A portál próbája a webaratás!

Magyar nyelvű hírportálok archívumainak vizsgálata a digitális örökség szemszögéből

Indig Balázs

Digitális Örökség Nemzeti Laboratórium
ELTE BTK TI Digitális Bölcsészet Tanszék

2021. november 24.

National Laboratory for Digital Heritage

Bevezetés

- A *Digitális Örökség Nemzeti laboratórium* (DH-Lab) feladata:
  - Az eleve digitális (born digital) kulturális örökségünk archiválása
- *N*agy tömegű, *magyar szöveget* is tartalmazó anyag
  - sajtóanyagok
  - médiatermékek
  - web 2.0-es források (blog, fórum, chat, stb.)
  - határon innen és túl
- bármilyen jellegű kutatás, illetve oktatás számára
  - bölcsészeti, társadalomtudományi
  - piaci
- elérhető, értelmezhető legyen
  - széles körben
  - szemantikus mélységben

Bevezetés (folyt.)

- Gépi feldolgozásra
- Jó minőségű be- és kimenetek előállításával
  - szabványos, nemzetközi projektekben is használható
  - a teljesség igényével, filológiai minőségben
- A legnagyobb volumenű filológiai projekt
  - Kezdetnek 6 millió cikk 25 hírportálról
  - Megfelel 20 évig 20 nyomtatott napilap minden számának (40 cikk per lapszám)
- A papír alapú forrásokhoz képest nagyságrendekkel bonyolultabb
  - A hitelesség kérdése
  - A módosulás/sérülés/eltűnés kérdése
  - A teljesség kérdése

A két évvel ezelőtti előadásom: `http://videotorium.hu/hu/recordings/35075`

Probléma

- A magyar hírportálok teljesen átlagosak: „Wordpress az egész világ”
  - A portálok 60%-a és az összes weboldal 40%-a ilyen (forrás)
  - „Széthekkelt Wordpress”
- A Wordpress mindent tud, de…
  - „Nyílt forrásom hatalom, nyílt forrásom eladom”
  - Ez nem egy kialakult világ, fejlődik, frissül
  - Hibák jönnek és hibák mennek
  - Nincs felelőse, nincs értéke
- Stratégiai kérdések
  - Titkolózni kell, mert a XXXXXX ellopja
  - Csak az új cikkek számítanak, mert azokból van reklámbevétel
  - Monetizálni kell a vagyont! (A régi/népszerű cikkek paywall mögé kerülnek.)
- Mi lesz a digitális örökséggel?

Élen járó példák

- Wikipedia!
  - Mindenkinek ugyanazt mutatja, nincs személyre szabott tartalom
  - Emberek által olvasható permalinkek
  - Verziókezelt, metaadatolt (GIT, mint minta)
  - Minden módosításnak van dátuma és szerkesztője (a visszavonásnak is)
  - Minden módosítás elérhető és hivatkozható
- Ezt mind tudja a Wordpress is! Csak be kell kapcsolni...
  - A crawlerek a Wordpress dolgaira tanulnak rá
- Vannak újdonságok, amik még a küszöbön állnak: *Blockchain-alapú* hitelesítés
  - A tartalom hitelességét egy olyan elosztott hálózat garantálná, mint ami a Bitcoin-ét
  - Ha a szereplők is akarják...

Portált crawljáról, archivistát kitartásáról!

Lehet akármilyen a portál kinézete, amíg

- Van napi, havi, éves cikkarchívum, szükség esetén ezeken belül lapozás
- Minden cikk egyszer szerepel az archívumban, de egyszer legalább szerepel
- Minden cikk egyedi azonosítóval rendelkezik vagy teljesen rendezhető
- Emberek által olvasható permalink minden oldalra
- Ha változott a link, átirányítás a régiről az új linkre
- Az új linken jelzés: „csak a link új, a tartalom megegyezik” (pl. egyedi cikkazonosító)
- Szabványos, géppel olvasható metaadatok
- A formázás és a tartalom szétválasztása: a formai változtatás nem ront el tartalmat
- Nincsenek végtelen közvetítések

