

SVEUČILIŠTE U ZAGREBU
SVEUČILIŠNI RAČUNSKI CENTAR



MJERENJE HRVATSKOG WEB PROSTORA
MWP6: ZAVRŠNI IZVJEŠTAJ

Zagreb, listopad 2008.

<http://www.srce.hr/mwp/>



Ovo mjerenje hrvatskog Web prostora pripremio je i proveo stručni tim Srca u sastavu:

- **Miroslav Milinović**, voditelj tima;
- **Marko Marušić**, član tima;
- **Dubravko Penezić**, član tima;
- **Nebojša Topolščak**, član tima.

U Zagrebu, 21. listopada 2008.

Ur. broj: 04-5738/003-08.

SADRŽAJ

A. OPIS SUSTAVA.....	5
A1 UVOD	5
A2 OPIS SUSTAVA I NAČINA MJERENJA	5
B. REZULTATI MJERENJA.....	6
B1. REZULTATI MWP6	6
<i>B1.1. Broj domena, web poslužitelja i resursa</i>	<i>6</i>
<i>B1.2. Obim i formati podataka.....</i>	<i>8</i>
<i>B1.3. Povezanost s drugim web sjedištima.....</i>	<i>11</i>
<i>B1.4. Tehnologija</i>	<i>12</i>
<i>B1.5. Metapodaci</i>	<i>13</i>
B2. USPOREDBA MWP6 S PRETHODNIM MJERENJIMA.....	14

A. OPIS SUSTAVA

A1 Uvod

U okviru aktivnosti vezanih uz istraživanje Web tehnologija i ispitivanje hrvatskog internetskog informacijskog prostora Srce, od 2002. godine, u pravilu jednom godišnje provodi mjerenje hrvatskog prostora weba (MWP). U ovom izvještaju donosimo rezultate mjerenja web prostora provedenog pod oznakom MWP6 u vremenu od 23.12.2007. do 25.3.2008. godine.

Terminologija koju rabimo u ovom izvještaju odgovara onoj koja je rabljena u ranijim izvještajima o mjerenjima. Svi su izvještaji dostupni u elektroničkom obliku na adresi <http://www.srce.hr/mwp/>.

Prije samog mjerenja testiran je te dodatno unaprijeđen programski sustav kojeg je Srce razvilo za potrebe mjerenja. Dorade sustava nužne su zbog neprestanog razvoja web tehnologija, a time i promjena uvjeta u kojima se provodi mjerenje.

A2 OPIS SUSTAVA I NAČINA MJERENJA

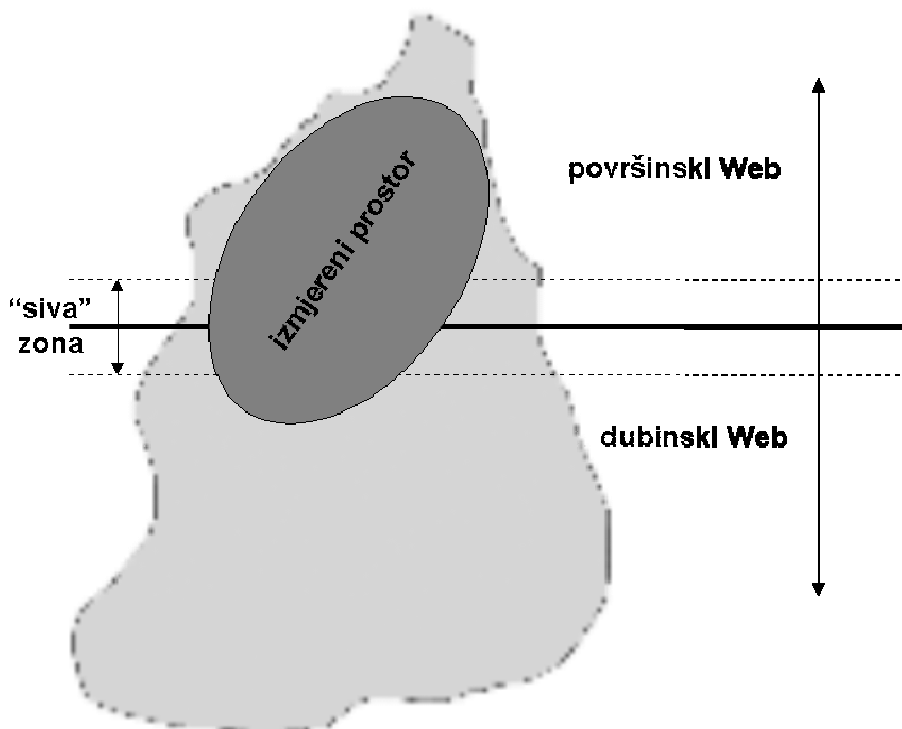
Sustav za mjerenje Web prostora (MWP) sastoji se od tri cjeline:

- relacijske baze podataka u koju se pohranjuju podaci
- programa za pobiranje Weba (gatherer)
- programa za upravljanje gathererima i bazom podataka (controller).

