An updated Portrait of the Portuguese Web

João Miranda, Daniel Gomes
{joao.miranda,daniel.gomes} @ fccn.pt
http://arquivo.pt

FCCN
Fundação para a Computação Científica Nacional
Foundation for National Scientific Computing

UMIC
Agência para
a Sociedade
do Conhecimento

POS_CONHECIMENTO
Programa Operacional Sociedade do Conhecimento

União Europeia
FEDER

Summary

- Introduction
- Methodology
- Metrics
- Conclusions

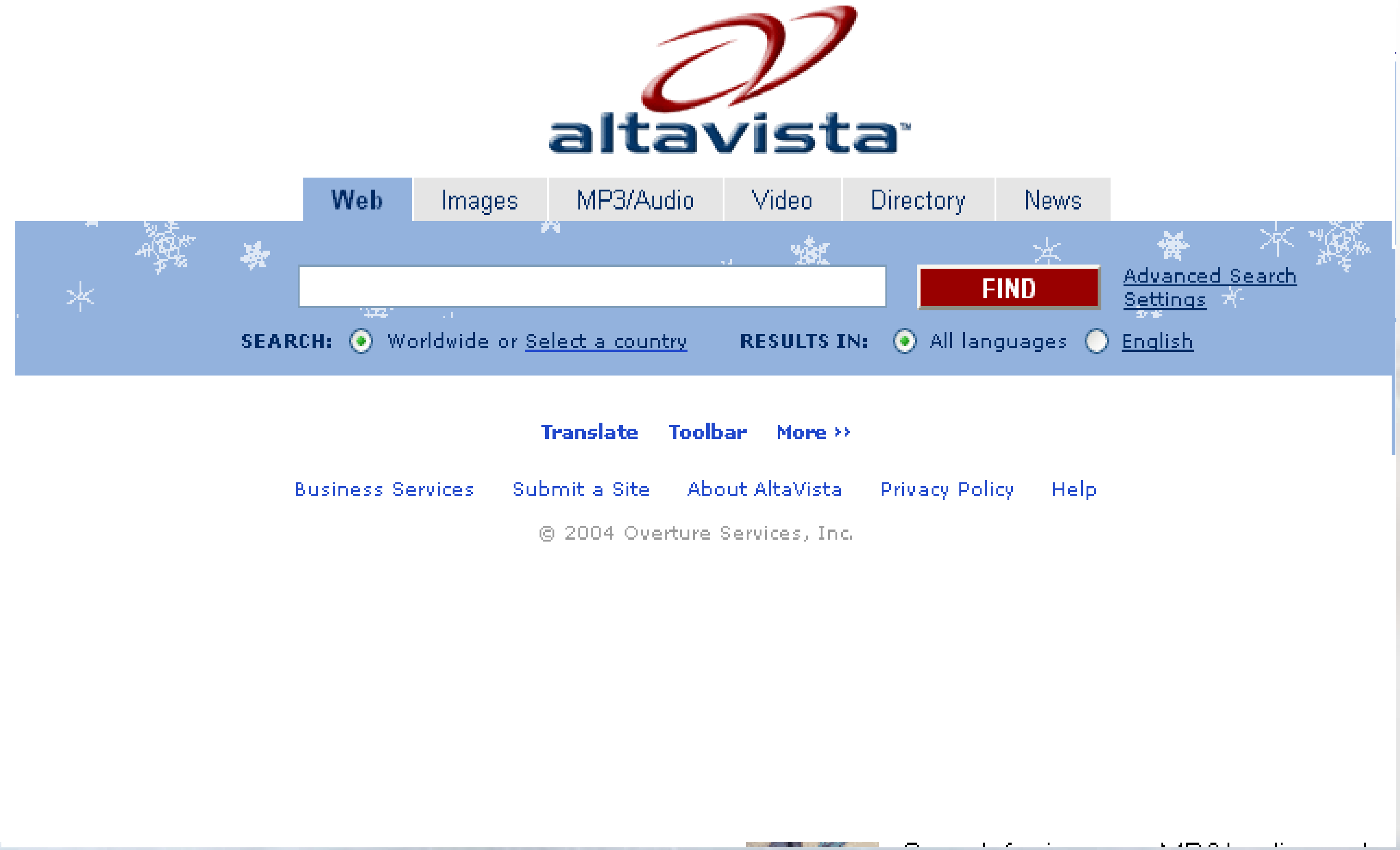
Introduction

The Web

- The Web is a huge source of information
  - Information published exclusively on the Web
  - Information disappears
- Preservation started by the Web Archives
  - Access for future generations
- First initiative: Internet Archive

