Harvestiranje hrvatskoga weba: arhitektura programskoga sustava za harvestiranje i iskustva stečena njegovom upotrebom

Celjak, Draženko; Milinović, Miroslav

Source / Izvornik: **15. seminar Arhivi, knjižnice, muzeji : Mogućnosti suradnje u okruženju globalne informacijske infrastrukture, 2012, 144 - 160**

Conference paper / Rad u zborniku

Publication status / Verzija rada: **Submitted version / Rukopis poslan na recenzijski postupak (preprint)**

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:102:141793

Rights / Prava: In copyright / Zaštićeno autorskim pravom.

Download date / Datum preuzimanja: **2024-09-09**

srce
Sveučilište u Zagrebu
Sveučilišni računski centar

Repository / Repozitorij:

Digital repository of the University Computing Centre (SRCE)

dabar
DIGITALNI AKADEMSKI ARHIVI I REPOZITORIJI

HARVESTIRANJE HRVATSKOGA WEBA: ARHITEKTURA PROGRAMSKOGA SUSTAVA ZA HARVESTIRANJE I ISKUSTVA STEČENA NJEGOVOM UPOTREBOM

CROATIAN WEB HARVESTING: ARCHITECTURE OF THE HARVESTING SYSTEM AND EXPERIENCES GAINED DURING HARVESTING PROCESS

Draženko Celjak
Sveučilišni računski centar Sveučilišta u Zagrebu
University Computing Centre, University of Zagreb

Miroslav Milinović
Sveučilišni računski centar Sveučilišta u Zagrebu
University Computing Centre, University of Zagreb
damp@srce.hr

UDK / UDC
Stručni rad / Professional paper
Primljeno / Received on: 27. 3. 2012.
Prihvaćeno / Accepted on: 10. 4. 2012.

Sažetak

U okviru suradnje u izgradnji i održavanju Hrvatskoga arhiva weba (HAW; http://haw.nsk.hr) tijekom srpnja i kolovoza 2011. godine Nacionalna i sveučilišna knjižnica (NSK) i Sveučilišni računski centar Sveučilišta u Zagrebu (Srce) proveli su prvo harvestiranje vršne .hr internetske domene. U ovom radu prezentiramo arhitekturu sustava koji je korišten za harvestiranje kao i iskustva stečena tijekom planiranja, provođenja i analize rezultata harvestiranja. Poseban naglasak stavljen je na tehničke izazove s kojima je zajednički tim Srca i NSK bio suočen.

Ključne riječi: harvestiranje hrvatskoga weba, Hrvatski arhiv weba, Heritrix, DAMP.

Summary

As a part of the joined project on development and maintenance of the Croatian Web Archive (HAW; http://haw.nsk.hr) National and university library (NUL) and University computing centre, University of Zagreb (SRCE) have perfomed first harvesting of the Croatian top level internet domain .hr during July and August 2011. In this paper we present the architecture of the harvesting system which has been used and experiences gained through the process of planning, execution and analisys of the results of the harvesting. We put an emphasis on the technical challengnes our team was confronted with.

Keywords: Croatian Web harvesting, Croatian Web archive, Heritrix, DAMP

Uvod

Harvestiranje je proces rekurzivnoga pobiranja i obrade elektroničkih resursa dostupnih internetom. Harvestiranje se najčešće provodi u svrhu indeksiranja sadržaja resursa za potrebe tražilica, radi arhiviranja resursa, odnosno čuvanja njihova izvornoga oblika, analize resursa u istraživačke svrhe (razna mjerenja ili provjere usklađenosti s normama) ili radi ciljanoga skupljanja informacija s weba (npr. skupljanje elektroničkih adresa radi slanja neželjene elektroničke pošte).

Harvester, odnosno robot računalni je program koji automatizirano pregledava web-stranice tako da slijedi poveznice (engl. *hyperlinks*) s jedne stranice na drugu. Često se za web-robote rabe i engleski nazivi kao što su *spider, crawler* ili *bot*. Način na koji funkcionira većina robota može se pojednostavljeno opisati u nekoliko koraka:

1. za početak rada potreban je inicijalni popis URL-adresa resursa (web-stranica) koje se želi obraditi (engl. *seed*),
2. s popisa URL-adresa odabire se, prema nekom kriteriju, jedan URL te se dohvaća resurs s tom adresom. Dohvat se obavlja HTTP ili HTTPS protokolom.
3. ovisno o svrsi harvestiranja dohvaćeni resurs na prikladan se način obrađuje. Obrada uključuje i pronalaženje novih poveznica, tj. URL-ova u resursu[1].

[1] Osim u HTML-u URL-ovi se mogu naći u raznim drugim formatima datoteka: CSS, JavaScript, Flash, PDF, MP3, tekstualne datoteke, datoteke uredskih alata kao što su MS Word ili Excel…

4. ako novootkriveni URL-ovi iz prethodnoga koraka zadovoljavaju predefinirane kriterije, dodaju se na popis URL-ova koji čekaju dohvat,
5. obrađeni URL uklanja se s popisa URL-ova koji čekaju dohvat i, ako taj popis nije prazan, robot se vraća na korak 2.

U nastavku ovoga rada donosimo iskustva stečena tijekom planiranja, provođenja i analize rezultata prvoga harvestiranja vršne .hr internetske domene te opisujemo arhitekturu sustava koji je pritom korišten. Harvestiranje je, u okviru višegodišnje suradnje u izgradnji i održavanju Hrvatskoga arhiva weba (HAW; http://haw.nsk.hr)[2], tijekom srpnja i kolovoza 2011. godine proveo zajednički tim Nacionalne i sveučilišne knjižnice (NSK) i Sveučilišnoga računskog centra Sveučilišta u Zagrebu (Srce).

U svim fazama tog projekta oslanjali smo se na iskustva Srca stečena razvojem alata DAMP[3] koji se rabi za selektivno pobiranje webova u okviru HAW-a.

U fazi planiranja harvestiranja najvažnije je bilo definirati opseg harvestiranja, odabrati alat za harvestiranje te procijeniti računalnu opremu potrebnu za uspješnu provedbu harvestiranja u predviđenom roku. Prilikom definiranja opsega, odnosno granica pobiranja razmatrana su pitanja kao što su dubina harvestiranja, tretiranje *robots.txt* datoteka, preusmjeravanje s početnih stranica u .hr domeni na .com ili .net domene, pobiranje oglasa koji se prenose s vanjskih poslužitelja (npr. Google Ads) te tretiranje Facebook dodataka (npr. „sviđa mi se" ili „preporuči"). U sklopu odabira alata za harvestiranje uspoređeni su alati DAMP koje je razvilo Srce i kojima se NSK koristi za selektivno pobiranje webova te *open source* alati Heritrix razvijeni u okviru inicijative *Internet Archive*. Da bi se prije punoga produkcijskog harvestiranja provjerili teorijski postavljeni parametri harvestiranja te utvrdila njihova konačna vrijednost, provedeno je pilot-harvestiranje na uzorku od 100 web-sjedišta.