Sustav je distribuirane naravi, i dizajniran da bude neovisan o računalnoj platformi. Kao računalna platforma koristi se UNIX/Linux operacijski sustav, a za bazu podataka MySQL.

Tijekom mjerenja posebna je pažnja posvećena ispitivanju i pravilnom utvrđivanju **kriterija zaustavljanja** procesa. Kako postupak pobiranja dokumenata s Web poslužitelja praktično može trajati vrlo dugo, a teorijski i beskrajno, nikada nije moguće tvrditi da je doista obrađen cjelokupni ciljani informacijski prostor weba. Nužno je dakle bilo postaviti kriterij zaustavljanja procesa mjerenja. U tu svrhu, tijekom mjerenja je redovito praćen broj registriranih Web poslužitelja te broj registriranih i obrađenih resursa. Postupak je prekinut kada je na ispitnom uzorku procesirano više od 75% registriranih resursa. Ispitni uzorak za zaustavljanje međutim sačinjavaju samo resursi s poslužitelja koji ne spadaju u jednu od posebnih kategorija koje pripadaju **dubinskom Webu** (za dodatne informacije vidjeti: *Mjerenje hrvatskog web prostora za potrebe projekta NISKA, SRCE svibanj 2002.* (<http://www.srce.hr/mwp/pdf/mwp-1-report.pdf>) i Bergman, Michael K. *The deep Web: Surfacing Hidden Value. White Paper. The Journal of Electronic Publishing, University of Michigan, July 2001*). Može se dakle reći da se kriterij zaustavljanja temelji na procjeni o obrađenom dijelu površinskog Weba.

Odnos izmjerenog prostora Weba, površinskog i dubinskog Weba ilustriramo slijedećom slikom. Istaknimo da je cilj svih mjerenja bio, u realnom vremenu, što točnije izmjeriti površinski Web.



Dinamičan rast i promjene informacijskog prostora weba kao i povećanje njegove složenosti uzrokovano pojavom novih tehnologija i usluga, čine mjerenja poput projekta MWP svakim danom sve složenijim. Posebno je složeno, temeljem dobivenih podataka procijeniti ukupnu veličinu ne samo dubinskog nego i površinskog dijela weba. Stoga kao rezultat MWP6 ne iznosimo procjenu ukupne veličine hrvatskog prostora weba već samo zapažanja o izmjerenom uzroku.

B. REZULTATI MJERENJA

Šesto mjerenje hrvatskog prostora Weba pod oznakom MWP6 provedeno je u vremenu od 23.12.2007. do 25.03.2008. godine. Nakon provedenog mjerenja izvršena je analiza dobivenih rezultata te usporedba s rezultatima prethodnih mjerenja.

B1. REZULTATI MWP6

B1.1. Broj domena, web poslužitelja i resursa

Domene

Prema podacima HR DNS službe, broj domena u .hr vršnoj Internet domeni na početku mjerenja bio je 44.951 od kojih je 40.358 bilo domena druge razine. Tijekom mjerenja u 37.007 domena (82,33%) uspješno je obrađen barem jedan resurs.

Naglasimo ovdje da tijekom mjerenja nisu dodavane nove domene već je mjerenje temeljeno na početnom broju i popisu domena.

Poslužitelji

Sustav je registrirao 307.142 poslužitelja. Za 44.952 poslužitelja utvrđeno je da su samo sinonimi za izvorni poslužitelj, te su isti dalje tretirani tako da se izbjegne višestruka evidencija i obrada istih resursa.

Od 262.190 izvornih poslužitelja, njih 252.017 je uspješno kontaktirano, a preostalih 10.173 tijekom mjerenja nije bilo dostupno, te se njihova egzistencija uopće ne može potvrditi. Na 249.581 poslužitelja uspješno je obrađen barem jedan resurs.

