Measurement of Croatian Web Space: Preliminary Results

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

Zagreb, September 2002

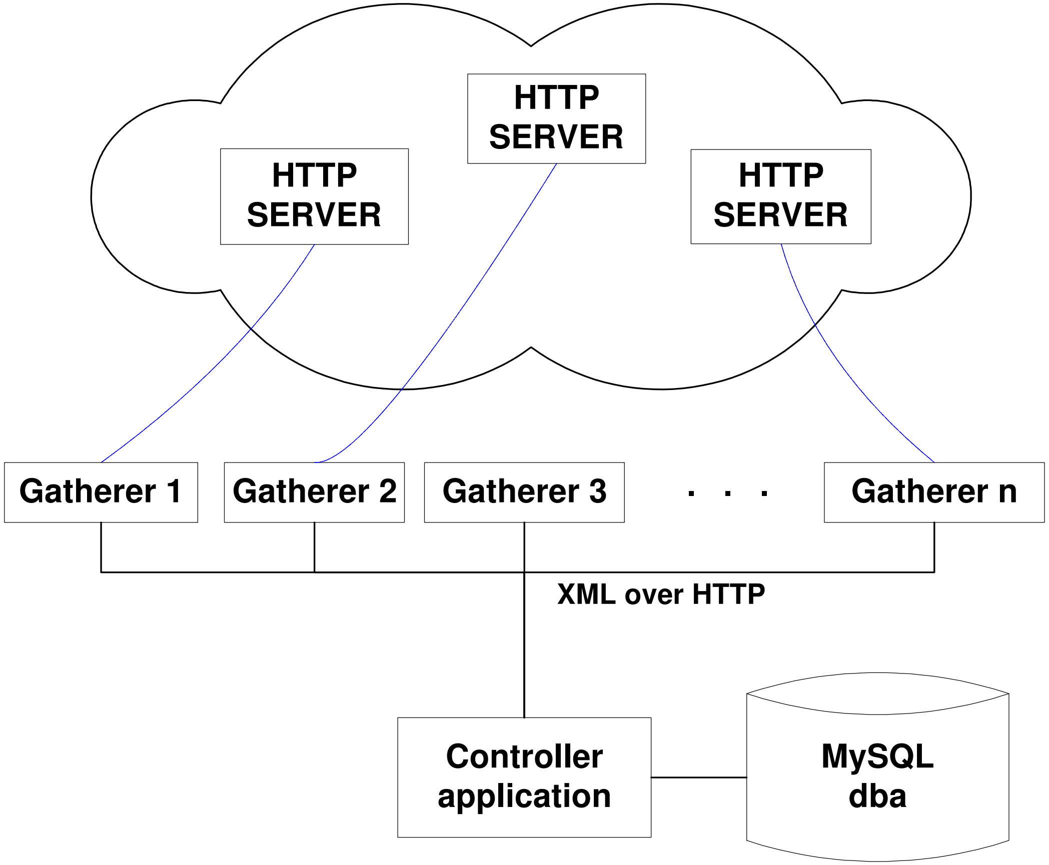
Content

- Goals and methods
- About the size of the Croatian Web space
- About content types (MIME types)
- About metadata
- Other interesting results
- Similar surveys in the world
- Conclusion

Scope & goals

- Scope:
  - all resources available via HTTP protocol from servers in .hr domain
- Main goals:
  - measure/estimate the size
  - examine diversity of used data formats
  - examine metadata
- Measurement carried out:
  - MWP project
  - with specially designed and developed system
  - March, 27th - May, 7th 2002

