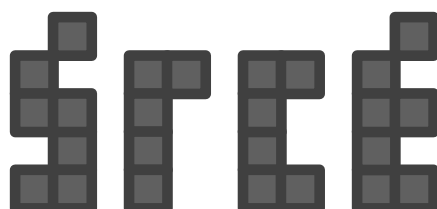


**SVEUČILIŠTE U ZAGREBU**  
**SVEUČILIŠNI RAČUNSKI CENTAR**



**Mjerenje hrvatskog web prostora  
za potrebe projekta NISKA**

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

**Zagreb, svibanj 2002.**

*Ovo izvješće nastalo je kao rezultat rada na projektu «Mjerenje hrvatskog Web prostora za potrebe projekta NISKA» čije je izvođenje ugovorom br.04-46/2002 Sveučilišni računski centar ugovorio s Nacionalnom i sveučilišnom knjižnicom, Zagreb.*

*Projekt je izveo tim u sastavu:*

- Miroslav Milinović, voditelj*
- Hrvoje Stipetić, stručni savjetnik*
- Dubravko Penezić, član tima*
- Nebojša Topolščak, član tima*
- Dražen Gemić, suradnik.*

*Zahvaljujemo svima koji su podržali i pomogli izvođenje ovog projekta.*

*U Zagrebu, 25. svibnja 2002. godine*

**SADRŽAJ**

<b>1. UVOD .....</b>	<b>5</b>
1.1. PREDMET MJERENJA .....	5
1.2. OGRANIČENJA.....	5
1.3. TERMINOLOGIJA.....	5
<b>2. REZULTATI MJERENJA .....</b>	<b>6</b>
2.1. HRVATSKI WEB PROSTOR.....	6
2.1.1. Broj Web poslužitelja i resursa .....	6
2.1.2. Obim i formati podataka .....	7
2.1.3. Dinamički sadržaji .....	11
2.1.4. Meta podaci.....	11
2.2. AKADEMSKA ZAJEDNICA, IZDAVAČI I PUBLIKACIJE U HRVATSKOM WEB PROSTORU .....	13
<b>3. KAKO SMO MJERILI .....</b>	<b>15</b>
3.1. SUSTAV ZA MJERENJE .....	15
3.2. POČETNI UVJETI.....	15
3.3. PROVEDBA MJERENJA .....	15
3.3.1. Kriterij zaustavljanja.....	16
3.3.2. Robot exclusion protocol.....	16
3.3.3. Deep Web .....	17
3.3.4. Posebne kategorije poslužitelja odnosno resursa .....	17
<b>4. ZAKLJUČAK.....</b>	<b>18</b>



## 1. UVOD

### 1.1. PREDMET MJERENJA

Provedeno mjerenje imalo je za cilj prikupiti informacije o veličini i sadržaju hrvatskog Web prostora. Mjerenjem se prije svega željelo ustanoviti:

- veličinu Web prostora,
- korištene formate datoteka (tipove) prema MIME standardu,
- omjer teksta, slike, audio i video zapisa,
- obim i sadržaj meta podataka.

Prikupljeni su temeljni podaci neophodni za svaku daljnju, složeniju analizu mrežno dostupne elektroničke građe u hrvatskom Web prostoru. U daljnjem tekstu pod elektroničkom građom podrazumjevat ćemo mrežno dostupne resurse jednoznačno identificirane URL adresom.

Ovim su mjerenjem obuhvaćeni resursi dostupni HTTP protokolom s poslužitelja u .hr vršnoj domeni. Time je definiran mjereni informacijski prostor.

### 1.2. OGRANIČENJA

Ovim mjerenjem nije obuhvaćen čitav hrvatski Internet informacijski prostor. Ono se odnosi samo na World Wide Web poslužitelje u .hr vršnoj domeni i ne obuhvaća druge mrežne izvore informacija. Važno je također istaknuti da nisu obuhvaćeni niti svi resursi koji su korisniku dostupni putem Internet preglednika (browsera). Preglednici naime podržavaju uporabu niza internetskih protokola dok su ovim mjerenjem obuhvaćeni samo resursi dostupni izvornim Web protokolom HTTP-om. Tako primjerice u mjerenje nisu uključeni resursi dostupni FTP-om.

Mjerenje je također ograničeno na resurse čija je uporaba javno dopuštena odnosno nije zaštićena autentikacijom.

Pri mjerenju su poštivana opće prihvaćena pravila za rad sustava za pobiranje mrežnih resursa (tzv. *robot exclusion protocol*) kakva poštuju svi sustavi za pronalaženje informacija na Internetu. Interesantnom nalazimo informaciju o tome koliko se i kako rabe *robot exclusion protocol* i ROBOT META tag u hrvatskom Web prostoru.

Nije bilo moguće mjeriti niti određene dinamičke resurse i to:

- resurse koji se dinamički generiraju u izravnoj interakciji s korisnikom,
- resurse čija se URL adresa generira dinamički,

kao ni resurse nepovezane s ostatkom Web prostora (tzv. orphan pages).

Konačno, u rezultate nisu uključeni i oni resursi koji su u trenutku mjerenja bili nedostupni.

### 1.3. TERMINOLOGIJA

Radi točnog razumijevanja navedenih podataka ovdje navodimo neke od termina koje smo koristili u daljem tekstu.

**Registrirani resurs** je svaki pronađeni mrežni entitet identificiran URL adresom. Treba naglasiti da svaka pronađena URL adresa ne mora nužno opisivati stvarno postojeći resurs.