- Altavista across time

2009

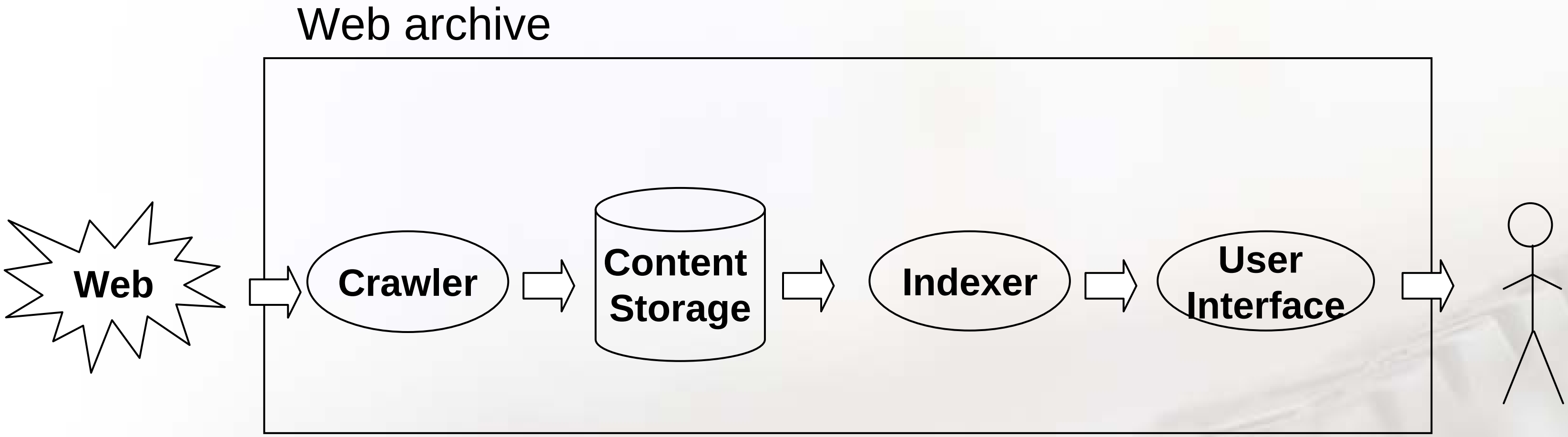


SEARCH: ● Worldwide or Select a country RESULTS IN: ● All languages ○ English

© 2004 Overture Services, Inc.

The Portuguese Web Archive

- The Portuguese Web Archive



Crawler

- What is a crawler?
  - Collects contents from the Web
  - Starts from an initial set of addresses
- How does it work?
  - Iteratively downloads contents
  - Extracts links to find new ones

Methodology

Methodology

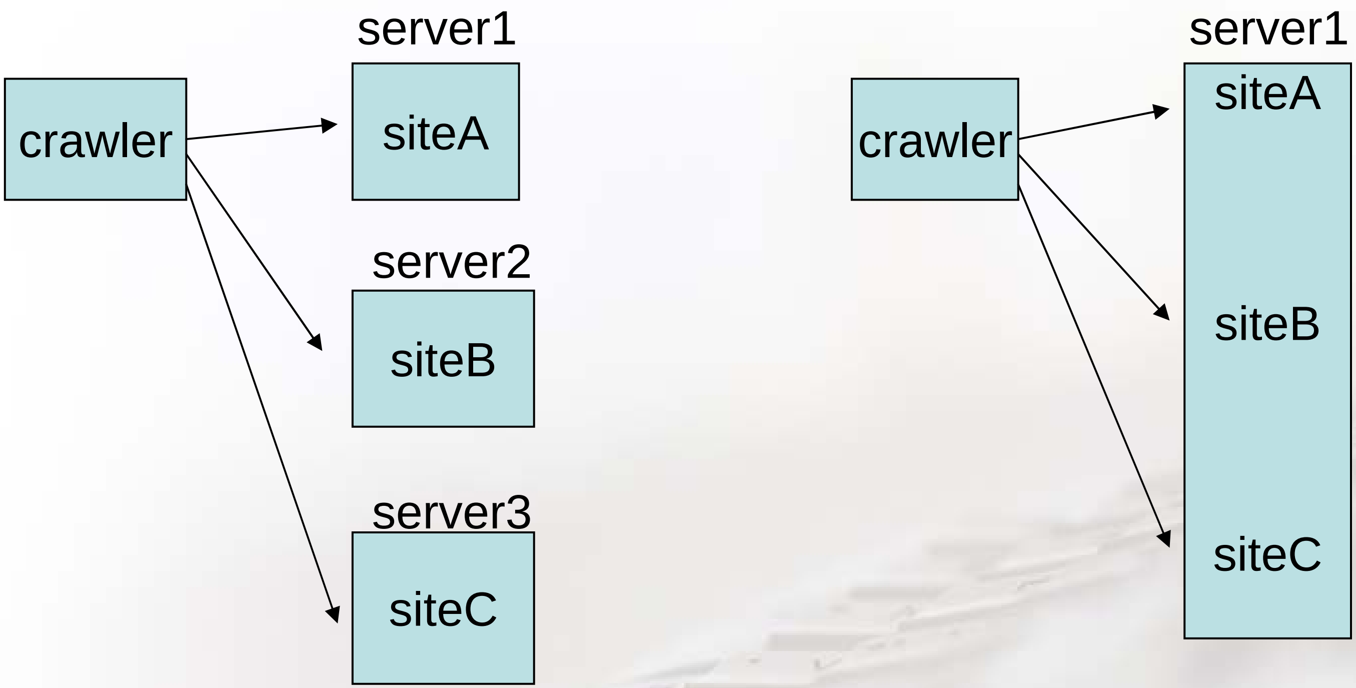
- Crawl of the Portuguese Web (March-May, 2008)
  - .PT domain
  - Heritrix crawler
  - 180 000 initial addresses
  - 48 million contents
  - 2.5 TB
- No content analysis, only log analysis