Tijekom harvestiranja pojavljivali su se problemi u vezi s restrikcijama postavljenim s pomoću *robots.txt* datoteka, zatim sa skupljanjem „beskonačnih sjedišta" i detektiranjem veza (*links*) u sadržajima koji nisu zapisani u formatu HTML. Također smo bili suočeni s problemom s kojim se standardno susreću roboti – kako se nositi s web-dućanima, forumima i

[2] Informacije o HAW-u dostupne su na http://haw.nsk.hr

[3] Sustav DAMP opisan je u Milinović, Miroslav; Topolščak, Nebojša. The architecture of DAMP. // Widwisawn, vol. 3, no. 3, 2005. < http://widwisawn.cdlr.strath.ac.uk/Issues/Vol3/issue3_3_1.html> i Milinović, Miroslav; Topolšćak, Nebojša. DAMP II: Digitalni arhiv mrežnih publikacija: nova funkcionalnost – novi planovi. // 9. seminar Arhivi, knjižnice, muzeji: mogućnosti suradnje u okruženju globalne informacijske infrastrukture: zbornik radova / uredile Mirna Willer i Ivana Zenić. Zagreb: Hrvatsko knjižničarsko društvo, 2006. Str. 40-55.

galerijama. Ta faza zahtijevala je i određenu interakciju s vlasnicima i održavateljima web-sjedišta.

Na kraju rada donosimo analizu rezultata harvestiranja koja, uz ostalo, obuhvaća distribuciju prikupljenih tipova podataka te broja i veličine datoteka po poslužiteljima. Te je podatke moguće usporediti s rezultatima mjerenja web-prostora (MWP; http://www.srce.hr/mwp/)[4] koje je prijašnjih godina provodilo Srce.

Harvestiranje hrvatskoga weba

Aktivnosti vezane uz harvestiranje možemo podijeliti u tri cjeline: planiranje, provođenje i analiza rezultata harvestiranja.

Planiranje harvestiranja

Osnovna pitanja na koja je bilo potrebno odgovoriti u fazi planiranja bila su u svezi s opsegom i parametrima harvestiranja, izborom alata za harvestiranje te procjenom računalnih resursa potrebnih za uspješno provođenje harvestiranja u prihvatljivom roku. Odgovori na ta pitanja povezani su i do njih se dolazilo u više iteracija, a cijeli postupak uključivao je i provođenje pilot-harvestiranja na uzorku od 100 web-sjedišta.

Opseg harvestiranja

Budući da je riječ o harvestiranju weba, podrazumijeva se da će se pobirati elektronički resursi dostupni HTTP i HTTPS protokolom. Pobiranje je ograničeno na javno dostupne resurse. Naime, zbog velikoga opsega harvestiranja nije bilo izvedivo baviti se autentikacijom i autorizacijom za pristup zaštićenim resursima. Mnogo je složenije bilo definirati što je to „hrvatski web" i koji je njegov dio realno obuhvatiti harvestiranjem. Za potrebe harvestiranja razmatrana su dva kriterija: pripadnost vršnoj domeni (engl. *top-level domain*) i jezik sadržaja web-stranica.

Vršna domena neke zemlje specifični je nastavak koji nazivu (adresi) računala daje nacionalni prizvuk te asocira na geografsku pripadnost.[5] Pripadnost vršnoj domeni jedan je od

[4] Informacije o mjerenjima web-prostora koje je od 2002. do 2008. provodilo Srce dostupne su na: http://www.srce.hr/mwp/

[5] Wikipedia – Vršna domena [citirano: 2012-01-23]. Dostupno na http://hr.wikipedia.org/wiki/Vršna_domena

kriterija koji za definiranje opsega harvestiranja rabe razne nacionalne knjižnice (npr. Bibliotheque nationale de France, National Library of New Zealand). Iako se hrvatska web-sjedišta nalaze na raznim vršnim domenama (hr, com, net, org…), odlučili smo pobiranje ograničiti na hrvatsku nacionalnu vršnu domenu, jer bi identifikacija hrvatskih web-sjedišta na drugim domenama značajno povećala kompleksnost harvestiranja te potrebne ljudske i računalne resurse.

Od jezika sadržaja kao kriterija za harvestiranje odustali smo iz dvaju razloga: prvi je taj što postoje alati koji mogu automatski prevoditi sadržaje na razne jezike, a drugi je što razna programska rješenja kao što su CMS-ovi, forumi i galerije imaju lokalizacije, tj. prijevode na razne jezike. Iz spomenutoga proizlazi da sadržaj na hrvatskom jeziku ne mora značiti da je riječ o „hrvatskom webu". Pritom ne treba zanemariti ni činjenicu da postoje web-stranice napisane na stranim jezicima iako im je sadržaj vezan uz Hrvatsku, pa bi se mogao uvrstiti pod „hrvatski web".

Hrvatska akademska i istraživačka mreža CARNet od 1993. godine upravlja nacionalnom domenom Republike Hrvatske (.hr)[6], stoga je od CARNetove DNS službe zatražen i dobiven (21. 4. 2011.) popis aktivnih domena koje pripadaju vršnoj nacionalnoj domeni. Popis je obuhvaćao i .from.hr (.name.hr, .iz.hr) domene namijenjene fizičkim osobama te komercijalne .com.hr domene. Na popisu su bile ukupno 85 764 aktivne domene, a njihova distribucija prikazana je u Tablici 1.

Tablica 1. Nacionalna hr domena na 21. 4. 2011.

.hr	70 672
.com.hr	10 595
.name.hr,.from.hr, .iz.hr	3 × 1 499
Ukupno .hr domena	**85 764**

Za pokretanje harvestiranja potrebni su ishodišni URL-ovi (engl. *seed*) od kojih pobiranje počinje. Da bismo iz popisa od 85 764 domene dobili ishodišne URL-ove, načinjen jer robot koji je provjeravao nalazi li se iza naziva domene doista neko web-sjedište. Pritom se polazilo od pretpostavke da su početne stranice web-sjedišta obično oblika http://www.domena.hr/ ili http://domena.hr/ pa je robot prvo pokušao HTTP upitom dohvatiti resurs dodavanjem www ispred naziva domene. Ako se kao ishod nije pojavio ispravan HTTP

[6] DNS – naslovnica [citirano: 2012-01-23]. Dostupno na: http://www.dns.hr/

odgovor, robot je pokušao dohvatiti URL bez www dijela. Rezultati tog postupka prikazani su u Tablici 2.

