O MJERENJU HRVATSKOG PROSTORA WEBA

ON MEASUREMENT OF THE CROATIAN WEB SPACE

Miroslav Milinović
Sveučilište u Zagrebu, Sveučilišni računski centar
miro@srce.hr

UDK / UDC 004.738.1
001.103.2
Istraživanje / Research paper
Primljeno / Received: 8. 7. 2004.

Sažetak

U radu se donosi pregled rezultata projekta *Mjerenje hrvatskog prostora weba* (MWP) koji je izveo stručni tim SRCA. Temeljni je cilj mjerenja procijeniti veličinu i složenost hrvatskog prostora weba te prikupiti osnovne informacije o njegovu sadržaju. U razdoblju od proljeća 2002. do kraja 2003. provedena su tri mjerenja. Dobiveni su rezultati uspoređeni međusobno, ali i s rezultatima sličnih istraživanja provedenih u svijetu.

Ključne riječi: informacijski prostor, web, mjerenje prostora weba, MIME, web-poslužitelj, metapodaci, domena

Summary

The paper gives an overview of the results of *Croatian Web Space Measurement Project* (MWP) conducted by the SRCE team. Basic goal of the project was to estimate the size and complexity of the Croatian Web space and to acquire the basic infromation about its content. In the period between spring 2002 and the end of 2003 the team has conducted three measurements, comparing the results of the three occasions. The paper also includes a comparison with other similar projects.

Keywords: information space, Web, web space measurement, MIME, web server, metadata, domain

1 O mjerenju prostora weba

Globalni informacijski prostor weba iznimno je velik, složen i bogat raznolikim sadržajima, ali istodobno i neuređen. Mrežni izvori informacija (resursi) odlikuju se distribuiranošću i dinamičnošću sadržaja pa je upravljanje internetnim informacijskim prostorom teško i uz ograničene mogućnosti.

Istraživanja kojima se nastoji izmjeriti globalni prostor weba svode se na procjene njegove stvarne veličine. Dinamičnost i distribuiranost sadržaja čini prostor weba potencijalno beskonačnim te se većina mjerenja obavlja na jasno definiranim uzorcima, odnosno uz jasna ograničenja. Osim toga, web-tehnologije, koje omogućuju dinamičku izradu web-stranica kao rezultata interakcije s korisnikom, te ograničavanje pristupa dijelovima prostora weba uporabom odgovarajućih autentifikacijskih i autorizacijskih mehanizama, dodatno otežavaju automatizirano pobiranje resursa, pa tako i mjerenje ukupnog prostora weba. Za takve "skrivene" resurse uveden je naziv "dubinski web" (*deep web, invisible web*). Postoje procjene[1] kako je dubinski web globalno 400-550 puta veći od "površinskog weba" (*surface web, indexable web*), dostupnog konvencionalnim tehnikama pobiranja resursa. Složenosti prostora weba doprinose i sustavi za *caching* i replikaciju (*mirroring*) mrežnih resursa jer uvođenje različitih kopija istog resursa dodatno komplicira informacijski prostor i time dovodi u pitanje metodologiju mjerenja, kao i dobivene rezultate.

Od provedenih istraživanja kojima se želi procijeniti veličina informacijskog prostora weba, izdvajamo istraživanje Lawrencea i Gilesa[2] objavljeno 1999. koje se bavi ne samo procjenom veličine prostora weba, nego i dostupnošću informacija putem weba. Prema tom istraživanju prostor weba u veljači 1999. činilo je oko 800 milijuna mrežnih stranica što dalje dovodi do procjene od 15 TB informacija, od čega je 6 TB čistog teksta. Novije procjene, s početka 2000., govorile su o više od milijarde mrežnih stranica, dok procjene iz srpnja 2001. govore o preko 3,5 milijarde mrežnih stranica. U 2003., prostor weba procijenjen je, vrlo okvirno, na gotovo 5 milijardi mrežnih stranica.

Preciznije su statistike koje govore o broju web-sjedišta. Tako Netcraft[3] navodi brojku od 40,936.076 web-sjedišta identificiranih u lipnju 2003. Eksplozivni rast prostora weba najbolje ilustriraju podaci iz istog izvora koji je u travnju 1997. zabilježio milijun, a već u veljači 2000. više od 11 milijuna web-sjedišta.

