Searching images in a web archive

André Mourão[1,2], Daniel Gomes[2]
Arquivo.pt[1,2], NovaLincs[1]
a.mourao@campus.fct.unl.pt

DSAA 2023: Thessaloniki, Greece

October 11th 2023

Fundação para a Ciência e a Tecnologia

Images of Lopez in the green dress were downloaded from the Grammy website 642,917 times in just 24 hours after the event.

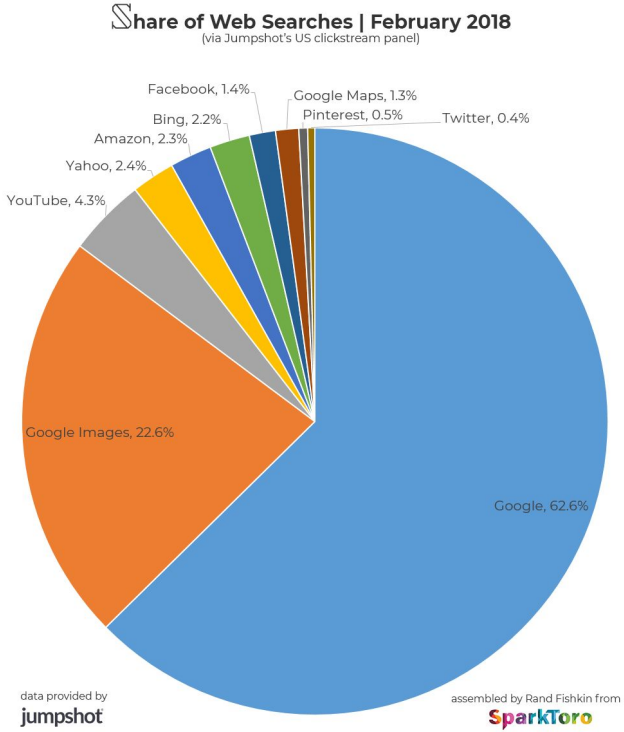
Lee, Michelle (February 11, 2003). *Fashion victim: our love-hate relationship with dressing, shopping, and the cost of style*. Broadway Books. p. 122. ISBN 978-0-7679-1048-4.

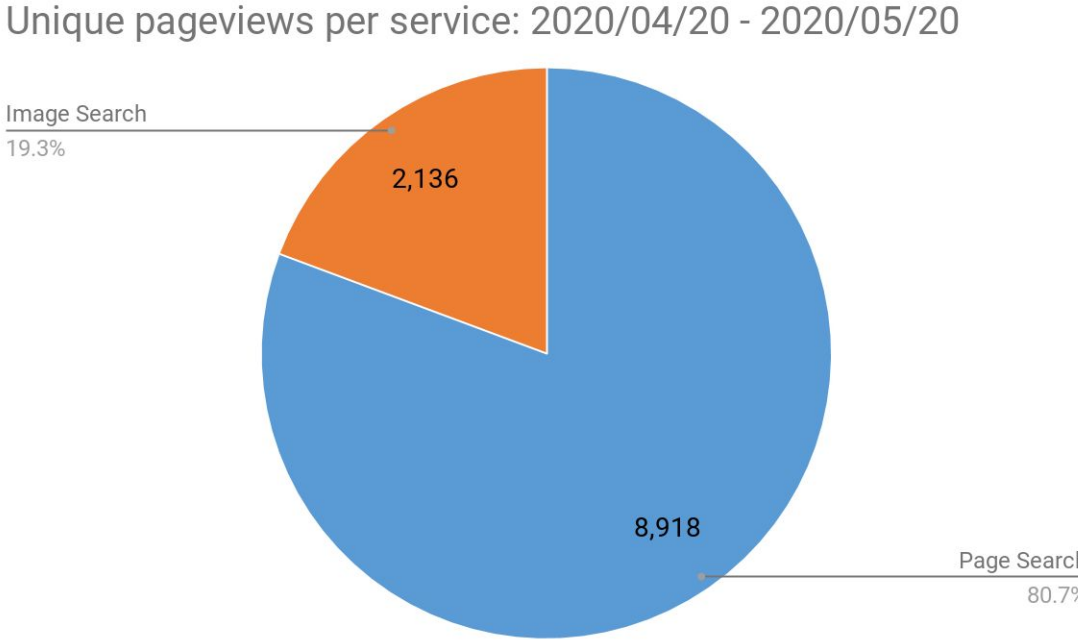
Google

But back in 2000, search results were still just a list of blue links. When the Search team realized they weren't able to directly surface the results that people wanted—a picture of Jennifer in the dress—they were inspired to create Google Images.

https://blog.google/products/search/18-years-after-google-images-versace-jungle-print-dress-back/

Does image search matter today?

