MWP Project: Measuring the Growth of Croatian Web Space

Miroslav Milinović, Dubravko Penzić, SRCE
Hrvoje Stipetić, Zagreb Fair
Nebojša Topolšćak, SRCE
mwp@srce.hr

Zagreb, September 2004.

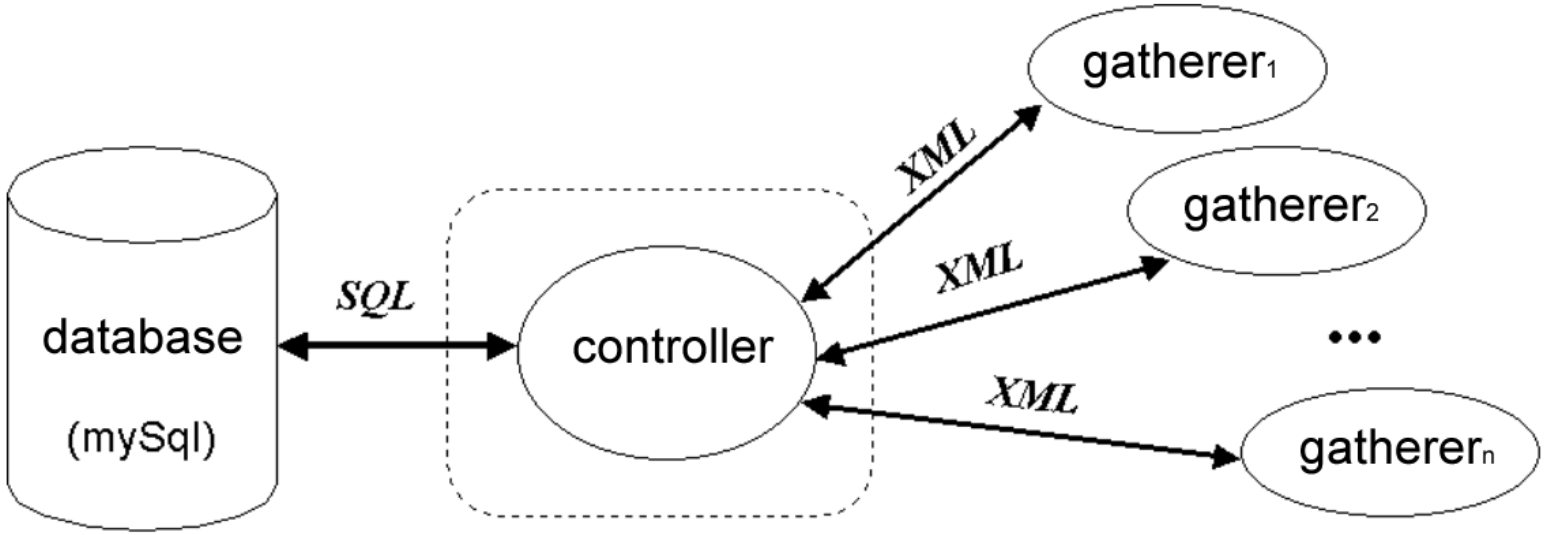
Content

- About measurement
- Size of Croatian Web space
- Data types and Formats
- About metadata
- Web servers (software)
- Comparison of measurements (MWP1-MWP2-MWP3)
- Measurability of Web space (surface vs. deep web)
- Conclusion

