O mjerenju hrvatskog prostora weba

Miroslav Milinović
Sveučilišni računski centar (Srce)
miro@srce.hr

Zagreb, 18. svibnja 2004.

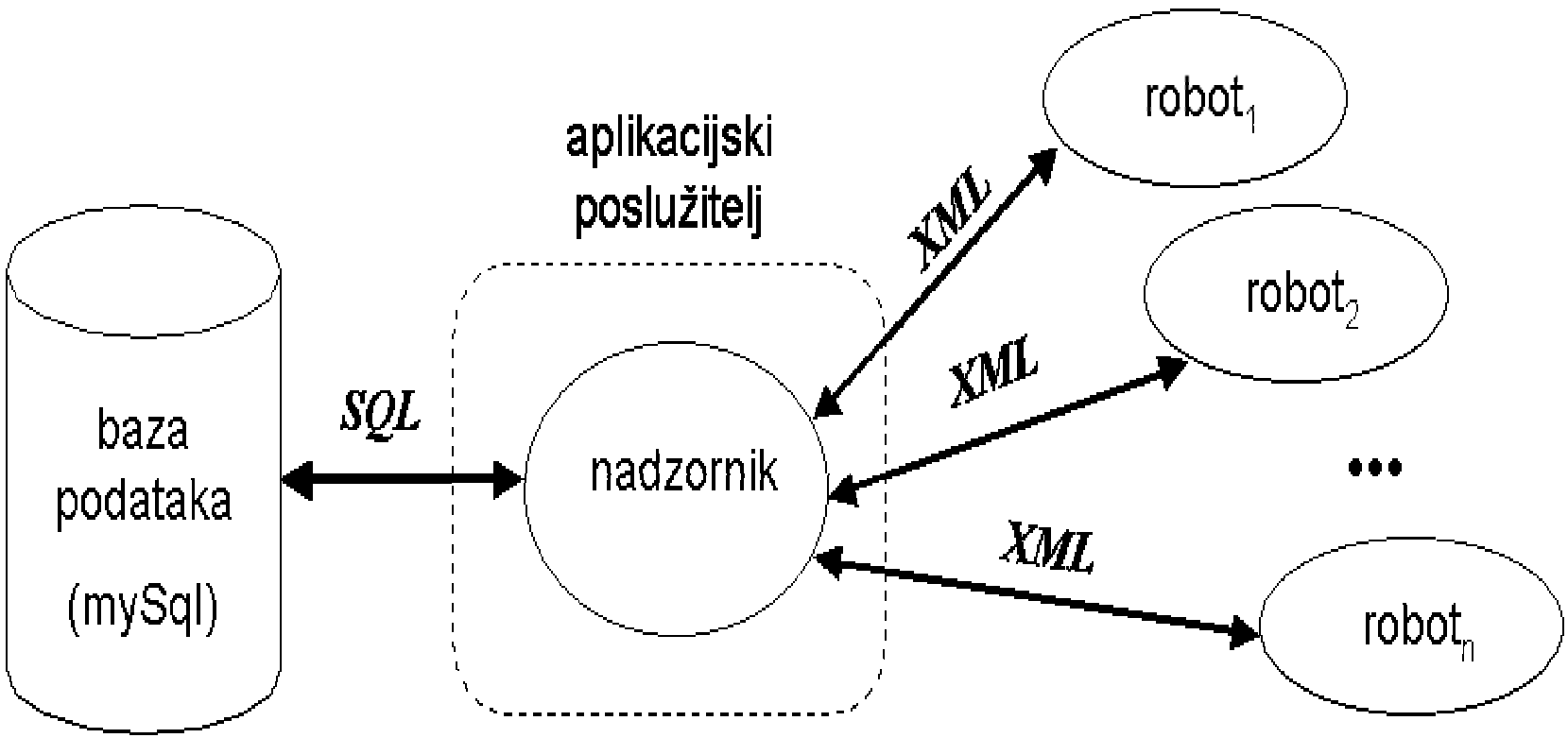
Sadržaj

- O mjerenju (što i kako ...)
- O veličini hrvatskog weba
- Tipovi / formati podataka
- O metapodacima
- Web poslužitelji (programska podrška)
- Usporedba mjerenja (MWP1 vs. MWP2 vs. MWP3)
- O mjerljivosti weba (površinski vs. dubinski web)
- Zaključak

Što smo i kako mjerili?

