Fundação para a Computação Científica Nacional
Foundation for National Scientific Computing

A Survey on Web Archiving Initiatives

Daniel Gomes, João Miranda and Miguel Costa

MCTES
MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR

UMIC
Agência para a Sociedade do Conhecimento

POS_CONHECIMENTO
Programa Operacional Sociedade do Conhecimento

União Europeia
FEDER

Fact

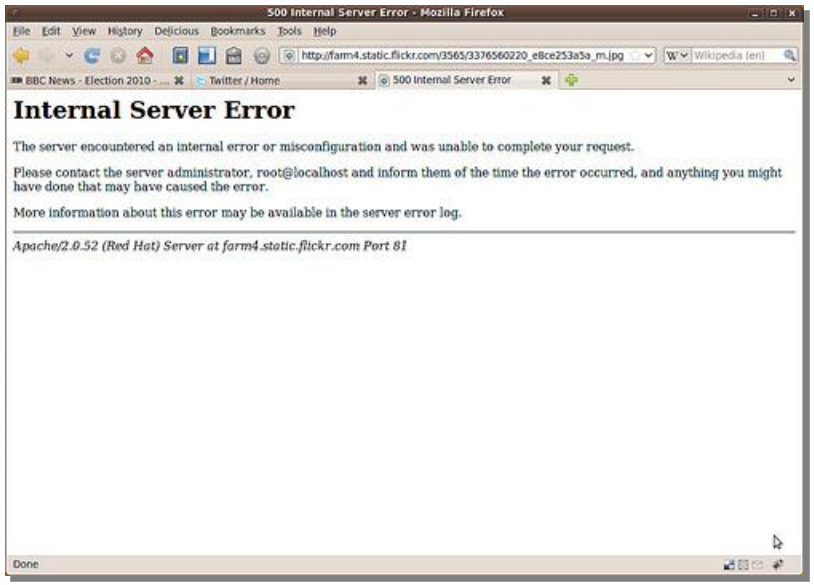
1 year from now,

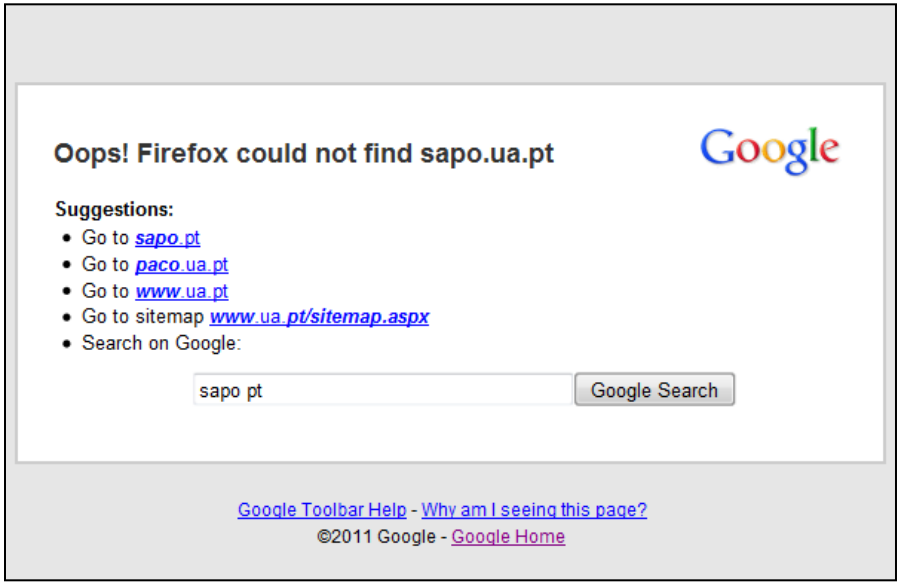
80% of the web pages available today,

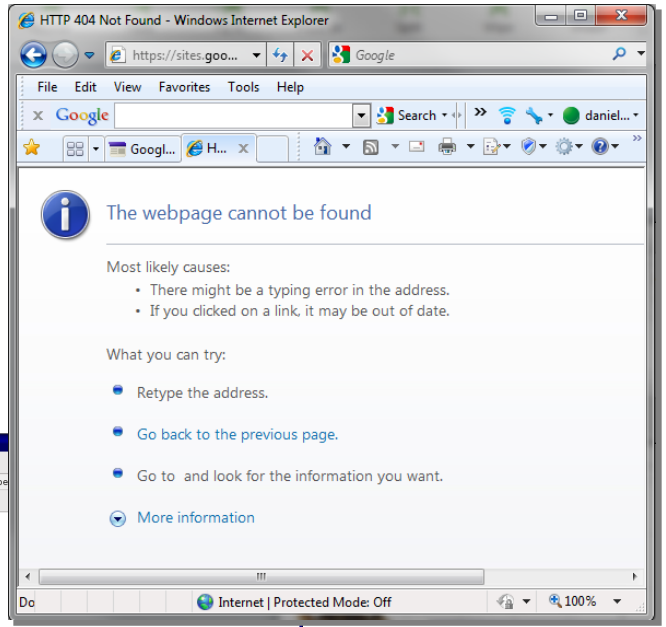
Will have **disappeared** or been **changed**

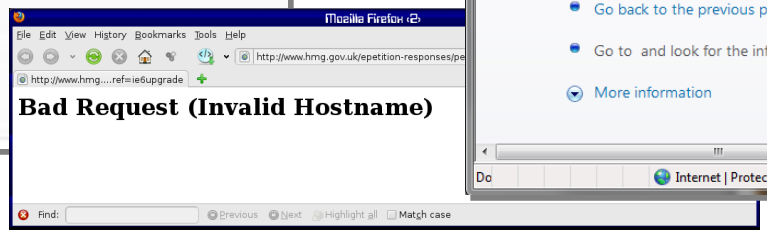
Fundação para a Computação Científica Nacional
Foundation for National Scientific Computing

Consequence









Reasons

"There are no reasons at all in theory for people to change URIs (or stop maintaining documents), but millions of reasons in practice."

Tim Berners-Lee, 1998

- The Web is a primary source of information
- But its information is ephemeral
- So what can you do about this problem?

The same that we did with printed media:

Archive it

Motivation for our survey about web archiving initiatives

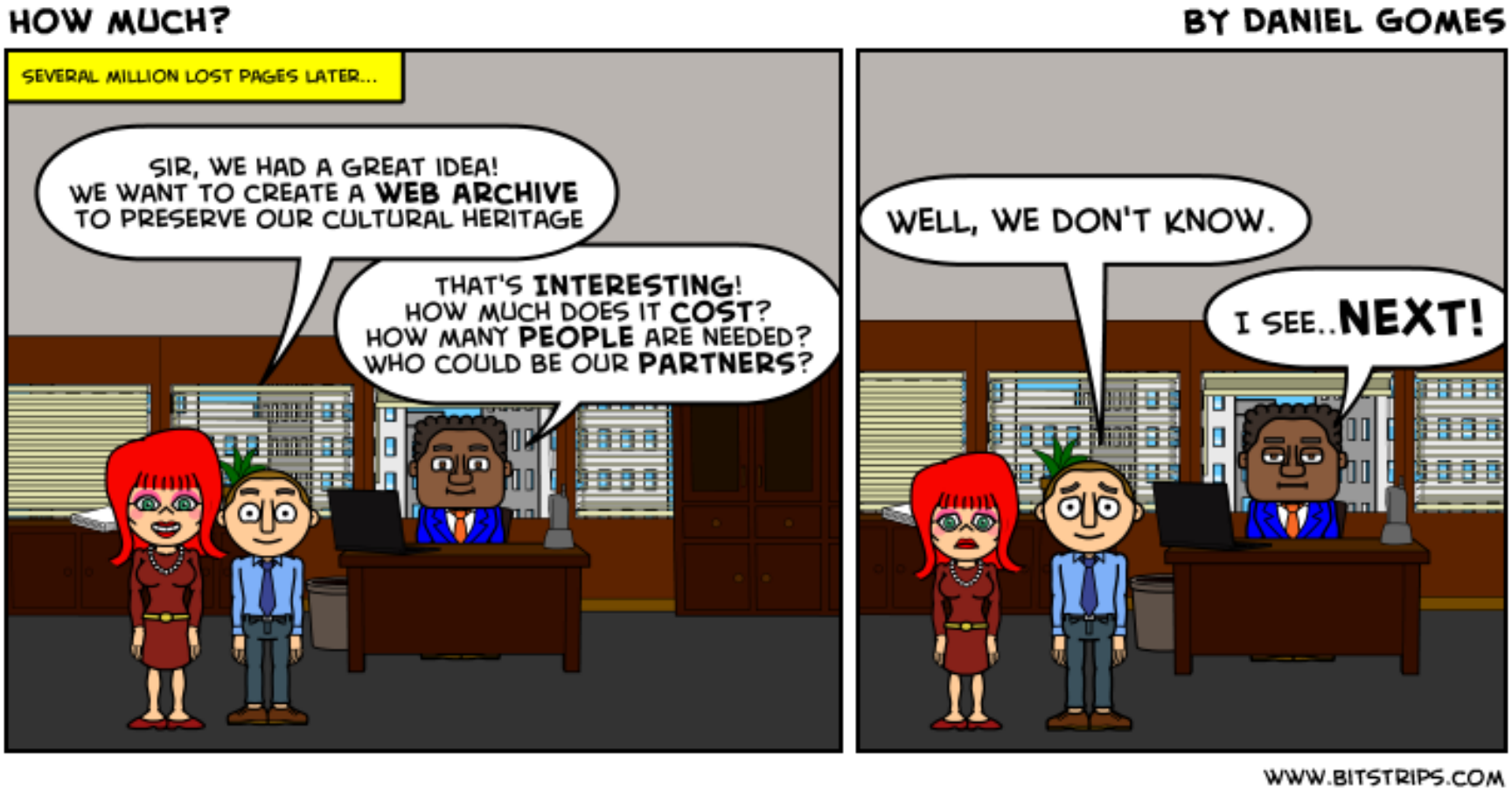
Based on a true story

Fundação para a Computação Científica Nacional
Foundation for National Scientific Computing

People need to know web archives

Fundação para a Computação Científica Nacional
Foundation for National Scientific Computing

Web archivists need to know each other