About measurements

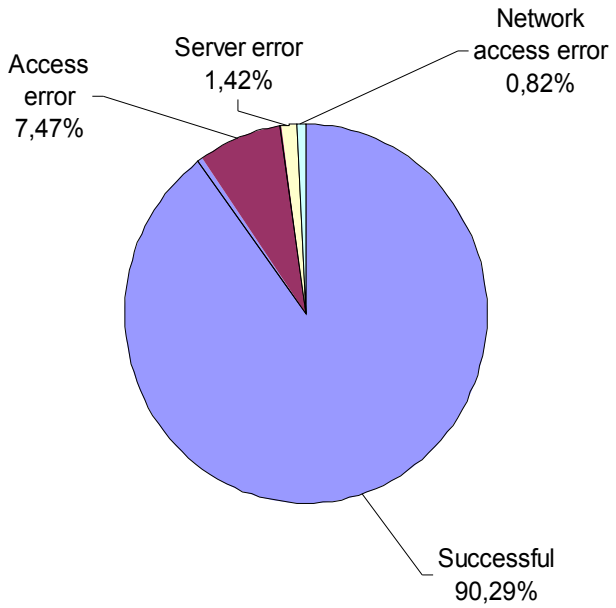
- MWP project was started in 2002.
  - 2002. – cooperation with National and University Library;
  - 2003. - I-project financed by Ministry of Science, Education and Sports
  - MWP system: was developed in 2002, improved during 2003.
  - http://www.srce.hr/mwp/
- Measurements:
  - MWP1: 29.03.-07.05.2002.
  - MWP2: 14.05.-22.07.2003.
  - MWP3: 08.09.-25.11.2003.
- Scope:
  - all resources available via HTTP/HTTPS protocols from servers in .hr top level domain
- We measured:
  - size
  - used data formats (according to MIME standard)
  - volume and content of metadata