Metrics

FCCN
Fundação para a Computação Científica Nacional
Foundation for National Scientific Computing

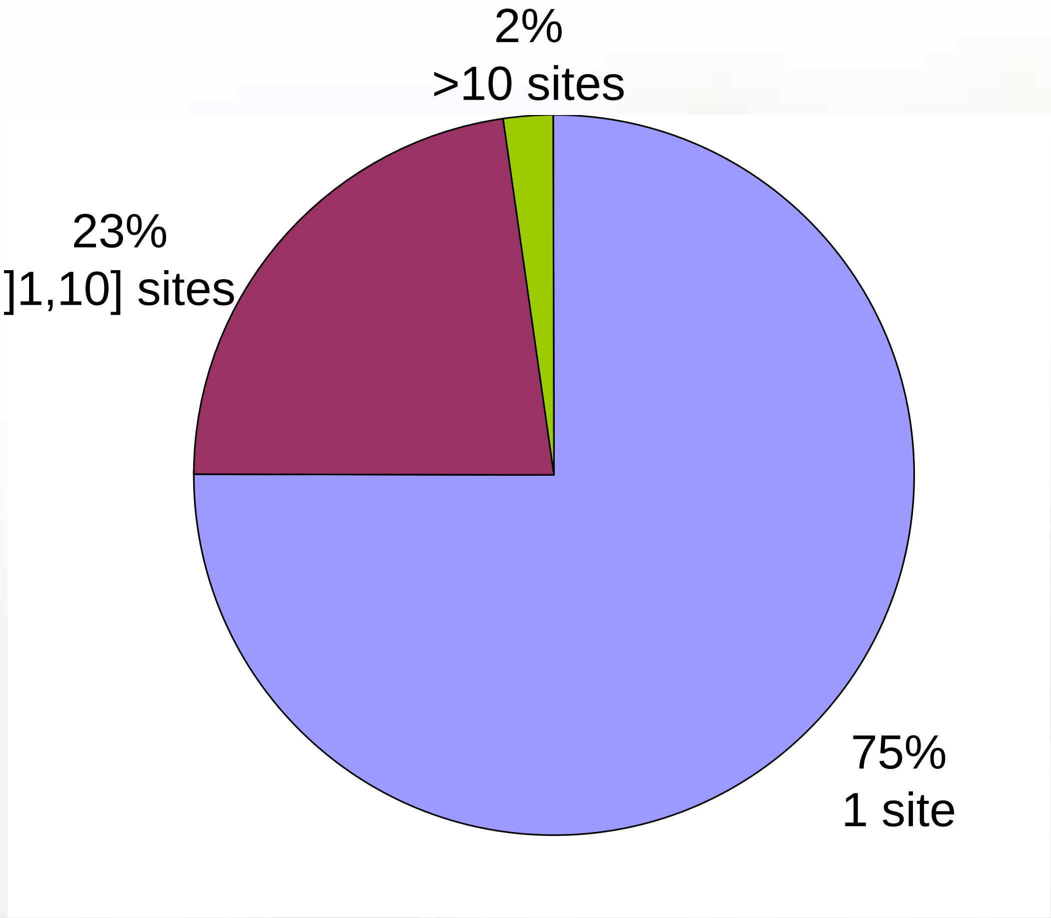
Sites hosted per IP address – Why?