Tablica 2. Broj aktivnih URL-ova

Ukupan broj domena	85 764
URL oblika http://www.domena.hr/	62 459
URL oblika http://domena.hr/	115
Nije dobiven HTTP odgovor	23 190
Ukupno URL-ova	**62 574**

Za 23 190 domena nije dobiven ispravan HTTP odgovor, a vjerojatni su razlozi:

- neispravne ili nepotpune postavke imenskoga poslužitelja (DNS),
- nedostatak ili neispravna konfiguracija web-poslužitelja,
- web-poslužitelj i lokalni DNS konfigurirani su tako da vraćaju resurse na adresi drugačijoj od tih dviju koje je robot testirao, npr. http://forum.domena.hr/. Iako takva web-sjedišta u ovom koraku ostaju neotkrivena, ona će biti otkrivena i harvestirana ako harvester bilo gdje nađe poveznicu na njih.

Tijekom opisanoga kreiranja inicijalnoga popisa URL-ova ustanovljeno je da 3 852 HTTP upita kao ishod imaju preusmjerenje[7] na druge domene. Od tog ukupnoga broja preusmjerenja 2 182 odnosi se na preusmjerenja na stranice koje nisu smještene unutar vršne .hr domene nego na drugim vršnim domenama (primjerice .com, .net, .biz, .org, .de, .eu). Odlučeno je da će se pobrati i takva web-sjedišta iako ne odgovaraju kriteriju da se nalaze u vršnoj hrvatskoj domeni[8].

Parametri harvestiranja

Izbor parametara harvestiranja određuje veličinu i kvalitetu kolekcije koja će se skupiti. Za harvestiranje hrvatskoga weba upotrijebljeni su sljedeći parametri:

- dubina *(max-hops):* 4. Iskustva iz selektivnoga harvestiranja weba kod kojega knjižničari rade validaciju arhivskih kopija i ovisno o kvaliteti kopije prilagođavaju parametar dubine pokazuju da se dubinom 4 može kvalitetno pobrati većina web-sjedišta[9]. S druge strane

[7] HTTP statusi 3xx, najčešće su to: 301 Moved Permanently, 307 Temporary Redirect.
[8] Opcija u Heritrixu koja omogućava takvo ponašanje harvestera jest *seed-redirects-new-seed: true.*
[9] Više od 98 % publikacija u Hrvatskom arhivu weba pobire se dubinom manjom ili jednakom 4.

takva relativno mala dubina osigurač je ako se harvester zapetlja u zamku za robote[10] i počne dohvaćati nepoželjne resurse.

- maksimalan broj resursa po poslužitelju *(queue-total-budget):* 50 000. Osim izbjegavanja skupljanja nepoželjnih resursa ako robot naiđe na zamku, tim ograničenjem željela se postići ravnomjerna zastupljenost velikih i malih sjedišta. Naime, takvim ograničenjem ne dopušta se da velika sjedišta potroše resurse raspoložive za harvestiranje i na taj način bude onemogućeno harvestiranje drugih sjedišta. U početku harvestiranja vrijednost za taj parametar bila je postavljena na 30 000, ali budući da je samo 475 sjedišta došlo do te granice, vrijednost parametra povećana je na 50 000.
- maksimalna veličina resursa *(max-length-bytes):* 100 MB (100 000 000 bajtova). Ograničavanjem veličine resursa koji se skupljaju izbjegava se skupljanje slika DVD i CD medija, distribucija operativnih i programskih sustava. Osim toga, izbjegavaju se i zamke za robote koje poslužuju beskonačno velike datoteke.
- naziv harvestera *(user-agent):* Mozilla/5.0 (compatible; heritrix/1.14.4; +http://haw.nsk.hr/harvestiranje). Harvester u zaglavlju svakog HTTP upita koji postavlja šalje User-Agent polje koje sadrži informacije kojima se harvester predstavlja drugoj strani. U tom polju naveli smo URL-adresu web-stranice koja je sadržavala objašnjenje što radi harvester i kontakt-adresu tima u Srcu koji je provodio harvestiranje i kojem su se vlasnici web-sjedišta mogli obratiti u slučaju da harvester ometa normalan rad njihovih sjedišta ili poslužitelja.
- pristojnost, odnosno obzirnost harvestera (engl. *politeness*). Kako ne bi previše ometao ili kompromitirao rad poslužitelja, poželjno je da robot čini stanke između dvaju uzastopnih upita prema istom poslužitelju. Za to harvestiranje određeno je da ta stanka bude u rasponu od 2 000 *(min-delay-ms)* do 8 000 *(max-delay-ms)* milisekundi.
- tretiranje robots.txt pravila *(robots-honoring-policy type): most-favored.* Standard za isključivanje robota (engl. *Robots Exclusion Standard, Robots Exclusion Protocol, robots.txt protocol*) standardan je način na koji vlasnici web-sjedišta mogu davati upute robotima o tome smiju li pristupati inače javno dostupnom sjedištu ili njegovim pojedinim

[10] Zamke za robote (engl. *crawler trap, spider trap*) jesu web-stranice koje namjerno ili nenamjerno *zarobe* harvester tako da on postavlja beskonačno mnogo upita prema poslužitelju ili dobiva od poslužitelja beskonačno velik odgovor. Primjer su nenamjerne zamke kalendari kod kojih harvester može beskonačno dohvaćati mjesece i godine u budućnost ili prošlost.

dijelovima.[11] Iskustva Australske nacionalne knjižnice[12] pokazuju da se pridržavanjem standarda za isključivanje robota smanjuje kvaliteta rezultata harvestiranja. Nadalje, radeći na Hrvatskom arhivu weba, primijetili smo da postoje web-sjedišta (npr. HRT) koja, da bi smanjila promet, dopuštaju pristup samo određenim robotima (uglavnom Googlebot), dok svim ostalim robotima zabranjuju pristup cijelom sjedištu. Zbog spomenutoga odlučili smo harvestirati sve što je dopušteno bilo kojem harvesteru. Nadzorom harvestiranja primijećeno je da je u slučaju umetnutih resursa[13] potrebno napraviti iznimku pa se umetnuti resursi skupljaju unatoč ograničenjima zadanim pravilima za robote.