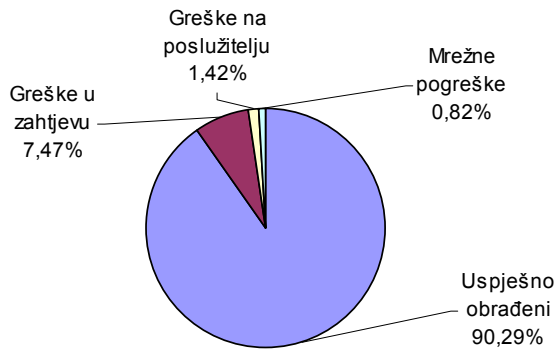
- MWP projekt započet je 2002. godine:
  - u 2002. - suradnja s NSK-om; u 2003. - i-projekt MZT
  - MWP sustav: razvijen 2002., usavršen tijekom 2003.
  - http://www.srce.hr/mwp/
- Provedena mjerenja:
  - MWP1: 29.03.-07.05.2002.
  - MWP2: 14.05.-22.07.2003.
  - MWP3: 08.09.-25.11.2003.
- Predmet mjerenja:
  - 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 metapodataka

Sustav MWP

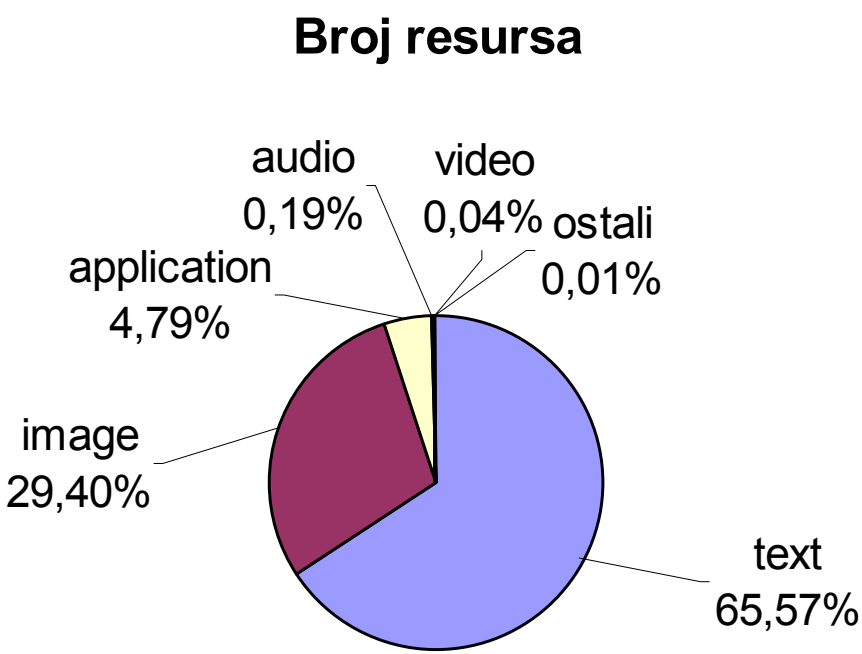

aplikacijski
poslužitelj
baza
podataka
(mySql)
SQL
nadzornik
XML
XML
XML
robot1
robot2
...
robotn

