Arquivo.pt

Improving the robustness of our service

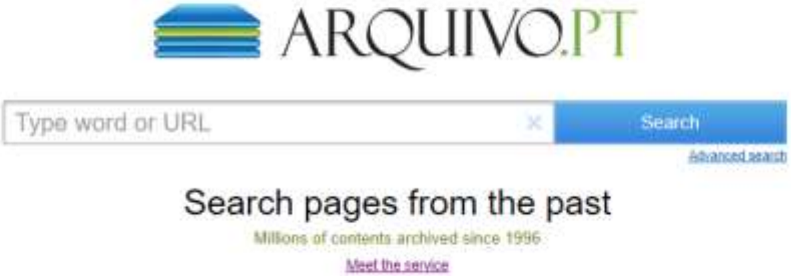
Daniel Gomes

What is Arquivo.pt?

Web pages preserved since 1996

Public search service

Information in several languages


ARQUIVO.PT

Type word or URL

Search pages from the past

Millions of contents archived since 1996

Meet the service

Brief history of Arquivo.pt

2007: Project launch

2010: Search prototype publicly available

9/2013: Service collapsed due to hardware malfunction

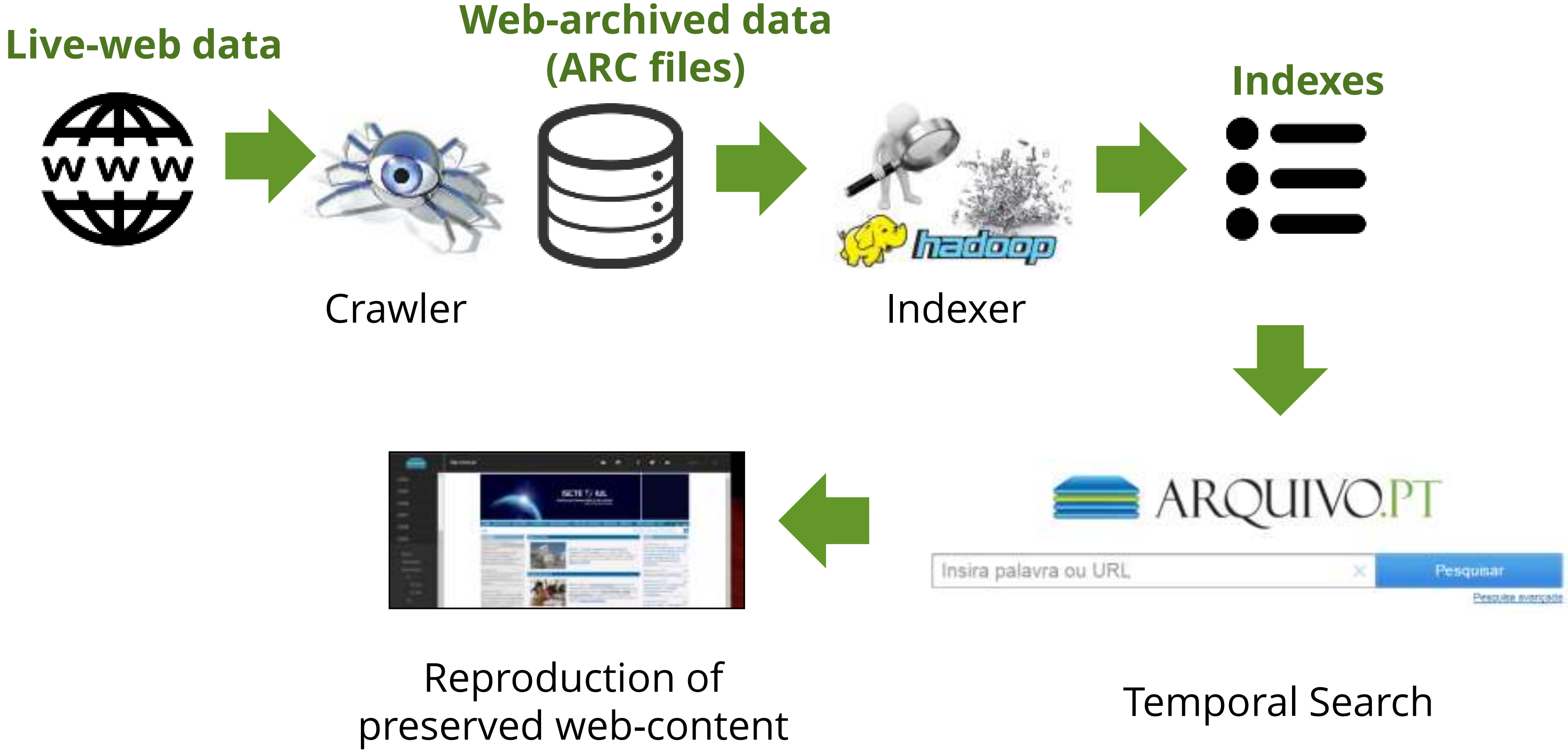
- Data loss of 17% (17 TB)
- Crawling interruptions
- Suspension of search service