Fundação para a Computação Científica Nacional
Foundation for National Scientific Computing

Our objective

Derive an **updated** portrait of web archiving initiatives worldwide

Methodology

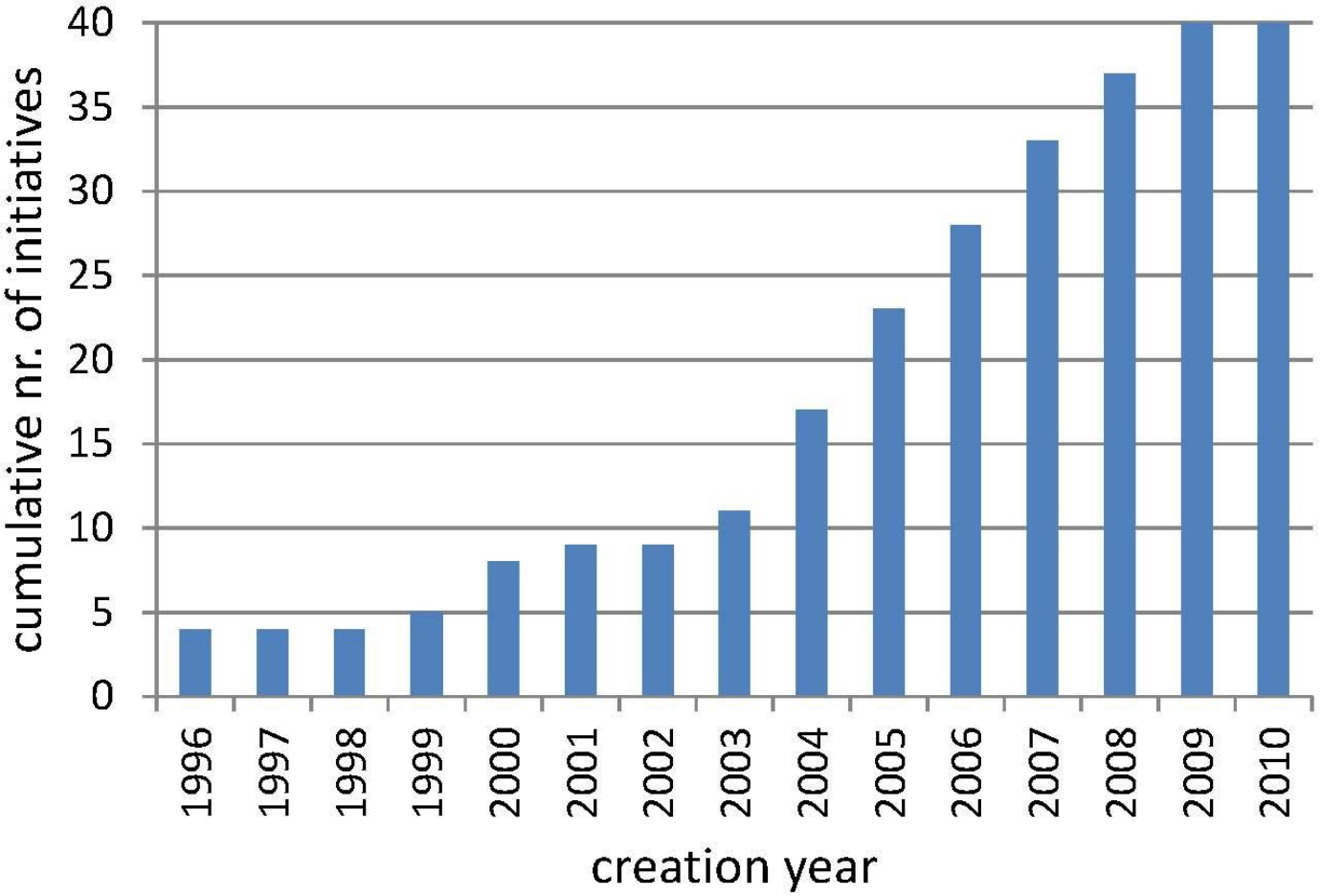
- We asked 3 questions:
  - What is the name of your web archive initiative?
  - Which is the amount of data that you have archived?
  - How many people work at your web archive?