MWP3: veličina hrvatskog Weba

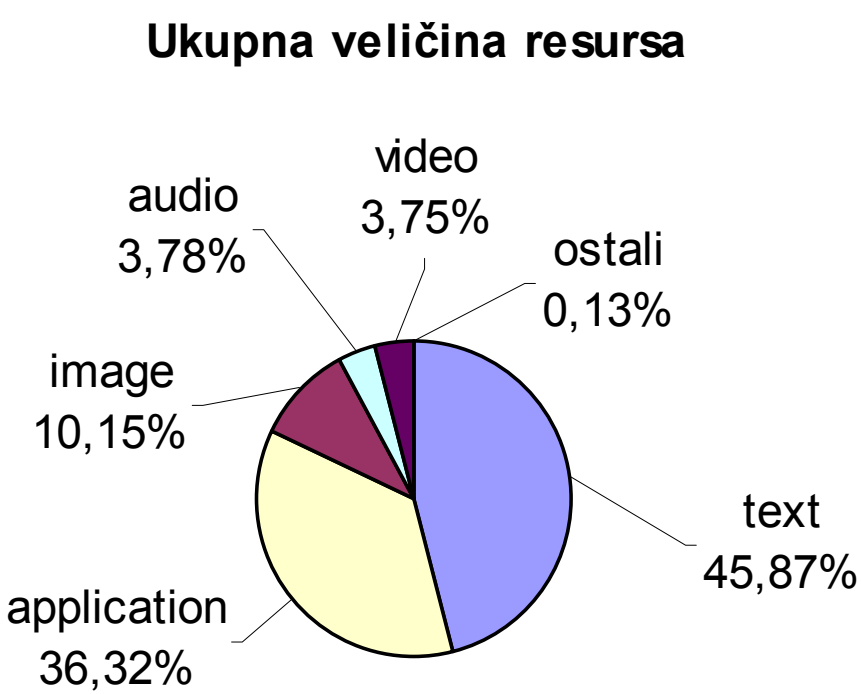
- **10.884 poslužitelja**
- 14.450.240 (10.347.959) resursa od kojih je 7.125.879 obrađeno
- 90,3% resursa je uspješno obrađeno (6.433.902)
- za 5.433.598 uspješno obrađenih resursa izmjerena je ukupna veličina od 269 GB
- **ukupna veličina uzorka je procjenjena na 548 GB**



MWP3: udio osnovnih tipova (1)



MWP3: udio osnovnih tipova (2)



MWP3: najčešći tipovi (1)

MIME tip	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%
ostali	84.259	441.351	1,55%

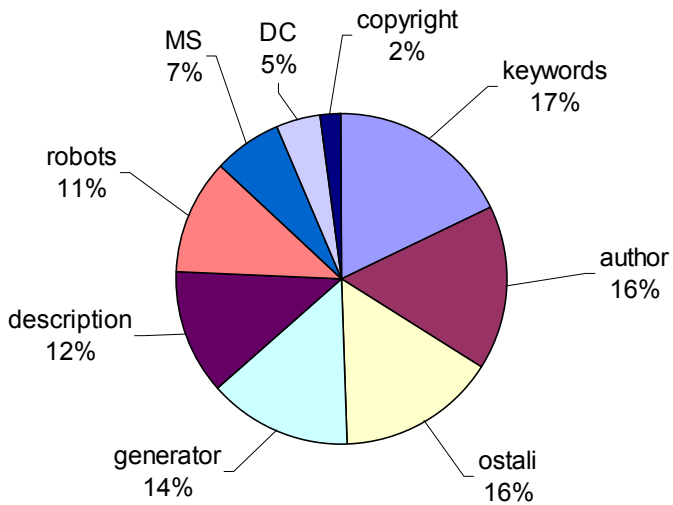


MWP3: najčešći tipovi (2)

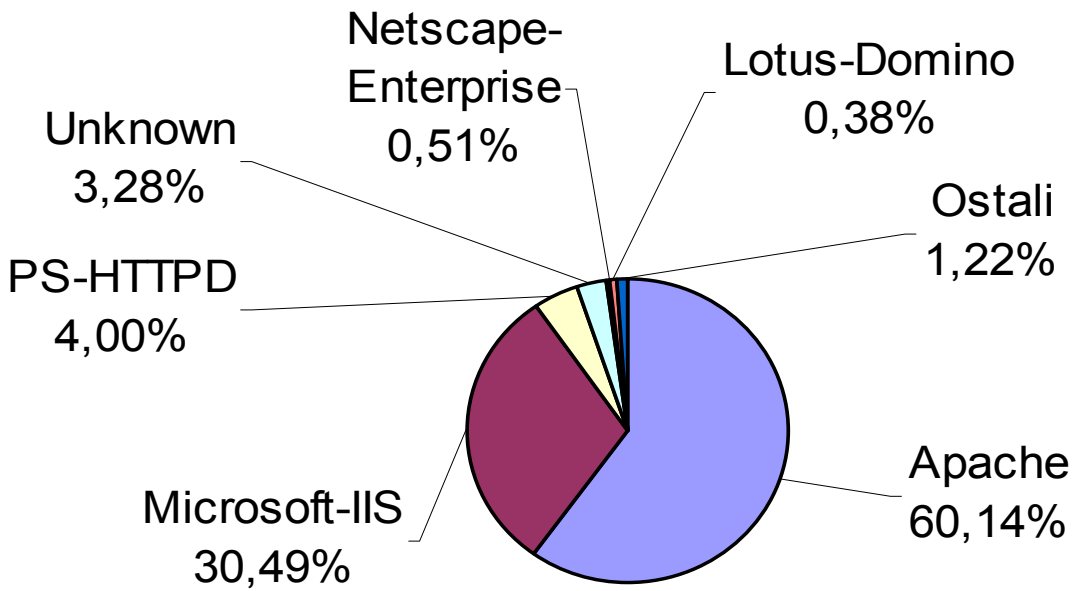
MIME tip	Sumarna veličina	Prosječna veličina	Udio
text/html	92.330.334.777	28.902	31,99%
application/x-tar	62.318.751.973	392.911	21,59%
text/plain	39.841.243.898	124.035	13,81%
image/jpeg	24.238.281.421	24.603	8,40%
application/pdf	15.448.075.009	420.470	5,35%
application/zip	10.713.258.978	694.089	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%
ostali	25.306.381.151	169.329	8,77%

