

Characterizing Search Behavior in Web Archives

Miguel Costa, Mário J. Silva

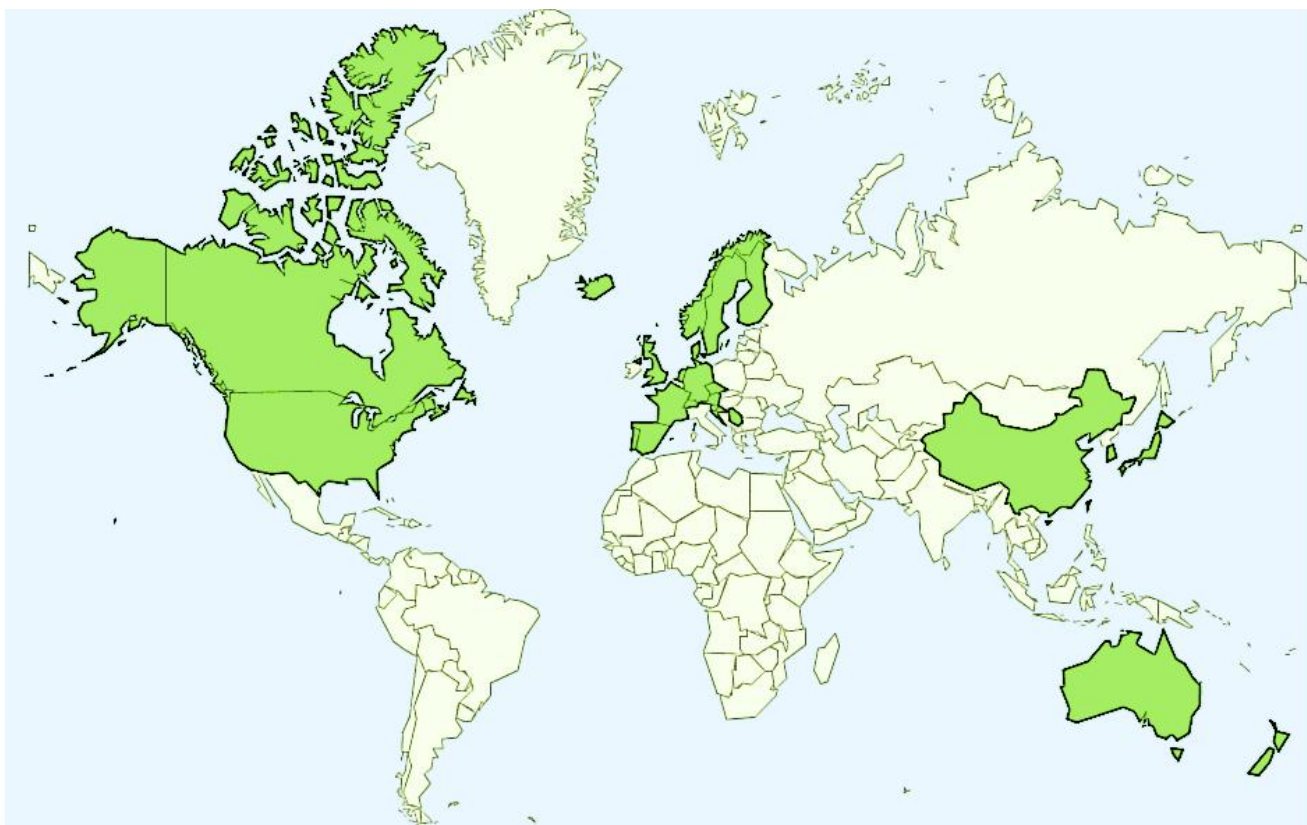
LaSIGE @ Faculty of Sciences, University of Lisbon