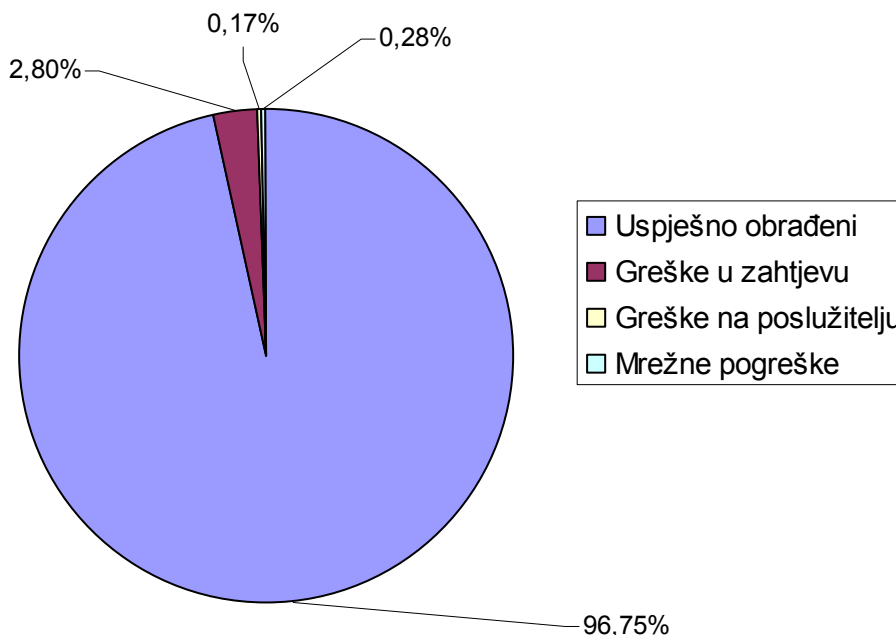
Resursi

Broj registriranih resursa je 77.711.447. Od toga broja 45.964.454 resursa smatrano je dijelom površinskog weba kojeg mjerimo dok je preostalih 31.746.993 resursa procijenjeno dijelom dubinskog weba. Obrađeno je 33.811.576 resursa, od kojih s uspjehom 32.712.621 ili 96,75%.

Koristeći podatak o udjelu uspješno obrađenih resursa u obrađenima, grubo se može procijeniti da bi obrada svih registriranih resursa rezultirala s 75.185.633 uspješno obrađena resursa. Na isti način procjenjujemo površinski dio uzroka na 44.470.496 resursa.

Pogreške

Od grešaka koje su se dogodile prilikom obrade resursa 2,80% su pogreške u zahtjevu (*client error*), 0,17% pogreške na poslužitelju (*server error*), a preostalih 0,28% čine pogreške u komunikaciji računalnom mrežom.



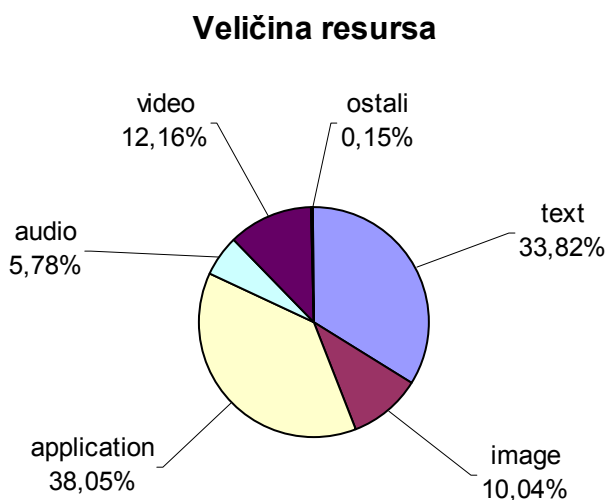
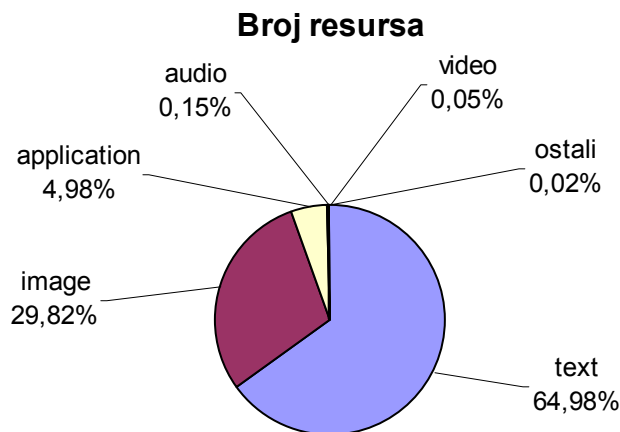
B1.2. Obim i formati podataka

Veličina

Sustav je prikupio podatke o veličini za 25.669.057 resursa, što iznosi 93,05% od 27.587.098 uspješno obrađenih resursa s HTTP statusom 200 (to su resursi koji pri dohvatit rezultiraju dokumentom/objektom). Ukupna veličina tih resursa iznosi 1.923.118.487.989 bytea ili **1.791,04 GB** ili približno 1,75 TB. Prosječna veličina resursa iznosi 74.919,72 bytea.

Procjena veličine površinskog dijela uzorka koji je procjenjen na 44.470.496 resursa iznosi **3.102,90 GB** (3.331.716.978.448 bytea) ili približno 3,03 TB. Ova je veličina dobivena kao umnožak prosječne veličine resursa (74.919,72 bytea) i procjene broja resursa u površinskom webu (44.470.496). Napominjemo da se ovdje radi isključivo o procjeni veličine mjenog uzorka, ali ne i ukupnoj veličini hrvatskog prostora weba.

MIME tip



Za 25.668.642 uspješno obrađena resursa prikupljen je i podatak o MIME tipu. U tablici je dan udio u broju resursa i ukupnoj veličini resursa za pet osnovnih MIME tipova.

MIME tip	Udio u broju resursa	Udio u veličini resursa
text	64,98%	33,82%
image	29,82%	10,04%
application	4,98%	38,05%
audio	0,15%	5,78%
video	0,05%	12,16%
ostali	0,02%	0,15%

U slijedećim tablicama dajemo pregled broja i udjela (unutar pojedinog tipa) za najčešće formate/podtipove (MIME subtype) za pet osnovnih MIME tipova.