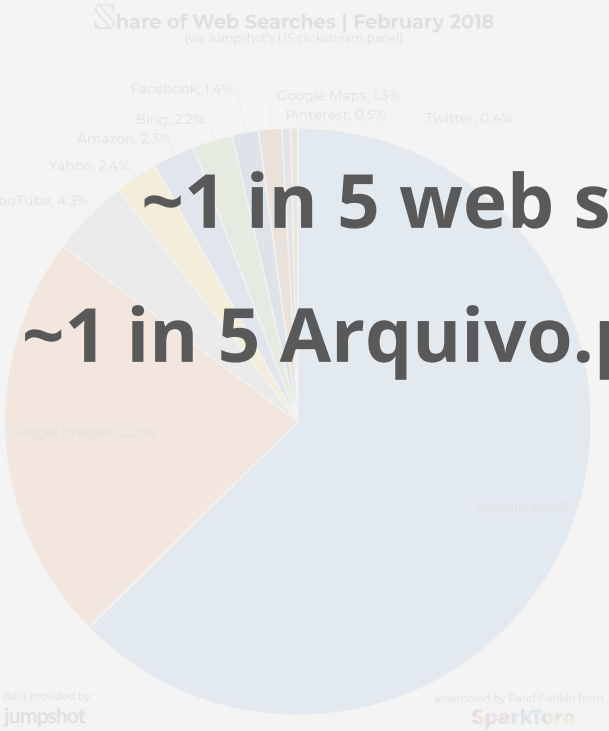


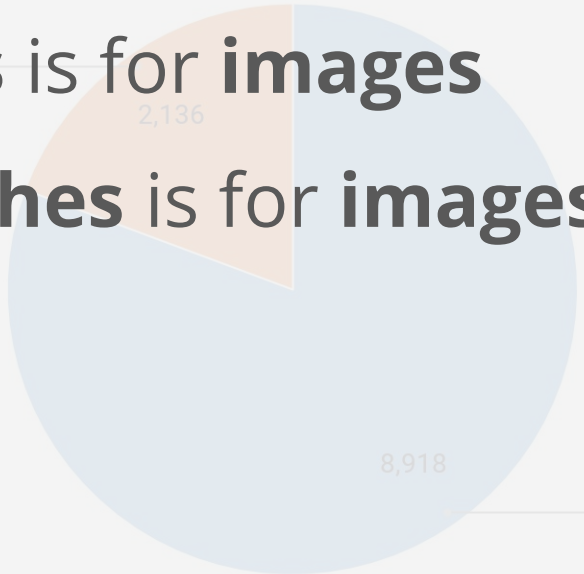


sparktoro.com/blog/new-jumpshot-2018-data-where-searches-happen-on-the-web-google-amazon-facebook-beyond/

Arquivo.pt Google Analytics

Does image search matter today?