Tehnike kojima se mjeri web u osnovi su gotovo identične onima kojima se koriste web-tražilice odnosno sustavi za pobiranje sadržaja s web-sjedišta. Podaci koji se prikupe mjerenjem dobra su osnova za daljnja istraživanja svojstava informacijskog prostora weba, kao i za planiranje i razvoj složenijih sustava za pretraživanje ili arhiviranje prostora weba.

2 Mjerenje hrvatskog prostora weba

2.1 O Projektu mjerenja hrvatskog prostora weba

Mjerenje hrvatskog prostora weba (MWP)[4] projekt je Sveučilišnog računskog centra (SRCE) Sveučilišta u Zagrebu. S namjerom da procijeni veličinu i pruži osnovne informacije o sadržaju hrvatskog prostora weba, SRCE je početkom 2002. započelo

[1]Bergman, Michael K. The deep web : surfacing hidden value. White paper. // The journal of electronic publishing, University of Michigan, July 2001.
[2]Lawrence, Steve; Lee Giles. Accessibility of information on the web. // Nature 400 (8 July 1999).
[3]Netcraft : web server survey archives. Dostupno na: http://news.netcraft.com/archives/web_server_survey.html
[4]SRCE. MWP projekt. Dostupno na: http://www.srce.hr/MWP

rad na razvoju sustava za mjerenje prostora weba. Izravni je poticaj bila suradnja s Nacionalnom i sveučilišnom knjižnicom na projektu NISKA. Upravo je za potrebe tog projekta, u vremenu od 29. 3. do 7. 5. 2002. obavljeno prvo mjerenje hrvatskog prostora weba (MWP1) kojim su obuhvaćeni resursi dostupni protokolom HTTP u *.hr* vršnoj internetnoj domeni. Cilj mjerenja bio je ustanoviti:

- veličinu mjerenog prostora weba,
- korištene formate datoteka (vrste) prema standardu MIME,[5]
- omjer teksta, slike, audio- i video-zapisa, te
- opseg i sadržaj metapodataka

te tako pružiti osnovne informacije o veličini i sadržaju hrvatskog prostora weba. Između ostalog, željelo se provjeriti i tezu o jednostavnosti weba na kojem prevladava nekoliko osnovnih vrsta odnosno formata podataka.

Detaljni rezultati MWP1 publicirani su na mrežnim stranicama projekta.[6] Uz rezultate, detaljno su objašnjena i ograničenja kojima je mjerenje bilo podložno.

S namjerom da dovrše razvoj te unaprijede postojeći sustav, stručnjaci SRCA nastavljaju rad na projektu MWP tijekom 2002. i 2003. Nastavak rada na razvoju sustava MWP financiralo je Ministarstvo znanosti i tehnologije RH (danas Ministarstvo znanosti, obrazovanja i športa).

Na temelju rezultata prvoga provedenog mjerenja (MWP1) i stečenih iskustava, na sustavu MWP učinjene su odgovarajuće promjene. Između ostalog, stvoreni su i uvjeti za periodično ponavljanje procesa mjerenja, kao i sustav s referentnim podacima o rezultatima svih mjerenja. Tijekom 2003. provedena su dva mjerenja:

- MWP2: od 14. 5. do 22. 7. 2003. i
- MWP3: od 8. 9. do 25. 11. 2003.

Za razliku od prvog, drugim i trećim mjerenjem, osim resursa dostupnih protokolom HTTP, obuhvaćeni su i resursi dostupni protokolom HTTPS (inačica protokola HTTP s povećanom zaštitom prijenosa podataka). Kao i sva ostala istovrsna mjerenja, i ova su bila podložna određenim ograničenjima koja su podrobnije objašnjena u samom projektu.[7]

Rezultati provedenih mjerenja dobra su podloga kako za daljnja detaljnija istraživanja hrvatskog prostora weba, tako i za razvoj složenijih sustava za pobiranje i pretraživanje ili arhiviranje sadržaja weba. Uspoređivanjem rezultata mjerenja moguće je pratiti promjene mjerenih parametara prostora weba te procijeniti njegov rast.