  Any additional comments are welcome.
- Disseminated them:
  - Archive.pt, mailing lists, FB, Twitter, RSS, blogs
- Interacted with respondents by email
  - Wanted to know more than numbers
  - 33 direct answers
- Web archiving initiatives sites and publications

Results…

Fundação para a Computação Científica Nacional
Foundation for National Scientific Computing

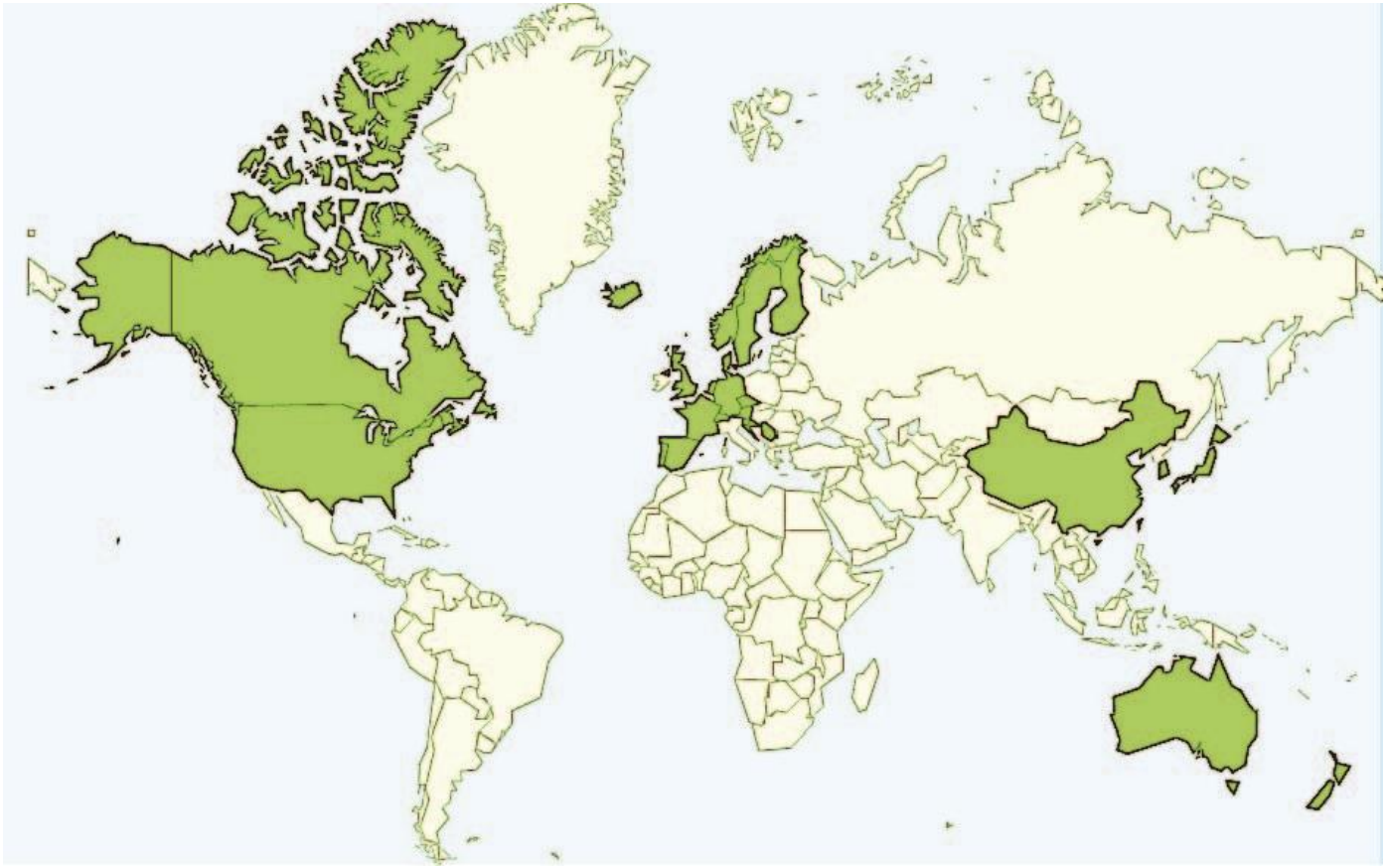
Web archiving is growing



- 42 initiatives
- 1996: Kulturarw3 and Internet Archive
- 2009: Web archive of Čačak (city of Serbia)

Fundação para a Computação Científica Nacional
Foundation for National Scientific Computing