- Politeness policies for crawling



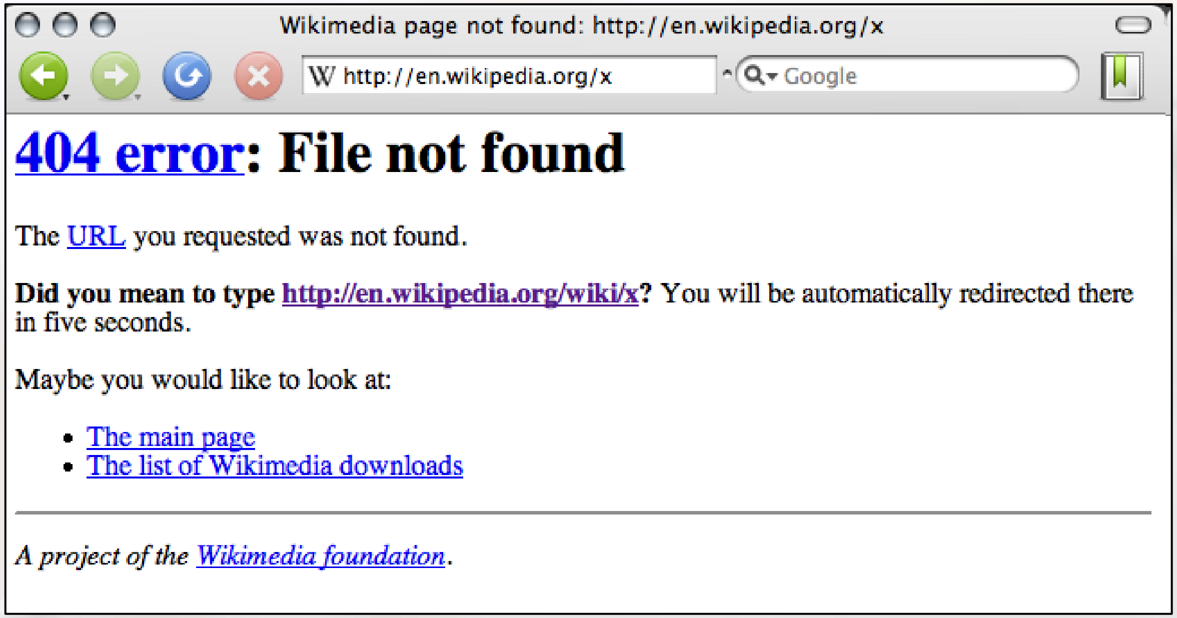
Sites hosted per IP address - Results

- 75% of the IP addresses host 1 site



Successful responses – Why?

- Quality indicator
- Large % of broken links mines trust of users



Wikimedia page not found: http://en.wikipedia.org/x
http://en.wikipedia.org/x
Google

404 error: File not found

The URL you requested was not found.

Did you mean to type http://en.wikipedia.org/wiki/x? You will be automatically redirected there in five seconds.

Maybe you would like to look at:

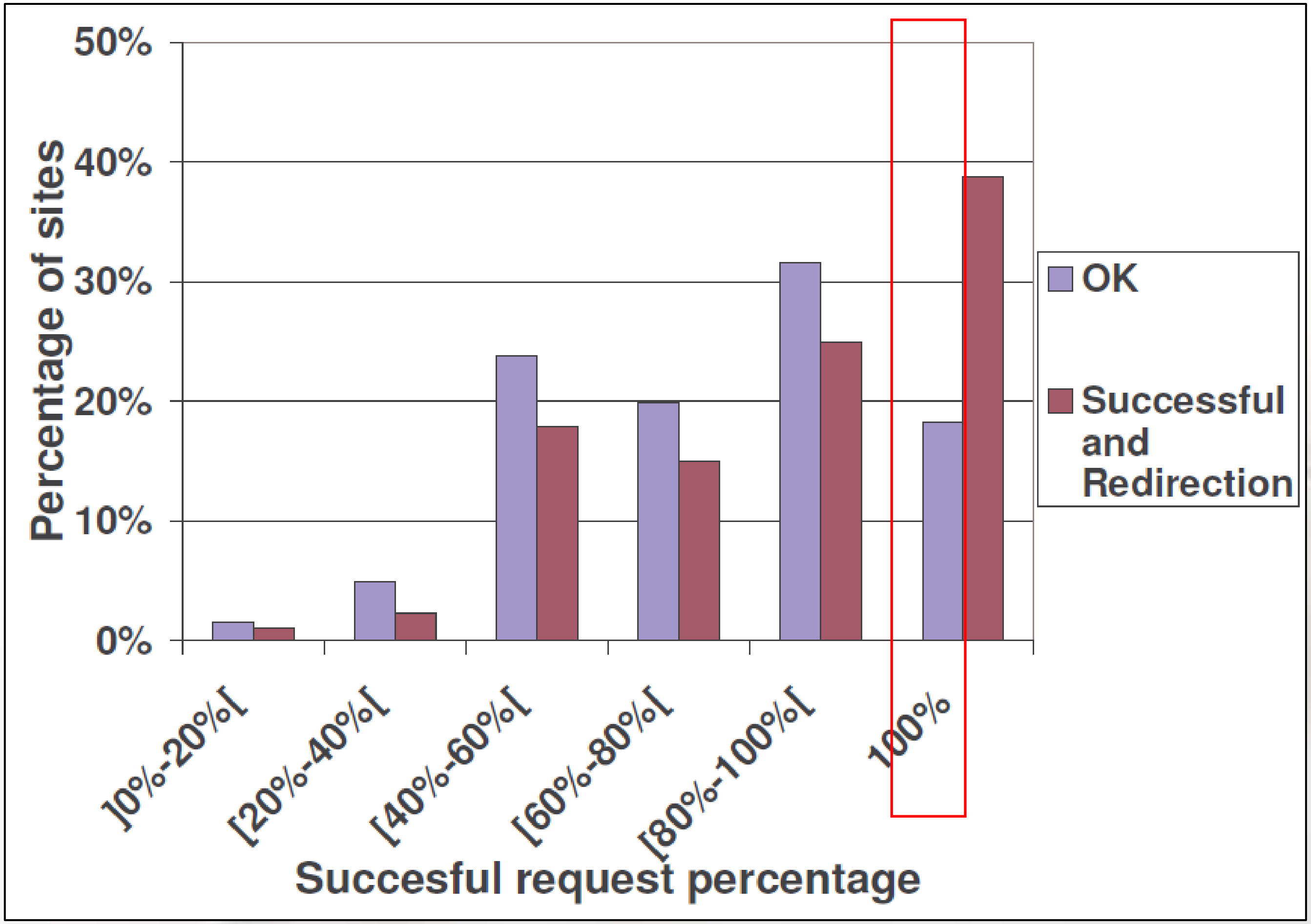
- The main page
- The list of Wikimedia downloads

A project of the Wikimedia foundation.

Fundação para a Computação Científica Nacional
Foundation for National Scientific Computing

Successful responses - Results

- 18% of the sites returned 100% OK responses



Media types – Why?

- Browsers or document viewers for cellphones
- Parsing and indexing for search engines

Media types - Results

- 90% of the number of contents are html, jpeg, gif

•69% of the amount of data are html, pdf, jpeg

	Media type	% contents
1	Text/html	57.8%
2	Image/jpeg	22.8%
3	Image/gif	9.4%
4	Text/xml	1.9%
-	Other	8.1%