- tretiranje umetnutih resursa *(max-trans-hops):* 3. Među opcijama odabrano je da se umetnuti resursi pobiru do dubine 3. Tom postavkom trebalo bi se omogućiti dobro pobranje ako npr. stranica u okviru (engl. *frame*) uključuje drugu stranicu koja uključuje CSS u kojem je URL neke slike koja je element dizajna. Već je spomenuto da se u slučaju umetnutih resursa posve ignoriraju pravila za robote. Tu ćemo odluku poslije potanko objasniti.
- reklame umetnute u web-stranice (DoubleClick, Google Ads, AdOcean,…) treba pokušati filtrirati. Reklame su većinom načinjene tako da ih harvester ne može spremiti, nego se prilikom prikaza arhivske kopije web-stranice umetnute reklame pokušavaju dohvatiti s izvornoga web-sjedišta. Posljedica je toga često prikaz aktualnih reklama unutar arhivskih kopija web-stranica, što zbunjuje korisnike. Ako izvorno web-sjedište više ne postoji ili je promijenilo format isporučivanja reklama, može se dogoditi da arhivske stranice javljaju grešku. Iz spomenutih razloga odlučeno je da se prilikom harvestiranja onemogući dohvat bilo kakvih resursa s poslužitelja koji distribuiraju reklame (doubleclick.net, pagead2.googlesyndication.com, googleadservices.com, adocean.pl, adserver.adtech.de, smartadserver.com, ads.24sata.hr, ad.vecernji.hr, ad.net.hr…).
- umetke društvenih mreža (prvenstveno Facebook) treba pokušati pobrati. To se odnosi na npr. sličice ljudi kojima se sviđa pojedina web-stranica (Slika 1.) ili komentare u svezi sa sadržajem stranice. Problem s harvestiranjem takvih umetaka sličan je opisanom problemu

[11]Robots exclusion standard [citirano: 2011-02-01]. Dostupno na: http://en.wikipedia.org/wiki/Robots_exclusion_standard.

[12] Koerbin, Paul. Report on the Crawl and Harvest of the Whole Australian Web Domain Undertaken during June and July 2005, stranica 14 [citirano: 01.02.2012-02-01]. Dostupno na: http://www.archive.org.au/documents/domain_harvest_report_public.pdf

[13] Umetnuti (engl. *embedded*) resursi koji su sastavni dio neke web-stranice, npr. slike, video, CSS datoteke, JavaScript datoteke, okviri (engl. *frame, iframe*).

s umetnutim reklamama, a to je da se zbog ograničenja harvestera umetci ne uspiju uvijek dobro dohvatiti i događa se da se prilikom prikaza arhivske kopije stranice umetnuti resursi dohvaćaju s izvornoga poslužitelja, a ne iz arhiva. Odlučeno je da će se takvi umetci ipak skupljati, a zbog velikoga broja sjedišta koja imaju takve umetke maksimalan broj resursa koji se dohvaćaju s facebook.com domene povećan je na 250 000.

Slika 1. Umetnuti resursi društvenih mreža. Izvor: vecernji.hr

- format u kojem se spremaju skupljeni resursi *(writers):* WARC. WARC format datoteke 2009. godine postao je službeni ISO standard (ISO 28500:2009[14]) za pohranu sadržaja i kontrolnih informacija aplikativnoga sloja mrežnih komunikacija, u što se ubraja i Hypertext Transfer Protocol, odnosno HTTP. Primjenom WARC (**W**eb **ARC**hive) formata moguće je zapisivanje većega broja elektroničkih resursa (npr. web-stranica) zajedno s pripadajućim zaglavljima i metapodatcima u jednu veliku datoteku, što smanjuje potrebe za diskovnim prostorom i olakšava održavanje harvestiranih resursa.

Odabir alata za harvestiranje

Za provedbu harvestiranja razmatrana su dva alata: DAMP i Heritrix. I jedno i drugo harvesteri su otvorenoga programskog koda čija je primarna svrha kvalitetno pobrati i sačuvati digitalne sadržaje s weba. Programski sustav DAMP razvijen je i održava se u Sveučilišnom računskom centru (Srce) te se od 2004. godine primjenjuje za selektivno

[14] ISO 28500:2009 Information and documentation -- WARC file format [citirano: 2012-02-01]. Dostupno na: http://www.iso.org/iso/catalogue_detail.htm?csnumber=44717

harvestiranje weba u sklopu Hrvatskoga arhiva weba. Razvoj Heritrixa započeo je 2003. godine na inicijativu neprofitne organizacije *Internet Archive*[15], a primjenjuje ga veći broj nacionalnih knjižnica *(French National Library, National and University Library of Iceland, The British Library, The Library of Congress, The Library of Congress,...)* te više komercijalnih projekata.[16]

Radi usporedbe DAMP-a i Heritrixa provedeno je pilot harvestiranje te su vizualno uspoređene arhivske kopije skupljenih sjedišta. Utvrđeno je da nema značajnijih odstupanja u kvaliteti prikaza arhivskih kopija skupljenih jednim i drugim alatom. Usprkos tome razlike između tih alata postoje, a karakteristike po kojima se najviše razlikuju prikazane su u Tablici 3.

Tablica 3. Usporedba alata Heritrix i DAMP

Heritrix	**DAMP**
Rezultati: ARC / WARC format (velike datoteke, manje zauzeće diskovnoga prostora, lakše održavanje), *mirror*,…	Rezultati: *mirror* (1 resurs = 1 datoteka, velik broj relativno malih datoteka).
Za prikaz arhiviranih sadržaja potrebna je dodatna programska podrška, npr. Wayback.	Prilikom harvestiranja prilagođava se sadržaj, pa nije potrebna dodatna programska podrška prilikom prikaza.
Fleksibilnost, velik broj opcija za fino prilagođavanje.	Jednostavnost primjene.
Sprema zaglavlja HTTP upita i odgovora.	Skalabilnost (posao se lako raspodijeli na dodatne poslužitelje).
Primjenjuje ga knjižnička zajednica.	

Nakon analize spomenutih dvaju alata za provođenje harvestiranja odabran je Heritrix. Glavni razlozi za njegov odabir bili su mogućnost spremanja resursa u WARC formatu i fleksibilnost. Fleksibilnost harvestera, tj. mogućnost finoga prilagođavanja raznih aspekata harvestiranja u slučaju harvestiranja domene vrlo je važna jer takvo harvestiranje uključuje

[15] Internet Archive (IA) neprofitna je organizacija osnovana radi izrade javne digitalne knjižnice interneta (Internet Digital Library).

[16] Detaljniji popis korisnika Heritrixa: Users of Heritrix [citirano:2012-02-01]. Dostupno na: https://webarchive.jira.com/wiki/display/Heritrix/Users+of+Heritrix

velik broj web-sjedišta koja se značajno razlikuju po broju i vrsti resursa koje sadrže, tehnologijama koje primjenjuju te načinu na koji su konfigurirane. Treba istaknuti da veća fleksibilnost Heritrixa ujedno znači i da se od korisnika očekuje veća razina tehničkih znanja, tj. moglo bi se reći da je Heritrix više namijenjen informatičarima nego knjižničarima, dok za DAMP vrijedi obrnuto.