MIME tip application	Broj	Udio
application/pdf	334.798	26,19%
application/x-debian-package	221.580	17,33%
application/x-javascript	183.698	14,37%
application/x-gzip	152.579	11,93%
application/atom+xml	111.686	8,74%
application/x-tar	78.123	6,11%
application/msword	72.442	5,67%
application/zip	20.713	1,62%
application/vnd.ms-excel	12.579	0,98%
application/xml	12.543	0,98%
application/vnd.ms-powerpoint	12.507	0,98%
application/octet-stream	10.300	0,81%
ostalo	54.991	4,30%

MIME tip audio	Broj	Udio
audio/mpeg	20.942	53,96%
audio/x-pn-realaudio	12.556	32,35%
audio/x-ms-wma	2.349	6,05%
audio/x-wav	1.193	3,07%
audio/midi	1.021	2,63%
ostalo	750	1,93%

MIME tip image	Broj	Udio
image/jpeg	6.118.910	79,93%
image/gif	1.372.819	17,93%
image/png	113.956	1,49%
image/x-png	23.882	0,31%
image/x-icon	7.995	0,10%
image/bmp	7.533	0,10%
image/pjpeg	5.277	0,07%
ostalo	4.730	0,06%

MIME tip text	Broj	Udio
text/html	16.012.817	96,01%
text/plain	444.896	2,67%
text/css	166.109	1,00%
text/xml	15.579	0,09%
text/x-diff	12.708	0,08%
text/x-csrc	10.349	0,06%
text/x-chdr	8.744	0,05%
ostalo	7.806	0,05%

MIME tip video	Broj	Udio
video/x-ms-wmv	8.200	60,73%
video/mpeg	2.493	18,46%
video/x-msvideo	1.668	12,35%
video/quicktime	547	4,05%
video/unknown	348	2,58%
video/msvideo	104	0,77%
ostalo	142	1,05%

Broj, prosječna veličina (u byteima) i udio u ukupnom broju resursa za deset najčešćih tipova sadržaja dani su u sljedećoj tablici. Pokazuje se da više od 90% svih resursa ima jedan od 5 najčešćih tipova.

MIME tip	broj resursa	prosječna veličina	udio
text/html	16.012.817	33.492	62,38%
image/jpeg	6.118.910	28.413	23,84%
image/gif	1.372.819	8.144	5,35%
text/plain	444.896	251.053	1,73%
application/pdf	334.798	512.299	1,30%
application/x-debian-package	221.580	736.861	0,86%
application/x-javascript	183.698	6.896	0,72%
text/css	166.109	5.265	0,65%
application/x-gzip	152.579	446.482	0,59%
image/png	113.956	25.336	0,44%
ostalo	546.480	1.248.121	2,13%

Sumarna veličina, prosječna veličina resursa i udio u ukupnoj veličini resursa za deset sumarno najvećih MIME tipova dani su u sljedećoj tablici. Ovdje se pokazuje da gotovo 90% veličine izmjenog uzorka zauzima 10 sumarno najvećih tipova.

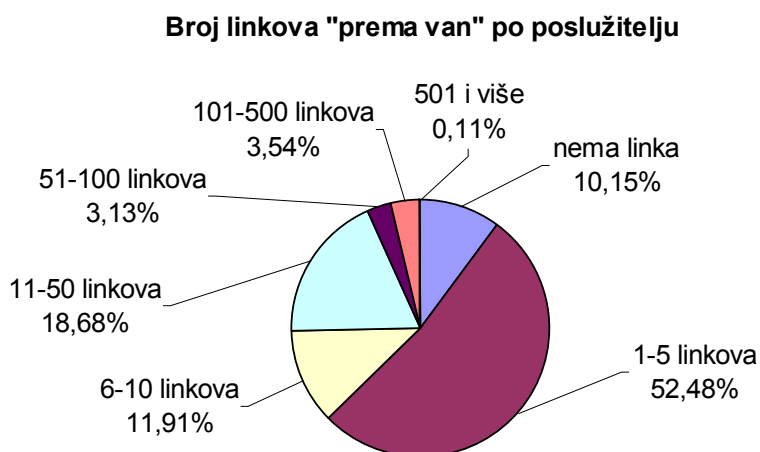
MIME tip	sumarna veličina	prosječna veličina	udio
text/html	536.305.822.722	33.492	27,89%
image/jpeg	173.857.863.660	28.413	9,04%
application/pdf	171.516.772.823	512.299	8,92%
application/x-debian-package	163.273.707.692	736.861	8,49%
video/x-ms-wmv	149.043.989.012	18.176.096	7,75%
application/x-tar	125.277.112.259	1.603.588	6,51%
text/plain	111.692.356.460	251.053	5,81%
audio/mpeg	103.137.331.893	4.924.904	5,36%
application/x-gzip	68.123.848.695	446.482	3,54%
video/x-msvideo	61.539.756.853	36.894.339	3,20%
ostali	259.283.652.122	114.014	13,48%

Tipovi text/plain i application/octet-stream su na web poslužiteljima najčešći odabir za MIME tip pri posluživanju dokumenata kojima je stvarni tip nepoznat, te su zato zastupljeniji no što je to stvarni slučaj.