2.2 Rezultati mjerenja

U nastavku, bit će riječi o rezultatima MWP3 koji će se usporediti s rezultatima MWP1. Kako su mjerenja MWP2 i MWP3 provedena u istoj godini, neće se posebno uspoređivati. MWP2 uzimamo kao kontrolno mjerenje.

Uzorak izmjeren u MWP3 obuhvatio je 10.884 web-poslužitelja. Prema očekivanju, rezultati mjerenja pokazali su da se kao programska podrška najviše, i

[5]MIME Media Types. Dostupno na: http://www.iana.org/assignments/media-types
[6]Isto.
[7]Isto.

to na 60 posto web-sjedišta, rabi neka od inačica web-poslužitelja Apache, dok se Microsoftov Internet Information Server koristi u 30 posto slučajeva.

Ukupna veličina javno dostupnih mrežnih stranica u *.hr*vršnoj internetnoj domeni procijenjena je na više od 548 GB, što iznosi povećanje od 41 posto s obzirom na prvo mjerenje iz 2002.

Broj, prosječna veličina (u bytima) i udio u ukupnom broju resursa za deset najčešćih vrsta sadržaja po MIME standardu (*MIME types*) prikazani su u Tablici 1. Vidljivo je da više od 95 posto svih resursa ima jednu od pet najčešćih vrsta.

MIME vrsta	**Broj resursa**	**Prosječna veličina**	**Udio**
text/html	3,194.548	28.902	58,80%
image/jpeg	985.171	24.603	18,13%
image/gif	562.218	6.964	10,35%
text/plain	321.210	124.035	5,91%
application/x-tar	158.608	392.911	2,92%
image/png	47.069	14.991	0,87%
application/pdf	36.740	420.470	0,68%
application/zip	15.435	694.089	0,28%
text/x-chdr	14.985	5.065	0,28%
application/msword	13.022	139.903	0,24%
ostalo	84.259	441.351	1,55%

Tablica 1

Ukupna veličina, prosječna veličina resursa i udio u ukupnoj veličini izmjerenog prostora za deset veličinom najzastupljenijih MIME vrsta prikazani su u Tablici 2. Pokazuje se da više od 90 posto weba zauzima deset sumarno najvećih vrsta.

MIME vrsta	**Ukupna veličina**	**Prosječna veličina**	**Udio**
text/html	92,330.334.777	28.902	31,99%
application/x-tar	62,318.751.973	392.911	21,59%
text/plain	39,841.243.898	124.035	13,81%
image/jpeg	24,238.281.421	24.603	8,40%
application/pdf	15,448.075.009	420.470	5,35%
application/zip	10,713.258.978	694.089	3,71%
application/octet-stream	5,938.983.260	883.120	2,06%
audio/mpeg	4,283.073.081	1,746.767	1,48%
video/mpeg	4,262.712.784	6,029.297	1,48%
image/gif	3,915.379.139	6.964	1,36%
ostalo	25,306.381.151	69.329	8,77%

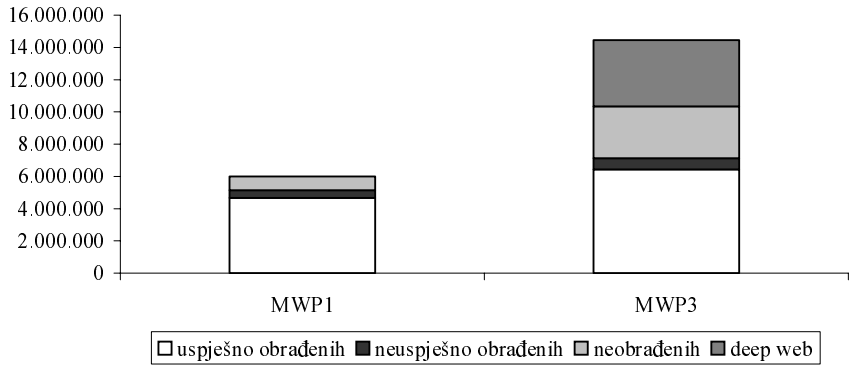
Tablica 2

Prema očekivanjima, najveći broj resursa, gotovo 60 posto, otpada na HTML koji, što se opsega tiče, zauzima nešto manje od 32 posto, uz prosječnu veličinu nešto veću od 20 KB. Na slikovne formate otpada gotovo 30 posto resursa po broju, odnosno samo deset posto po veličini.