~1 in 5 web searches is for **images**

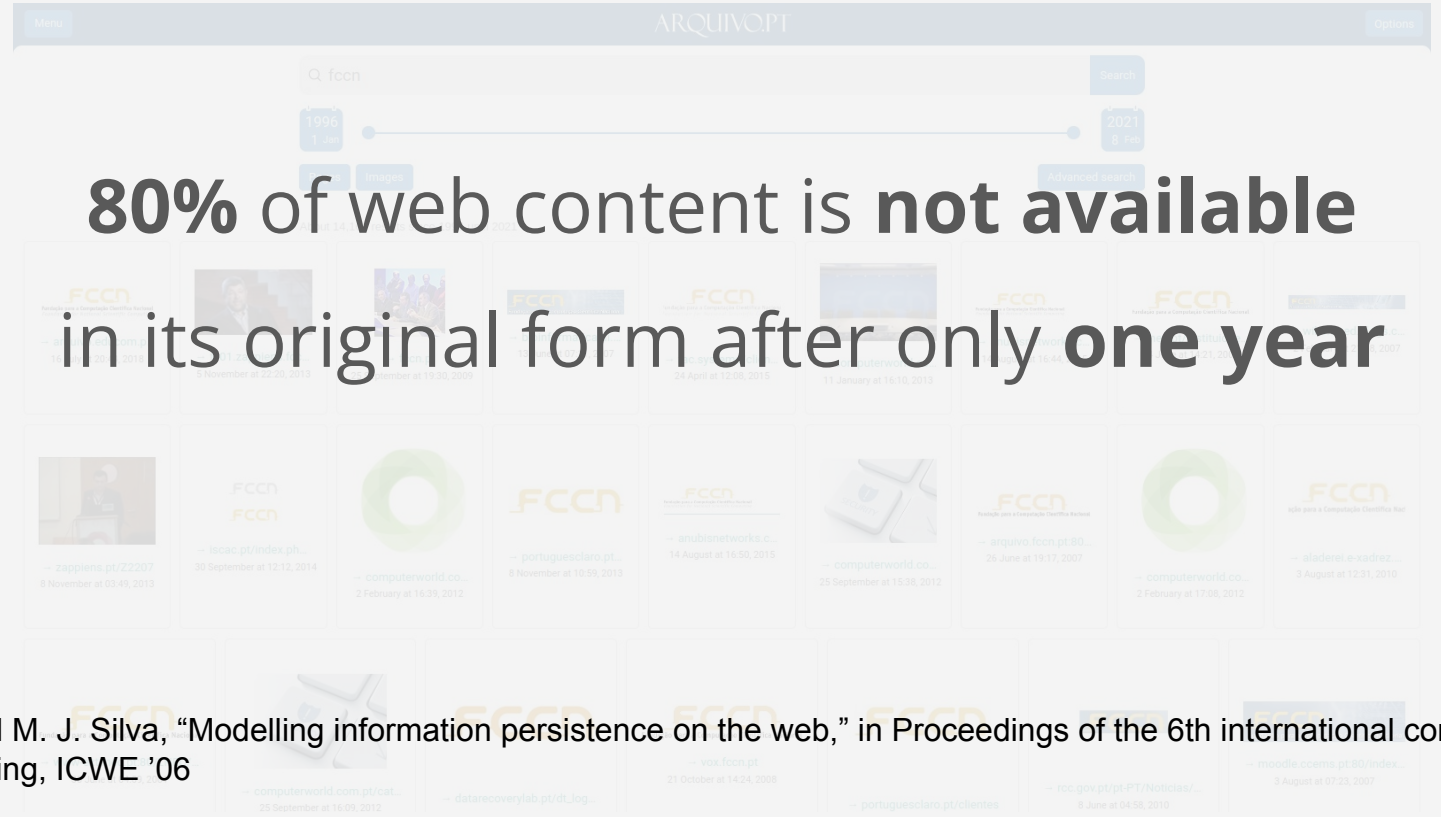
~1 in 5 Arquivo.pt searches is for **images**

sparktoro.com/blog/new-jumpshot-2018-data-where-searches-happen-on-the-web-google-amazon-facebook-beyond/

Arquivo.pt Google Analytics

Arquivo.pt Image Search

Arquivo.pt Image Search



80% of web content is **not available**

in its original form after only **one year**

D. Gomes and M. J. Silva, “Modelling information persistence on the web,” in Proceedings of the 6th international conference on Web engineering, ICWE ’06

Archived image indexing challenges

- Metadata extraction for web content over decades
- Multiple versions of the same image over time
  - Captured more than once
  - Shows up on multiple pages
- Processing and indexing large amounts of data for real-time search

(W)ARC sizes	520 TB
Total collected files	8,500 million
Total collected images	2,408 million
Oldest image date	1994/04/15
Newest image date	2020/11/14