Sustav za provedbu harvestiranja

U sklopu priprema harvestiranja provedena su dva pilot-harvestiranja na uzorku od 100 web-sjedišta. Pilot-harvestiranja provedena su u svrhu testiranja parametara harvestiranja, usporedbe alata DAMP i Heritrix te procjene potrebnih računalnih resursa za uspješno provođenje harvestiranja (prilikom prvoga pilot-harvestiranja) i testiranja hoće li alocirani resursi biti dovoljni (prilikom drugoga pilot-harvestiranja).

Za potrebe harvestiranja alocirani su sljedeći računalni resursi Srca: virtualni poslužitelj s četirima jezgrama, 24 GB radne memorije i 10 TB diskovnoga prostora.

Na poslužitelju je instalirana programska podrška otvorenoga koda: operativni sustav Debian GNU/Linux, Java SDK 1.6, Apache Tomcat 6.0, MySQL 5.1, dorađeni Heritrix 1.14.4 i Wayback 1.6.0. Za potrebe harvestiranja na Heritrixu verzije 1.14.4. tim Srca učinio je sljedeće dorade:

- implementiran je dio koda koji za dohvaćene web-stranice forsira harvestiranje svih umetnutih (engl. *embedded*) resursa bez obzira na zabrane u pravilima za robote,
- proširen je skup izraza za odbacivanje pogrešno detektiranih URL-ova unutar JavaScipt koda te Flash datoteka,
- implementirana je zakrpa koja sprječava ponavljanje znaka „/“ u URL-ovima[17].

Provedba harvestiranja

Harvestiranje je provedeno od 18. 7. do 18. 8. 2011. godine. Da bi se na vrijeme uočili eventualni problemi, harvestiranje je redovito nadzirano. Poslovi nadzora uključivali su nadzor rada poslužitelja koji je provodio harvestiranje, nadzor funkcionalnosti Heritrixa te kontrolu kvalitete harvestiranja. Kontrola kvalitete s jedne se strane odnosila na provjere pobire li se sve što treba da bismo u konačnici imali arhivske kopije web-sjedišta što sličnije izvornom obliku i s druge strane pobire li se nešto što nije potrebno i na taj način nepotrebno

[17] Riječ je o dokumentiranoj grešci koja je ispravljena u radnoj verziji Heritrixa. Opis greške: Double Forward-Slashes [citirano: 2011-02-01]. Dostupno na: http://tech.groups.yahoo.com/group/archive-crawler/message/6983

troši resurse harvestera. Niz izvještaja[18] te razne mogućnosti pregledavanja i filtriranja logova koje pruža Heritrix značajno olakšavaju posao kontrole kvalitete tijekom i nakon završetka harvestiranja. U nastavku će biti spomenuti neki od izazova na koje smo naišli tijekom harvestiranja.

Umetnuti resursi i pravila za isključivanje robota

Ubrzo nakon pokretanja harvestiranja ustanovljeno je da se za relativno velik broj sjedišta ne skupljaju umetnuti resursi kao što su slike i CSS datoteke. Nakon analize otkriveno je da je razlog tome u inicijalno restriktivnim pravilima za robote Joomla CMS-a[19] koje održavatelji web-sjedišta nakon instalacije uglavnom ne promijene. Iz isječka inicijalne robots.txt datoteke Joomla CMS-a vidljivo je da je svim robotima zabranjen pristup direktorijima sa slikama i multimedijskim sadržajima:

User-agent: *

Disallow: /images/

Disallow: /media/

U fazi planiranja harvestiranja odlučeno je da će se harevester pridržavati samo zabrana koje se odnose na sve robote (*most-favored* politika), stoga se nisu pobirali umetnuti resursi iz direktorija kojima je bio zabranjen pristup s pomoću pravila za robote. Web-stranice arhivirane bez umetnutih slika i stilova (CSS) izgledaju osiromašeno i kvaliteta takvog harvestiranja bila bi upitna, pa smo odlučili da se za umetnute resurse dohvaćenih stranica posve ignoriraju pravila za robote. Drugim riječima, ako je dohvaćena stranica, tada se dohvaćaju i svi umetnuti resursi bez obzira na restrikcije u pravilima za robote. Takav pristup zahtijevao je i doradu programskoga koda Heritrixa.

Zamke za robote

Zbog relativno male dubine harvestiranja (*max-hops*: 4) nisu očekivani značajniji problemi sa zamkama za robote, ali se nekoliko takvih situacija ipak dogodilo. Otkrivena su tri[20] web-sjedišta koja u interne poveznice (engl. *hyperlinks*) dodaju neku vrstu brojača ili

[18] Web-sučelje Heritrixa sadrži sljedeće izvještaje: Crawl report, Seed report, Frontier report, Processors report, ToeThread report. Više o njima možete naći u dokumentaciji Heritrixa.

[19] Joomla je sustav za uređivanje sadržaja (engl. *Content Management System*, tj. CMS) za izradu web-sjedišta i web-aplikacija. Više o Joomli na: http://www.joomla.org/

[20] Sjedišta su: http://www.kostrena.hr/, http://mali-losinj.hr/, http://www.textum-dekor.hr/.

oznake vremena (engl. *timestamp*). Primjer je takve prakse parametar *uui* u sljedeća dva URL-a:

http://www.mali-losinj.hr/Content.aspx?oid=f4f22fbd-0aec-4a4f-ab94-9cf7384854c2&uui=129666339453545757

http://www.mali-losinj.hr/Content.aspx?oid=f4f22fbd-0aec-4a4f-ab94-9cf7384854c2&uui=129666343265342515

Riječ je o istoj poveznici iz izbornika sjedišta (konkretno „AKTI GRADA") koja prilikom svakog dohvata bilo koje stranice sjedišta ima drugačiji URL zbog parametra *uui*. Budući da je URL drugačiji, harvester ne zna da je riječ o istom resursu i dodaje takve nove URL-ove u red čekanja za skupljanje. Broj tako detektiranih novih URL-ova raste eksponencijalno sa svakom razinom dubine pa, primjerice, jedna web-stranica s 40 internih poveznica u izborniku na dubini 4 generira 40^4 odnosno 2 560 000 različitih URL-ova. Kako harvestiranje odmiče, tako redovi čekanja takvih sjedišta postaju sve veći i po tome ih se može prepoznati u Heritrixovu graničnom izvještaju (engl. *frontier report*). Problemi sa spomenutim trima web-sjedištima rješavani su brisanjem redova čekanja takvih sjedišta da se spriječi daljnje nepotrebno trošenje resursa harvestera. U budućim bi pobiranjima trebalo implementirati pravilo da se za ta tri web-sjedišta izbacuje parametar „uui" iz URL-ova.