Countries that host initiatives



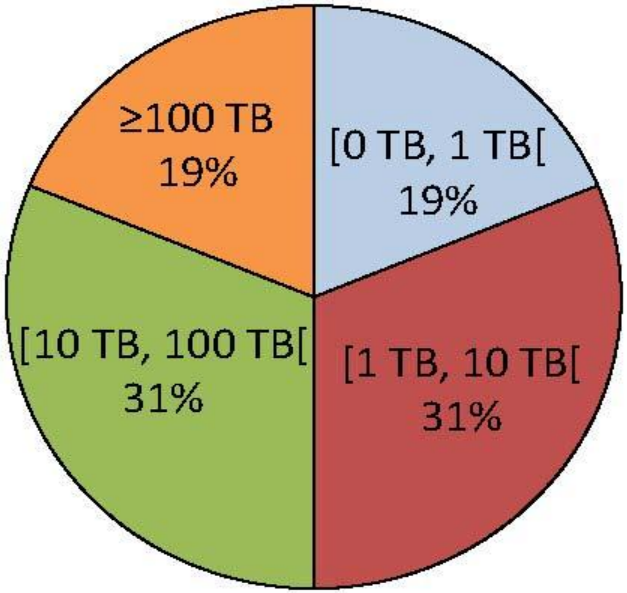
- 96% of the countries host at most 2 initiatives
- 62% are hosted on a country member of the OECD
- 80% hold content exclusively related to their country

Web archiving teams

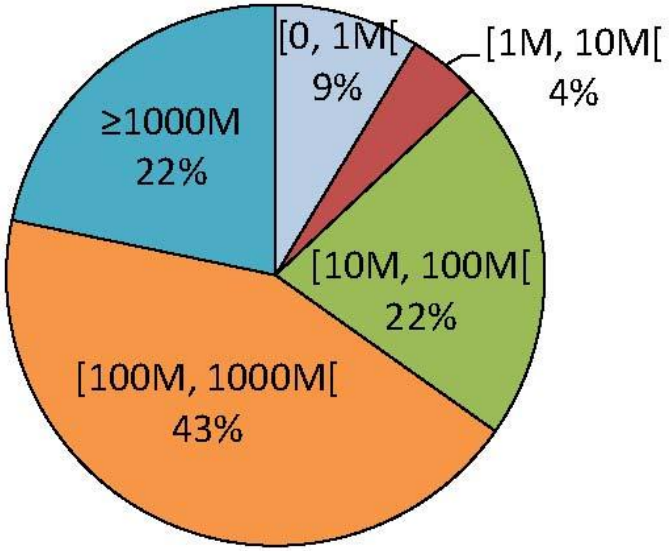
- Small staff
  - Median: 2.5 full-time, 2 part-time
  - 26% of the initiatives don’t have any full-time employee
  - Librarians and engineers
- 277 people worldwide to preserve the Web since its inception
  - 112 Full-time
  - 166 Part-time
- Google employs more than 24 000 people to provide access to the current web