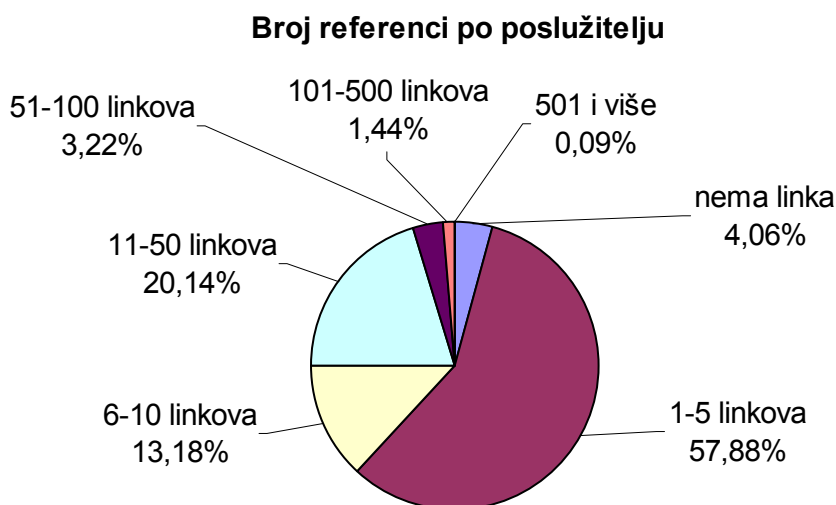
B1.3. Povezanost s drugim web sjedištima

U MWP6 je, po treći puta praćena međusobna povezanost web poslužitelja unutar .hr vršne domene. Imajući na umu mogućnost uporabe kontrole pristupa za robote (vidjeti B1.4) valja pretpostaviti da dio linkova s nekog web poslužitelja na druge web poslužitelje neće biti evidentiran.

Donosimo najprije grafikon na kojem je prikazan, pod nazivom *Broj linkova "prema van" po poslužitelju*, broj veza koje pojedini poslužitelj ima prema drugim poslužiteljima u .hr domeni. Zanimljivo je uočiti da od 249.581 poslužitelja na kojima je uspješno je obrađen barem jedan resurs samo njih 10,15% nema link prema nekom drugom poslužitelju u .hr domeni.



Na slijedećem grafikonu prikazan je, pod nazivom *Broj reference po poslužitelju*, broj evidentiranih veza koje drugi poslužitelji u .hr domeni imaju na odabrani poslužitelj. Ovaj podatak svojevrsna je mjera citiranosti web poslužitelja. Zanimljivo je uočiti da od 249.581 poslužitelja na kojima je uspješno je obrađen barem jedan resurs samo njih 4,06% nije referencirano ni na jednom drugom poslužitelju u .hr domeni.



Za pretpostaviti je da je ovakva dobra povezanost web sjedišta u mjerenom uzroku uzrokovana fenomenom blogova i njihovom uzajamnom povezanošću.

B1.4. Tehnologija

Kontrola pristupa za robote

Na 9.139 izvornih poslužitelja uspješno je dohvaćena robots.txt datoteka, od kojih je 8.136 bilo sintaktički ispravno i uspješno procesirano. Uzevši u obzir ukupni broj uspješno kontaktiranih poslužitelja koji iznosi 252.017 izlazi da samo 3,23% poslužitelja uspješno rabi *Robots Exclusion Protocol* (REP).

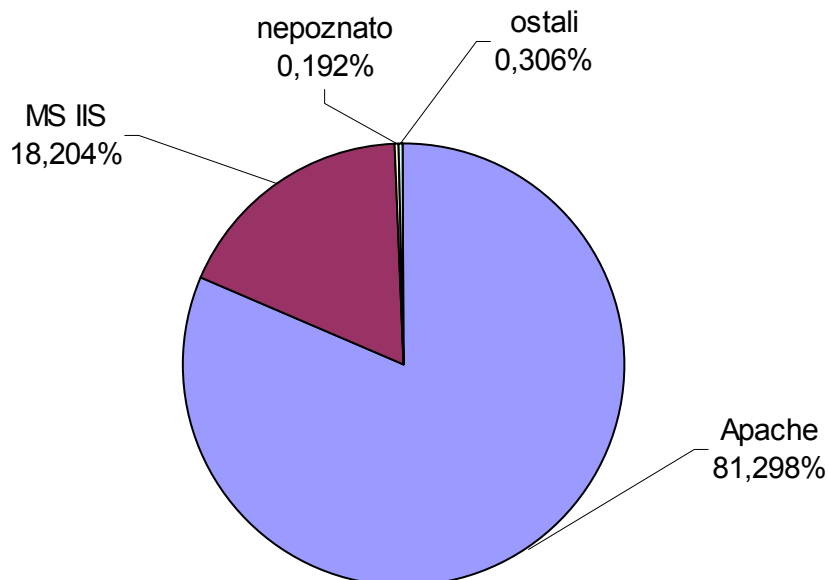
Obrada 1.996.008 registriranih resursa (2,57%) nije pokušana jer im je pristup bio zabranjen putem REP. Vjerojatno je putem REP pristup zabranjen bitno većem broju resursa, no za očekivati je i da se hiperlinkovi na te resurse nalaze upravo na resursima kojima je pristup zabranjen.