Foundation for National Scientific Computing

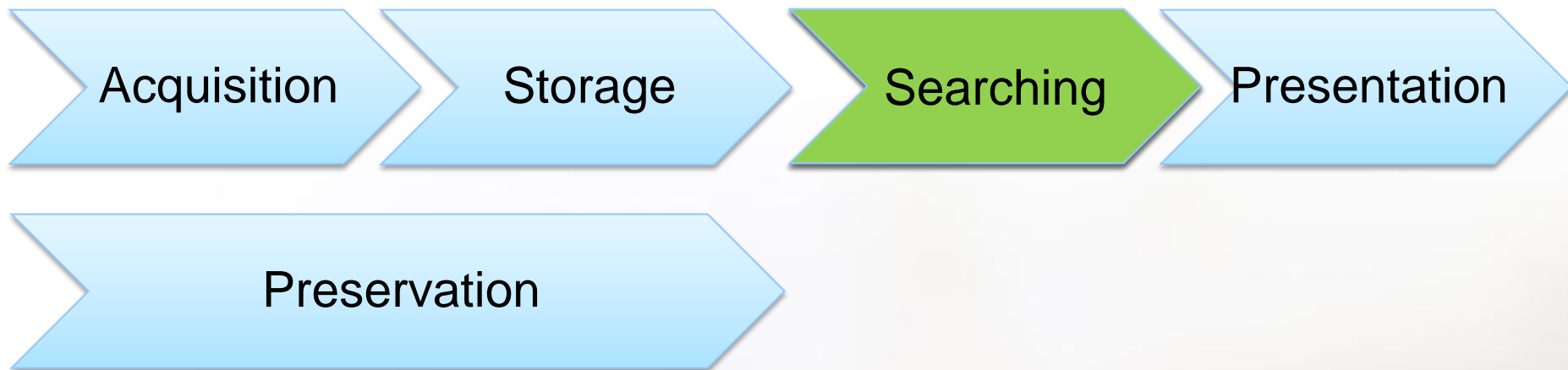
TWAW2011, Hyderabad, India

- The web contains unique and valuable information
 - news, interviews, opinions, feelings
- 80% of the web documents are unavailable after 1 year.
- Knowledge gap for future generations





- 42 web archiving initiatives in 26 countries.
- +180 billion documents archived since 1996.



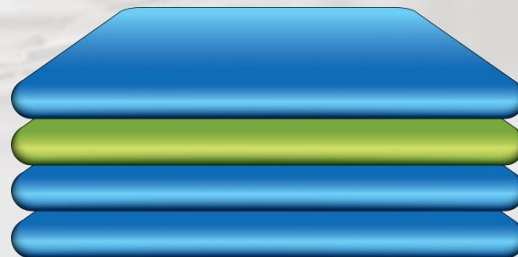
- Search technology based on web search engines
 - ignores the temporal dimension
 - don't understand the end users

- **Why** do users search? (information needs)
- **What** do users search for? (topics)
- **How** do users search? (search behavior)
 - this study: 1st characterization

- Response time
 - e.g. cache, special indexes
- Quality of results
 - e.g. better ranking, suggest queries
- Web design
 - e.g. make most used functionalities stand out

- Archives the Portuguese Web \approx .PT domain
- \approx 182M documents:
 - searchable by full-text and URL.
 - range between 1996 and 2009.
- Search available since 2010.

<http://archive.pt>



Interface: full-text search

Português | Help

fccn

between 01/01/1996 and 31/12/2009 Search

Advanced Search

ARQUIVO DA WEB PORTUGUESA

Experimental

Results 1 - 10 of 231,373

[FCCN - Fundação para a Computação Científica Nacional](#) - 12 March, 2008 - [view history](#)

FCCN - Fundação para a Computação Científica Nacional Login... Localização Contacte-nos FCCN ... do Concurso Público n.º 2/2008

FCCN lança concurso público internacional n.º 2/2008, para ... novos serviços aos seus utilizadores. ServerSign EDI FCCN celebra contrato com a TERENA tendo em ...

<http://www.fccn.pt/>

[Help us improve!](#)
It only takes 30s

[FCCN - Fundação para a Computação Científica Nacional](#) - 12 December, 1998 - [view history](#)

FCCN - Fundação para a Computação Científica Nacional | FCCN | RCCN | RCTS | DNS-PT | PIX | A EQUIPA | LOCALIZAÇÃO | DOCUMENTOS | CRC'98 | RECRUTAMENTO | ...