Architecture of MWP application

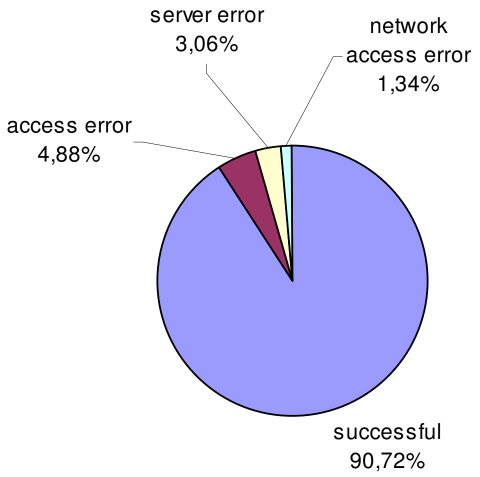


The size of Croatian Web

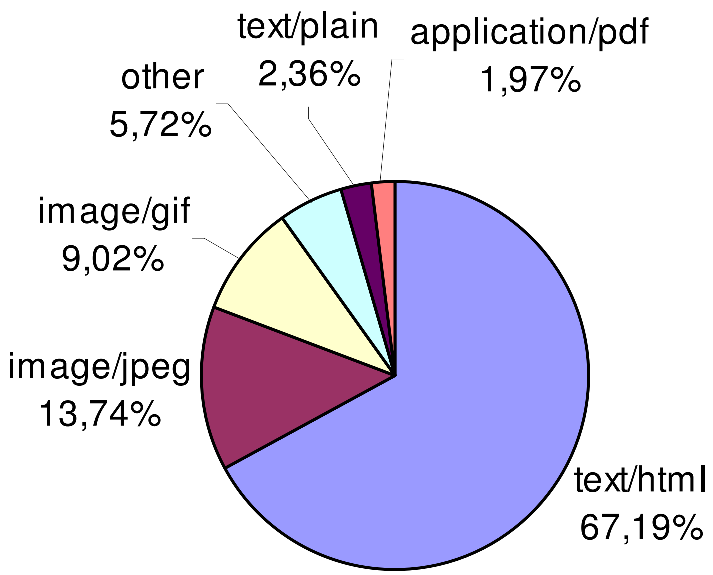
- **6.564 servers**

- 6.006.105 resources
- 5.145.383 have been processed
- 4.667.920 (91%) successfully processed
- The size of 79% of successfully processed resources was found to be 263.4 GB
- The rest was estimated to be 55,3 GB

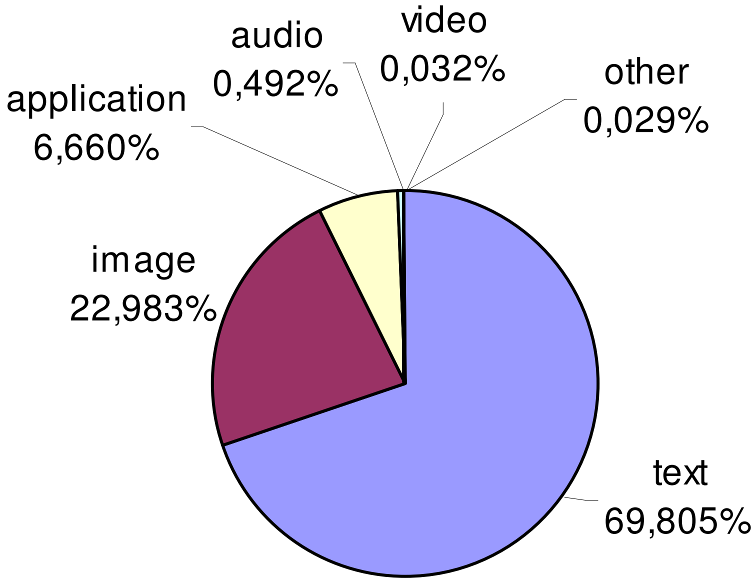
- **The size of the sample was estimated to be 318,7 GB**