MWP System

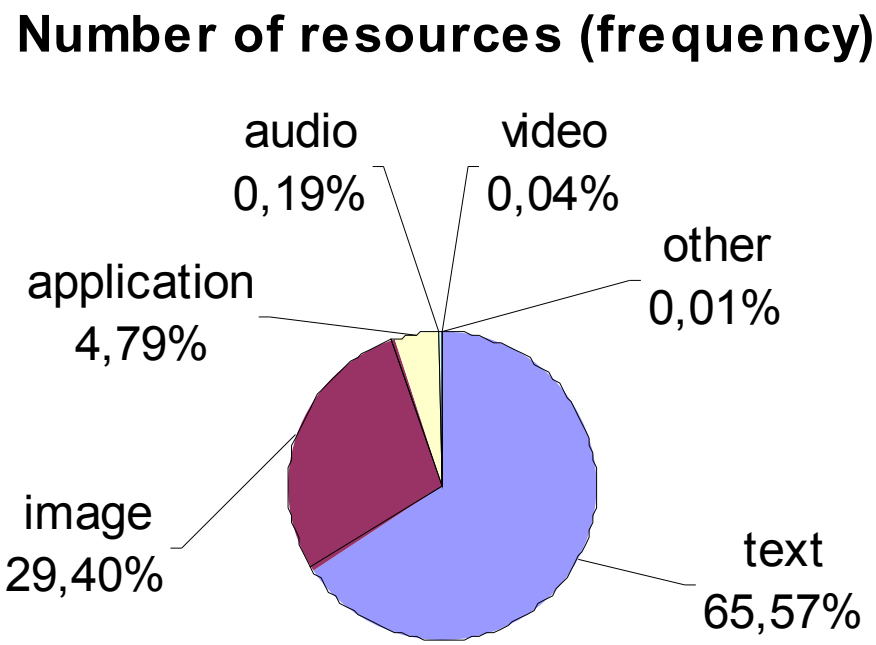

gatherer1
XML
gatherer2
XML
database
(mySql)
SQL
controller
...
XML
gatherern

MWP3: The Size of Croatian Web

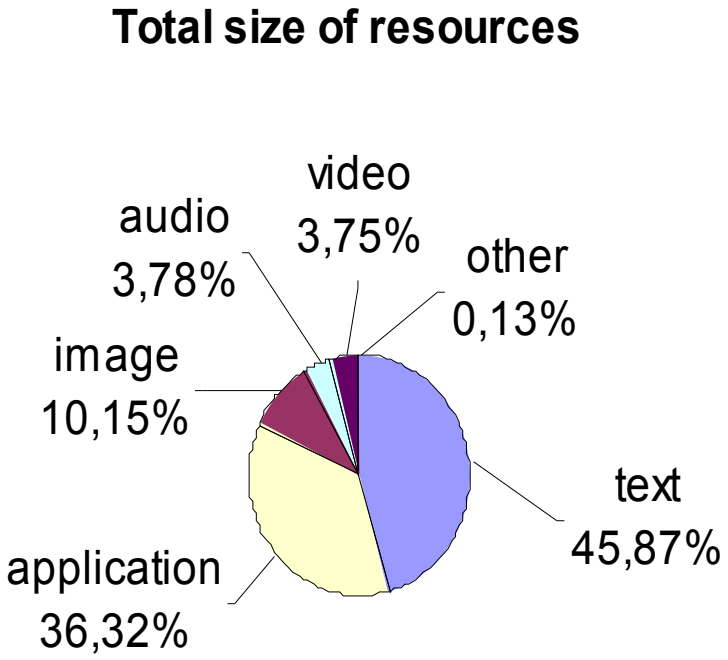
- **10.884 servers**
- 14.450.240 (10.347.959) resources 7.125.879 processed
- 90,3% resources were successfully gathered (6.433.902)
- The total size of 5.433.598 successfully gathered resources amount to 269 GB
- **The size of the measured Web was estimated to be 548 GB**



MWP3: Distribution of data types (1)



MWP3: Distribution of data types (2)



MWP3: Most common types (1)

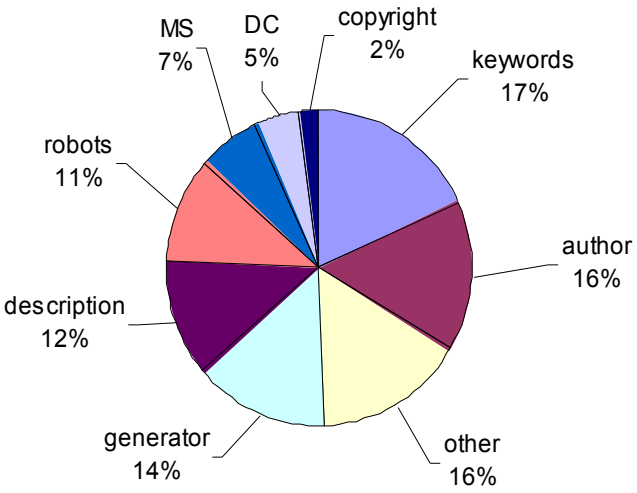
MIME type	Number of resources	Average size	Percentage
text/html	3.194.548	28.902	58,80%
image/jpeg	985.171	24.603	18,13%
image/gif	562.218	6.964	10,35%
text/plain	321.210	124.035	5,91%
application/x-tar	158.608	392.911	2,92%
image/png	47.069	14.991	0,87%
application/pdf	36.740	420.470	0,68%
application/zip	15.435	694.089	0,28%
text/x-chdr	14.985	5.065	0,28%
application/msword	13.022	139.903	0,24%
other	84.259	441.351	1,55%

MWP3: Most common types (2)