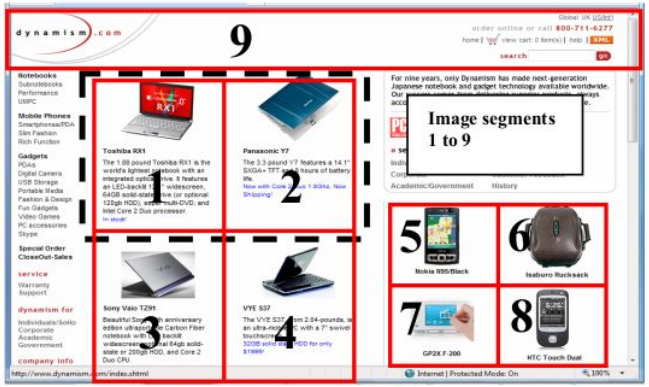
Finding images in pages results

- <img> tag attributes
- <a> tag attributes
- Inline CSS background image
- Inline base64 images
- Images set by JS
- <figure>, <picture>

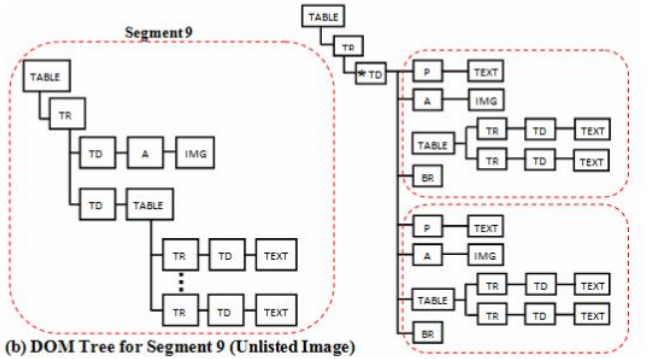
	Percentage of references
<img>	90.6%
<a>	8.7%
css	0.7%

	Percentage of references
imgAlt or imgTitle	49%
URL only	51%

Finding an image caption

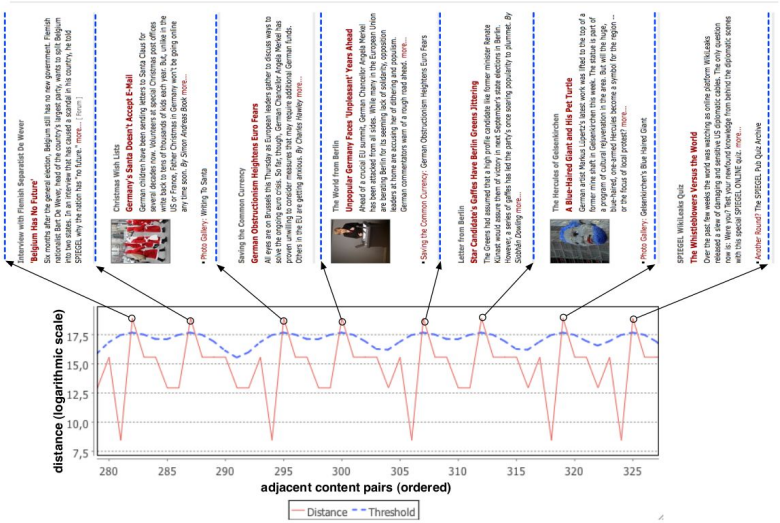


(a) Image segments 1 - 9



(b) DOM Tree for Segment 9 (Unlisted Image)

Fauzi, Fariza & Hong, Jer Lang & Belkhatir, Mohammed. (2009). Webpage segmentation for extracting images and their surrounding contextual information. 649-652. 10.1145/1631272.1631379.



Sadet, Alcic & Conrad, Stefan. (2011). A Clustering-based Approach to Web Image Context Extraction. MMEDIA - International Conferences on Advances in Multimedia.

Image caption extraction

First parent with text

- Default method
- Works well for images in boxes or *reasonably* structured pages