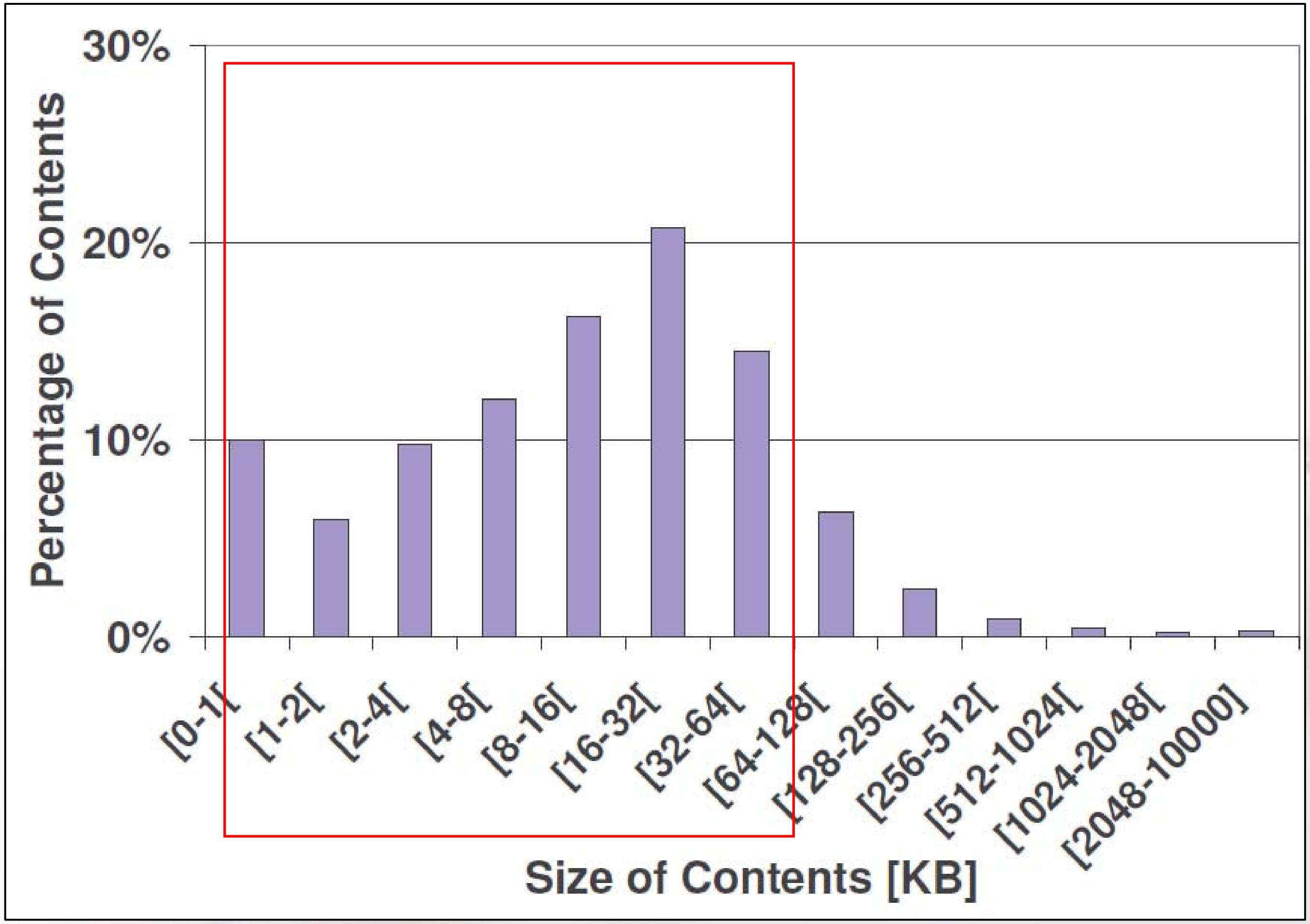
	Media type	% amount data
1	Text/html	35.4%
2	App’n/pdf	17.9%
3	Image/jpeg	16.1%
4	Text/plain	4.2%
-	Other	26.4%

Content size – Why?

- Estimate the storage resources required to create Web data repositories

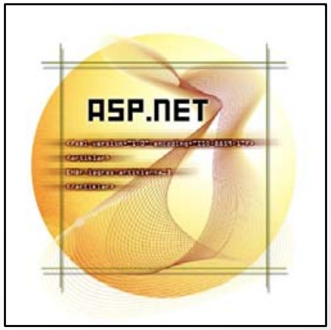
Content size - Results

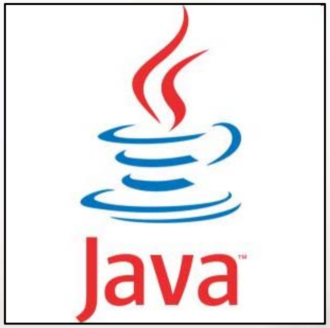
- 96% lower than 128 KB



Dynamically generated contents – Why?