**Obrađeni resurs** je resurs koji je sustav:

- uspješno obradio, tj. dohvatio, ili
- nije pokušao obraditi radi zabrane putem *robot exclusion protocola*, ili
- nije uspio obraditi uslijed trajne pogreške pri dohvat (nepostojeći, zabranjeni ili nedostupni resurs).

**Poslužitelj** je u ovom tekstu HTTP poslužitelj u smislu programskog procesa koji se jednoznačno identificira shemom, imenom računala i TCP portom na kojem navedeni proces prima zahtjeve. Tako je primjer poslužitelja npr. "http://www.srce.hr". Ovakva definicija poslužitelja je napravljena temeljem HTTP 1.1 standarda.

**Sinonim** je naziv za poslužitelj koji poslužuje iste entitete (resurse) kao i drugi (originalni) poslužitelj. Najčešće se radi o situaciji u kojoj je jedan te isti programski proces dostupan kroz dva ili više imena računala, npr. http://www.srce.hr je dostupan i kao http://regoc.srce.hr. Nije rijetka i situacija u kojoj isti proces prima zahtjeve na više TCP portova.

## 2. REZULTATI MJERENJA

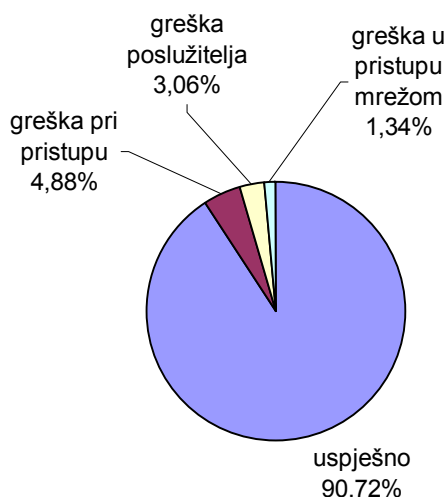
### 2.1. HRVATSKI WEB PROSTOR

#### 2.1.1. Broj Web poslužitelja i resursa

Do trenutka zaustavljanja sustav je registrirao 20.282 poslužitelja od kojih za 6.149 utvrđeno da su samo sinonim originalnog poslužitelja, te su tretirani tako da se izbjegne višestruko mjerenje istih resursa. Nakon eliminacije sinonima, dolazi se do podatka od ukupno registriranih **14.133** poslužitelja.

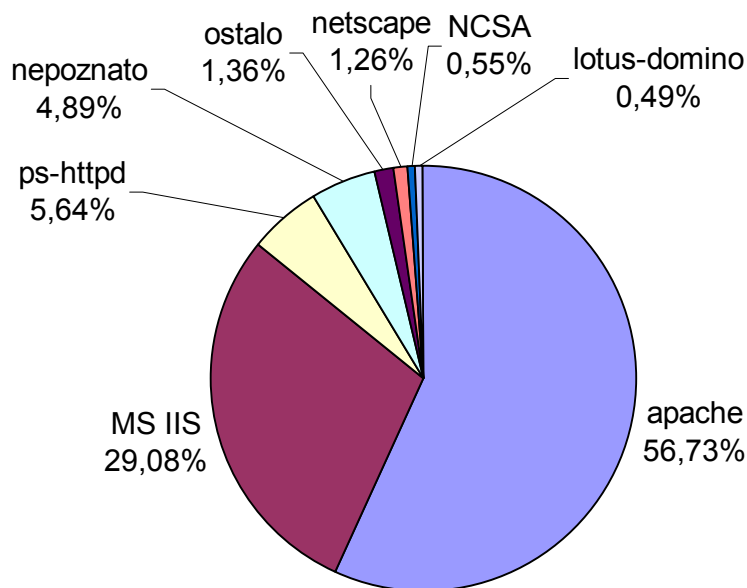
Na **6.564** od tih poslužitelja je uspješno obrađen barem jedan resurs. Preostali poslužitelji nisu bili dostupni, bili su u kvaru ili jednostavno više ne postoje. Jedan dio registriranih poslužitelja, slično kao i kod registriranih resursa, nikad nije niti postojao. Dio od 6.564 poslužitelja u trenutku zaustavljanja mjerenja nije u potpunosti obrađen, odnosno nisu obrađeni svi resursi s tih poslužitelja.

U procesu mjerenja registrirano je **6.006.105** resursa od kojih je **5.145.383** obrađeno. Pri tome je s uspjehom obrađeno **4.667.920** ili 91% resursa kao što je prikazano na slijedećoj slici:



Od ukupno 250.925 grešaka u pristupu, 170.782 otpada na HTTP status 404 – "Not found", odnosno u trenutku pristupa nepostojeće resurse, a 74.125 na status 403 – "Forbidden", odnosno resurse kojima je pristup ograničen. Na resurse kojima je pristup ograničen valja gledati i kao na ulaz u zaštićene dijelove Web prostora čiju veličinu nije jednostavno procijeniti.

Od 6.564 Web poslužitelja za 321 nije bilo moguće ustanoviti o kojoj se programskoj podršci radi zbog nedostatka informacija (poslužitelj ne šalje zaglavlje s identifikacijom servera ili ono sadrži oznaku Unknown). Među preostalim 6.243 registrirane su 384 različite inačice poslužiteljskog programa. Na slijedećoj slici dan je udio najčešće rabljenih Web poslužitelja:



Interesantno je primjetiti da su kod MS Internet Information Servera (MS IIS) zastupljene točno tri inačice 3.0., 4.0 i 5.0 od kojih je ova posljednja uvjerljivo najzastupljenija (89%) dok je Apache zastupljen s bitno većim brojem različitih inačica s različitim kombinacijama dopunskih modula (najčešći su PHP, perl, MySQL, SSL). Uvjerljivo je najzastupljenija inačica Apache 1.3., preciznije od 1.3.11. na više. Među 189 servera koji čine blizu 1,4% ukupne populacije su između ostalog i IBM i Oracle Web poslužitelji, razvijeni na bazi Apachea.