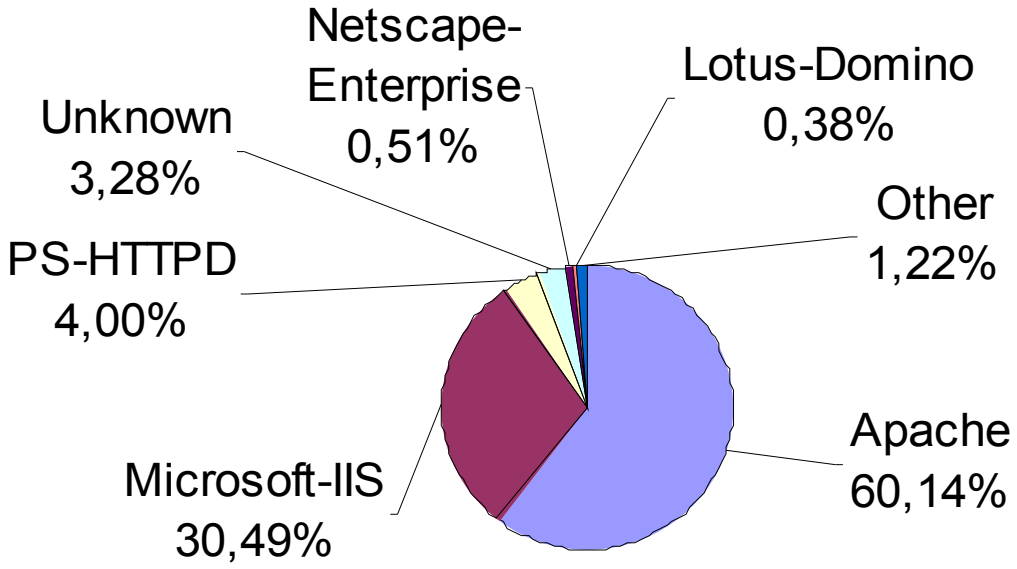
MIME type	Total size	Average size	Percentage
text/html	92.330.334.777	28.902	31,99%
application/x-tar	62.318.751.973	392.911	21,59%
text/plain	39.841.243.898	124.035	13,81%
image/jpeg	24.238.281.421	24.603	8,40%
application/pdf	15.448.075.009	420.470	5,35%
application/zip	10.713.258.978	694.089	3,71%
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%
other	25.306.381.151	169.329	8,77%

MWP3: metadata

- 41% HTML files (57,42% servers) have META tag
- 666 distinct values of NAME attribute in META tag
- Authors still don't take enough care about metadata (but situation is better then in MWP1)



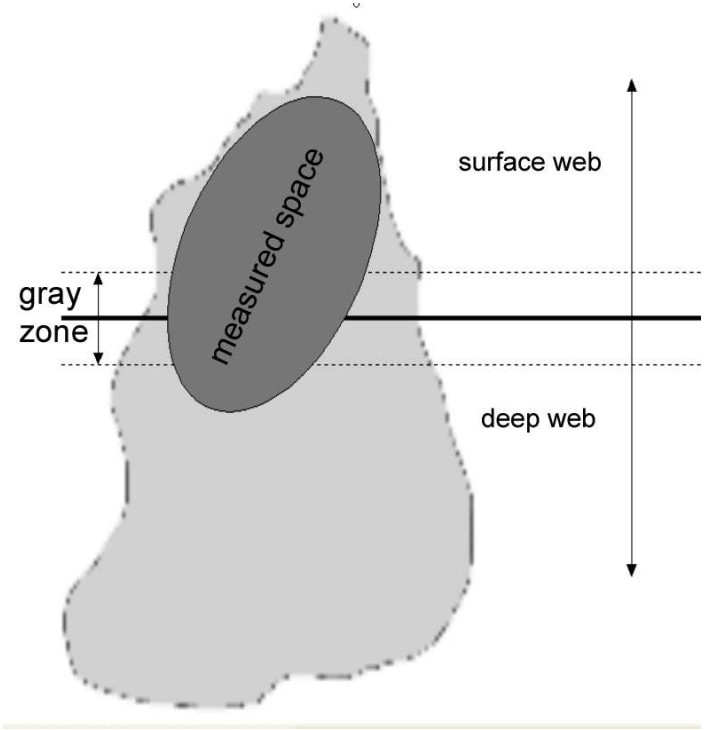
MWP3: Web servers



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

MWP1 vs. MWP3

- MWP1 vs. MWP3:
  - MWP1 (2002.)
    - HTTP resources
    - ∀ ≈ 320 GB (389 GB)
    - 5.145.383 processed resources
  - MWP3 (2003.)
    - HTTP/HTTPS resources
    - ∀ ≈ 548 GB
    - 7.125.879 processed resources

