Hrvatski web prostor
-
15 godina mjerenja

Miroslav Milinović
Sveučilište u Zagrebu, Sveučilišni računski centar
<Miroslav.Milinovic@srce.hr>

Zagreb, 15. veljače 2016.

srce
Sveučilište u Zagrebu
Sveučilišni računski centar

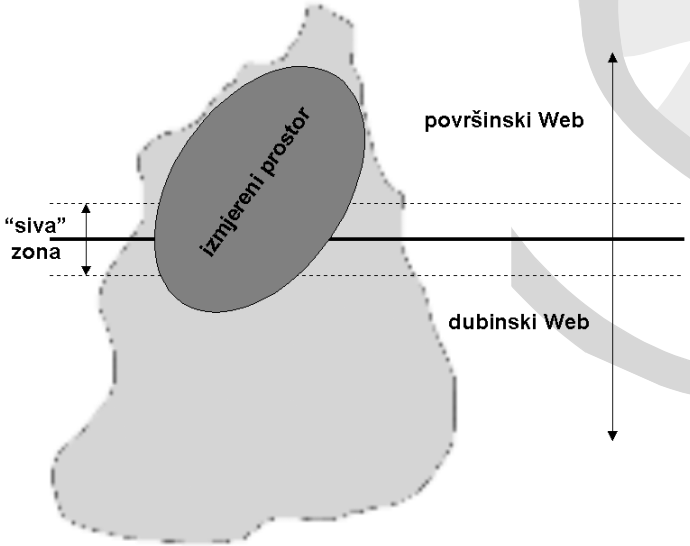
Web – kako je počelo?

Information Management: A Proposal

Tim Berners-Lee, CERN, March 1989, May 1990

"The Web will become a
robust, scalable, adaptive infrastructure,
framework for computation of knowledge,
communication medium."

O mjerljivosti weba



Postoje procjene kako je dubinski web 400-550 puta veći od površinskog weba (Bergman, Michael K. The deep Web: Surfacing Hidden Value. White Paper. The Journal of Electronic Publishing, University of Michigan, July 2001.)

Iskustva Srca

- projekt mjerenja web-prostora (MWP)
  - od 2002. do 2008. godine
  - programski sustav MWP

- suradnja s NSK-om na razvoju i održavanju Hrvatskog arhiva weba (HAW)
  - kontinuirano od 2003. godine
  - http://haw.nsk.hr
  - programski sustav DAMP

- suradnja sa Središnjim državnim uredom za razvoj digitalnog društva (ranije: Hrvatska informacijsko-dokumentacijska referalna agencija - HIDRA)
  - kontinuirano od 2004. godine
  - sustav AMD (arhiv mrežnih dokumenata) koji je temelj sustava DAMIR (Digitalni arhiv mrežnih izvora Republike Hrvatske)

Mjerenje hrvatskog weba – kako smo počeli?

- osnovni cilj:
  - prikupiti informacije o veličini i sadržaju hrvatskog prostora weba
- poticaj:
  - suradnja Srca i NSK u okviru II. faze projekta „Nacionalni informacijski sustav knjižnica Republike Hrvatske - NISKA" (zadatak pod nazivom „Izgradnja nacionalne digitalne knjižnice")
- prvo mjerenje:
  - provedeno u vremenu od 29. 3. do 7. 5. 2002.
  - razvijen vlastiti programski sustav - MWP
  - stručni tim Srca: Miroslav Milinović, Hrvoje Stipetić, Dubravko Penezić, Nebojša Topolščak i Dražen Gemić
- daljnji razvoj sustava MWP:
  - I-projekt 2002-066 Mjerenje hrvatskog web prostora (financiralo Ministarstvo znanosti i tehnologije RH)
  - proveden u periodu od 5. 11. 2002. do 5. 11. 2003.

Što smo i kako mjerili?

- predmet mjerenja:
  - elektronički resursi dostupni HTTP / HTTPS protokolom s poslužitelja u .hr vršnoj internetskoj domeni
- mjerili smo:
  - veličinu
  - korištene formate zapisa (prema MIME standardu)
  - obim i sadržaj meta podataka
  - (veze među poslužiteljima)
- mjerenje je provedeno:
  - na računalnoj opremi Srca
  - korištenjem vlastitog programskog sustava - MWP
    - baze podataka, program(i) za pobiranje (*gatherer*) i program za upravljanje gathererima i bazom podataka (*controller*)

Provedena mjerenja (2002. - 2008.)