### 2.1.2. Obim i formati podataka

#### Procjena veličine

Od uspješno obrađenih 4.667.920 resursa sustav je nezavisno prikupio podatke o veličini na 2 načina:

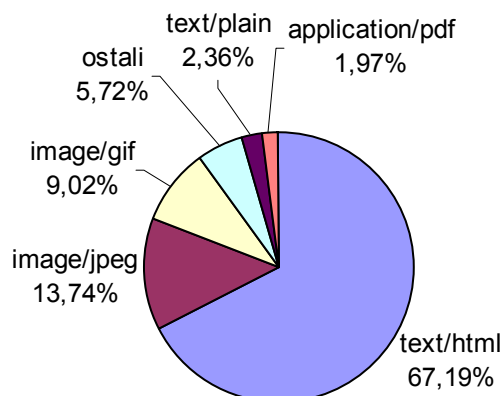
- putem HTTP zaglavlja (Content-Length header) za 2.631.023 resursa
- dohvaćajući HTML stranice HTTP GET metodom za 2.269.458 resursa.

Utvrđeno je da su podaci o veličini ustanovljeni za 3.687.160 resursa (odnosno 79% ukupnog broja resursa koji su uspješno obrađeni). Ukupna veličina tako dobivenog uzorka iznosi 263,4 GB (282.790.680.088 bytea). Veličinu preostalih 21% resursa koji su uspješno obrađeni, ali pripadaju poslužiteljima koji ne šalju Content-Length zaglavlje moguće je procijeniti primjerice tako da se utvrdi njihov tip (MIME) te im pridjeli prosječna veličina izmjerenih resursa za taj tip ili grublje, tako da se rabi prosječna veličina svih izmjerenih resursa koja iznosi 60.582 bytea. Tom grubom procjenom izlazi da je ostatak od 21% obrađenih resursa velik 55,3 GB (59.416.402.320 bytea) te konačno i **procjena ukupne veličine obrađenih resursa od 318,7 GB.**

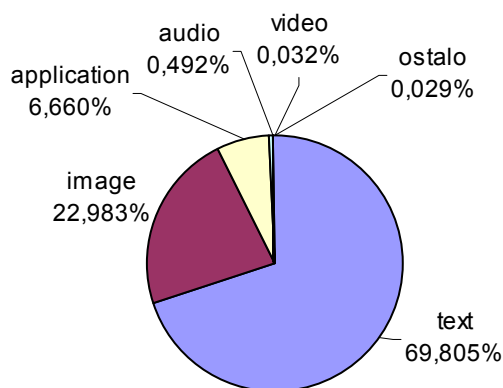
### Broj resursa po tipovima (formatima zapisa)

Od uspješno obrađenih 4.667.920 resursa sustav je putem HTTP zaglavlja (Content-Type header) prikupio podatke o tipu podatka za 4.320.581 resurs (93% ukupnog broja uspješno obrađenih resursa). Interesantno je uočiti kako usprkos više od 150 različitih tipova koji su evidentirani (neki od njih, poput “tekst/html” i kao rezultat odstupanja od standarda) 5 tipova (text/html, image/jpeg, image/gif, text/plain, application/pdf) čini brojem više od 90% resursa iz čega se daje zaključiti da glavnina Weba ipak nije toliko raznolika formatima. Uočiti ipak valja da je tip text/plain u pravilu *default* što znači da ga Web poslužitelj pridjeljuje resursima čiji tip nije u mogućnosti prepoznati, čak i kad im je sadržaj binaran.

Odnos je prikazan na slijedećoj slici.



Distribucija resursa po osnovnim MIME tipovima dana je na slijedećoj slici:



Najčešći podtipovi unutar osnovnih tipova text, image, audio, video i application prikazani su u slijedećim tabelama:

tip (text)	broj resursa	% udio
text/html	2903104	96,257%
text/plain	101822	3,376%
text/css	5940	0,197%
text/xml	2968	0,098%
text/vnd.wap.wml	1502	0,050%
ostali	655	0,022%

<b>tip (image)</b>	<b>broj resursa</b>	<b>% udio</b>
image/jpeg	593469	59,767%
image/gif	389840	39,260%
image/png	8280	0,834%
image/bmp	784	0,079%
image/tiff	302	0,030%
ostali	303	0,031%

<b>tip (audio)</b>	<b>broj resursa</b>	<b>% udio</b>
audio/x-pn-realaudio	10210	48,011%
audio/x-pn-realaudio-plugin	7177	33,749%
audio/mpeg	1664	7,825%
audio/basic	617	2,901%
audio/midi	552	2,596%
ostali	1046	4,919%

<b>tip (video)</b>	<b>broj resursa</b>	<b>% udio</b>
video/mpeg	460	33,774%
video/x-ms-wmv	441	32,379%
video/quicktime	162	11,894%
video/x-msvideo	115	8,443%
video/unknown	95	6,975%
video/msvideo	72	5,286%
ostali	17	1,248%