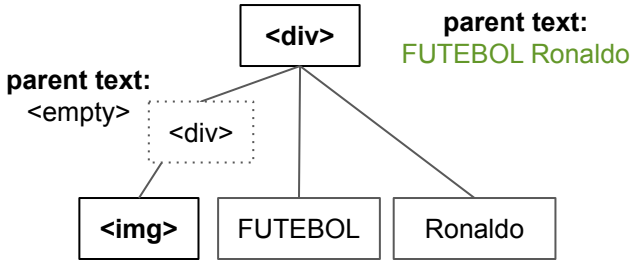


Image caption extraction

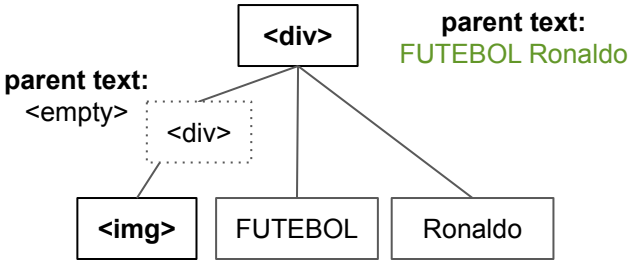
First parent with text

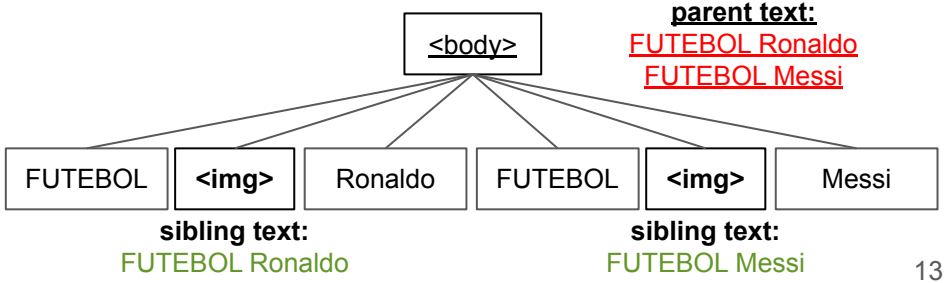
- Default method
- Works well for images in boxes or *reasonably* structured pages

Previous and next node text

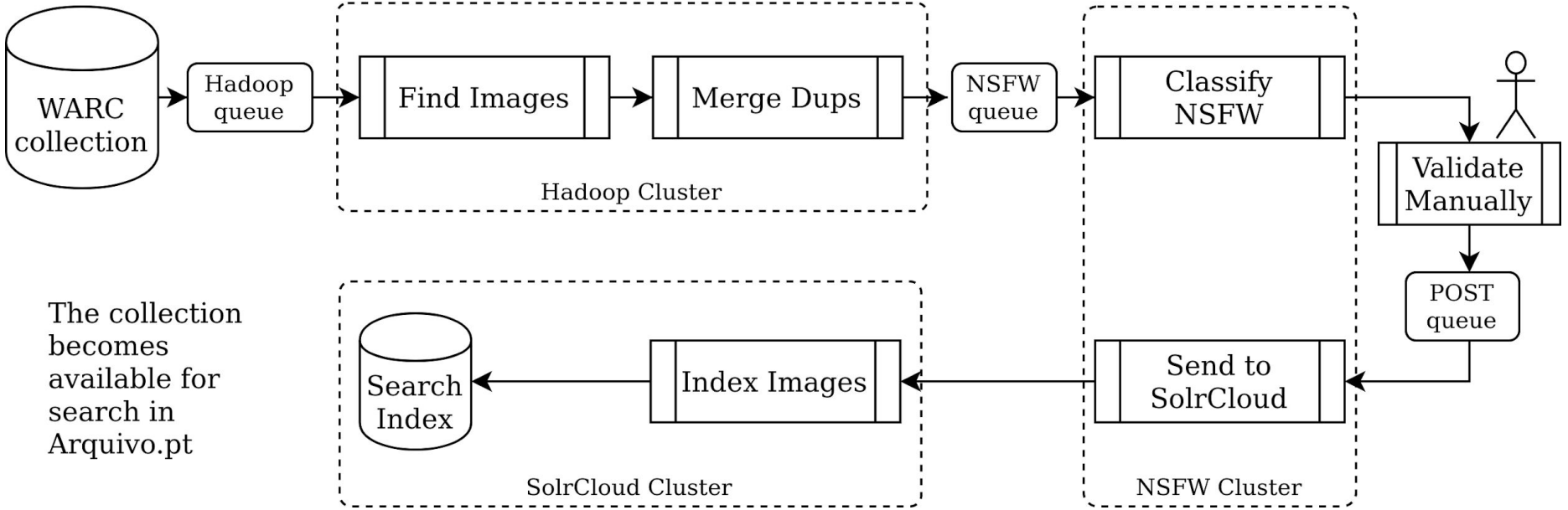
- Used if the first parent with text is at the level of the page with more siblings
- List of images as in a blog