Putem META elementa ROBOTS zabranjen je pristup na 32.362 ili svega 0,04% resursa.

Programska podrška poslužitelja

Od 252.017 uspješno kontaktiranih poslužitelja njih 184.588 (73,24%) dalo je informaciju o programskoj podršci koju koristi.

Tako od spomenutih 184.588 web poslužitelja najveći broj još uvijek za programsku podršku koristi Apache (81,298%), a zatim Microsoft IIS poslužitelj (18,204%), što zajedno čini više od 99%. Po svemu sudeći ovakav odnos u korist programske podrške Apache uvjetovan je njenom popularnošću kod davatelja usluga udomljavanja web sjedišta te uporabom za vrlo brojna *blogerska* web sjedišta.



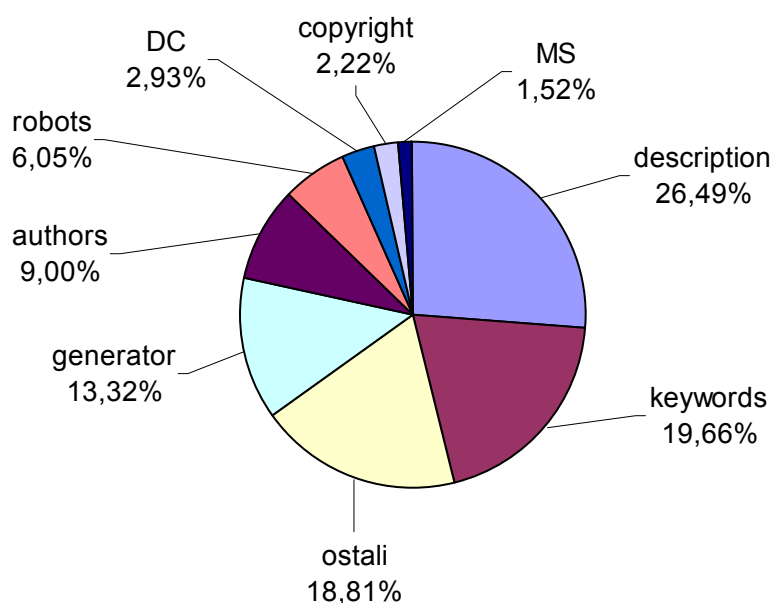
B1.5. Metapodaci

Kao i kod ranijih mjerenja analiziran je način uporabe i obim metapodataka zapisanih u Web stranice putem HTML META oznake. Ponovno je zamjećen nemali broj pogrešaka nastalih:

- neispravnom sintaksom META oznake
- pogreškama u pisanju vrijednosti atributa.

U ovom mjerenju evidentirano je 987 različitih vrijednosti atributa *name* u META oznaci (743 u MWP1, 645 u MWP2, 666 u MWP3, 813 u MWP4, 945 u MWP5). META oznaka s *name* atributom prisutna je na 7.075.703 obrađena resursa što je 44,19% od 16.012.817 koliko je evidentirano resursa tipa *text/html*. META oznaku s *name* atributom u barem jednom svom resursu ima 110.966 poslužitelja što je 44,46% od ukupnog broja od 249.581 poslužitelja u uzorku.

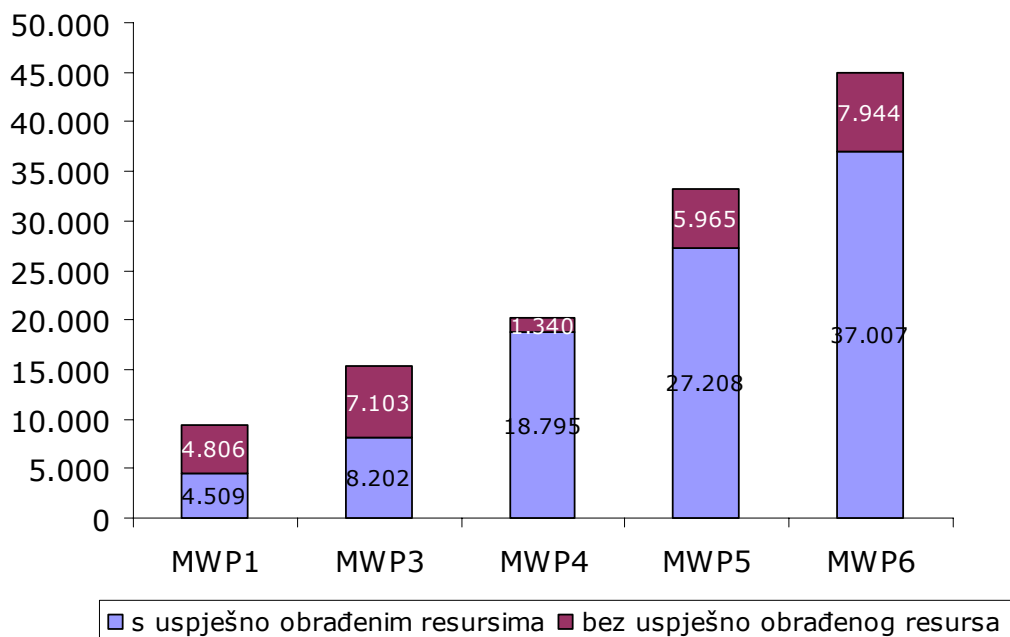
Slijedeća slika daje podatke o udjelu najčešćih vrijednosti *name* atributa u odnosu na ukupni broj pojava *name* atributa koji za ovo mjerenje iznosi 15.199.243.