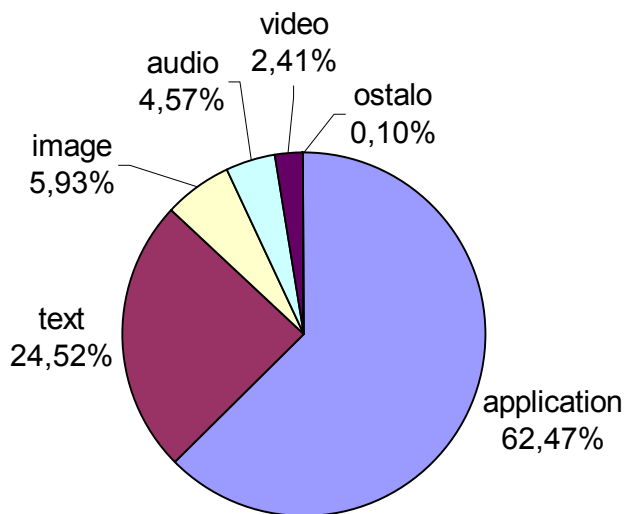
<b>tip (application)</b>	<b>broj resursa</b>	<b>% udio</b>
application/pdf	85209	29,614%
application/x-debian-package	68755	23,896%
application/x-tar	61549	21,391%
application/zip	23573	8,193%
application/octet-stream	17989	6,252%
application/x-zip-compressed	6951	2,416%
application/msword	5002	1,738%
application/x-msdos-program	3820	1,328%
application/x-javascript	3075	1,069%
application/x-gzip	2733	0,950%
application/x-shockwave-flash	1706	0,593%
application/postscript	1479	0,514%
ostali	5889	2,047%

### Odnosi tipa i veličine resursa

Glede pak veličine resursa pojedinog tipa dajemo prvo 5 tipova resursa čiji je % udio u ukupnoj veličini najveći:

<b>tip</b>	<b>% udio</b>
application/octet-stream	16,29%
text/html	14,93%
application/pdf	12,06%
text/plain	9,58%
application/x-tar	9,40%

Slijedeći grafikon prikazuje odnos osnovnih tipova glede veličine pripadnih resursa:



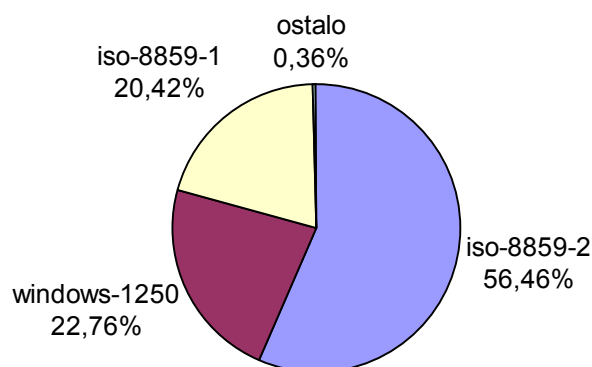
Naredna tablica sadrži usporednu informaciju o najzastupljenijim tipovima brojem i veličinom resursa:

MIME tip	broj resursa		Veličina		
	%	rang	%	rang	prosjek (KB)
application/octet-stream	0,42%	9	16,29%	1	2500,37
text/html	67,19%	1	14,93%	2	14,20
application/pdf	1,97%	5	12,06%	3	390,89
text/plain	2,36%	4	9,58%	4	259,89
application/x-tar	1,42%	7	9,40%	5	421,71
image/jpeg	13,74%	2	4,70%	10	21,89
image/gif	9,02%	3	0,98%	15	6,91

Konačno donosimo i usporednu tablicu za 5 osnovnih tipova.

osnovni tip	broj resursa		veličina		
	%	rang	%	rang	prosjek (KB)
application	6,660%	3	62,467%	1	599,56
text	69,805%	1	24,523%	2	22,45
image	22,983%	2	5,929%	3	16,49
audio	0,492%	4	4,575%	4	594,08
video	0,032%	5	2,410%	5	4885,74

Osvrnimo se na kraju i na uporabu hrvatskih grafema odnosno izbor odgovarajućeg standarda zapisivanja znakova (character set). Uz ISO 8859-2 standard primjeren našem pismu i jeziku postoji i MS "standard" Windows 1250. Mjerenjem je utvrđeno da samo 401.682 resursa ima eksplicitno definiran charset i to ili izravno, podešavanjem Web poslužitelja ili posredno HTTP-EQUIV META oznakom. Odnos korištenih standarda dan je na slijedećoj slici.



### 2.1.3. Dinamički sadržaji

Tijekom mjerenja evidentirane su i informacije o uporabi različitih tehnika za izradu dinamičkih Web stranica. Donosimo nekoliko osnovnih pokazatelja uz napomenu da je temeljem prikupljenih podataka moguće sprovesti i detaljniju analizu.

HTML oznaka Script je često rabljena. S atributom LANGUAGE pojavljuje se na 3.296 poslužitelja u 929.816 obrađenih resursa. Različite inačice JavaScripta (najčešće bez oznake inačice ili 1.1) referenciraju se pri tome u 98% slučajeva. Ostatak otpada uglavnom na VBScript.

Java Applete rabi 576 poslužitelja, preciznije oznaka APPLET evidentirana je u skromnom broju od 10.202 obrađenih resursa.

Tehnologiju Cookiea rabi 1.758 poslužitelja, a evidentirana je u HTTP zaglavlju od 1.642.387 uspješno obrađenih resursa (35,2% od ukupnih 4.667.920).