Vrste "text/plain" i "application/octet-stream" na web-poslužiteljima najčešći su odabir za MIME vrstu pri posluživanju dokumenata kojima je stvarna vrsta nepoznata, te su zato zastupljeniji no što je to stvarni slučaj. Odstupanje od realnosti može se uočiti kod prosječne veličine vrste "text/plain", za koju se sa sigurnošću može pretpostaviti da za stvarne tekstualne dokumente nije čak 124 KB.

Tijekom mjerenja, resursi se evidentiraju (bilježenjem njihova URL-a) te potom mjere (obrađuju). Mjerenje se prekida kad se dostigne odgovarajući kriterij zaustavljanja[8] odnosno omjer između evidentiranih i obrađenih resursa. Dio obrađenih resursa nije dohvatljiv u trenutku mjerenja pa govorimo o uspješno ili neuspješno obrađenim resursima.

U odnosu na MWP1, u MWP3 broj mjerenjem evidentiranih (registriranih) resursa povećao se za čak 141 posto, a broj obrađenih za 38 posto. Sličan porast pokazao je i broj uspješno obrađenih resursa. U MWP2 i MWP3 uvedena je i automatizirana detekcija resursa koji imaju svojstva dubinskog weba, a koji se dalje ne obrađuju. Pri kraju MWP3, više od 4 milijuna resursa bilo je prepoznato kao dio graničnog područja dubinskog weba. Preostali neobrađeni resursi registrirani su resursi čija obrada još nije započela ili kojima je pristup zabranjen metodama kontrole pristupa za robote. Usporedba broja resursa u MWP1 i MWP3 prikazana je na Slici 1.

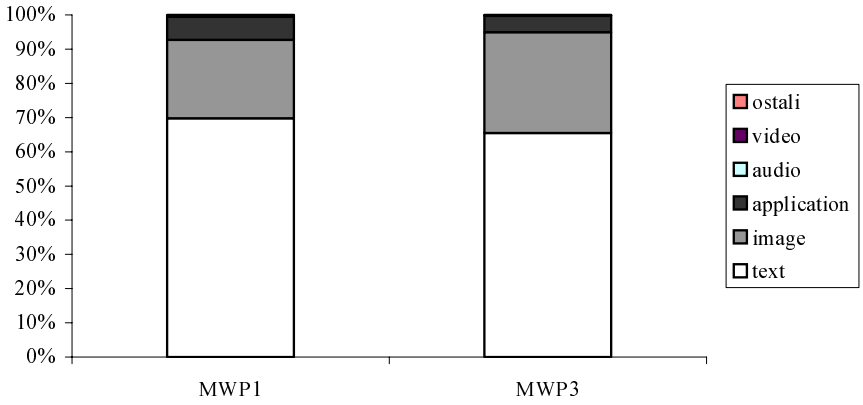


Slika 1

Usprkos smanjenju prosječne veličine resursa, povećana je procjena veličine svih registriranih resursa jer je broj registriranih resursa bitno veći. Koristeći istu metodu, procjena za MWP1 je 389 GB, a za MWP3 548 GB pa povećanje veličine iznosi 41 posto.

Odnos vrsta sadržaja (prema MIME vrsti) izražen brojem obrađenih resursa nije se bitno promijenio u rezultatima mjerenja MWP1 i MWP3, no u MWP3 nešto je zastupljenija MIME vrsta "image". Usporedni prikaz dan je na Slici 2.