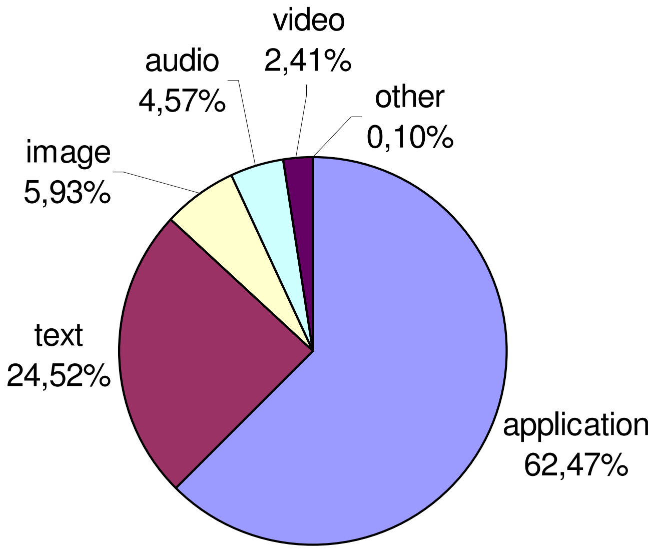
Most common content types (frequency)



Distribution of data types (frequency)



Distribution of content types (size)



content type	%
application/octet-stream	16,29%
text/html	14,93%
application/pdf	12,06%
text/plain	9,58%
application/x-tar	9,40%

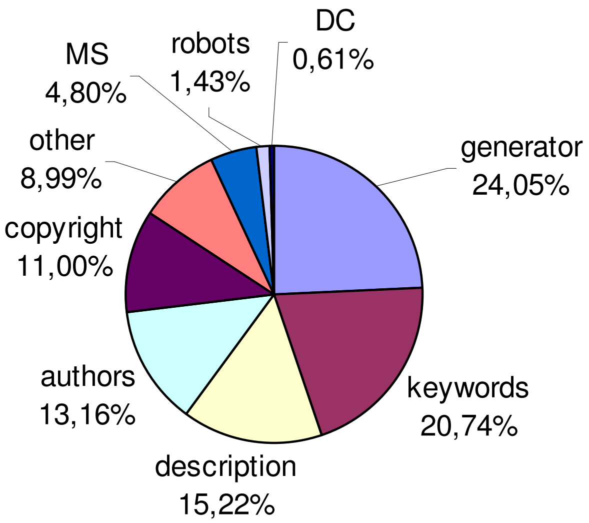
Size compared with frequency

content type	no. of resources		size		
	%	rank	%	rank	average (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

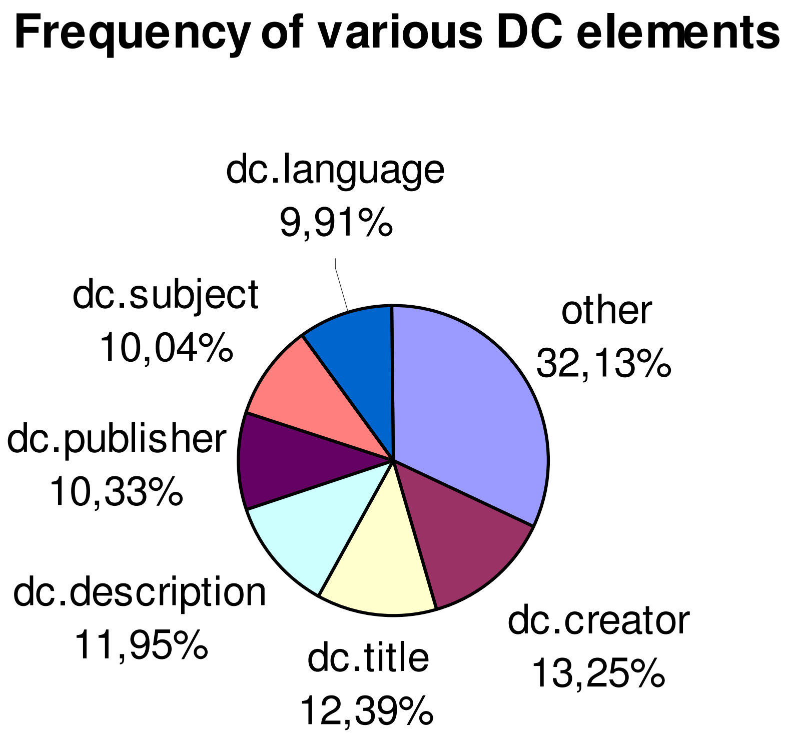
type	no. of resources		size		
	%	rank	%	rank	average (KB)
application	6,660%	3	62,467%	1	599,56
text	69,805%	1	24,523%	2	22,45
image	22,983%	2	5,929%	3	16,49
audio	0,492%	4	4,575%	4	594,08
video	0,032%	5	2,410%	5	4885,74