### 2.1.4. Meta podaci

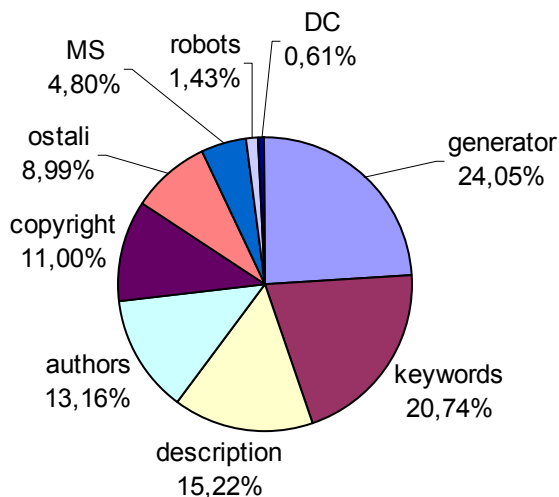
Analiziran je način uporabe i obim meta podataka zapisanih u Web stranice putem HTML META oznake. Valja na početku reći da je zamjećen nemali broj grešaka nastalih:

- neispravnom sintaksom META oznake
- greškama u pisanju naziva atributa, posebno njihovih vrijednosti.

Kako takve greške ne izazivaju probleme korisniku prilikom pregledavanja Web stranice to ih se, po svemu sudeći, slabo ili nikako otklanja.

Evidentirane su 744 različite vrijednosti atributa NAME u META oznaci. META oznaka s NAME atributom prisutna je na 906.090 obrađenih resursa što je 31% od 2.903.104 koliko je evidentirano resursa tipa text/html odnosno 19,4% od ukupnog broja obrađenih resursa.

Slijedeća slika daje podatke o udjelu najčešćih vrijednosti NAME atributa u odnosu na ukupni broj pojava NAME atributa koji iznosi 2.729.585. Pri računanju je svaka vrijednost atributa brojena samo jednom po resursu, iako koncept meta podataka dopušta ponavljanje istog atributa NAME s različitim vrijednostima odgovarajućeg atributa CONTENT, prozvoljan broj puta.



Uočena je uporaba različitih metapodatkovnih standarda. Izdvajamo:

- Dublin Core (DC) - vrijednost atributa NAME ima prefiks "DC."
- metapodaci koje standardno upisuju alati za izradu Web stranica kao što su MS Front Page, Netscape - najčešće vrijednosti atributa NAME su author, generator, copyright
- metapodaci koje rabe tražilice - vrijednosti atributa NAME su keywords i description
- meta podaci koje generiraju MS programski proizvodi
- ROBOTS META oznaka.

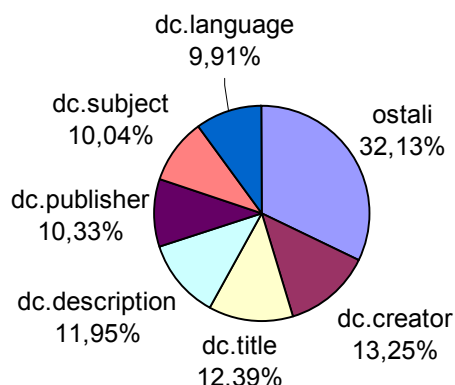
Uporaba nabrojenih standarda nije isključiva već je prema očekivanjima uočeno njihovo kombinirano korištenje.

Daljnjom analizom je ustanovljeno:

- DC se pojavljuje u 2.524 resursa (0,09% resursa tipa text/html)
- metapodatke koje standardno upisuju alati za izradu Web stranica ima 725.554 resursa (25% resursa tipa text/html)
- keywords i/ili description ima 572.259 resursa (19,7% resursa tipa text/html)
- ROBOTS META oznaku ima 39.046 resursa (1,35% resursa tipa text/html).

U 2.524 resursa u kojima je utvrđena uporaba DC-a evidentirano je ukupno 16.565 pojava DC elemenata što znači prosječno više od 6 elemenata po resursu. Izbrojeno je 37 različitih vrijednost NAME atributa, odnosno DC elementa s kvalifikatorima. Ukupni udio DC-a je skromnih 0,28% od ukupnog broja resursa koji sadrže metapodatke.

### Frekvencija uporabe različitih DC elemenata



## 2.2. AKADEMSKA ZAJEDNICA, IZDAVAČI I PUBLIKACIJE U HRVATSKOM WEB PROSTORU

### Poslužitelji

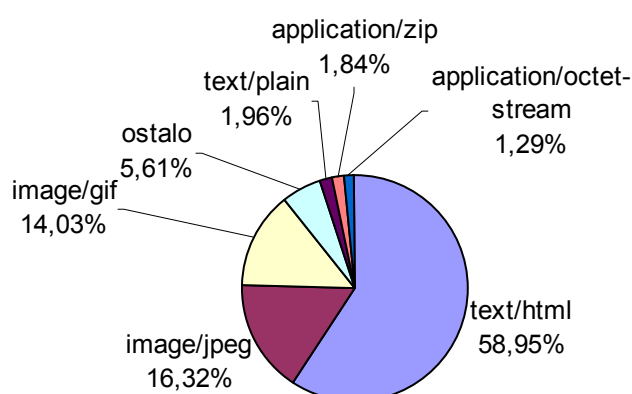
Među 6.564 poslužitelja obuhvaćenih ovim mjerenjem 1.183 pripadaju akademskoj zajednici, točnije ustanovama spojenim u CARNet mrežu, 71 pripada izdavačima izvan akademske zajednice, a 58 elektroničkim publikacijama. Ova je podjela dakako uvjetna i vezana uz vlasništvo nad poslužiteljem i njegov procjenjeni pretežni sadržaj.