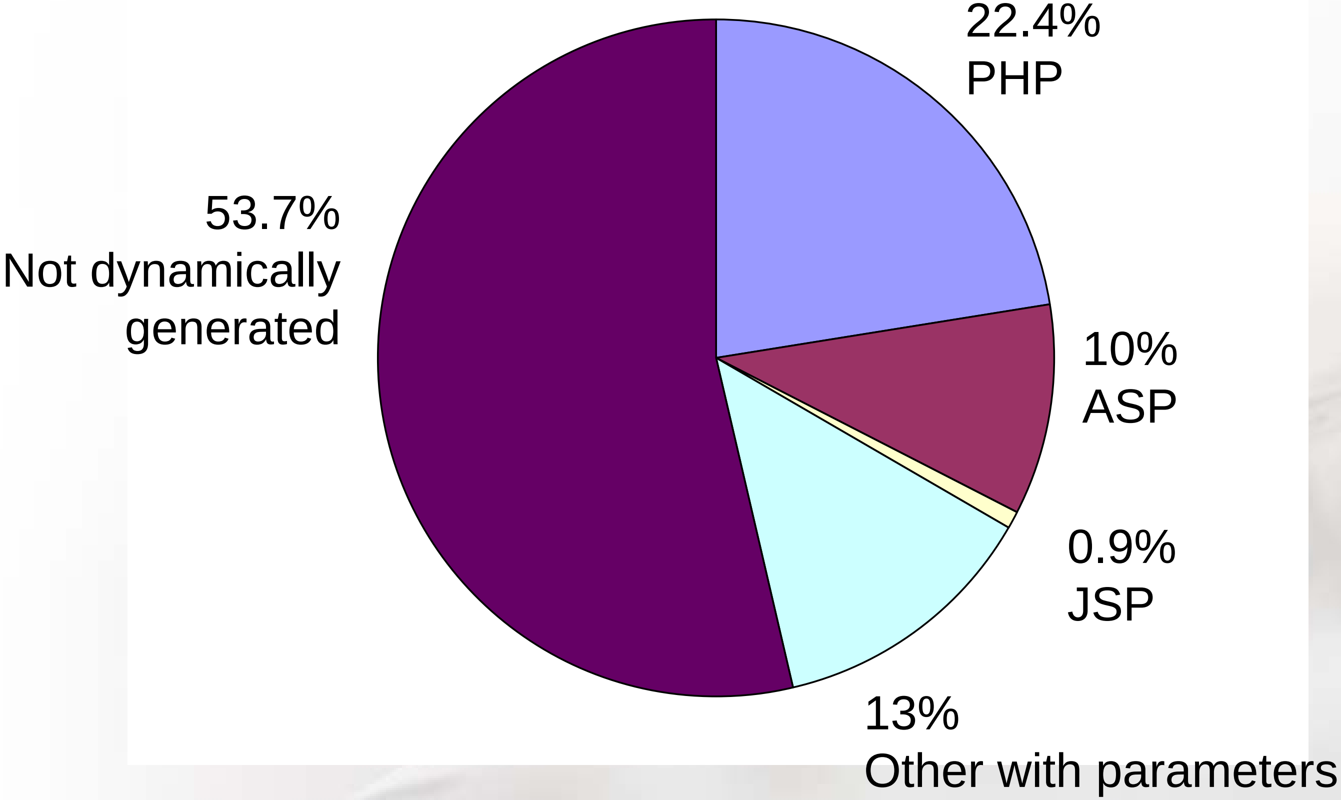
- Identify technological trends in Web publishing





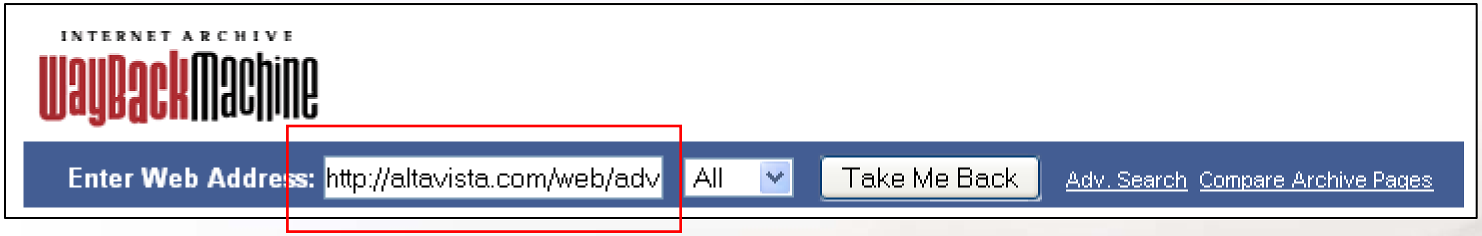
Dynamically generated contents - Results

- At least 46.3 % of the contents were dynamically generated



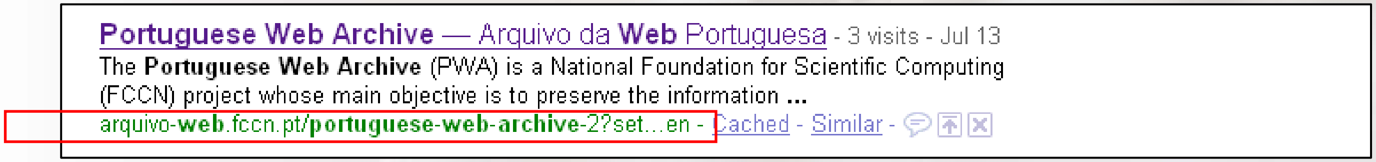
URL length – Why?

- Influences interaction design
- Determine adequate length for input boxes that receive URLs