Treba primjetiti kako rezultati ne pokazuju bitnije promjene glede uporabe metapodataka u odnosu na prethodna mjerenja.

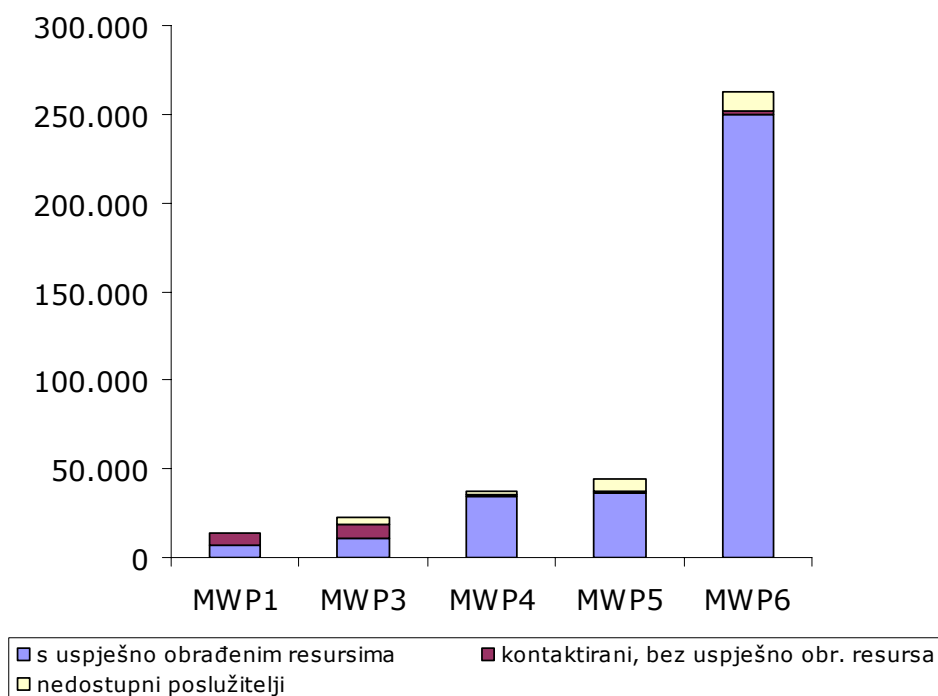
B2. USPOREDBA MWP6 S PRETHODNIM MJERENJIMA

1) broj domena



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%

2) broj poslužitelja



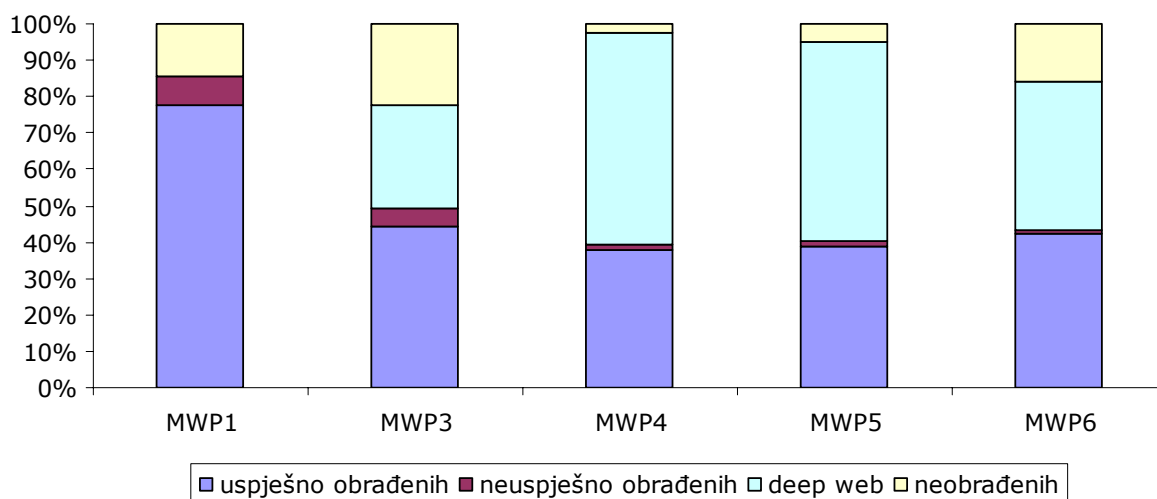
broj poslužitelja	MWP1	MWP3	MWP4	MWP5	MWP6
s uspješno obrađenim resursima	6.565	10.884	33.972	36.391	249.581
bez uspješno obrađenog resursa	7.568	11.670	3.394	7.783	12.609
ukupni broj poslužitelja	14.133	22.554	37.366	44.174	262.190