Metadata

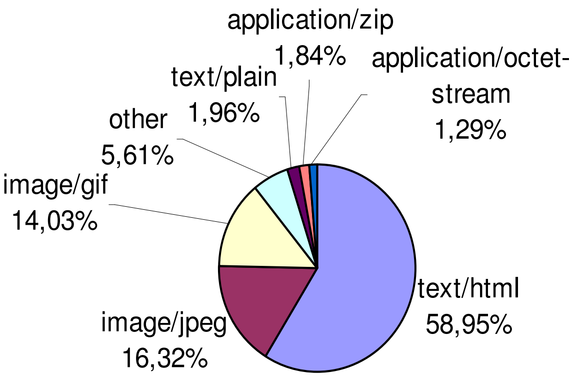
- 31% of HTML files have META tag
- 744 distinct values of NAME attribute in META tag
- Distribution of “standards”:
  - Dublin Core – 0,09%
  - HTML editors – 25%
  - Search engines – 19,7%
  - ROBOTS META tag – 1,35%



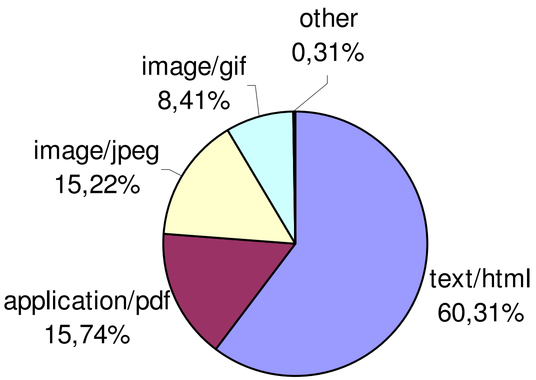
Dublin Core


Frequency of various DC elements
dc.language 9,91%
other 32,13%
dc.subject 10,04%
dc.publisher 10,33%
dc.description 11,95%
dc.title 12,39%
dc.creator 13,25%

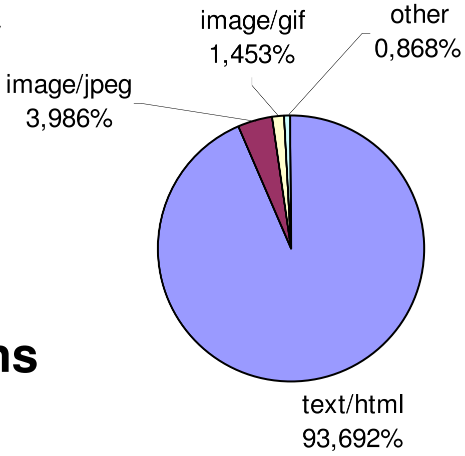
Academic community, publishers and e-publications