Enter Web Address: http://altavista.com/web/adv

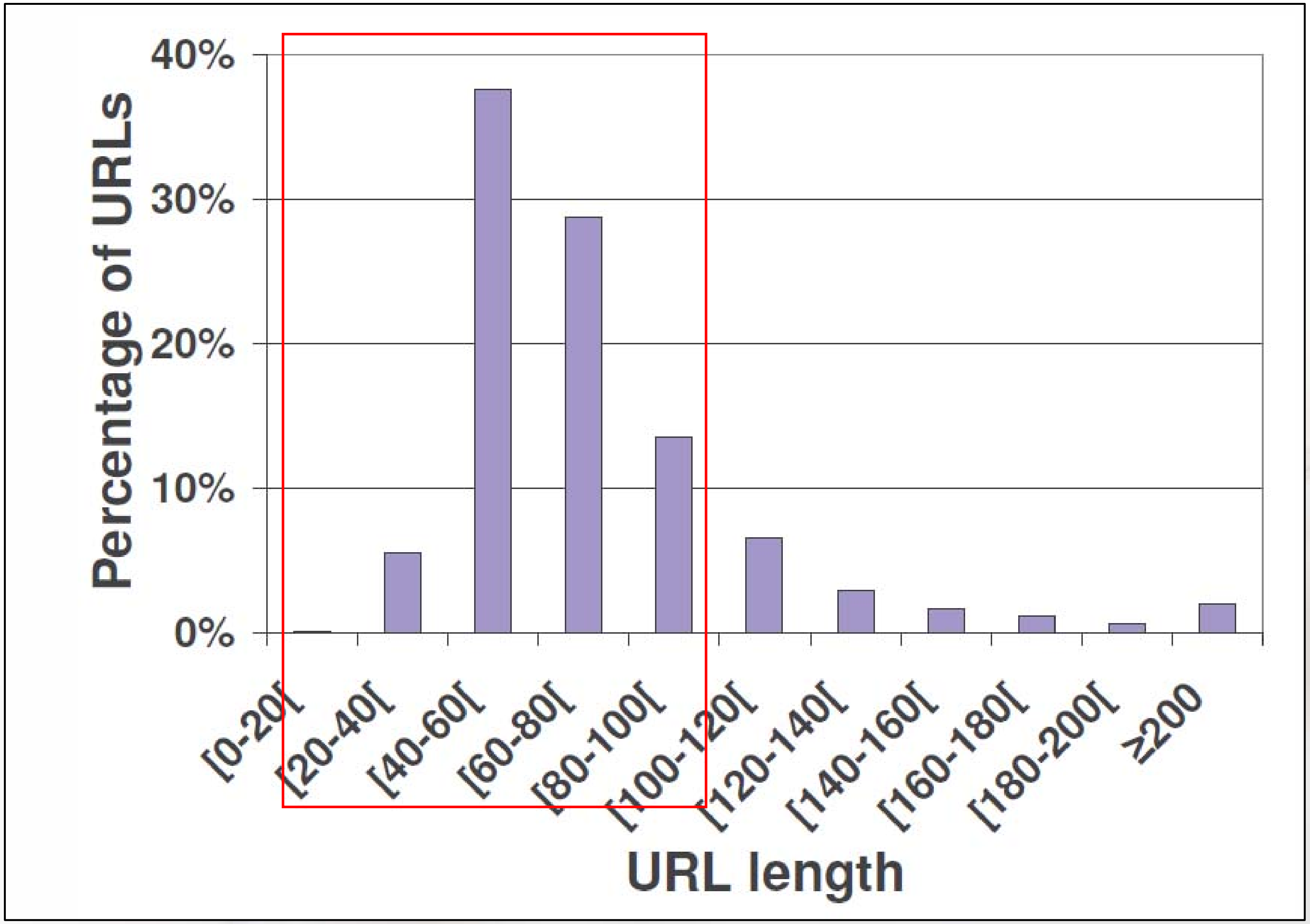
- How many characters should be presented on a search engine results page



Portuguese Web Archive — Arquivo da Web Portuguesa - 3 visits - Jul 13
The Portuguese Web Archive (PWA) is a National Foundation for Scientific Computing (FCCN) project whose main objective is to preserve the information ...
arquivo-web.fccn.pt/portuguese-web-archive-2?set...en - Cached - Similar -

URL length - Results

- 84% lower than 100 characters



Conclusions

Conclusions I

- Long URL addresses
- Half of the contents are dynamically generated (mainly PHP)
- 90% of the contents are HTML, JPEG and GIF
- 69% of the amount of data are HTML, PDF and JPEG

Conclusions II

- 96% of the contents are smaller than 128 KB
- Half of the sites present a successful response rate below 80%
- Most IP addresses host a single site

"Future" work

- Study trends in the evolution of web characteristics
  - João Miranda, Daniel Gomes, Trends in Web characteristics, 7th Latin American Web Congress
- Analyze metrics extracted from content and link analysis

Contribute to preserve the Web

- Anyone can contribute to preserve the Web
- Lend disk space to keep backup copies
  - Just need to install rARC
  - http://arquivo.pt/rarc
- Help required to test beta version

PORTUGUESE WEB ARCHIVE

Thank you.

Logs used in this study are available for research purposes. Please contact us.

http://arquivo.pt