Collection sizes

Volume of archived data



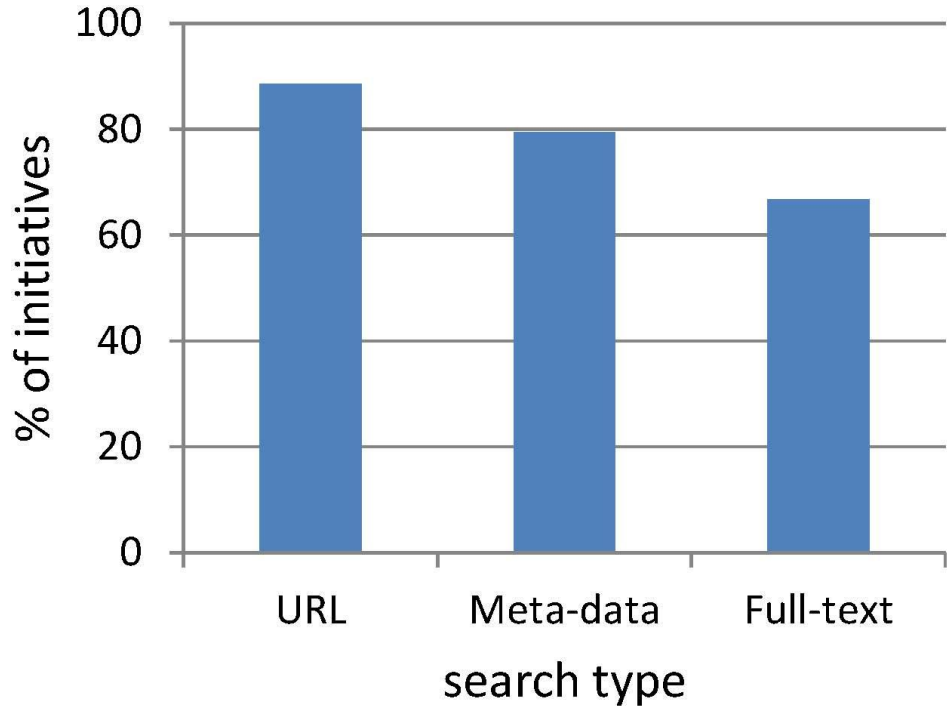
Number of archived contents (e.g. images, pages, videos)



- 81% host <100 TB of data
- 78% host <1 million contents
- Average content size varies according to collection characteristics: 14 KB to 119 KB

Fundação para a Computação Científica Nacional
Foundation for National Scientific Computing

Access to archived data



- 67% support full-text search
- 50% provide full online access to contents
- 38% **restrict access** to archived data
  - Partial access: only to meta-data
  - On-demand only for research
  - Special rooms in their facilities