- provedeno je ukupno 6 mjerenja:
  - MWP1: 29. 3. 2002. - 07. 5. 2002.
  - MWP2: 14. 5. 2003. - 22. 7. 2003. (kontrolno mjerenje)
  - MWP3: 8. 9. 2003. - 25. 11. 2003.
  - MWP4: 15. 2. 2005. - 22. 3. 2005.
  - MWP5: 22. 4. 2006. - 29. 6. 2006.
  - MWP6: 23. 12. 2007. – 25. 3. 2008.
- Sustav MWP je konstantno nadograđivan

Usporedba rezultata provedenih mjerenja (2002. – 2008.)

broj domena	MWP1	MWP3	MWP4	MWP5	MWP6
s uspješno obrađenim resursima	4.509	8.202	18.795	27.208	37.007
bez uspješno obrađenog resursa	4.806	7.103	1.340	5.965	7.944
ukupni broj domena	9.315	15.305	20.135	33.173	44.951
udio domena s uspješno obrađenim resursima	48,41%	53,59%	93,34%	82,02%	82,33%

broj poslužitelja	MWP1	MWP3	MWP4	MWP5	MWP6
s uspješno obrađenim resursima	6.565	10.884	33.972	36.391	249.581
bez uspješno obrađenog resursa	7.568	11.670	3.394	7.783	12.609
ukupni broj poslužitelja	14.133	22.554	37.366	44.174	262.190

Usporedba rezultata provedenih mjerenja (2002. – 2008.)

broj resursa	MWP1	MWP3	MWP4	MWP5	MWP6
text	69,81%	65,57%	61,65%	63,24%	64,98%
image	22,98%	29,40%	29,26%	25,65%	29,82%
application	6,66%	4,79%	8,70%	10,94%	4,98%
audio	0,49%	0,19%	0,32%	0,09%	0,15%
video	0,03%	0,04%	0,06%	0,05%	0,05%
ostali	0,03%	0,01%	0,01%	0,02%	0,02%

veličina resursa	MWP1	MWP3	MWP4	MWP5	MWP6
text	24,52%	45,87%	52,11%	52,43%	33,82%
image	5,93%	10,15%	5,23%	5,46%	10,04%
application	62,47%	36,32%	36,92%	36,52%	38,05%
audio	4,58%	3,78%	1,79%	1,19%	5,78%
video	2,41%	3,75%	3,93%	4,38%	12,16%
ostali	0,09%	0,13%	0,01%	0,03%	0,15%

Zaključci nakon 6 godina mjerenja

- rezultati odgovaraju očekivanjima i sličnim istraživanjima provedenim u svijetu
- (površinski) web je relativno jednostavan: rabimo mali broj različitih formata
- dinamički web, inventivne, ali i nestandardne uporabe web tehnologija čine mjerenje sve složenijim

- stečena iskustva iskoristili smo za razvoj sustava DAMP i AMD

- nastavljamo istraživati hrvatski prostor weba

- mjerenje → harvestiranje

Harvestiranje (pobiranje) weba

- harvestiranje = pobiranje
- svrha harvestiranja:
  - indeksiranje (za potrebe tražilica)
  - **arhiviranje (cilj: sačuvati izvorni izgled)**
  - istraživanje (npr. mjerenje ili provjera usklađenosti s normama)
- harvester (*crawler, gatherer, robot)* – program koji obilazi web
  - započinje početnim popisom URL adresa (*seed*)
  - iterativni postupak:
    - dohvat web stranice odnosno svih resursa koji joj pripadaju
    - obrada i pronalaženje poveznica (linkovi na stranice, slike, video, skripte, dokumente, css,…)
    - dodavanje novootkrivenih URL-ova na popis

Hrvatski arhiv weba (HAW)

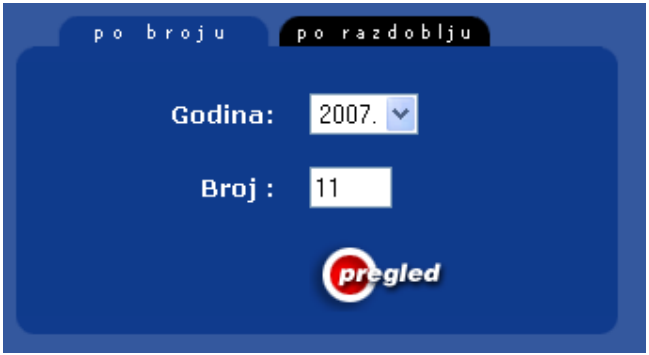
- dva temeljna dijela:
  - **sustav za selektivno pobiranje/arhiviranje**
    - tehnička pozadina je sustav **DAMP** razvijen u Srcu
      - modularan, proširiv i prilagodljiv
      - utemeljen na otvorenom kodu
      - u produkciji od 2004. godine
  - **sustav za arhiviranje nacionalne domene i tematska arhiviranja**
    - modificirana programska podrška **Heritrix / Wayback**
    - harvestiranje .hr domene provodi se redovito jednom godišnje od 2011. godine
    - prvo harvestiranje (arhiviranje) .hr domene provedeno je u vremenu od 18. srpnja do 18. kolovoza 2011.
      (ukupno je pobrano više od 56 milijuna datoteka ukupne veličine od preko 3.1 TB)