<http://www.fccn.pt/>

Results Page

same text box



www.fccn.pt

between 01/01/1996 and 31/12/2009

dd/mm/yyyy dd/mm/yyyy

Search

[Português](#) | [Help](#)

Experimental

[Advanced Search](#)

430 Results

Did you want to find results containing the text: "<http://www.fccn.pt>" ?

Search Results between 1 January, 1996 and 5 February, 2011

1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
1 page	1 page	3 pages	7 pages	25 pages	12 pages	8 pages	7 pages	57 pages	116 pages	89 pages	102 pages	2 pages	0 pages	0 pages	---
13 October	10 December	15 February 3 December 12 December	16 January 25 January 28 January 22 February 17 April 23 April 28 April	1 March 2 March 10 May 10 May 20 May 20 May 28 May	18 January 2 February 7 February 24 February 1 March 2 March 1 April	28 March 3 June 20 July 2 August 27 September 29 September 2 October	10 February 6 June 12 June 9 August 18 October 23 October 24 November	21 January 15 April 9 May 26 May 6 June 11 June 12 June	6 January 7 January 12 January 16 January 20 January 22 January 29 January	1 January 6 January 15 January 18 January 18 January 27 January 2 February	1 January 2 January 11 January 16 January 21 January 26 January 27 January	12 March 12 March			Available soon

Versions Page

Methodology

- Pros

- Large and varied
- Less bias
- Cheaper
- Non-intrusive
- Real information needs