2014 - 2016: Recovery and improving robustness

Now, we can share our experience.

Arquivo.pt system overview

Our web archiving workflow is mainly automatic



Live-web data

Crawler

Web-archived data (ARC files)

Indexer

Indexes

Temporal Search

Reproduction of preserved web-content

Arquivo.pt is a medium-size web archive

Hardware

- 85 servers

Archived data

- 4 billion files
- 468 TB (ARC files, indexes, replication)

Estimated data growth

- 72 TB/year

5 measures to improve the robustness of Arquivo.pt

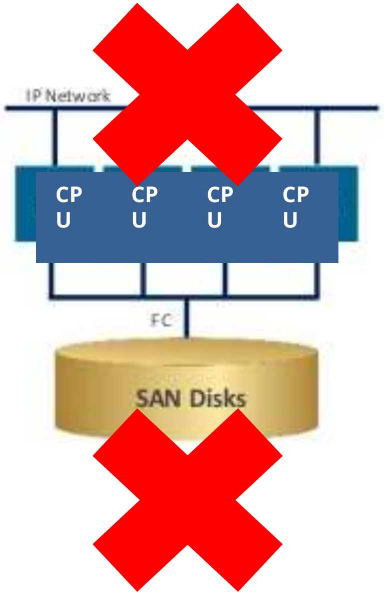
Hardware and software architecture shifted to *Shared-Nothing* (#1)

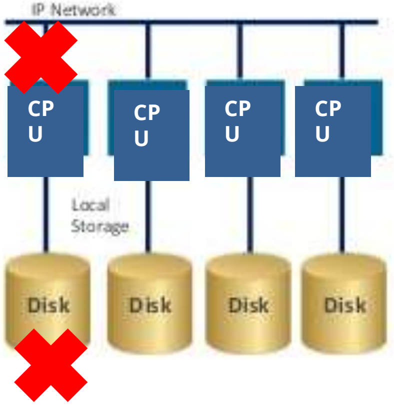
Design-to-fail: the failure of a single equipment cannot jeopardize the service

Centralized: blade server enclosures + storage arrays

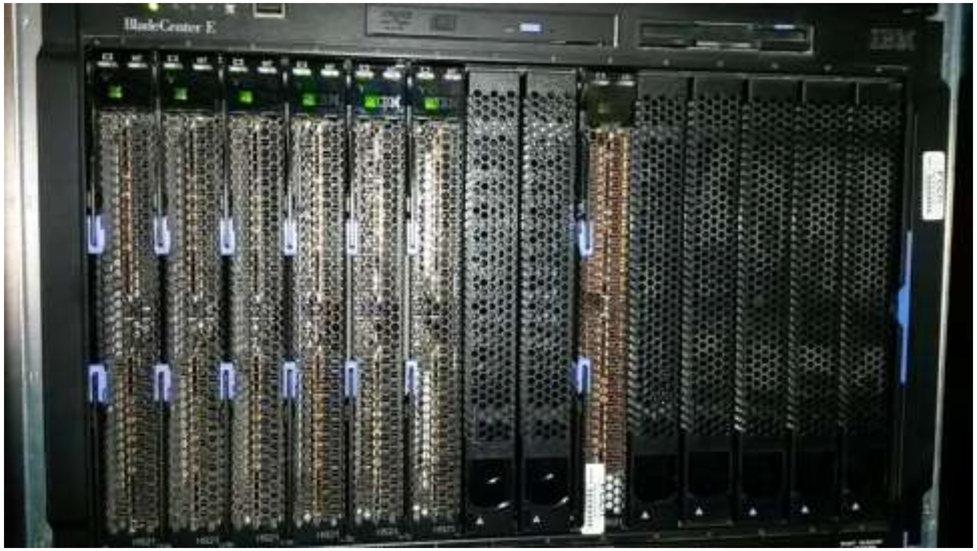
vs.