Web-aplikacije (web-dućani, forumi, galerije…)

Očekivani problemi pojavili su se prilikom harvestiranja web-aplikacija kao što su web-dućani, forumi i galerije slika. Web-dućani standardno za svaki proizvod imaju poveznice tipa *dodaj u košaricu*, *dodaj na listu želja* i *dodaj za usporedbu* (Slika 2.). Harvester (Heritrix) pronalazi i dohvaća te URL-ove i na taj način u web-aplikaciji dodaje sve te proizvode u košaricu, na listu želja i na popis proizvoda za usporedbu. Nakon tih akcija web-aplikacija generira nove poveznice za njihovo poništavanje kao što su *obriši iz košarice*, *obriši iz liste želja*, *obriši s popisa za usporedbu* (Slika 3.). Sve te poveznice koje okidaju neku akciju u web-aplikaciji vraćaju većinom HTML stranicu gotovo istovjetnu prethodnoj (već skupljenoj) s tom razlikom da nova stranica sadrži i poruku korisniku, npr. „*Proizvod je dodan u košaricu*".

Slika 2. Poveznice koje harvester pronalazi u web-dućanu (izvor: http://shop.emgd.hr/)

Slika 3. Posljedice harvesterova okidanja akcija (izvor: http://shop.emgd.hr/)

Opisanim postupkom prikupi se mnogo resursa koji ne pridonose kvaliteti skupljenoga sadržaja i većinom znače nepotrebno trošenje resursa harvestera.

Kod foruma se često uza svaki pojedini odgovor nalaze poveznice kao što su *citiraj korisnika*, *prijavi administratoru* i *pošalji korisniku privatnu poruku*. Svaka ta poveznica predstavlja harvesteru novi URL koji treba dohvatiti i obraditi, a u konačnici svi ti URL-ovi vode na isti resurs – na stranicu za prijavu, tj. autentikaciju korisnika. Slična je situacija i s raznim oglasnicima (Njuškalo, Plavi oglasnik,…): uza svaki oglas nalaze se poveznice kao što su *prijavi oglas administratoru* i *spremi oglas*. Iako te poveznice imaju različite URL-ove, svi oni vode na stranicu za prijavu korisnika.

Kod galerija slika često se svaka slika može ocijeniti ocjenom od jedan do pet. Harvester svaku tu poveznicu za ocjenjivanje prepoznaje kao novi URL i na taj način generira nepoželjan promet i skupljeni sadržaj.

Radi optimalne uporabe resursa harvestera, smatramo da je poželjno filtrirati, odnosno odbaciti URL-ove koji predstavljaju akcije u web-aplikacijama. U sklopu harvestiranja takve smo URL-ove detektirali i odbacivali s pomoću regularnih izraza[21] koje možemo podijeliti u dvije skupine: jednu koja se odnosi na programska rješenja (npr. često rabljeno programsko rješenje za web-dućane Magento[22]) i drugu koja se odnosi na pojedinačne, specifične web-aplikacije, odnosno web-sjedišta (npr. spomenuta sjedišta Njuškalo i Plavi oglasnik). Sa skupinom regularnih izraza koji se odnose na specifična sjedišta uspoređuju se samo URL-ovi koji su pronađeni na tim sjedištima, dok se sa skupinom regularnih izraza koji se odnose na općenita programska rješenja uspoređuje svaki URL koji Heritrix pronađe na bilo kojem sjedištu. Programska rješenja za koja smo implementirali jedan ili više regularnih izraza za odbacivanje URL-ova u konfiguraciji Heritrixa *(decide-rules)* jesu:

- web-dućani: Magento, VirtueMart,
- forumi: phpBB, vBulletin, Fireboard, SMF, Kunena, Web Wiz,
- galerije: Coppermine Photo Gallery.

To je primjer regularnoga izraza koji za sve web-dućane koji rabe Magento programsko rješenje odbacuje poveznice za rad s košaricom (dodavanje stavke u košaricu, brisanje stavke iz košarice, promjena parametara,...):

.*/cart/(?:add|delete|configure|remove|update_quantity)/.*

Preporuka je za vlasnike web-sjedišta da, kad je god to moguće, s pomoću robots.txt datoteka zabrane robotima okidanje akcija u web-aplikacijama.

Pogrešna detekcija URL-ova (JavaScript, Flash)

Heritrix, kao i većina robota, relativno jednostavno pronalazi nove URL-ove u HTML-u jer HTML ima jasno definirane oznake (engl. *tag*) i pripadajuće atribute u kojima se može pojaviti URL. Međutim, kad je riječ o JavaScript kodu ili Flash datotekama, taj zadatak nije tako jednostavan. Heritrix tada, kao i većina robota, pokušava s pomoću regularnih izraza u sadržaju prepoznati niz znakova koji bi mogao biti URL. Tako otkriveni potencijalni URL-ovi često nisu pravi URL-ovi i harvester nakon postavljanja HTTP upita prema poslužitelju

[21] Regularni izraz (engl. *regular expressions*) (još i pravilni izraz, ispravni izraz – engl. skr. *regexp* ili *regex*, u množini *regexps*, *regexes* ili *regexen*) niz je znakova koji opisuje druge nizove znakova u skladu s određenim sintaksnim pravilima [citirano: 2012-02-02]. Dostupno na: http://hr.wikipedia.org/wiki/Regularni_izraz

[22] Primjer je sjedišta koje za web-dućan rabi Magento http://shop.emgd.hr/ [citirano 2012-02-01]. Više o samom programskom rješenju za web-dućane na: http://www.magentocommerce.com/ [citirano 2012-02-01].

dobiva odgovor sa statusom 404 Not Found[23]. Primjeri nizova znakova koje Heritrix tretira kao relativne URL-ove jesusu „multipart/form-data" i „video/quicktime", iako je jasno da nije riječ o URL-ovima. Na takvo ponašanje Heritrixa bilo je i pritužbi vlasnika web-sjedišta.

Kad bismo tijekom nadzora ustanovili da se među URL-ovima koji kao posljedicu imaju status 404 učestalo pojavljuje niz znakova za koji je očito da nije URL, dodali bismo novo pravilo za odbacivanje tako detektiranih URL-ova u izvorni kôd Heritrixa (STRING_URI_DETECTOR_EXCEPTIONS). Do kraja harvestiranja dodana su ukupno 34 takva pravila. Osim toga, u konfiguraciju Heritrixa dodano je pravilo (Refinements) da se ne otkrivaju URL-ovi u standardnim JavaScript bibliotekama kao što su jQuery ili Mootools biblioteka, koje sigurno ne sadrže navigacijske poveznice relevantne za harvestiranje hrvatskoga weba.