MWP3: metapodaci

- 41% HTML resursa (57,42% poslužitelja) ima META oznaku
- 666 različitih vrijednosti NAME atributa META oznake
- nebriga autora (smanjena u odnosu na MWP1)



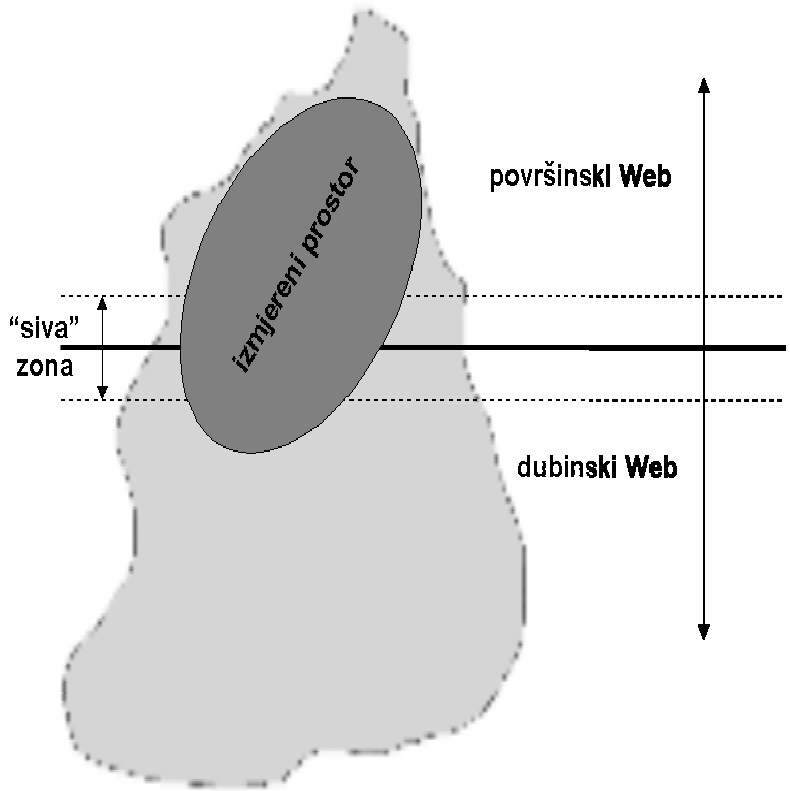
MWP3: web poslužitelji



- (MWP1: Apache 56,73%; MS IIS 29,08%, ...)