Distributed (*shared-nothing*): independent rack servers

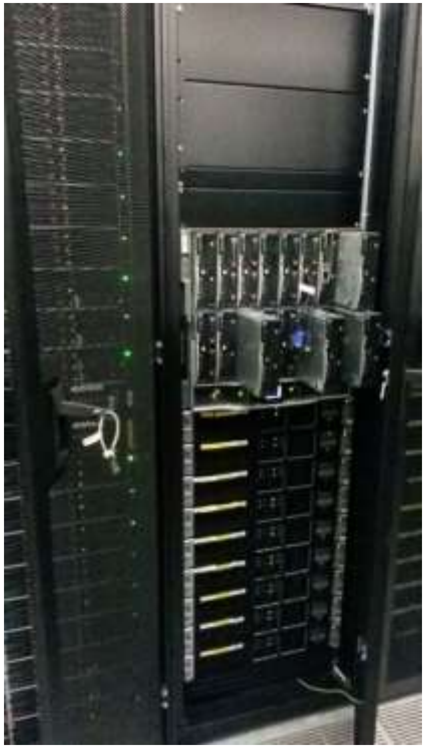




Inefficient physical space management at the data center with blade systems



Space that was never used



Space still occupied after servers disabled

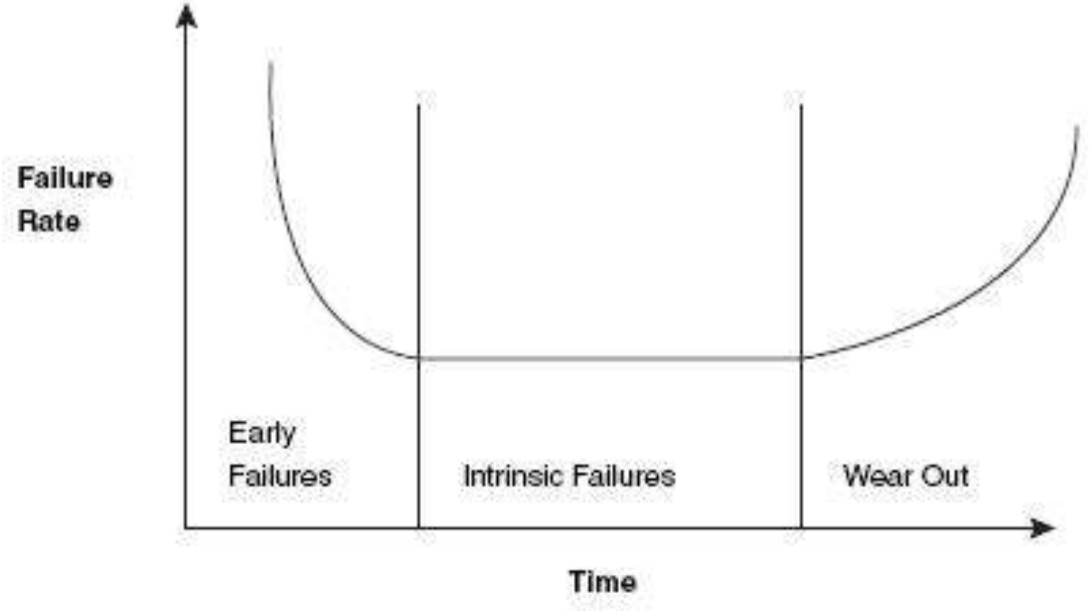
Independent rack servers

Only operational servers occupy physical space

Physical space is released as servers break

Perform load tests immediately after buying to induce failures

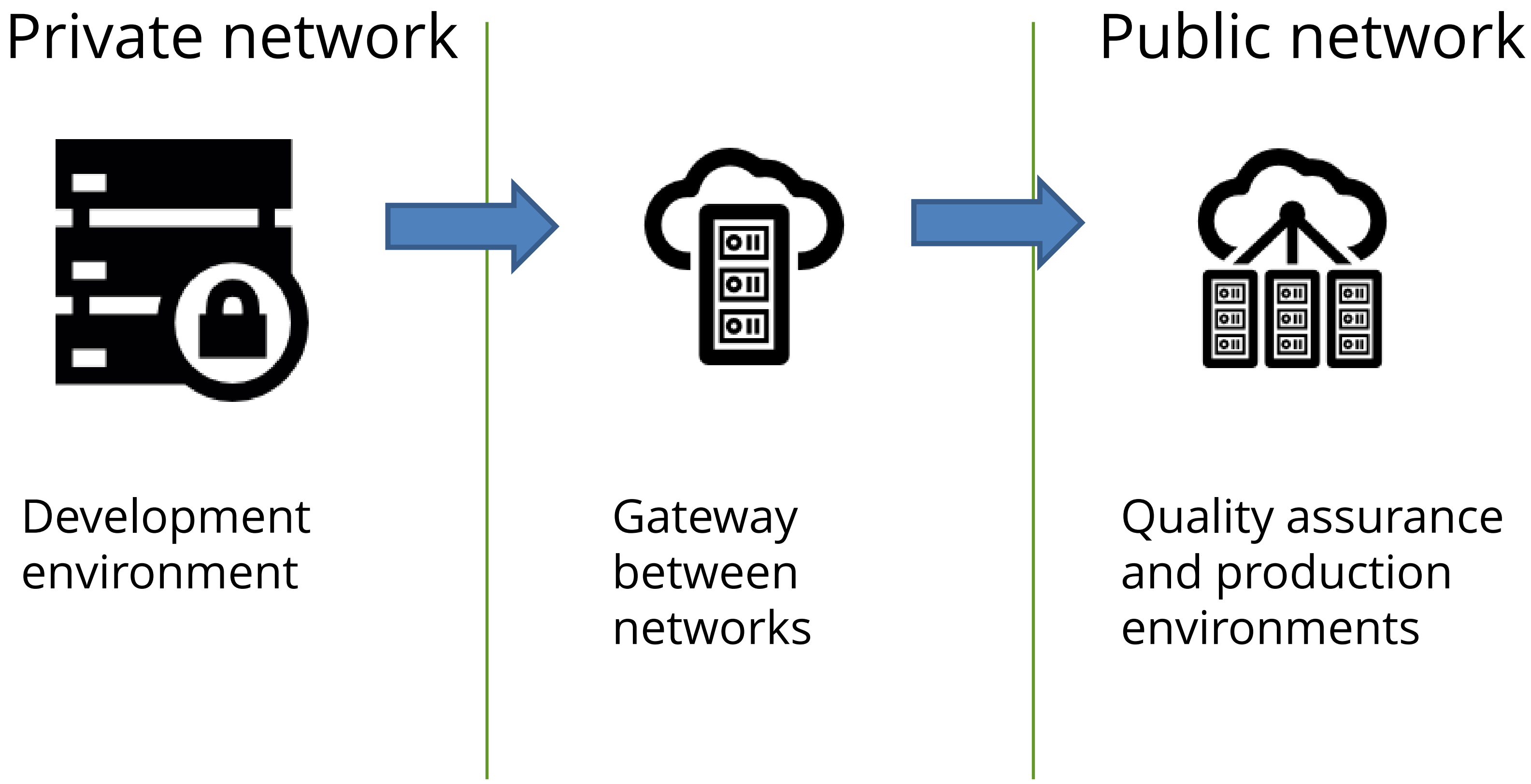
Figure 6-1. Bathtub Curve



Open source tools: *bonnie (disk)*, *stress (CPU)*, *memtest (Memory)*

Bathtub curve: identify Early Failures during the warranty period

Segregate development from production networks



Private network

Public network

Development environment

Gateway between networks

Quality assurance and production environments

Icons designed by Freepik

Reinforced replication policies (#2)

Tape

Offline backup

Bundle backup to tape every 4 months

- ARC files, indexes

Random test recoveries from t

- Data recovery from tape is very slow

Hard disks

Online backups

Redundant server disks (RAID-5)

All data is replicated across 2 independent servers

- ARC files, indexes, software

Daily backup during crawl on live hard disks

- Lose at most 1 day of crawled data

Distant location backups

Tapes moved to distant geographical location

- Lisbon to Porto: 275 KM

ARC files copied to the Internet Archive through the Internet

- Lisbon to California: 9 000 KM

Monitor the service (#3)