Food for thought

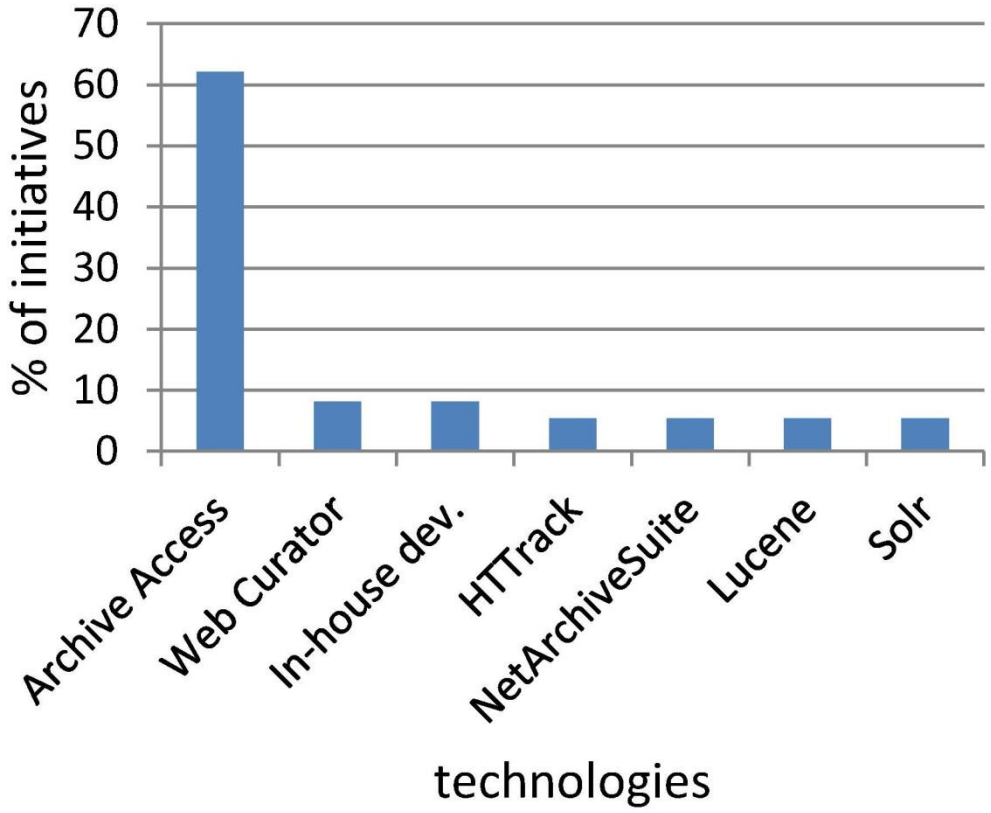
If a content was **published** on the web,

Should its access be **restricted** when it becomes part of **History**?

Maintaining the accessibility level of the contents archived from the web is mandatory to make **web archives useful to all citizens**.

Fundação para a Computação Científica Nacional
Foundation for National Scientific Computing

Technologies



- 62% use Internet Archive Archive-access tools
  - Heritrix for harvesting
  - NutchWax and Wayback for search and access
    - Full-text search is not satisfactory

But all these results will soon become obsolete

Our most important contribution…

Fundação para a Computação Científica Nacional
Foundation for National Scientific Computing

A new Wikipedia page

Log in / create account

WIKIPEDIA
The Free Encyclopedia

Article | Discussion | Read | Edit | View history | Search

List of Web archiving initiatives

From Wikipedia, the free encyclopedia
(Redirected from List of Web Archiving Initiatives)

This page contains a list of Web archiving initiatives worldwide. For easier reading, the information is divided in three tables: web archiving initiatives, archived data and access methods.

Map of Web archiving initiatives worldwide in June, 2011.

Contents [hide]
1 Web archiving initiatives
2 Archived data
3 Access methods
4 References

- Main page
- Contents
- Featured content
- Current events
- Random article
- Donate to Wikipedia
- Interaction
  - Help
  - About Wikipedia
  - Community portal
  - Recent changes
  - Contact Wikipedia
- Toolbox
- Print/export

Web archiving initiatives [edit]

Name	Country	Creation Year	Technologies	Number of Employees		Comments
				Full-time	Part-time	
Australia's Web Archive[1]	Australia	1996	PANDORA Digital Archiving System (PANDAS), NLA Trove, HTTrack	4	>4.25	It is a collaborative program of 11 agencies that provide an estimate average monthly staffing equivalent to 4 FTE. IT outsourced support: 0.25 person-month. Whole Domain Harvests are conducted by the Internet Archive using Heritrix, Wayback Machine.
Our digital island, a Tasmanian Web Archive[2]	Australia	1996	HTTrack, Experimentally: Web Curator, Heritrix and Wayback Machine		1	
			Archive-access tools and			

Fundação para a Computação Científica Nacional
Foundation for National Scientific Computing

Impact of the Wikipedia page

- Positive feedback from the web archiving community
- Information about 25 initiatives was **collaboratively** updated (after the paper was submitted)
- 49 web archiving initiatives (August, 2011)
  - 7 more than we identified in our study
  - 7 web archiving commercial services
    - 3 new services created in 2010

Fundação para a Computação Científica Nacional
Foundation for National Scientific Computing

Conclusions

- Web archiving is useful to everyone
- Web archiving is worldwide
- We must support web archiving initiatives

Fundação para a Computação Científica Nacional
Foundation for National Scientific Computing

Contact us

PORTUGUESE
WEB ARCHIVE

www.archive.pt

daniel.gomes@fccn.pt

It will be a pleasure to hear from you