### Resursi

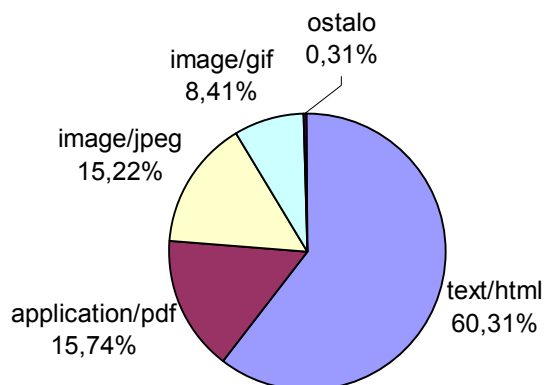
Na poslužiteljima iz akademske zajednice registrirano je 931.652 (20%) obrađenih resursa, na izdavače je otpalo 491.459 (10,5%), a na elektroničke publikacije 687.900 (14,8%).

Slijedeće slike ilustriraju udio pojedinih tipova u ukupnom broju resursa.

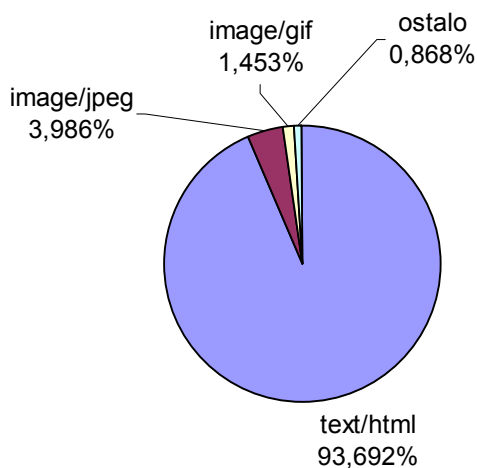
### Udio pojedinih tipova u ukupnom broju resursa – akademska zajednica



### Udio pojedinih tipova u ukupnom broju resursa – izdavači



### Udio pojedinih tipova u ukupnom broju resursa – elektroničke publikacije



Što se veličine obrađenih resursa tiče ona iznosi:

- za akademsku zajednicu: 76,32 GB (29%)
- za izdavače: 32,87 GB (12,5%)
- za elektroničke publikacije: 13,92 GB (5,3%)

### Meta podaci

Od 906.090 obrađenih resursa koji sadrže META oznaku s NAME atributom:

- na akademsku zajednicu otpada: 135.226 od kojih 406 rabi DC
- na izdavače otpada: 168.833 resursa od kojih se u 381 rabi DC
- na elektroničke publikacije otpada: 18.500 resursa od kojih se samo u 6 rabi DC standard.

### 3. KAKO SMO MJERILI

#### 3.1. SUSTAV ZA MJERENJE

Za potrebe mjerenja razvijen je poseban programski sustav koji se sastoji od tri cjeline:

- baze podataka u koju su bilježeni podaci
- programa za pobiranje podataka s Weba (gatherera)
- programa za kontrolu rada gatherera i baze podataka (controllera).

Kao računalna platforma korištena su računala s UNIX operacijskim sustavom. Za bazu podataka odabran je MySQL. Controller je izveden kao aplikacijski poslužitelj temeljen na Java platformi dok je gatherer realiziran kao PERL aplikacija. Sustav je realiziran tako da controller odjednom opslužuje više gatherera: dodjeljuje im resurse koje trebaju analizirati i prima podatke o obrađenim resursima. Gathereri su pak Web roboti koji standardnim tehnikama analiziraju mrežne resurse i prosljeđuju dobivene podatke controlleru.

Tijekom mjerenja sustav je dotjerivan kako bi se postigla što veća brzina pobiranja informacija. Određene preinake također su bile nužne slijedom uočenih problema u radu sustava koji su pak izazvani nestandardnom uporabom Web tehnologija na pojedinim Web poslužiteljima.

Posebna je pažnja posvećena detekciji različitih sinonima istog Web poslužitelja kako bi se izbjeglo višestruko mjerenje istih resursa.

#### 3.2. POČETNI UVJETI

Za početak mjerenja nužno je bilo osigurati odgovarajuće početno stanje baze podataka. U tu svrhu korišteni su podaci HR DNS službe, CROSS tražilice (<http://cross.carnet.hr>), Web kataloga [WWW.HR](http://www.hr/wwwhr/) (<http://www.hr/wwwhr/>) i CARNetovog caching sustava. Iz navedenih je izvora definiran popis aktivnih domena u .hr vršnoj domeni te kao rezultat posebne analize i početni popis Web poslužitelja kao i pojedinih resursa. Pri tome se posebno pazilo kako na konzistentnost tako i na ograničenja nabrojena u točki 1.2. ovog izvješća.

Tijekom mjerenja popis aktivnih domena nije mijenjan. Podaci o Web poslužiteljima i resursima nadopunjavani su sukladno podacima koje su prikupili gathereri. Omjer između novootkrivenih i obrađenih resursa iskorišten je kao kriterij zaustavljanja.

#### 3.3. PROVEDBA MJERENJA

Uz teoretski pretpostavljena ograničenja nabrojena u točki 1.2. samo je mjerenje bilo ograničeno još i:

- predviđenim vremenom,
- raspoloživim mrežnim i računalnim kapacitetima,
- nestandardnom uporabom Web tehnologija na pojedinim Web poslužiteljima,
- kriterijem zaustavljanja.

Mjerenje je provedeno u periodu **od 29.03. do 07.05.2002.** godine. Količina informacija prikupljenih u tom periodu nije ovisila samo o računalnim resursima SRCE-a nego i o stanju i dostupnosti pojedinih Web poslužitelja u trenutku kad su ih gathereri analizirali.

Nestandardna uporaba Web tehnologija također je uticala na rezultate mjerenja. Primjerice, raznoliki načini identifikacije korisnikove posjete (sesije) Web poslužitelju zahtjevali su dijelom prilagodbu programskog sustava, a dijelom i onemogućili potpunije pobiranje podataka.

