattal látják el, kifejezve, hogy a projekt résztvevői együtt kívánnak működni a webszolgáltatókkal.

Az UKWAC rendszer felvállalt néhány innovatív fejlesztést. A PANDAS túlterhelés esetén hibákra hajlamos; osztott környezetben nehéz azonosítani, hogy a rendszer mikor mit tölt, van-e éppen kapacitás a begyűjtött honlap feldolgozására. A projekt számára „jelzőlámparendszert" készítettek, hogy jelezze, van-e a rendszernek szabad kapacitása. Ezzel az egyszerű alkalmazással csökkent a rendszer túlterheltsége. Az UKWAC a *robots.txt* segítségével kizárja a keresőmotorok használatát, amire azért volt szükség, hogy a honlaptulajdonosok el tudják különíteni az archivált és a működő honlapokat. A használókban tudatosítják, hogy archívumban vannak, ahol régi anyagokat találnak, miközben a működő oldalakat a keresőmotorok segítségével lehet megjeleníteni.

A webarchiválás nehézségei

A projekt kivitelezése mindig rejt magában nehézségeket. A PANDAS szoftveren kívül volt még néhány archiválásra használható alkalmazás, ez volt viszont az egyetlen, amely irányított környezetben a teljes munkafolyamatot átölelő szolgáltatást nyújtott. A PANDAS mégsem ideális rendszer, mivel a jelenlegi verzió nem használ katalogizálási szabványokat, illetve megbízható tárgyi ellenőrzést. Az elosztott architektúra megtagadja a partnerek kapcsolódását a rendszer logfájljaihoz, és a kódokhoz, amelyek a problémák és más alkalmazási kérdések megállapítását végezné. Az eredmény az, hogy a rendszer a PANDAS külső tárolószerverétől függ, ami szokványos rendszerirányításnak tekinthető.

Az internet gyors fejlődést és változást mutató médiummá vált. A statikus HTML oldalakat egy évtizede leváltották a nagyon dinamikus, adatbázis-vezérelt környezetekkel. A PANDAS képzett szakembert igényel az ilyen oldalak archiválására.

Az UKWAC a nemzetközi tapasztalatokra épít

A projekt eredményei a nemzetközileg irányított jelenlegi és korábbi kezdeményezésekre épülnek. Az infrastruktúrát és a szoftvert az Ausztrál Nemzeti Könyvtártól vették át. A PANDAS-ban rejlő szelektív és minőségi megközelítés beleillett a projekt ideológiájába, amely két projektpartner megbízásából készült tanulmányokon alapul.

Néhány munkatárs szerzett már korábban tapasztalatokat a webarchiválás területén az Egyesült Királyságban, Ausztráliában, illetve Új-Zélandon, ami meggyorsította a projekt elindítását. A PANDAS eredeti fejlesztői szembesültek néhány akadállyal, amelyeket sikeresen leküzdöttek folyamatos ismétlés segítségével. Az UKWAC ezért tudott az archiválásra koncentrálni, a fejlesztés és felhasználási előírások helyett. Mind a konzorcium, mind a partnerek egyenként igyekeznek szoros kapcsolatokat kiépíteni más webarchiváló kezdeményezésekkel a kölcsönös tapasztalatcsere érdekében.

Az UKWAC archívum gyakorlati haszna

Az archívum interfésze könnyen kezelhető, a használók információt kapnak mind a projektről, mind magáról az archivált tartalomról. Az elmentett oldalakat kereső és böngésző funkciókkal lehet megtalálni, az utóbbi hierarchikus tárgyszórendszer segítségével működik. Mindkét módszer a Google-hoz és a Yahoo-hoz hasonló elveken működik. Az archívum a *Lucene* nevű keresőmotort használja, amely az egyes oldalak tartalmára tud keresni. Jelenleg még csak az egyszerű kere-

sés működik, de tervezik, hogy menet közben kifejlesztik az összetettebb keresés lehetőségét.

A honlap és archívum bárki számára ingyenesen hozzáférhető, továbbá bármilyen szervezet saját anyagait is díjmentesen felveszik. A konzorciumi tagok remélik, hogy az archívum széles közönség számára kínál értékes információkat, az akadémikusoktól az egyszerű érdeklődőkig. Ezzel nemcsak a használók széles rétegét elégítik ki, hanem elősegítik a digitális megőrzés fontosságának tudatosítását a társadalomban.

Az együttműködés máris várakozáson felüli haszonnal járt. A projekt résztvevői összehívták az idősebb szakképzett munkatársakat, és bemutatták a webarchiválással kapcsolatos bonyolult problémákat. Informatikai szakemberek is csatlakoztak a kuratóriumi és archiváló csoporthoz, hogy mindenki a saját képességeit kamatoztassa, illetve bővítse.

A legfőbb érdem, ami megkülönbözteti ezt a projektet a többitől, hogy a brit intézmények szakemberek segítségével végeznek szelektív webarchiválást. Az eredmény magas színvonalú archívum, amelynek a tartalma világos és megállapodáson alapuló elvek szerint válogatott, szilárd és megbízható.

Az UKWAC hosszú távú előnyökkel szolgál a digitális megőrzés terén

Nyilvánvaló, hogy minden digitális megőrző kezdeményezés korai fázisban van, a siker igazi tesztje évtizedek múlva fog bekövetkezni.
Az UKWAC biztonsággal kijelentheti, hogy a projekt kitűnő alapokat fektetett le az archívum tartalmának hosszú távú megőrzéséhez. A projekt nem függ a PANDAS szoftvertől, hanem technológiától független hosszú távú digitális megőrzést kínál. Ha a jövőben más webarchiváló megoldás mellett döntenek, a váltás könnyen megoldható, és minimálisan veszélyeztetné az archivált oldalakat.

Következtetések

A webarchiválás nem egzakt tudomány. A nehézségek ellenére az UKWAC jelentős projekt a digitális megőrzés terén. Bizonyította, hogy a szelektív webarchiválás kivitelezhető az Egyesült Királyságban konzorciumi keretek között. Rávilágított a webalapú anyagok sérülékenységére, miközben használható megoldást kínált a megőrzésükre.

/BAILEY, Steve–THOMPSON, Dave: UKWAC: building the UK's first public web archive. = D-Lib Magazine, 12. köt. 1. sz. 2006.
http://www.webarchive.org.uk/

(Szalóki Gabriella)