49% -> 99% images with specific metadata

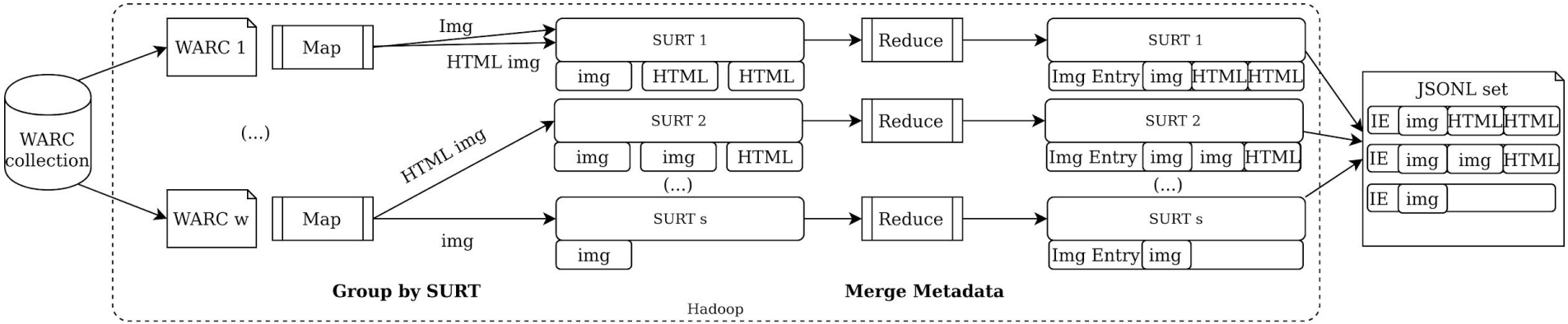




Indexing architecture



Map Reduce: Extract images and metadata



Dealing with duplicate information at scale

- The amount of data produced by this step is huge!
- But most of this information is duplicate
  - Images and pages that were crawled at different times but have not changed
  - References to the images that have the same caption/metadata

Deduplication selected solution

- We arrived at three deduplication scenarios:
  a. every page-image pair is a document
  b. the oldest page that references the image is the canonical document
  c. oldest page information and image specific information from all pages
     - keep reference to oldest page
     - Add all new image specific information (title, alt, caption) to the document
     - replace oldest page reference if a new oldest document shows up

Impact of deduplication

	Number of documents
a	1,862 million image-page pair documents
b	971 million before deduplication across collections
c	**584 million** documents, containing information from all 1,862 million image-page pairs

