Capacity management study of Arquivo.pt

Ivo Branco
ivo.branco@fccn.pt

July 7, 2020

Contents

List of Figures

List of Tables

1 Executive Summary

The Arquivo.pt preserves 8 925 millions of web files collected since 1 996 and provides a free public research service over this information.

The system needs increasing resources for the crawling, the indexing and the online service. On 7 July of 2020, it has 84 servers, 17,2 Terabytes of Random Access Memory, 1 896 of Virtual Central Processing Units and 3 948,1 Terabytes of gross raw storage.

The Arquivo.pt have 562 Terabytes of compressed ARC and WARC files on two different servers, each server data is preserved on disks partitions configured on RAID 5 or 6.

It is estimate that there won't be any available free disk during the first four months of 2022. It was calculated that the growth ratio of the data storage capacity to preserve the collected web files of the Arquivo.pt collections is 31,30% yearly and on four months period the value is 10,85%.

The online service has two Broker servers that act has intermediate to access the search services and the archived files. Those servers contains the CDXJ indexes used to replay an archived page. Those index grows on a ratio of 31,56% annually and 9,78% on a four months period. It's projected that on January 2022 the Broker servers would require more available disk space to put in production the collections of 2020. This includes a conservative projection for other newly services like the Patching and the Save Page Now.

Related to the Page full text search system it's estimate that they would reach its limit on January 2024 of memory RAM and disk space on the current equipment's.

On the new version of the Image search service it's forecast using an empirical analysis that there won't be any enough memory RAM for the Solr servers by January 2022.

2 Introduction

Arquivo.pt preserves 8 925 millions of files collected from the web since 1 996 and provides a free public research service over this information. Due to the preservation nature of Arquivo.pt, the necessary resources for its operation is constantly growing. For the continuous operation of Arquivo.pt, it is necessary to obtain information about current resources and futures needed for operation, in order to plan equipment's purchase.

3 Current Equipment's

Arquivo.pt has available several servers and storage equipment. For storage it has equipment with the Direct Attach Storage (DAS) typology and for servers it has *Blade* servers and independent servers. The table 1 shows the number of servers per model and their main characteristics. Namely, per server the amount of memory in Gigabytes, the number of Virtual Central Processing Units (vCPUs) and the amount of raw storage in *Terabytes*. Therefore, Arquivo.pt has a total of 17,2 Terabytes of Random Access Memory (RAM), 1 896 of Virtual Central Processing Units (vCPUs), 3 948,1 *Terabytes* of raw storage (excluding storage for the Operating System), spread across 84 physical servers.

Currently the models *IBM BladeCenter HS21* only have a mounted local disk. The Storage Area Network (SAN) have been removed last year. This equipment is considered to be at the end of its

Table 1: Description of Arquivo.pt servers by model with memory in Gigabytes, number of Virtual Central Processing Units (vCPUs), internal gross storage for data in Terabytes and the number of servers.

Model	Memory GB	vCPU	Gross Disk TB	Number of servers
IBM BladeCenter HS21	16	8	0,2	1
	32	8	0,2	23
Dell PowerEdge R710	32	16	12,0	12
Dell PowerEdge R620	448	32	26,4	1
	512	32	26,4	8
Dell PowerEdge R730xd	256	24	49,2	3
	256	24	145,2	18
	512	40	48,6	4
Dell PowerEdge R630	256	40	10,0	2
Dell PowerEdge R740xd	256	40	48,9	10
	512	40	48,5	2
Total	17 200	1 896	3 948,1	84

life, with some servers have ceased to function and therefore were not included in this analysis. At the moment there are 24 servers in operation and its limited disk storage accounts with Operating System. These machines have been used for load balancing, indexing, tests and development tasks. For this reason, despite being considered end-of-life, they were included in the description of current equipment.

There are 6 virtual machines that were not considered in this analysis because they have little capacity and contains auxiliary services. Two to serve Content management system (CMS) with the website sobre.arquivo.pt, other for its pre-production and another for the development of the CMS. There is other virtual machine to be as secondary node for the Load Balancing service and another virtual machine to serve the partner service conta-me histórias.

Each server is configured to have a local redundancy using RAID. Different versions of RAID are used depending on the type of information that is stored. RAID 1 is used for redundancy of the Operating System. The RAID 5 and 6 are used to persist relevant information. RAID 6 have been in use for servers and/or disk boxes acquired after 2018 year. The main storage servers, namely Document Servers, have a mix of disks in RAID 5 and 6.

Not all storage is useful for storing information. In addition to the RAID mentioned above there is a loss of useful space to store the Operating System, Linux swap partition, log files or temporary files.

4 Architecture

Arquivo.pt architecture can be divided on three logical parts: crawling, indexing and online service. The table 2 shows all the server roles, its quantity and the usable disk space in *Terabytes* per type. Next, a brief description of the three logical parts is made.

The crawling is majorly done on specialized servers that run Heritrix, Brozzler and/or arquivo-patcher. Currently there are three machines with Heritrix: one for Four-month crawl of the .pt top domain (AWP-PT); other for the Daily crawls (FAWP); and a last server for Special crawls (EAWP) and Four-month crawl outside of the .pt domain (AWP-ForaPT). There are two with Brozzler: one to be used on the Memorial service and another one for the experimental Monthly

Table 2: Type of servers, its quantity and the usable disk space in *Terabytes*.

Server type	**Number of servers**	**Usable disk space in TB**
Broker server	2	14
Query server	10	194
Solr server	2	74
Document server (indexing w/ Hadoop version 0.14)	18	2034
Crawler server	6	222
Indexing server (Hadoop version 3)	11	197

crawl (MAWP).