Academic community

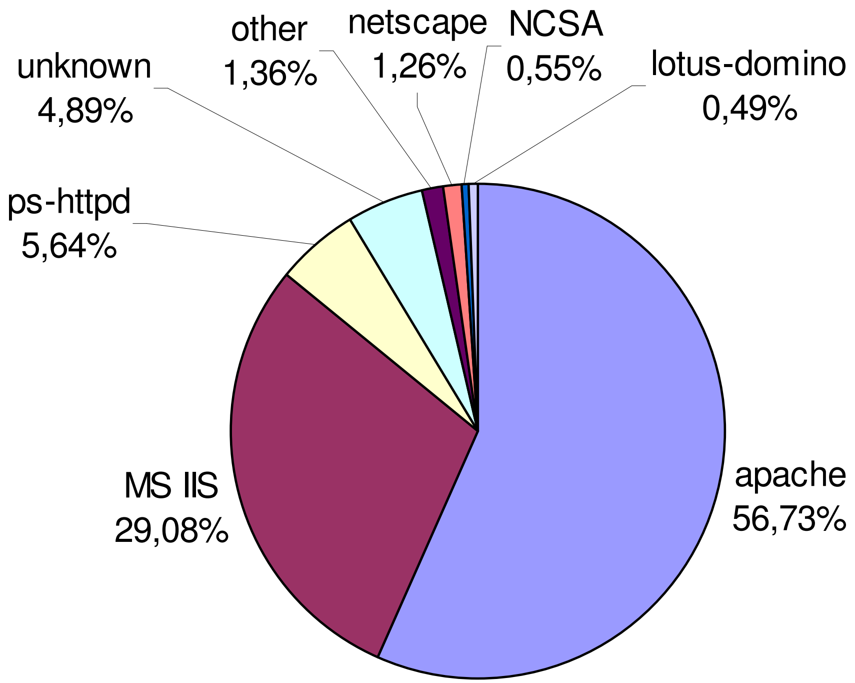


Publishers



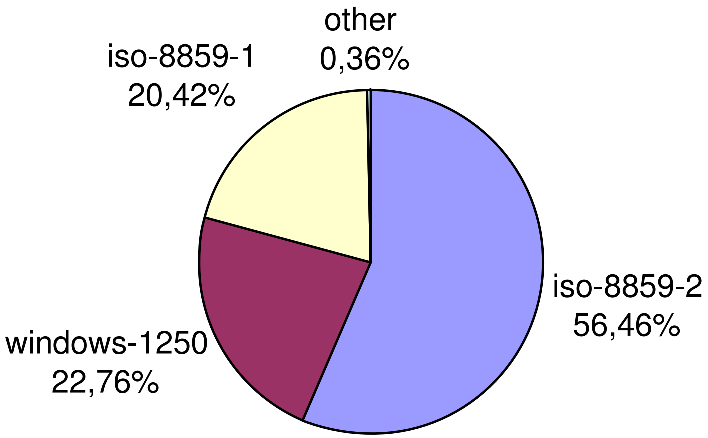
E-publications

Web servers



Robots exclusion protocol **is used by 1.476 (22,49%) servers**

Usage of various charater sets (code pages)



Only 401.682 resources are using explicitly defined standard

Dynamic content

- Scripting languages:
  - 3.296 servers, 929.816 are using some scripting language (98% JavaScript)
- Java applets:
  - 576 servers, 10.202 resources
- Cookies:
  - 1.758 servers, 1.642.387 (35,2%) resources

Similar surveys in the world

- ***Lawrence and Giles, NEC Institute, February 1999.***
  - There are about 800 millions web pages; 15TB (6TB) of data
  - 34% of web pages have HTML META tag
  - 0,3 % of web pages are using Dublin Core standard
  - *Wide range of different META tags (123 types)*
- ***Harvesting and archiving the Web, J. Hakala, August 2000.***
  - “Web is small and simple”
  - In 1999 Swedish web comprised of 7,5 millions files, with overall volume of 300 GB; 4 main content types cover 97% of Web
- ***The Deep Web: Surfacing Hidden Value; BrightPlanet.com, Jull 2000.***
- ***Netcraft Web server survey (http://www.netcraft.com/survey/)***

Conclusion

- Results meet our expectations and correspond to similar surveys in the world
- Web is simple: we use small number of different formats
- Authors don’t take enough care about metadata
- Inventive but non-standard use of web technologies makes gathering of data difficult

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