Néhány érdekes hiba, amivel eddig találkoztunk

- A rovatok eltűnnek, átneveződnek, elérhetetlenné válnak
- A cikkek URL-jébe valami hiba kerül (pl. URL-kódolás és az Unicode karakterek)
- Rovatonként van csak archívum, az alrovatok duplikálják az archívumot
- Angol- és magyarnyelvű cikkek vegyesen
- Dátum scripttel van csak generálva (pl. tegnap, múlthéten)
- Ismétlődő reklám „cikkek” az archívumban
- Nem működő formázási elemek (pl. lapozások)
- Latin-1 vagy Latin-2 kódolás 2021-ben (!)
- Import hibák (pl. másodpercre azonos időben megjelent cikkek, kódolási hibák)
- Különféle (teljesen másként működő) portálok összedrótozása
- Hiányos, rossz, ellentmondásos metaadatok (meta tag vs. ember számára látható)

Az említett hibák kezelése

- A saját crawlerünk a fenti hibákat kezelni tudja (Indig és tsai. 2020)
- A formázást a saját megoldásunkkal hibátlanul egységesítjük (HTML2TEI)
- Automatikusan összevetjük, kiegészítjük az archívumainkat
  - Archive.org
  - CommonCrawl
- Vizsgáljuk a konkurens megoldásokat, amit lehet, automatikusan összehasonlítunk

- Felajánljuk a segítségünket a portál üzemeltetőknek!
  - A legrosszabb, ami történhet egy oldallal, ha az emberek nem látogatják
  - A hibák eltántorítják a felhasználókat, a crawlereket viszont nem
  - Nem elég csak jó tartalmat szolgáltatni, a hosszútávú megbízhatóság is fontos
- Haladni kell a web fejlődésével, de nem a minőség rovására

A hitelesség kérdése

- Miért bízna bárki bennünk?
- Mi történik, ha feltörik a mi archívumainkat?
- Ugyanez igaz a portálüzemeltetőkre vagy az Archive.org-ra!

- A hibajelző- és hibajavító-kódoknak külön szakterülete van
- Blockchain: teljesen digitális pénzeket alapoznak rá
- Hibajavító-kódok + Blockchain = WarChain (Lendák, Indig és Palkó 2021)
- Jelenleg kísérleti stádiumban van, a crawler(ek) kimenetének validálására tervezve
- Ha be tudnánk építeni a Wordpressbe...

Összefoglalás

- A DH-LAB feladata a kutatás, fejlesztés a digitális örökség területén
- A memóriaintézmények feladata az anyagok megőrzése
- Szabványos, filológiai minőségű kimenet, összevethető más gyűjteményekkel
- Hatalmas adatmennyiség, hibák és hibalehetőségek tárháza
- A portálok nagy része Wordpress (és a hibák ebből fakadnak)
- A saját megoldásaink kezelik az eddig szóbajött hibákat
- A hitelesség kérdése felmerül a portáloknál és a webarchívumoknál is
- Blockchain-re alapuló hitelességellenőrzés, kísérleti stádiumban
- Közeljövő: Eszközök és archívumok automatikus összevetése, napi webaratás

Köszönöm a figyelmet!

https://dh-lab.hu/
https://elte-dh.hu/
https://github.com/elte-dh
https://zenodo.org/communities/elte-dh

Hivatkozások I

- Indig Balázs és tsai., „The ELTE.DH Pilot Corpus – Creating a Handcrafted Gigaword Web Corpus with Metadata”, English, *Proceedings of the 12th Web as Corpus Workshop*, Marseille, France: European Language Resources Association, 2020. máj., 33–41. old., isbn: 979-10-95546-68-9, url: https://aclanthology.org/2020.wac-1.5.
- Lendák Imre, Balázs Indig és Gábor Palkó, „WARChain: Blockchain-Based Validation of Web Archives”, *Socio-Technical Aspects in Security and Trust*, szerk. Thomas Groß és Luca Viganò, Cham: Springer International Publishing, 2021, 121–134. old., isbn: 978-3-030-79318-0, doi: https://doi.org/10.1007/978-3-030-79318-0_7.