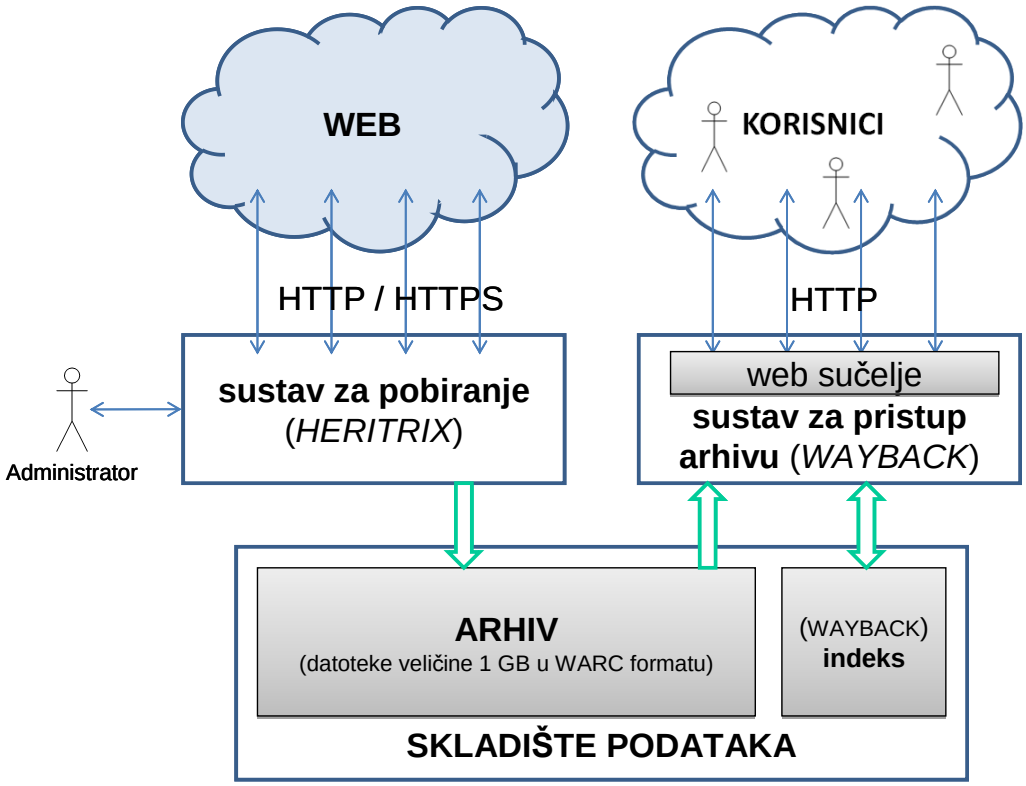
Preusmjerenja na Facebook domenu

Tijekom planiranja odlučeno je da će se harvestiranjem obuhvatiti i sjedišta do kojih se odlazi preusmjerenjem s neke domene koja pripada hrvatskoj vršnoj domeni. U inicijalnom harvestiranju otkrivene su 2 182 takve domene. U 8 slučajeva hrvatska domena služila je za preusmjeravanje na odgovarajuću Facebook stranicu. S obzirom na predviđenu dubinu harvestiranja 4 i veličinu facebook.com domene, broj URL-ova koji se trebao pobrati s nje drastično je rastao, a sadržaj tih URL-ova prelazio je okvire ovoga harvestiranja. Stoga smo smanjili dubinu kojom se harvestira facebook.com domena na 2 (*max-hops*: 2).

Arhitektura sustava

Izvedena arhitektura sustava za harvestiranje prikazana je na Slici 4.

[23] Status 404 Not Found znači da na poslužitelju nema resursa s traženim URI-jem.



Slika 4. Arhitektura sustava za harvestiranje

Skupljeni resursi spremani su u kompresirane (gzip) datoteke u WARC formatu veličine 1 GB. Za pregledavanje resursa iz takvih datoteka potreban je dodatni alat – Wayback. Wayback sučelje omogućava pristup resursima u arhivi s pomoću njihovih izvornih URL-ova[24]. Da bi to funkcioniralo, Wayback detektira nove WARC/ARC datoteke i indeksira njihov sadržaj. Indeksiranje se ne odnosi na indeksiranje teksta, nego povezivanje resursa s pojedinom WARC/ARC datotekom kako bi u trenutku posluživanja bilo poznato u kojoj se WARC/ARC datoteci resurs nalazi. Prilikom posluživanja resursa kao što su HTML, Javascript, Flash i CSS datoteka, Wayback u njihovu sadržaju mijenja poveznice u *lokalne* poveznice arhiva kako bi se resursima pristupalo iz arhiva, a ne s izvornih web-sjedišta. Treba znati da katkad taj postupak nije posve uspješan, pa se događa da web-preglednik dio sadržaja ipak dohvati s izvornih web-sjedišta.

Analiza rezultata

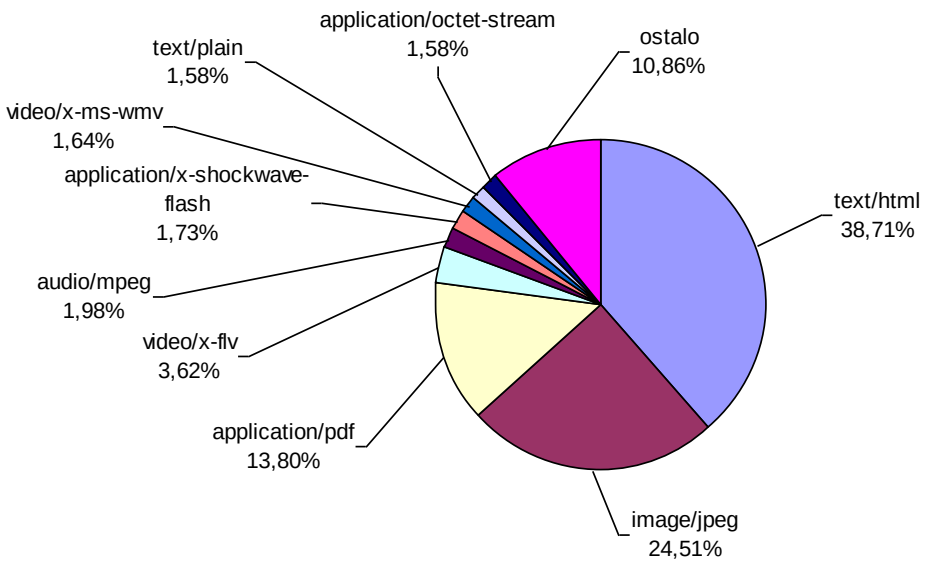
Harvestiranjem nacionalne internetske domene prikupljeni su i arhivirani javno dostupni sadržaji sa svih aktivnih web-sjedišta u domenama .hr, .com.hr i .iz.hr.

[24] Wayback [citirano: 2011-02-01]. Dostupno na: http://archive-access.sourceforge.net/projects/wayback/

Harvestiranje je provedeno od 18. 7. do 18. 8. 2011. godine. Ukupno je prikupljeno i arhivirano više od 56 milijuna datoteka ukupne veličine više od 3.1 TB.