- Cons

- Lack of context
- Lack of control

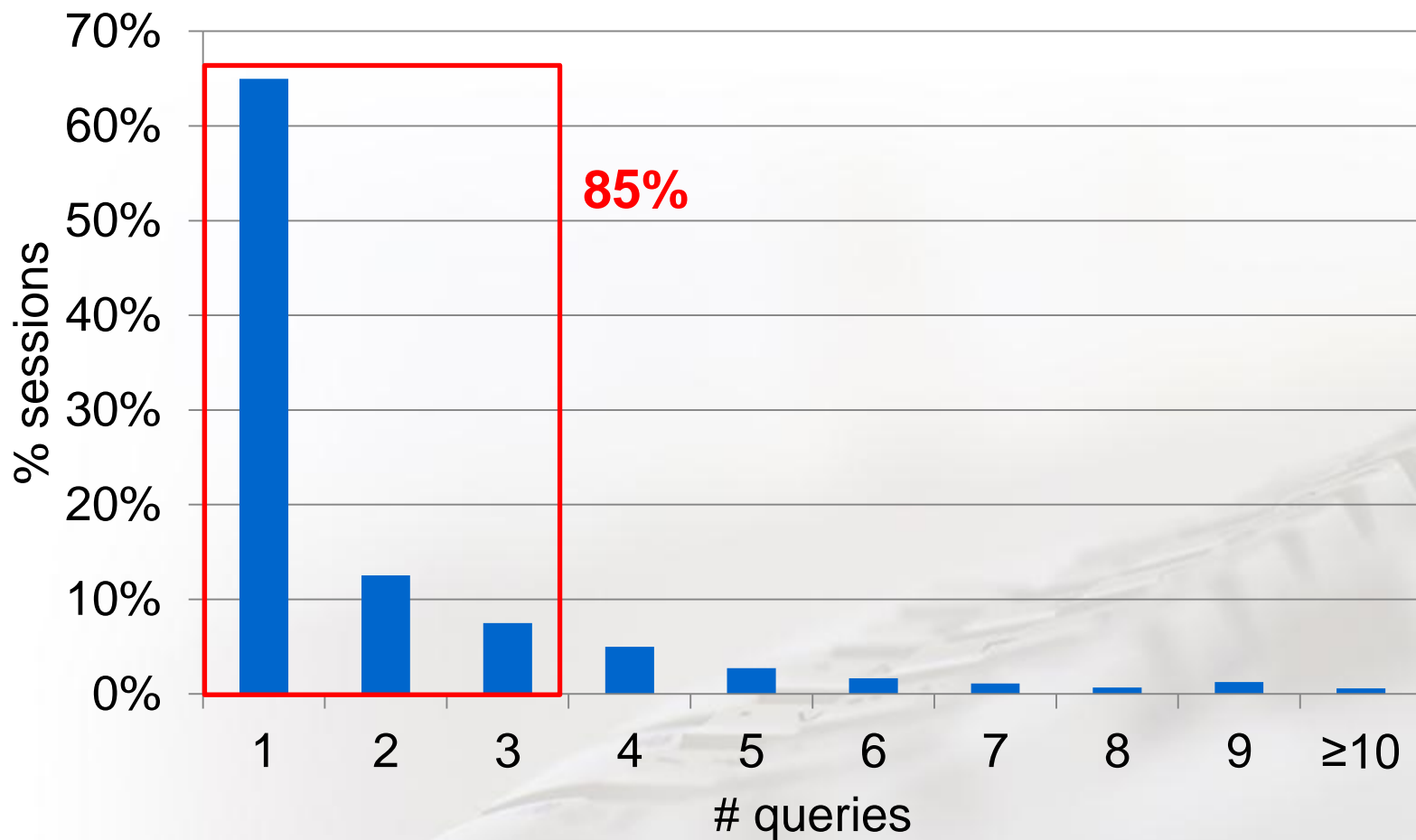


- \approx 10K sessions - 7 months of 2010
- Procedure
 - cleansing
 - normalized and excluded invalid sessions & queries
 - session delimitation