MWP1 vs. MWP3

- MWP1 vs. MWP3:
  - MWP1 (2002.)
    - “HTTP resursi”
    - ≈ 320 GB (389 GB)
    - 5.145.383 obrađena resursa
  - MWP3 (2003.)
    - “HTTP/HTTPS resursi”
    - ≈ 548 GB
    - 7.125.879 obrađenih resursa



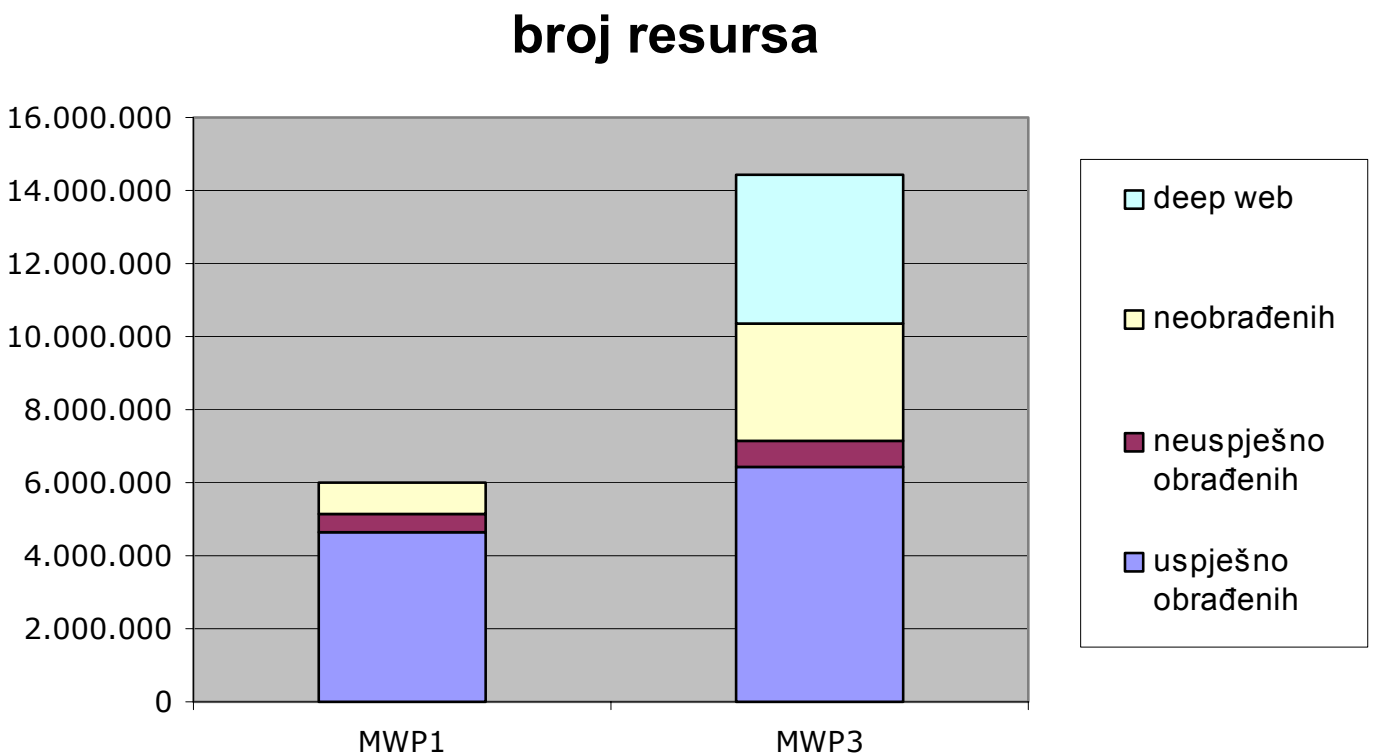
MWP1 vs. MWP2 vs. MWP3 (1)

broj domena	MWP1	MWP3	Indeks
Broj domena s uspješno obrađenim resursima	4.509	8.202	181,90
Ukupni broj domena	9.315	15.305	164,30
Udio domena s uspješno obrađenim resursima	48,41%	53,59%	

broj poslužitelja	**MWP1**	**MWP3**	**Indeks**
S uspješno obrađenim resursima	6.565	10.884	165,79
Kontaktirani poslužitelji	13.382	18.555	138,66
Ukupni broj poslužitelja	14.133	22.554	159,58



MWP1 vs. MWP2 vs. MWP3 (2)



MWP1 vs. MWP2 vs. MWP3 (3)

MIME type	broj resursa		veličina resursa	
	MWP1	MWP3	MWP1	MWP3
text	69,81%	65,57%	24,52%	45,87%
image	22,98%	29,40%	5,93%	10,15%
application	6,66%	4,79%	62,47%	36,32%
audio	0,49%	0,19%	4,58%	3,78%
video	0,03%	0,04%	2,41%	3,75%
ostali	0,03%	0,01%	0,09%	0,13%