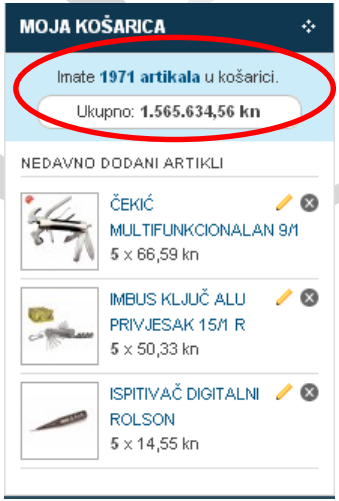
Izazovi pri arhiviranju: planiranje

- izbor resursa koji se arhiviraju
  - protokoli (HTTP / HTTPS),
  - izbor internetske domene
  - početni popis adresa (*seed*)
  - dubina, broj resursa, veličina po sjedištu
  - kriterij završavanja
- umetnuti (*embedded*) resursi
  - slike, video, skripte (Facebook umetci, Google Ads,…), elementi dizajna tj. stilovi, *frame*-ovi…
- preusmjeravanja
  - http://www.nesto.hr/ -> http://nesto.com/
- pridržavanje robots.txt pravila

Izazovi pri arhiviranju: provedba

- konfiguracije CMS alata - robots.txt datoteke koje zabranjuju pristup nekim / svim sadržajima
- „beskonačna” web sjedišta - isti sadržaj na svakoj stranici ima drugačiji link
- Web katalozi / prodaja
- forumi – uz svaki odgovor standardni linkovi
- galerije – uz svaku sliku link na ocjeni sliku
- pogrešna detekcija linkova koji to nisu (javascript, …)





Usporedba rezultata provedenih harvestiranja

godina	broj datoteka (milijuni)	veličina (TB)
2011.	56	3,1
2012.	60	4,1
2013.	69	4,6
2014.	79	5,7
2015.	74	6,1
2016.	77	7,7

Usporedba rezultata provedenih harvestiranja (postotni udio popularnih formata)

	2011.	2012.	2013.	2014.	2015.	2016.
html	55,2	51,8	51,6	51,9	51,2	51,3
jpeg	30,2	33,2	33,6	33,7	34,2	33,8
png	3	3,9	4,5	4,6	4,7	4,5
gif	4,7	3,7	3	2,5	2	1,5
...						
pdf	1,1	1,2	1	0,9	1	1
javascript	1	1,1	1,4	1,5	1,6	1,8
flash	0,5	0,3	0,3	0,2	0,2	0,1

Zaključci nakon 6 godina harvestiranja

- pobiranje / arhiviranje je moguće
  - uz određene ograde
  - moguće je izraditi *snap-shot* (sliku) „nepoberivih” sjedišta
- (površinski) web je i dalje jednostavan
  - rabimo mali broj različitih formata
- autori ne brinu dovoljno o standardima i mogućnosti arhiviranja
- dinamički web, inventivne, ali i nestandardne uporabe web tehnologija čine pobiranje sve složenijim
- izazovi:
  - definiranje opsega pobiranja
  - uobičajeni izazovi za tražilice (forumi, interaktivni katalozi, ...)
  - uloženi resursi (oglasi, društvene mreže, ...)
  - nove web-tehnologije i njihova primjena

MWP1 vs. harvestiranje 2016.

- MWP1 (29.03.2002. - 07.05.2002.)
  - 4.667.920 resursa
  - procjenjena veličina > 300 GB podataka
  - brojčani udio:
    - HTML 67%
    - slikovni formati 23%

- harvestiranje 2016. (25.12.2016. - 02.01.2017.)
  - > 77 milijuna datoteka
  - ukupna veličina ≈ 7.0 TB
  - brojčani udio
    - HTML 51.3%
    - slike u JPEG formatu 33.8%

Hrvatski web prostor
-
15 godina mjerenja

damp@srce.hr

srce
Sveučilište u Zagrebu
Sveučilišni računski centar

Srce politikom otvorenog pristupa široj javnosti osigurava dostupnost i korištenje svih rezultata rada Srca, a prvenstveno obrazovnih i stručnih informacija i sadržaja nastalih djelovanjem i radom Srca.

www.srce.unizg.hr

creativecommons.org/licenses/by-nc/4.0/deed.hr

www.srce.unizg.hr/otvoreni-pristup

srce
otvoreni pristup