 - used IP, user session and a 30 minute gap
- Users
 - 72% of IP addresses \rightarrow Portugal
 - 89% of interactions \rightarrow PT language interface

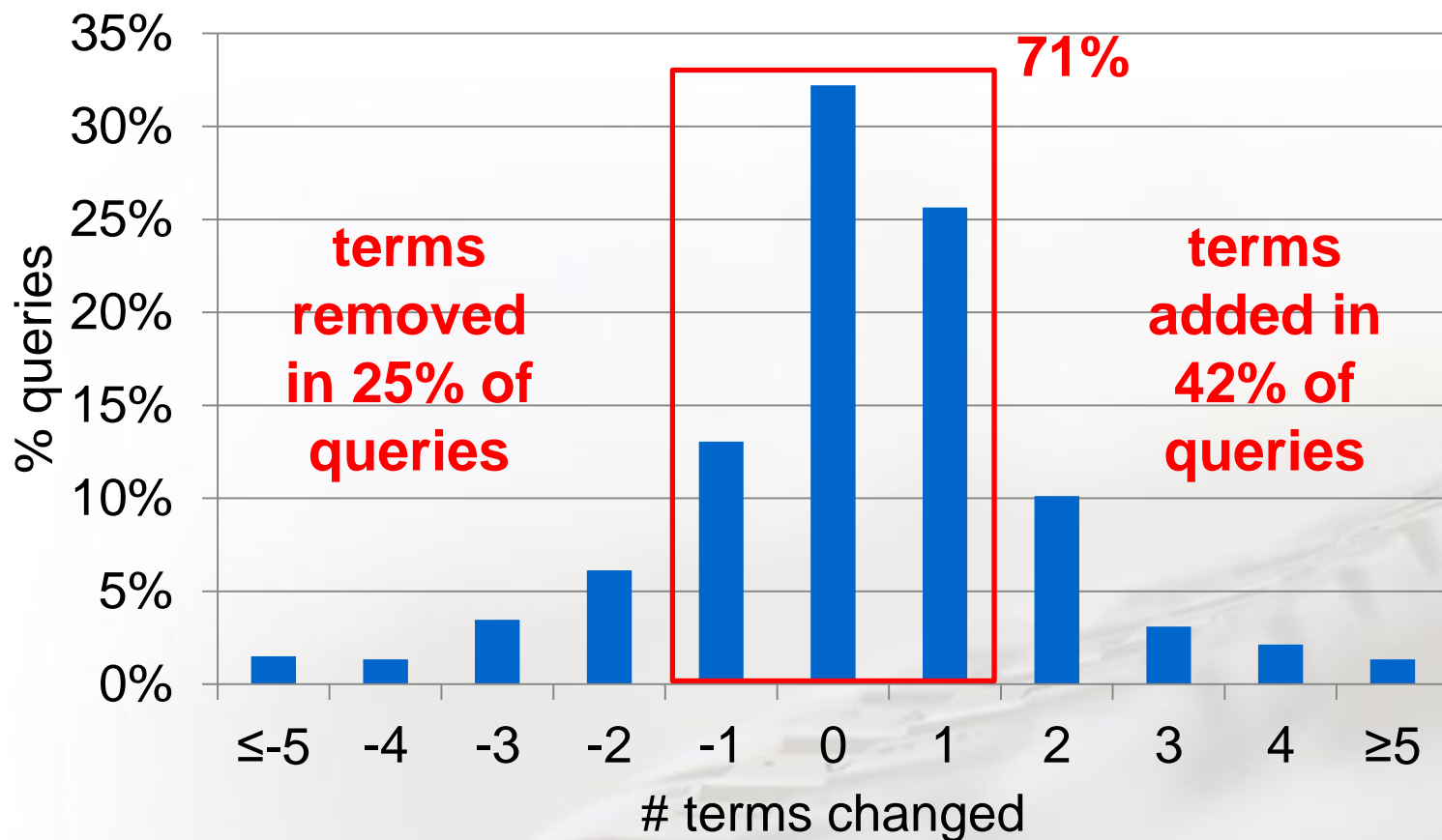
How do users search?