Monitoring tools fail

The service is broke
but we didn’t know

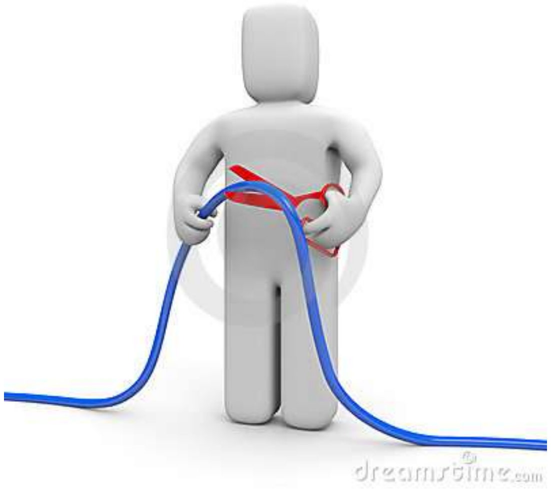
So we did not fix it

Who monitors the monitoring tools?

Use redundant monitoring tools

- Hardware failures
  - Vendor tools are not enough
- Hardware resources
  - Cacti and Ganglia
- Service availability
  - Nagios and Uptime Robot (external)
- Access statistics
  - Awstats and Google Analytics (external)

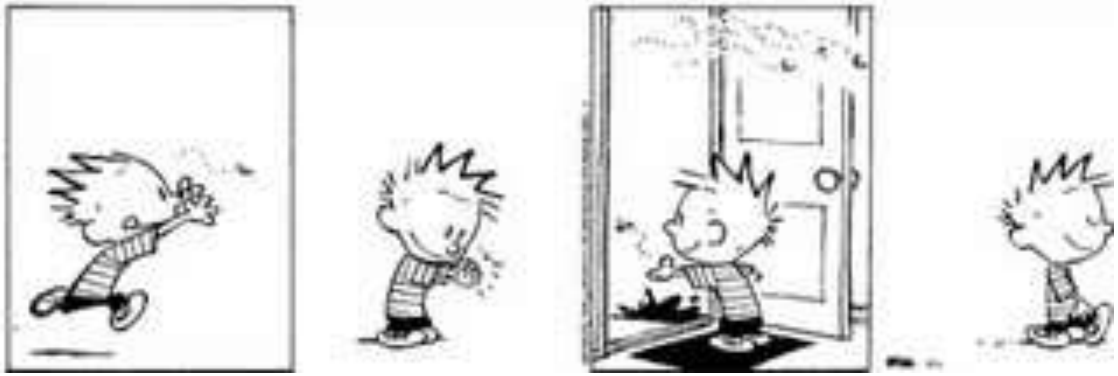
Induce faults to test monitoring!



It’s better to identify problems when you are ready for them

Quality Assurance for software development (#4)

Regression:
"when you fix one bug, you
introduce several newer bugs."



People get tired from doing repeatedly the same (testing).
Computers don’t.

Code testing: automatize

Compilation: the code is well written!

Unit: does what it supposes to do!

Functional: makes the service work

- Simulate user workflows (e.g. search for an archived page)

Many free and powerful tools to automatize testing

- SeleniumHQ, SauceLabs, Jenkins, SonarCube

```
        username: null,
        password: null
    })
},{
    init: function() {
        var self = this;
        this.element.html(can.view('//app/src/views/sign
        this.element.parent().addClass('login-screen');

        App.db.getSettings().then(function(settings) {
            App.attr('settings', settings);
            self.element.find('#login-remember').prop('c

            App.db.getLoggedAccount().then(function(acco
                if(account) {
                    self.options.attr('username', accou
                    self.options.attr('password', accou

                    //account.attr('username') && acco
```

Workload capacity testing: automatize

Establish minimum thresholds for new service release

Jmeter

Workload average: 3 responses/second

Speed average: 5 seconds per response

Security testing: automatize

It’s not “**if** we get attacked”,
it’s “**when** we get attacked”

OWASP Zed Attack Proxy (ZAP)

Expert reviews

Usability testing: conducted by skilled professionals

What is the **use** of a service that **users** cannot **use**?