[8]Podrobnije vidjeti u: SRCE. MWP projekt. Dostupno na: http://www.srce.hr/MWP



Slika 2

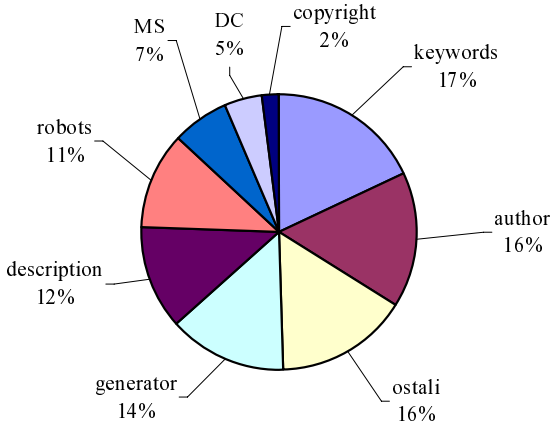
U MWP3 registrirana je uporaba metapodataka na 41 posto mrežnih stranica (HTML datoteka) što znači povećanje u odnosu na 31-postotnu zastupljenost metapodataka dobivenu mjerenjem 2002. Zabilježen je i porast uporabe standarda Dublin Core.

Usporedimo sada rezultate svih triju mjerenja. HTML oznaku META (*META tag*) ima:

- u MWP1: 31 posto HTML resursa / 53,9 posto poslužitelja
- u MWP2: 43 posto HTML resursa / 56,5 posto poslužitelja
- u MWP3: 41 posto HTML resursa / 57,42 posto poslužitelja

Broj različitih vrijednosti NAME atributa HTML oznake META za MWP1 iznosi čak 743, za MWP2 645, a za MWP3 666.

Glede udjela metapodatkovnih standarda odnosno shema, u MWP1 izmjereno je da Dublin Core sudjeluje s 0,09 posto, a metapodatkovni zapisi sukladni zahtjevima tražilica s 19,7 posto. U MWP2 zamijećen je porast uporabe standarda Dublin Core na 2,31 posto, dok je postotak tražilicama prilagođenih zapisa smanjen na 14,62 posto. Odnos korištenih metapodatkovnih shema dobiven u MWP3 prikazan je na Slici 3.



Slika 3

Drugo i treće mjerenje daju nešto bolje rezultate vezane uz uporabu metapodataka, ali valja istaknuti kako su oni dijelom uvjetovani malim brojem novih odnosno obnovljenih web-sjedišta s kataloškim sadržajem i velikim brojem resursa s metapodatkovnim zapisima. U tom smislu valja tumačiti i kolebanja između MWP2 i MWP3.[9] Ponovimo ovdje kako dobiveni rezultati, usprkos pozitivnom pomaku, pokazuju nedovoljnu brigu autora mrežnih stranica za metapodatkovne tehnologije.

Podrobnije informacije o sva tri provedena mjerenja mogu se naći na web-sjedištu MWP projekta.[10]

3 Usporedba sa sličnim mjerenjima u svijetu

Usporedimo rezultate projekta MWP sa sličnim mjerenjima provedenim u svijetu. Za usporedbu rabit ćemo istraživanje Lawrencea i Gilesa[11] te osvrt Juhe Hakale.[12]

Hakala navodi da je švedski prostor weba u proljeće 1999. procijenjen na 7,5 milijuna dokumenata ukupne veličine 300 GB. Iako je ukupno registrirano oko 200 različitih vrsta dokumenata, on je jednostavan jer četiri vrste dokumenata pokrivaju 97 posto svih resursa. Valja uočiti da ove brojke, kao i ocjena o jednostavnosti prostora weba odgovaraju onima koje su dobivene MWP-om. Naime, hrvatski prostor weba u prvom je mjerenju, 2002., procijenjen na više od 300 GB s preko 6 milijuna resursa, od kojih gotovo 95 posto pripada jednoj od pet MIME vrsta.

[9]Detaljnu analizu rezultata mjerenja vidjeti u: Milinović, Miroslav. Analiza uporabe metapodataka u omeđenom informacijskom prostoru weba : magistarski rad. Zagreb : Sveučilište u Zagrebu, Fakultet elektrotehnike i računarstva, 2003.

[10]http://www.srce.hr/MWP/

[11]Lawrence, Steve; Lee Giles. Nav. dj.

[12]Hakala, Juha. Harvesting and archiving the web. Nordunet2000 Conference, Helsinki, Finland, September 2000.