- Full-text sessions + URL sessions $\approx 90\%$
- Full-text sessions / URL sessions $\approx 2:1$
- A typical full-text session:
 - 1 or 2 queries
 - 1 to 3 terms per query
 - 1 or 2 result pages seen per query
 - 1 click per query
- A typical URL session:
 - 1 or 2 queries
 - 1 or 2 clicks per query

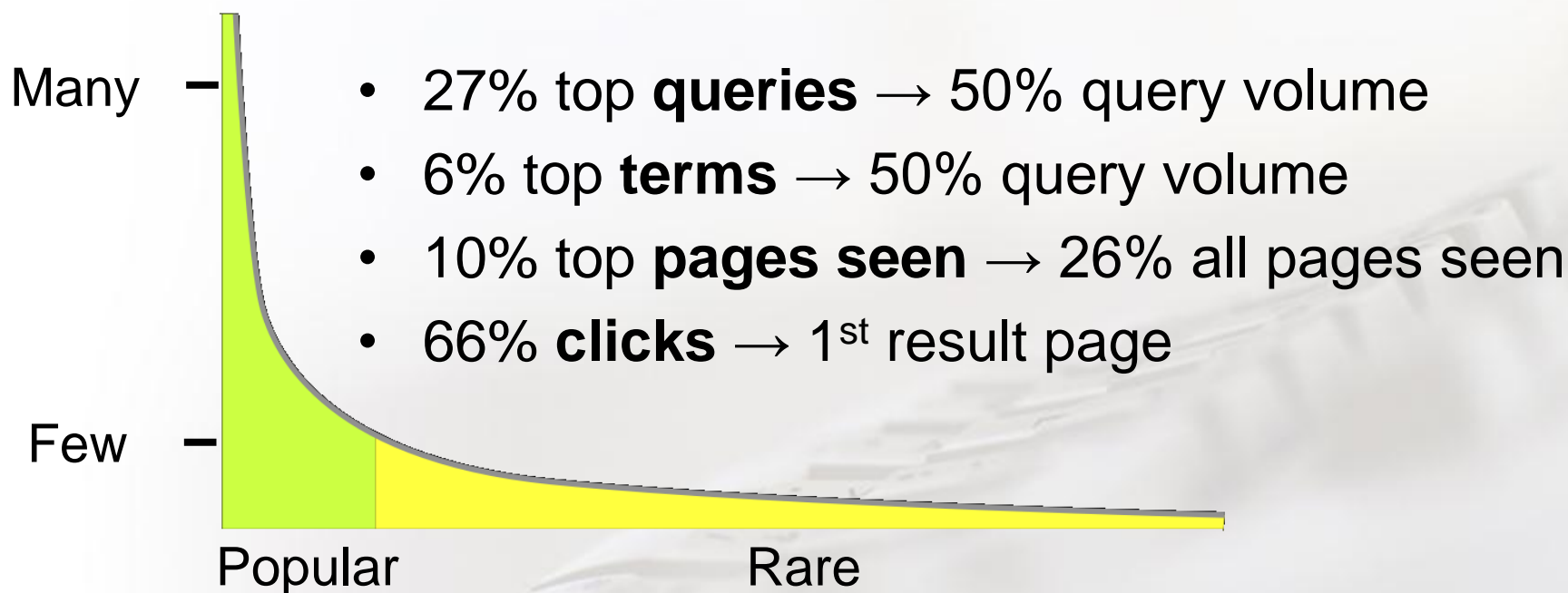
full-text queries per session



full-text terms changed



- Queries, terms, clicks and archived pages seen
 - follow a **power law distribution**



- Spend **little time and effort** on individual searches
- Search and explore following **power law** distributions
- **Search in web archives as in web search engines**
 - Excite (U.S.), Fast (Europe), Tumba! (Portugal)
 - A little less queries, but a little longer

But, what about time?