Identify the problems that **really affect** the service

Most technical problems are reflected on usability obstacles

Help from Human Computer Interaction group from University of Lisbon and UX training

Document and test procedures (#5)

Different types of documentation for different purposes

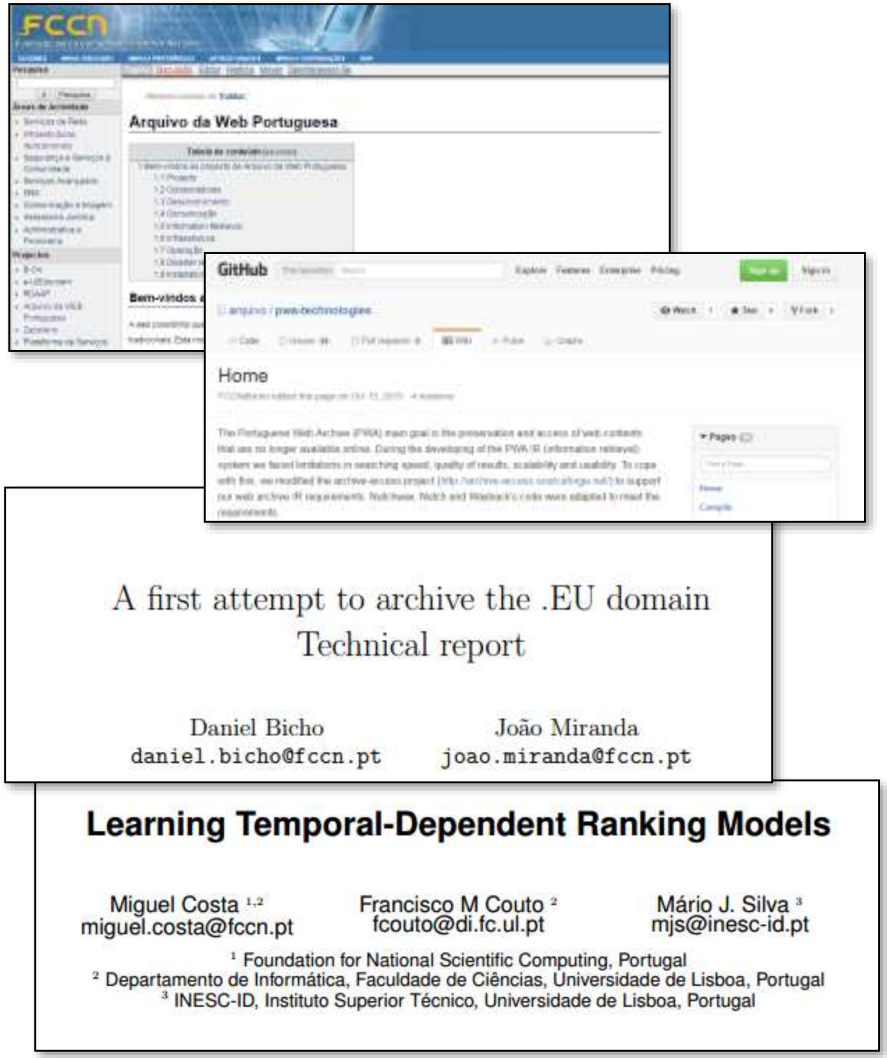
Wiki: internal procedures

GitHub: software

Reports: analysis

Internal and external presentations: collaborations

Scientific and technical publications: peer-review



FCCN

Arquivo da Web Portuguesa

Bem-vindos a

GitHub

arquivo / pwa-technologies

Home

The Portuguese Web Archive (PWA) main goal is the preservation and access of web contents that are no longer available online. During the developing of the PWA IR (information retrieval) system we faced limitations in searching speed, quality of results, scalability and usability. To cope with this, we modified the archive-access project (http://archive-access.sourceforge.net/) to support our web archive IR requirements. Nutchwax, Nutch and Wayback's code were adapted to meet the requirements.