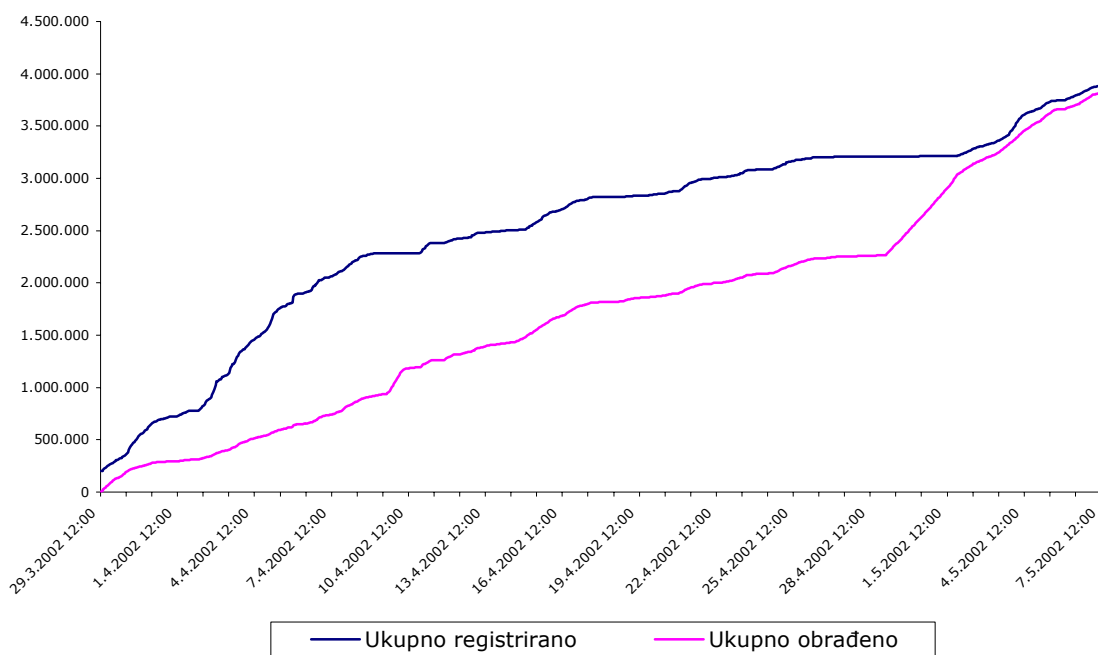
### 3.3.1. Kriterij zaustavljanja

Postupak pobiranja dokumenata s Web poslužitelja praktično može trajati vrlo dugo, a teoretski i beskrajno, pa zato nikada nije moguće tvrditi da je doista obrađen cjelokupni ciljani Web informacijski prostor. 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. Odlučeno je da se postupak prekine kada je na ispitnom uzorku procesirano više od **95%** registriranih resursa.

Ispitni uzorak sačinjavaju resursi s poslužitelja koji ne spadaju u jednu od posebnih kategorija koje se navode u točki 3.3.4.

U ispitni uzorak za zaustavljanje ušlo je 6.564 poslužitelja (vidjeti 2.1.1.) s kojih je evidentiran barem jedan uspješno obrađen resurs. Međutim, od tih 6.564 poslužitelja, 58 je uvršteno u kategoriju posebnih (vidjeti točku 3.3.4.), pa je konačno u ispitni uzorak za zaustavljanje mjerenja ušlo 6.506 poslužitelja. Ukupno je registrirano 6.006.105 resursa od kojih je na 58 ili 0,88% posebnih Web poslužitelja otpalo 2.115.124 resursa ili 35,22%. Odluka o zaustavljanju donesena je dakle na temelju informacija sa 99,12% poslužitelja koji su pokrivali 64,78% registriranih resursa u ispitnom uzorku. U trenutku zaustavljanja 98% resursa u tako definiranom ispitnom uzorku bilo je obrađeno. Na slijedećoj je slici dan grafički prikaz kretanja broja registriranih i obrađenih resursa u uzorku tijekom mjerenja.

**Broj registriranih i obrađenih resursa na ispitnom uzorku**



Valja naglasiti da je navedeni ispitni uzorak formiran samo za potrebe kriterija zaustavljanja. Ostali rezultati u ovom izvješću temelje se na svim prikupljenim podacima tijekom mjerenja.

### 3.3.2. Robot exclusion protocol

Ustanovljeno je da *robot exclusion protocol* rabi 1.476 poslužitelja (**22,49%** od 6.564). Udio registriranih resursa koji nisu obrađeni radi zabrane pristupa "robotima" u ukupnom broju registriranih resursa je **0,57%** ili brojčano 34.497 od 6.006.105. Određeni broj resursa, koji je dostupan samo s tih, za robote zabranjenih resursa, nije niti registriran, a smatramo da je nemoguće realno procijeniti o kolikom se broju resursa radi.

### 3.3.3. Deep Web

*Deep Web* (ili *Invisible Web*) je naziv koji se u pravilu rabi za resurse koji se generiraju isključivo kao rezultat upita, tj. interakcijom korisnika (najčešće) kroz HTML obrasce (forms). U najvećem broju slučajeva radi se o bazama podataka koje se pregledavaju putem upita koji sadržavaju ključne riječi ili termine.

Na resurse iz Deep Weba ne postoji niti jedan hiperlink, te oni ne mogu biti pronađeni konvencionalnim tehnikama pobiranja Web resursa, koje su primjenjene pri ovom mjerenju. Postoje procjene da je Deep Web globalno 400-550 puta veći od Weba dostupnog konvencionalnim tehnikama, ali da bi se ovakva tvrdnja mogla postaviti za hrvatski Web prostor, bilo bi potrebno koristiti drugačije sustave i dijelom manualno obrađivati pojedinačne slučajeve.