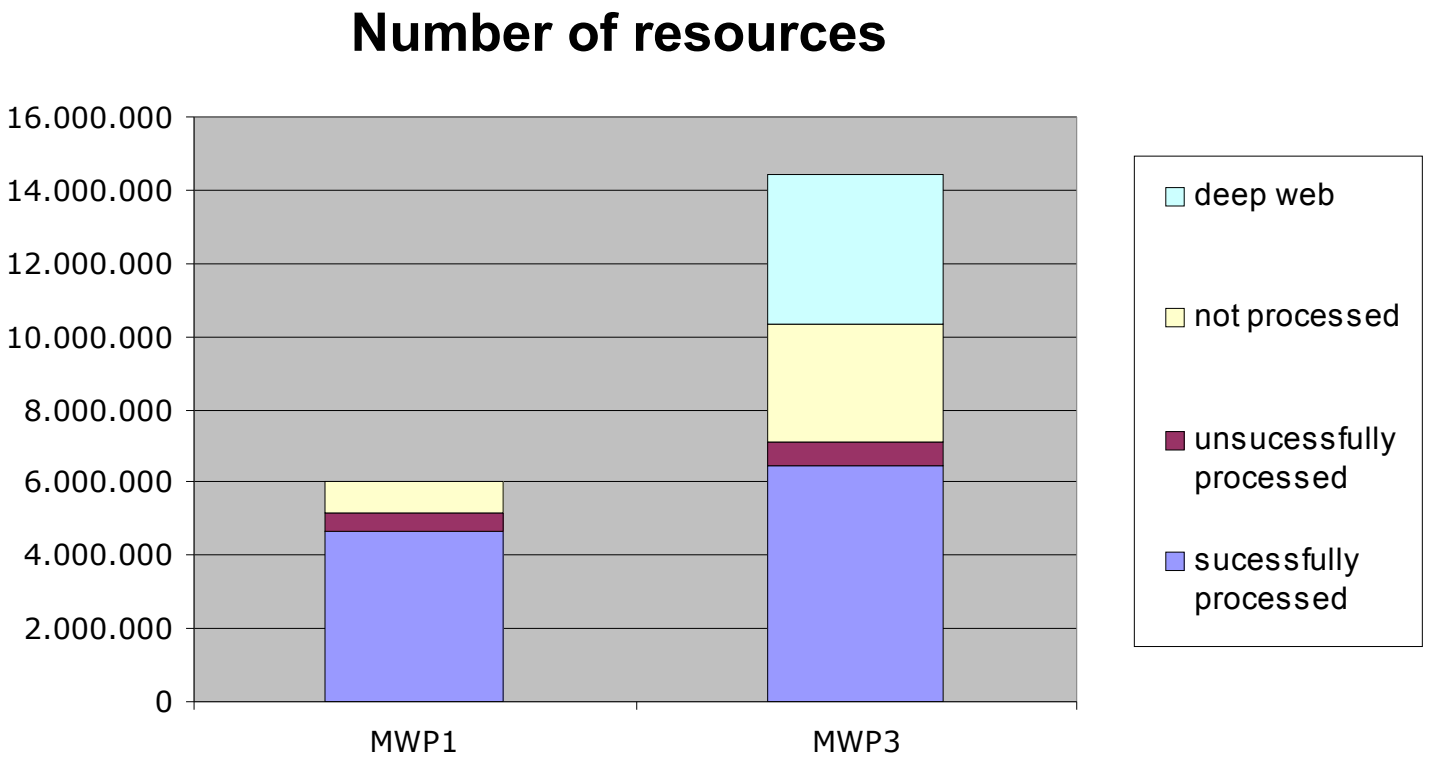


MWP1 vs. MWP2 vs. MWP3 (1)

Number of domains	MWP1	MWP3	Index
Number of domains with successfully processed resources	4.509	8.202	181,90
Total number of domains	9.315	15.305	164,30
Percentage of domains with successfully processed resources	48,41%	53,59%	

Number of servers	**MWP1**	**MWP3**	**Index**
With successfully processed resources	6.565	10.884	165,79
Accessed servers	13.382	18.555	138,66
Total number of servers	14.133	22.554	159,58

MWP1 vs. MWP2 vs. MWP3 (2)



MWP1 vs. MWP2 vs. MWP3 (3)

MIME type	Number of resources		Size of resources	
	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%
other	0,03%	0,01%	0,09%	0,13%