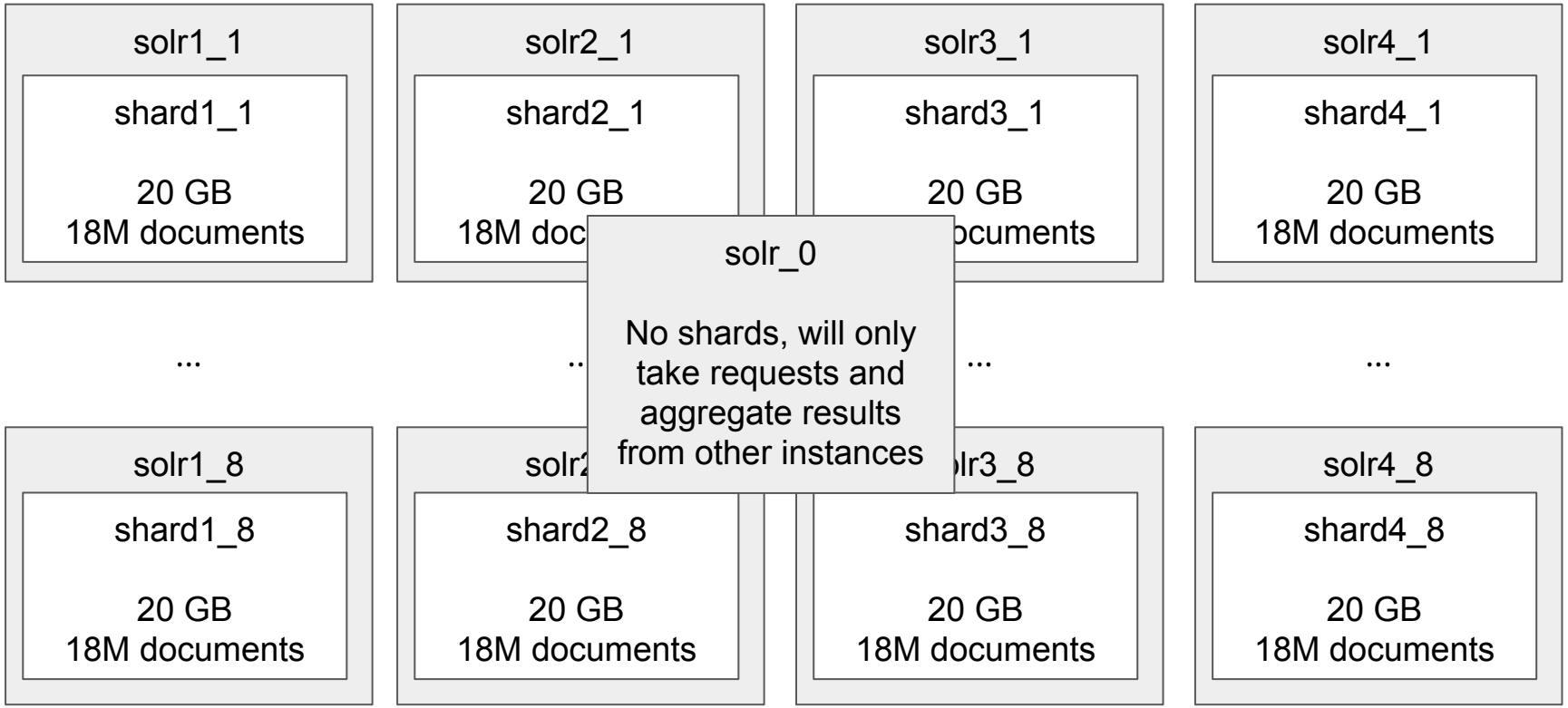
How will we index these **584 million** documents?

Planning SolrCloud resource allocation

- Expected index size: **~720GB**
- SolrCloud servers:
  - 8 servers, 4 per branch
    - **512GB**: p87, p91 (20/40 cores/threads)
    - **256GB**: p82, p83 (12/24 c/t), p93, p94, p98, p99 (20/40 c/t)
  - **2560GB** total, **1280GB** per branch

- No SSD, only HDD, but we have more RAM than indexed data

How we configured SolrCloud?



Realistic query performance test

# requests	Avg.	Med.	$P_{95\%}$	$P_{99\%}$	Throughput
1	115 ms	74 ms	235 ms	769 ms	8 q/sec
3	120 ms	76 ms	259 ms	872 ms	24 q/sec
5	136 ms	85 ms	304 ms	1059 ms	36 q/sec
10	211 ms	128 ms	501 ms	1718 ms	46 q/sec
25	489 ms	266 ms	1297 ms	4334 ms	50 q/sec
50	970 ms	593 ms	2694 ms	6699 ms	50 q/sec

- Random pairs of Portuguese words
- Warmup the index using 50 queries
- Query for 5 minutes and parse the results

Summary

- More images processed
  - 22M -> 1,862M images processed, 584M after deduplication
- More metadata per image
  - 99%+ have image metadata (imgAlt, imgTitle, imgCaption)
- Improved NSFW image processing
  - 7x faster processing (40 -> 280 images per second)
- Improved processing and indexing architecture

Contributions

- Fast image metadata and caption extraction for web images from all over the history of the web
- Impact of the deduplication for web archived data
- Detailed technical system architecture for a live system in the scale of the hundreds of millions

Arquivo.pt

- Arquivo.pt makes **8,000+ million pages** and **1,800+ million images** available for visualization and search:
  - Archived web pages -> **Text Search API**/Memento/CDX Server
  - Text and metadata search -> **Text Search API**
  - Image search -> **Image Search API**
- Available to the general public without registration
- Open Source
- https://github.com/arquivo/pwa-technologies/wiki/APIs