### 3.3.4. Posebne kategorije poslužitelja odnosno resursa

Za ovdje navedene kategorije resursa može se reći da "graniče s Deep Web-om", i to u smislu da se ti resursi otežano obrađuju konvencionalnim tehnikama.

Kao što je navedeno u točki 3.3.1., pri procjeni završenosti postupka pobiranja podataka ignorirali smo poslužitelje koji spadaju u ove kategorije, kojih je bilo 0,88% ali su sadržavali čak 35,22% registriranih resursa.

#### Poslužitelji s konačnim, ali vrlo velikim brojem dinamički generiranih resursa

Resurse na ovim poslužiteljima (ili jednom njihovom dijelu) generiraju Web aplikacije (to su tzv. dinamičke Web stranice). Primjeri ovakvih resursa su:

- veliki skupovi tekstualnih resursa, npr. arhive lista e-pošte, forumi, i sl.
- baze podataka, odnosno Web sučelja za (najčešće) relacijske baze podataka
- Web sučelja za druge informacijske servise, npr. mrežne novine (network news, USENET), imeničke servise i dr.

Zajedničke karakteristike su:

- potencijalno goleme količine podataka i Web stranica, a naročito jer se podaci iz baze najčešće mogu prikazati na više načina kako bi bili pregledniji posjetitelju, te podijeljeni u stranice od 10-ak zapisa, kako bi se pojedine Web stranice brže učitavale
- sporiji odziv nego kod statičkih Web stranica

Ograničavajući faktor pri pobiranju ovih resursa je brzina. Gotovo nemoguće je obraditi sve resurse u realnom vremenu. Nedovoljna brzina može biti posljedica brzine mrežne veze ili brzine poslužitelja (tj. Web aplikacije i/ili baze podataka).

#### Poslužitelji s beskonačnim brojem resursa

Pojedine Web aplikacije, uslijed programske arhitekture ili pogrešaka, generiraju beskrajno veliki broj URL-ova, koji striktno uzevši, predstavljaju beskonačni izvor resursa.

#### Praćenje aktivnosti korisnika putem URL-a

Karakteristika je Web aplikacija koje generiraju resurse na ovim poslužiteljima da kao dio URL-a (najčešće dio *query string*-a) prenose informacije koje koriste za praćenje (*tracking*) aktivnosti posjetitelja odnosno definiranja jedne posjete (*session*). Kako je URL promjenjiv sa svakim HTTP zahtjevom (ili pak svakom "posjetom") striktno uzevši, poslužitelj generira beskrajno mnogo resursa.

Za identifikaciju ovakvih resursa potreban je manualni nadzor i upravljanje postupkom pobiranja, pa ovakvi poslužitelji stvaraju dodatni napor i zato je njihova obrada sporija ili u nekim slučajevim i nemoguća. Resursi koji su informacijski jednaki, ali se razlikuju u dijelu URL-a koji služi za praćenje aktivnosti posjetitelja, naknadnom obradom su uklonjeni i ne prikazuju se u rezultatima ovog mjerenja.

## URL petlja

URL petljom (*loop*) se naziva pogreška u izvedbi Web aplikacije koja generira dinamičke WWW stranice, pri čemu se nenamjerno pogreškom više puta ponavlja pojedini segment URL-a. Svakim novim dohvatom ovakvog resursa, on proizvodi hiperlink na samog sebe, ali se u hiperlinku pogrešno nalazi po jedan dodatni segment URL-a (atribut-vrijednost par iz *query string*-a ili folder u *path info*).

Sustav je djelomično imun na "URL petlje", ali su svejedno identificirani ovakvi resursi, te su naknadnom obradom uklonjeni i ne prikazuju se u rezultatima mjerenja.

## Restriktivna politika

Prilikom procesiranja resursa s pojedinih poslužitelja registrirane su neke restriktivne politike.

Nekoliko poslužitelja konfigurirano je tako da ne poslužuju dokumente HEAD HTTP metodom, radi čega se nije mogla točno ustanoviti veličina inače dostupnih resursa. Ova se metoda koristi za procjenu veličine resursa svih tipova osim HTML-a.

## 4. ZAKLJUČAK

Mjerenje hrvatskog Web prostora, preciznije HTTP protokolom dostupnih resursa u .hr vršnoj Internet domeni, pripremljeno je i provedeno u relativno kratkom roku. Kao i sva ostala mjerenja tog tipa i ovo je mjerenje bilo podložno određenim ograničenjima.

Posebni su izazov bile inventivne, ali nestandardne uporabe Web tehnologija koje onemogućuju rad standardnih mehanizama za pobiranje podataka.

Dobiveni rezultati odgovaraju našim očekivanjima, ali i rezultatima sličnih istraživanja provedenih u svijetu. Ovo se posebno odnosi na procjene veličine i složenosti prostora. Dobiveni su očekivani rezultati vezani uz tipove resursa koji potvrđuju tezu o relativnom malom broju različitih formata koji se rabe na Webu. Analiza metapodataka također nije ponudila neočekivane rezultate, makar brojnost grešaka ukazuje na nebrigu autora za taj segment Web tehnologija.

Kao poseban rezultat mjerenja valja istaći procjenu da uz razvoj kvalitetnih alata za ciljano pobiranje Web resursa nužno treba prirediti i preporuke za autore i izdavače kako bi njihovi resursi bili spremniji za pobiranje.

Prikupljeni podaci svakako su osnov za daljnje, ozbiljnije analize i istraživanja hrvatskog Web prostora.