MWP1 vs. MWP2 vs. MWP3 (4)

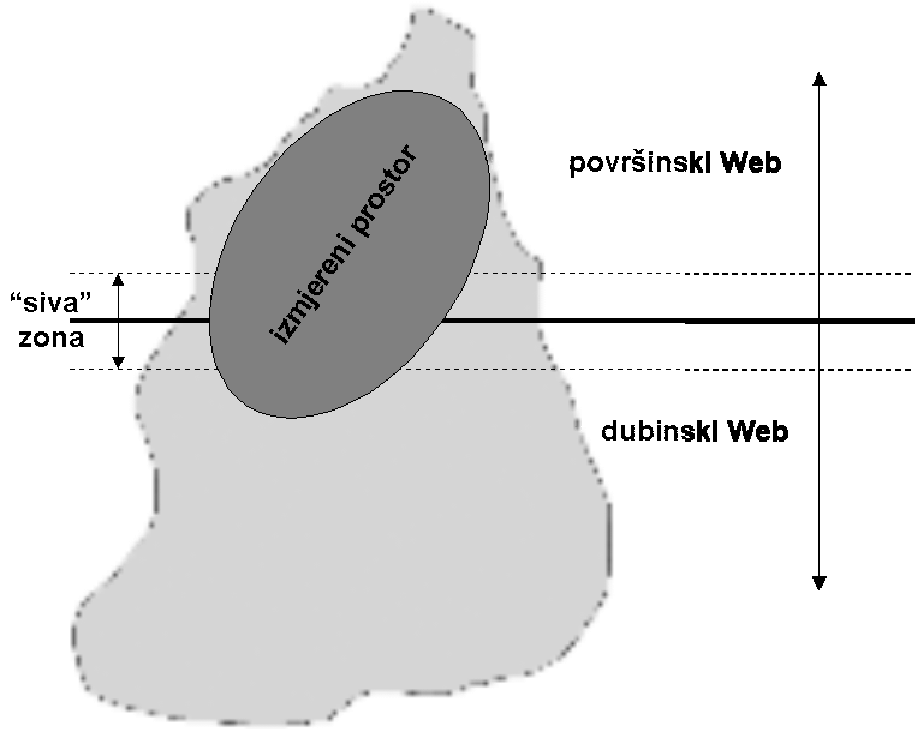
- Meta oznaka:
  - MWP1: 31% HTML resursa / 53,9% poslužitelja
  - MWP2: 43% HTML resursa / 56,5% poslužitelja
  - MWP3: 41% HTML resursa / 57,42% poslužitelja
- broj različitih vrijednosti NAME atributa META oznake
  - MWP1: 743 / MWP2: 645 / MWP3: 666
- udio standarda:
  - MWP1:
    - Dublin Core – 0,09%
    - tražilice – 19,7%
  - MWP2:
    - Dublin Core – 2,31%
    - tražilice – 14,62%



MWP vs. svijet

- Web je "*small and simple*":
  - MWP1(2002.):
    - 320 GB i ≈ 6 milijuna resursa
    - 5 MIME tipova čini više od 90% obrađenih resursa
  - švedski Web (Hakala, 1999.):
    - 300 GB i ≈ 7,5 milijuna resursa
    - 4 MIME tipa pokrivaju 97% resursa
- Uporaba metapodataka:
  - Lawrence & Giles (1999.):
    - 34% Web resursa ima META oznaku / 0,3% rabi DC
  - MWP:
    - MWP1: 31% Web resursa ima META oznaku / 0,09% rabi DC
    - MWP2: 43,1% Web resursa ima META oznaku / 2,31% rabi DC

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.)

Zaključak

- rezultati odgovaraju očekivanjima i sličnim istraživanjima provedenim u svijetu
- (površinski) Web je i dalje jednostavan: rabimo mali broj različitih formata
- autori ne brinu dovoljno o metapodacima
- dinamički web, inventivne, ali i nestandardne uporabe Web tehnologija čine mjerenje sve složenijim

- **Srce će nastaviti mjeriti hrvatski prostor weba (jednom godišnje)**
- **Pozivamo zainteresirane na suradnju ...**

http://www.srce.hr/mwp/
mwp@srce.hr