Jednostavnom analizom formata zapisa prikupljenih resursa provjerene su osnovne odlike hrvatskoga web-prostora koji je, u skladu s očekivanjima, i dalje „jednostavan". Tako je još uvijek, kao i prilikom prijašnjih istraživanja[25], 90 % resursa zapisano u desetak osnovnih formata. Kao i prije,[26] tekstualni zapisi u HTML formatu zauzimaju vodeće mjesto i ukupnim brojem i ukupnom veličinom. Od slikovnih formata najzastupljeniji je JPEG. Zanimljivo je zatim primijetiti značajnu zastupljenost PDF formata te, posebno ukupnim brojem, mali postotni udio audiosadržaja i videosadržaja.

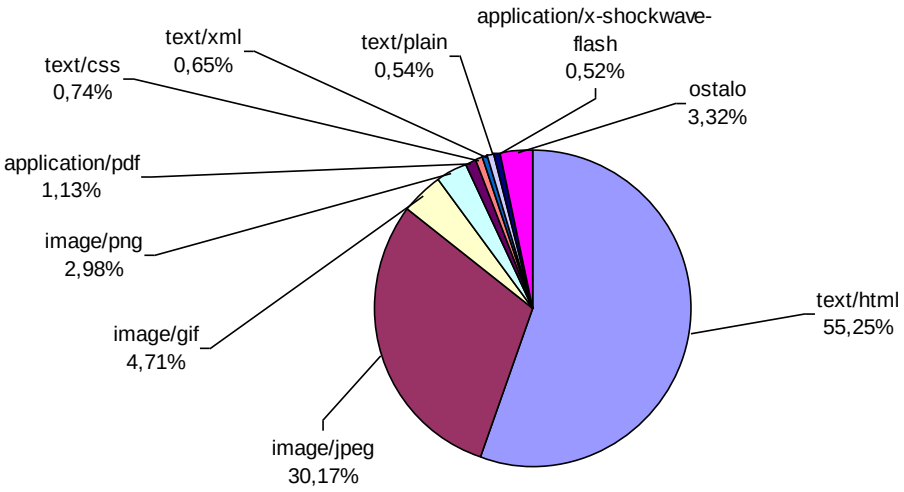
Sljedeće slike prikazuju udio najzastupljenijih formata zapisa veličinom (Slika 5.) i brojem (Slika 6.).



Slika 5. Udio formata zapisa – veličinom

[25] Srce: Mjerenje web-prostora. Dostupno na http://www.srce.hr/mwp/

[26] Rezultati MWP6 iz 2008. (http://www.srce.hr/mwp/) pokazivali su da je udio teksta 33,82 % u ukupnoj veličini te 64,98 % u ukupnom broju resursa.



Slika 6. Udio formata zapisa – brojem

Zaključak

Harvestiranje hrvatske vršne domene provedeno je u predviđenom roku, a rezultati harvestiranja dostupni su korisnicima s pomoću alata Wayback. Smatramo kako je harvestiranje provedeno uspješno te ga je moguće po potrebi ponoviti. Sadržaj skupljenih resursa moguće je indeksirati i omogućiti pretraživanje rezultata harvestiranja po punom tekstu.

Smatramo da je izbor parametara harvestiranja bio je dobar te ga za eventualna buduća harvestiranja nije potrebno značajnije mijenjati. Tijekom harvestiranja kreiran je skup pravila s pomoću kojih se izbjegava okidanje akcija u web-aplikacijama, čime je bilo obuhvaćeno 9 standardnih rješenja za web-dućane, forume, galerije te nekoliko specifičnih web-sjedišta kao što su oglasnici. Kreiran je i skup od 34 pravila za izbjegavanje pogrešnoga detektiranja URL-ova u JavaScript kodu te Flash datotekama. Spomenuti bi rezultati, kada se bude provodilo sljedeće harvestiranje, mogli značajno smanjiti skupljanje neželjenih resursa te tako ubrzati harvestiranje i u konačnici osigurati kvalitetniju kolekciju.

Za buduća harvestiranja smatramo kako bi trebalo dodatno unaprijediti:

- poberivost umetnutih videosadržaja, prvenstveno YouTube filmova koji se u ovom harvestiranju nisu uspješno pobrali,

- pokušati izbjeći skupljanje i spremanje dupliciranoga sadržaja (engl. *deduplication*),
- preporuke vlasnicima web-sjedišta kako konfigurirati robots.txt datoteke da se izbjegne nepotrebno okidanje akcija u web-aplikacijama,
- preispitati opseg harvestiranja jer kriterij da se pobire hrvatska vršna domena izvan opsega harvestiranja ostavlja velik broj hrvatskih webova smještenih na drugim vršnim domenama.

LITERATURA

Fielding, R. RFC 2616 Hypertext Transfer Protocol HTTP/1.1 - Status Code Definitions [citirano: 2012-02-01]. Dostupno na: http://www.w3.org/Protocols/rfc2616/rfc2616-sec10.html

Lasfargues, France; Oury, Clément; Wendland, Bert. Legal deposit of the French Web: harvesting strategies for a national domain [citirano: 2011-02-01]. Dostupno na: http://iwaw.europarchive.org/08/IWAW2008-Lasfargues.pdf

Koerbin, Paul. Report on the Crawl and Harvest of the Whole Australian Web Domain Undertaken during June and July 2005 [citirano: 01.02.2012]. Dostupno na: http://www.archive.org.au/documents/domain_harvest_report_public.pdf

Milinović, Miroslav; Topolščak, Nebojša. The architecture of DAMP. // Widwisawn, 3, 3 (2005) [citirano: 01.02.2012-02-01]. Dostupno na: http://widwisawn.cdlr.strath.ac.uk/Issues/Vol3/issue3_3_1.html

Milinović, Miroslav; Topolšćak, Nebojša. DAMP II: Digitalni arhiv mrežnih publikacija: nova funkcionalnost – novi planovi. // 9. seminar Arhivi, knjižnice, muzeji: mogućnosti suradnje u okruženju globalne informacijske infrastrukture: zbornik radova / uredile Mirna Willer i Ivana Zenić. Zagreb: Hrvatsko knjižničarsko društvo, 2006. Str. 40-55.

Sigurđsson, Kristinn et al. Heritrix User Manual [citirano: 2012-02-01]. Dostupno na: http://crawler.archive.org/articles/user_manual/index.html

Srce: Mjerenje web-prostora [citirano: 2012-02-01]. Dostupno na http://www.srce.hr/mwp/

The Web Robots Pages [citirano: 2012-02-01]. Dostupno na: http://www.robotstxt.org/

Wayback [citirano: 2012-02-01]. Dostupno na: http://archive-access.sourceforge.net/projects/wayback/