A first attempt to archive the .EU domain
Technical report

Daniel Bicho
daniel.bicho@fccn.pt

João Miranda
joao.miranda@fccn.pt

Learning Temporal-Dependent Ranking Models

Miguel Costa [1,2]
miguel.costa@fccn.pt

Francisco M Couto [2]
fcouto@di.fc.ul.pt

Mário J. Silva [3]
mjs@inesc-id.pt

[1] Foundation for National Scientific Computing, Portugal
[2] Departamento de Informática, Faculdade de Ciências, Universidade de Lisboa, Portugal
[3] INESC-ID, Instituto Superior Técnico, Universidade de Lisboa, Portugal

Test the documentation

Installations of software components from scratch

Procedures executed by colleagues based on existing documentation without help

Open source

everything we do

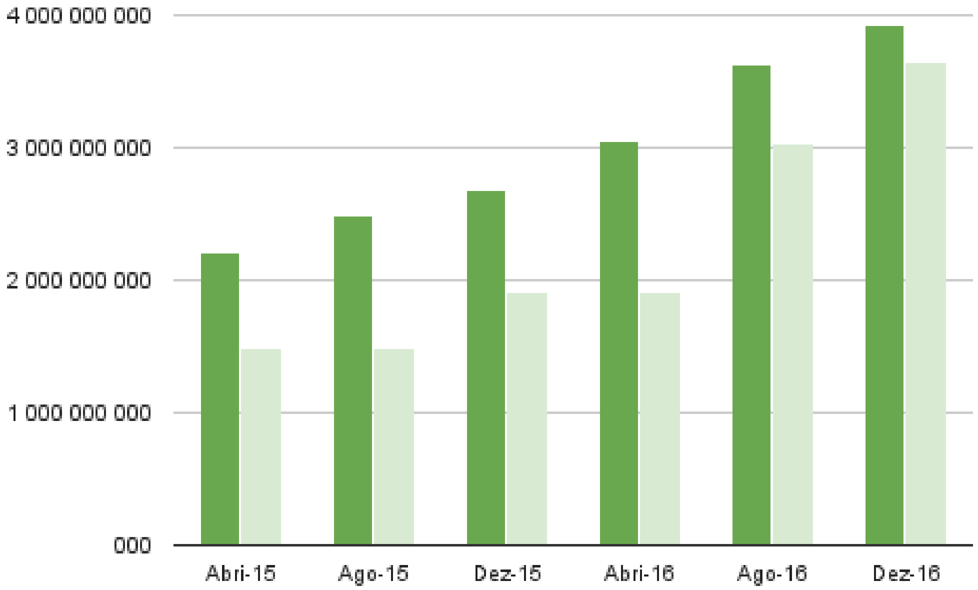
github.com/arquivo

Increases responsibility

Increases software quality

Results

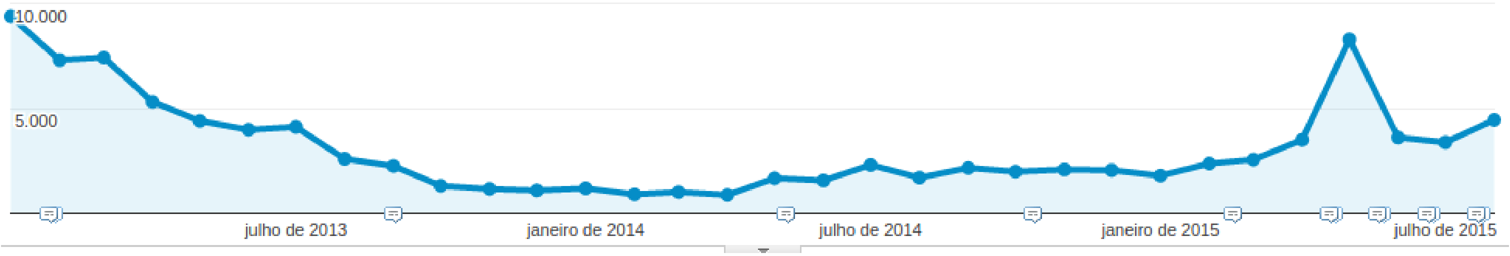
Crawling and indexing are stable



Search availability in 2016

100%

Recovering our users



4 090 users per month (average)

Gaining new users

- 90% are new users

Lessons learned

Strict *Shared-nothing* architecture for hardware and software

Replicate data on multiple distinct media

Software development without proper Quality Assurance leads to waste of resources

Test everything, every time, automatically.

Accept staff rotation and proactively prepare for it

ARQUIVO.PT

daniel.gomes@fccn.pt