1/3 Queries are Restricted by Date

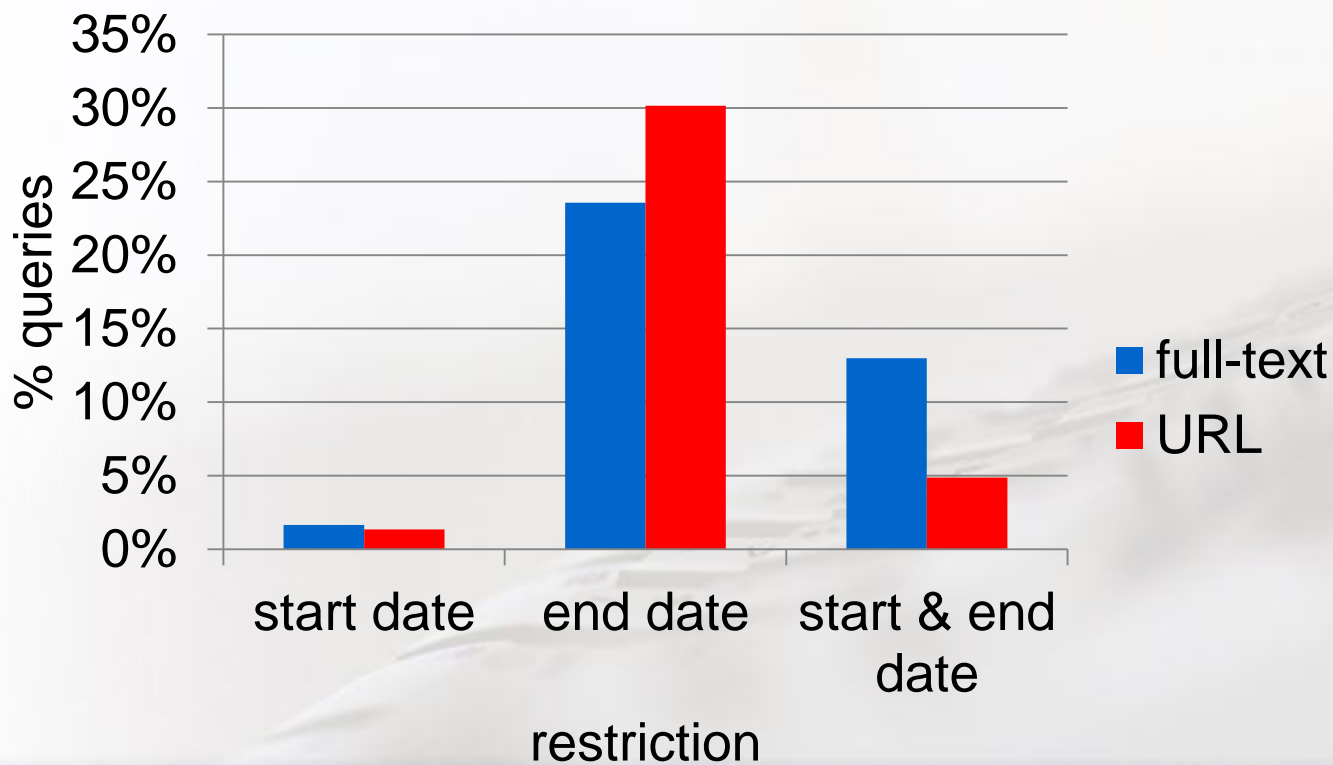
fccn

between 01/01/1996 and 31/12/2009

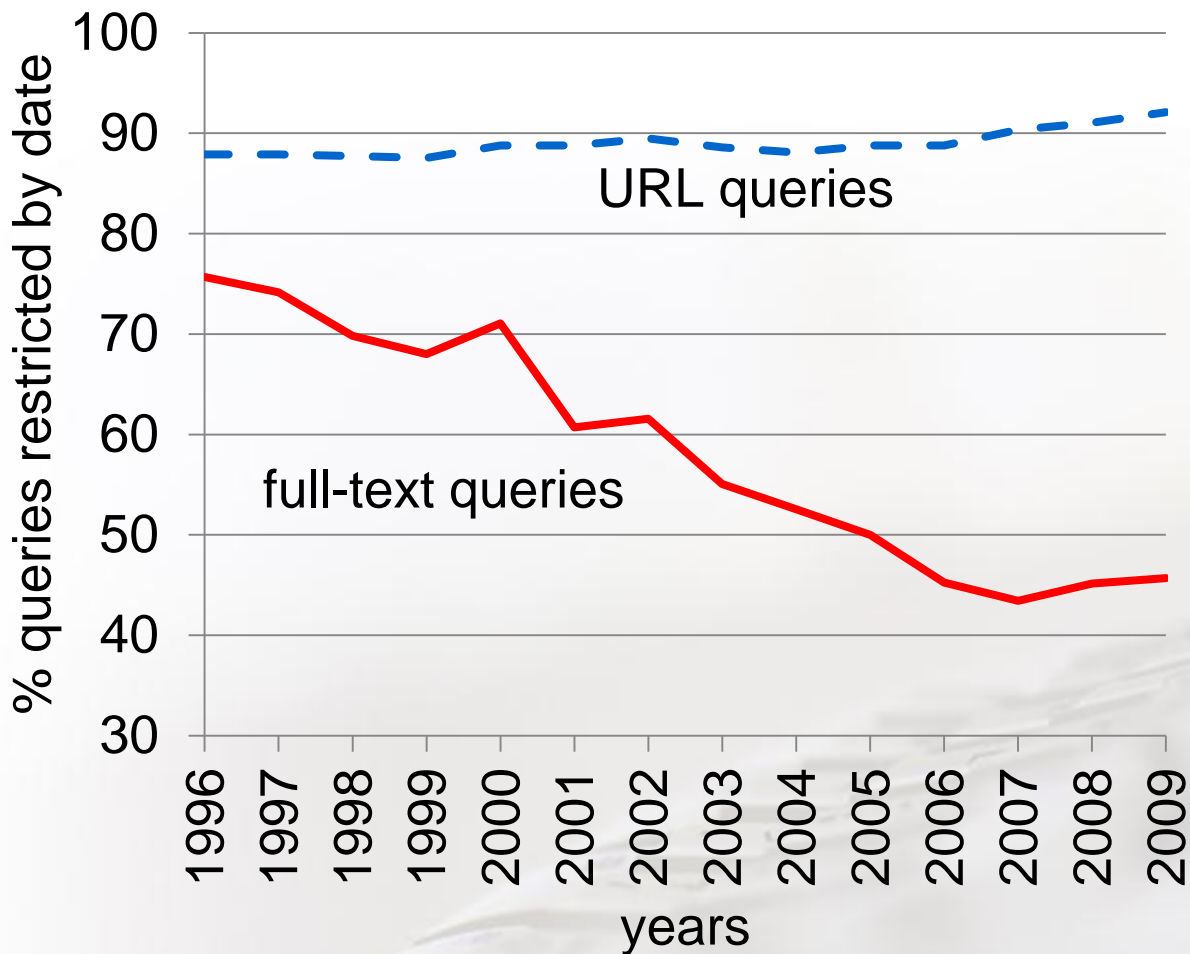
dd/mm/yyyy dd/mm/yyyy

Search

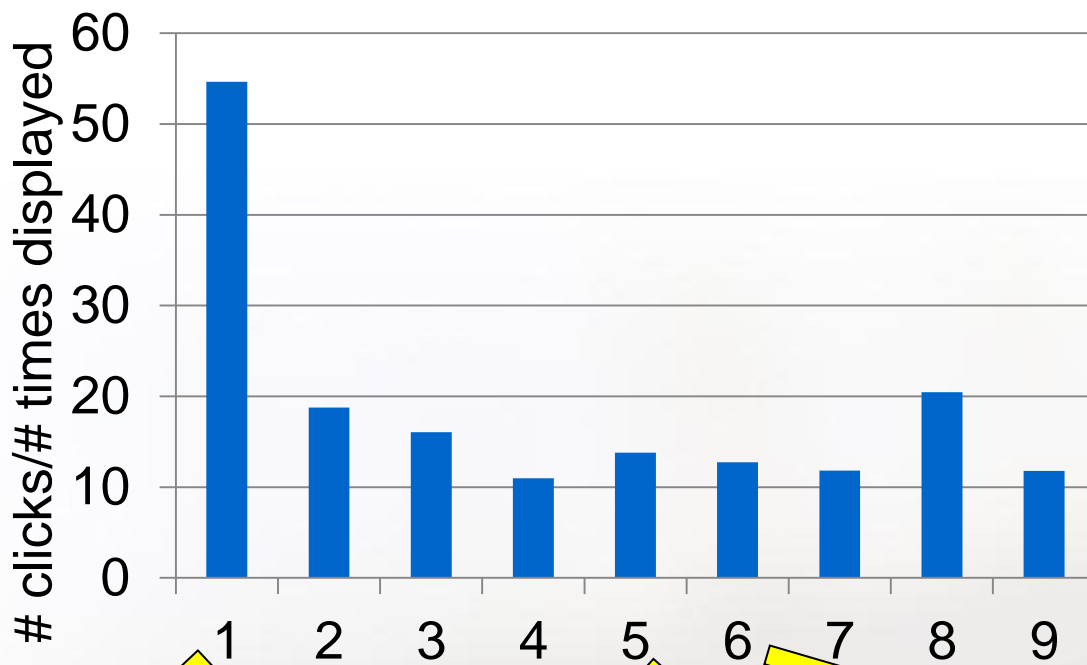
% queries restricted by date



Oldest Versions are more Searched



Oldest Versions are more Clicked



1996	1997	1998	1999	2000	2001	2002	2003	2004	2005
0 pages	0 pages	3 pages	7 pages	32 pages	10 pages	1 page	0 pages	0 pages	73 pages
		2 December 12 December 12 December	25 January 25 January 8 February	29 February 29 February 1 March	9 March 5 April 13 April	14 September			13 October 13 October 16 October

Conclusions

- Web archive users:
 - search as in web search engines
 - prefer full-text search over URL search
 - prefer the oldest documents over the newest

- Validate results
 - larger datasets, other sources, throughout time
- Use results to improve:
 - ranking considering time
 - throughput and response speed
 - user interface

Thank you.



<http://archive.pt>