Promatrajući samo izvorne poslužitelje, dakle ne i njihove sinonime, broj poslužitelja je u porastu, kako ukupni broj tako i broj poslužitelja koji su uspješno kontaktirani.

3) broj resursa

U MWP3 uvedena je i automatizirana detekcija resursa koji imaju svojstva dubinskog weba, te se takvi resursi dalje ne obrađuju. Preostali neobrađeni resursi su resursi čija obrada još nije započeta ili kojima je pristup zabranjen metodama kontrole pristupa za robote.

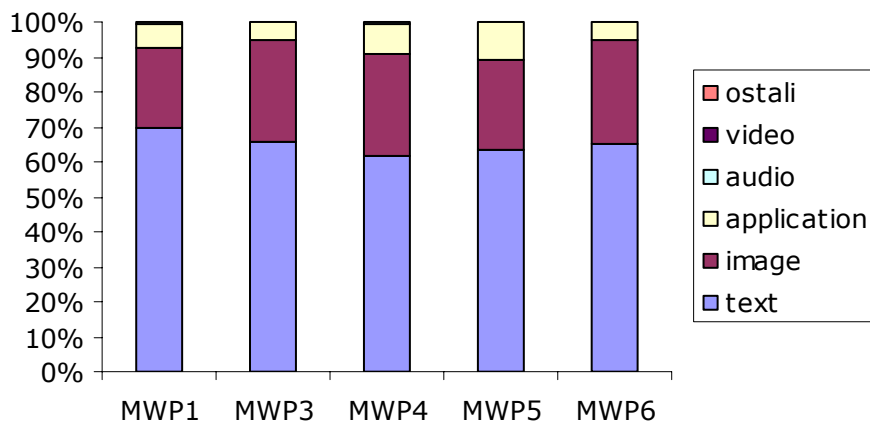
broj resursa	MWP1	MWP3	MWP4	MWP5	MWP6
uspješno obrađenih	4.667.920	6.433.902	10.531.057	18.419.824	32.712.621
neuspješno obrađenih	477.463	691.977	355.144	603.928	1.098.960
neobrađenih	860.873	3.222.080	632.968	2.422.830	12.206.330
deep web	-	4.102.281	16.352.675	25.989.476	31.746.993
registriranih	6.006.105	14.450.240	27.860.215	47.435.716	77.711.447
obrađenih	5.145.383	7.125.879	10.886.201	19.023.752	33.811.581



5) MIME

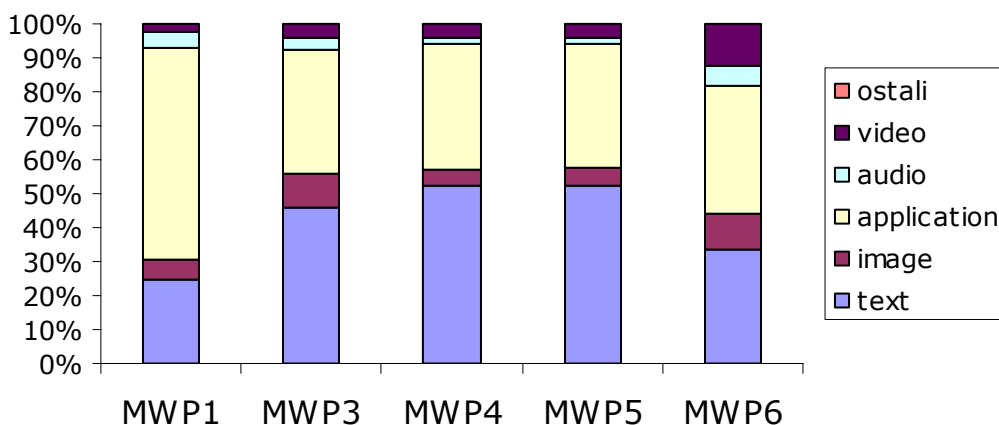
Odnos vrsta sadržaja (prema MIME tipu) izražen brojem obrađenih resursa nije se bitno promijenio niti u MWP6. Valja tek zamjetiti značajniji udio video materijala. Podaci su ilustrirani grafikonom u nastavku teksta. Donosimo i tablice s odgovarajućim brojčanim pokazateljima.

MIME tipovi / 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%

MIME tipovi / 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%