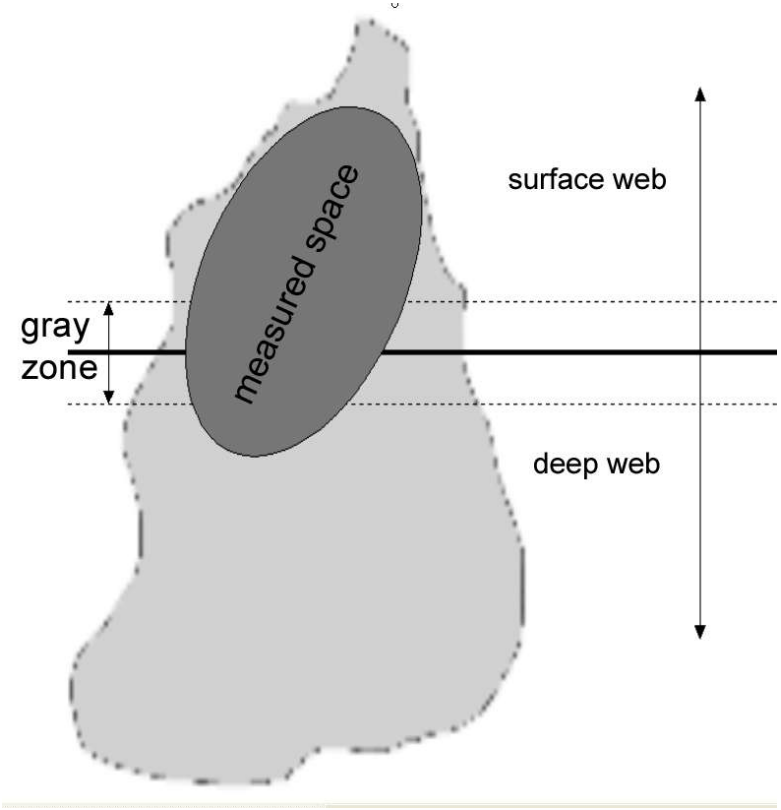
MWP1 vs. MWP2 vs. MWP3 (4)

- Meta tags:
  - MWP1: 31% HTML resources / 53,9% servers
  - MWP2: 43% HTML resources / 56,5% servers
  - MWP3: 41% HTML resources / 57,42% servers
- Number of distinct values of NAME attribute in META tag
  - MWP1: 743 / MWP2: 645 / MWP3: 666
- Percentage of different standards:
  - MWP1:
    - Dublin Core – 0,09%
    - Search engines – 19,7%
  - MWP2:
    - Dublin Core – 2,31%
    - Search engines – 14,62%

MWP vs. world

- Web is "*small and simple*":
  - MWP1(2002.):
    - 320 GB and ≈ 6 millions of resources
    - 5 MIME types cover more than 90% processed resources
  - Swedish Web (Hakala, 1999.):
    - 300 GB and ≈ 7,5 millions of resources
    - 4 MIME types cover 97% of resources
- Usage of metadata:
  - Lawrence & Giles (1999.):
    - 34% of Web resources have META tag / 0,3% use DC
  - MWP:
    - MWP1: 31% of Web resources have META tag / 0,09% use DC
    - MWP2: 43,1% of Web resources have META tag / 2,31% use DC

About measurability of Web



There is an estimation that deep Web is 400-550 times bigger than surface Web
(Bergman, Michael K. The deep Web: Surfacing Hidden Value. White Paper. The Journal of Electronic Publishing, University of Michigan, July 2001.)

Conclusion

- Results meet our expectations and correspond to similar surveys in the world
- (surface) Web is still simple: we use small number of different formats
- Authors (still) don't take enough care about metadata
- Dynamic web, inventive but non-standard use of web technologies make gathering of data more and more difficult

- **SRCE will continue with measurements of Croatian web**

  (once per year; MWP4 will start in October 2004)
- **Your cooperation is wellcome ...**

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

mwp@srce.hr