Istraživanje Lawrencea i Gilesa[13] pokazuje da jednostavnu HTML metaoznaku rabi 34 posto web-sjedišta, dok ih samo 0,3 posto rabi Dublin Core. Naše prvo mjerenje (MWP1) daje slične vrijednosti: 31 posto resursa vrste "text/html" ima HTML metaoznaku, ali svega 0,09 posto njih rabi Dublin Core. U drugom su pak mjerenju dobiveni bolji rezultati (43,1 posto resursa ima metaoznaku; 2,31 posto rabi Dublin Core), no pri tome svakako treba imati na umu da oni ne znače ekstremno poboljšanje stvarnog stanja jer su uzrokovani uporabom metapodatka na nekoliko veoma velikih web-sjedišta. Lawrence i Giles u svom istraživanju navode kako su registrirali 123 različita oblika HTML oznake META. MWP-om je, međutim, u prvom mjerenju registrirano 743, a u drugom 645 različitih vrijednosti atributa NAME u oznaci META.

4 Zaključak

Provedenim mjerenjima prikupljeni su temeljni podaci neophodni za svaku daljnju, složeniju analizu, prikupljanje kao i istraživanje mrežno dostupne elektroničke građe u hrvatskom prostoru weba. Rezultati mjerenja važni su i kao povratna informacija izdavačima, dizajnerima i drugim stručnjacima, posebno u akademskoj zajednici.

Provedena mjerenja dala su rezultate u okviru očekivanja, ali i konkretnu podlogu za preporuke kojima se želi postići unapređenje kvalitete hrvatskog prostora weba.

Analiza metapodataka dala je rezultate mjerljive sa svjetskim iskustvima, ali brojnost pogrešaka u uporabi ukazuje na nebrigu autora za taj segment mrežnih tehnologija. Uz nebrigu autora, uzrokom utvrđenog stanja zasigurno je i nepostojanje odgovarajućih programskih alata. Razvoj programa za izradu internetnih resursa svakako mora osigurati i kvalitetnu podršku za izradu i održavanje metapodatkovnih zapisa.

Temeljna preporuka ponajprije se odnosi na intenzivniju uporabu metapodataka, ali uz striktno pridržavanje postojećih metapodatkovnih standarda. Time se stvaraju osnovni preduvjeti za povećanje kvalitete hrvatskog prostora weba i njegovu veću dostupnost svekupnoj internetnoj javnosti, što u prvom redu znači mogućnost efikasnog pronalaženja kvalitetnih sadržaja u hrvatskom prostoru weba.

Iskustva stečena u dosadašnjim mjerenjima poslužila su i kao temelj u razvoju sustava za pobiranje i arhiviranje mrežnih publikacija odnosno mrežom dostupnih dokumenata.

Istaknimo na kraju kako SRCE planira nastaviti aktivnosti vezane uz istraživanje hrvatskog prostora weba i redovito, jednom godišnje, provoditi mjerenje o čijim će rezultatima obavještavati javnost.[14] Sljedeće mjerenje (MWP4) planirano je za prvo tromjesečje 2005., a kao posebnost toga mjerenja valja istaknuti planiranu analizu međusobne povezanosti web-sjedišta u hrvatskom prostoru weba.

[13]Lawrence, Steve; Lee Giles. Nav. dj.

[14]Kontakt adresa za sva pitanja, prijedloge i komentare vezane uz istraživanje prostora weba koje provodi Srce: mwp@srce.hr

LITERATURA

Bergman, Michael K. The deep web : surfacing hidden value. White paper. // The journal of electronic publishing, University of Michigan, July 2001.

Hakala, Juha. Harvesting and archiving the web. Nordunet2000 Conference, Helsinki, Finland, September 2000.

Lawrence, Steve; Lee Giles. Accessibility of information on the web. // Nature 400 (8 July 1999).

Milinović, Miroslav. Analiza uporabe metapodataka u omeđenom informacijskom prostoru weba : magistarski rad. Zagreb : Sveučilište u Zagrebu, Fakultet elektrotehnike i računarstva, 2003.

MIME Media Types. Dostupno na: http://www.iana.org/assignments/media-types

Netcraft : web server survey archives. Dostupno na: http://news.netcraft.com/archives/web_server_survey.html

SRCE. MWP projekt. Dostupno na: http://www.srce.hr/MWP