The indexing have been running on a mixed of old hardware and machines that run the online service. During indexing the server resources like CPU and memory RAM are consumed very intensively. To minimize the impact on the online service, only the Document Servers are shared between the indexing and the online service. Currently there are two clusters during indexing: one with Hadoop version 0.14 and another with Hadoop version 3. The older version for the full text page search and newest version for image seacrh. It's planned on the next major upgrade of the full text page search the merge of the two cluster on a recent version of Hadoop version 3.

The online service has more components. It have two branches that act has a mirror of each one. Each branch consist of one Broker server, five Query servers for full text page search, one Solr server for image search and nine Document servers to store the archived information on ARC and WARC files. The web requests to search or view an archived page or image is load balanced on an active-passive cluster, then the request goes to one of the Brokers servers, and then, depending of the request, to one or multiple back-end servers: Query, Solr or Document server.

5 Indicators for analysis

Each server and logical component requires an independent capacity analysis. However some aren't just as stressed as others or couldn't be predicted has others.

The page full text search is one of core services of the Arquivo.pt it permits the search on any archived page using a text term. A further analysis of this system is made on section 8 specifically the growth of the Arquivo.pt Query servers.

The current image search system is still in beta, so it's expected to have major changes on it. A new version of the system is being in development that would detect many more images on the same archived information. Nevertheless a detailed analysis of the current system and an empirical analysis on the new version have been made on the section 9.

About the crawlers, as long the available time to crawl is enough to collect everything on the current servers, we could maintain the same architecture. The network, CPU and memory RAM currently aren't an issue because the servers didn't spend all available time on crawling. However, if necessary, the crawling jobs of FAWP, AWP-ForaPT and EAWP collections could be split-ed on different servers, dividing the seeds URLs on multiple crawling jobs, this without a major configuration change of software. The AWP-PT crawling job couldn't be splitted without a major configuration change of the Heritrix software. The current limit is bandwidth and time available for the next crawl. Although a performance and tuning of each crawler job is out of the scope of this document, a simpler analysis could be made. The bigger crawling job that the Arquivo.pt have is the Four-month crawl of the .pt top domain. On 2019 year the crawler server only have been in use during 6 months, about 2 months for each AWP-PT, so there is available hardware for growth. Currently, all the collected ARC and WARC files of any collection could still fit on any Crawler server, however if needed they could be moved to other servers when they are full (currently those files contains about 100 Megabytes).

The Document servers, like described before, contain the collected ARC and WARC files. They are also used for indexing using the Hadoop software. It takes time to index a collection, but the time is much lower than the available time to index for the collections that are being downloaded. The metric that is stressed is the available data storage capacity required to preserve the crawled collections, a further analysis is done on the section 6.

To analyse the capacity management of the Broker servers a full load test for each software component should be made but, for brevity, this report is only to inspect the data disk on this servers. The analysis is in section 7.

6 Data storage capacity for collections

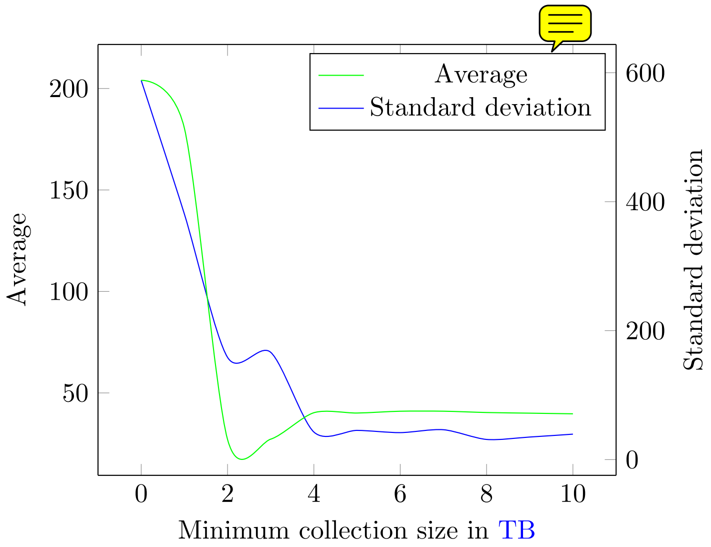
The Arquivo.pt preserves 562 TB of useful files collected from the web since 1 996. More information is downloaded daily by Arquivo.pt, requiring more available disk space. This section investigate the requirements of the disk to save the collected information. It's captured on the crawler servers and it's copied to the document servers waiting to be indexed and read when its indexes are put on live. The collected information is stored on compressed files using ARC or WARC formats. The sum of all the usable disk space of the Document servers to store information is 2 034 Terabytes spread on 18 physical servers, but because each file is stored on two mirrored servers, then it's only possible to save 1 017 Terabytes of unique data.

The crawled information is aggregated on a Arquivo.pt collection, with a unique identification prefixed with a code like AWP, FAWP or EAWP and suffixed with a incremental number. For donated collections is used a short name associated the information that it contains. Most collections contains captured files related to a single crawl job. Nevertheless there are many collections that have a small size with a long time duration.

To predict the requirements of disk space, an analyze of the collected data needs to be done. Ideally an analytic could be done for each record stored on Arquivo.pt, because each record has a crawled date and size. A possibility approach is to process all the 4,8 Terabytes CDXJ indexes. On this report it was used a simpler approximation, it was considered the begin of the crawling job. On figure 1 is visible that when the minimum size in Terabytes of a collection increases, then the average and standard deviation stabilize with values around 30 and 75 days. Those values demonstrate that is good enough to use this approach to estimate the growth of the collections.

Using the described approach, it's possible to calculate the growth of the disk space in Terabytes required to persist the collected files. On a four months period, it has an median of 10,85% and an average of 12,13%, yearly the median increases to 31,30% and the average increases to 30,74%. The figure 2 shows the evolution of the percentage of growth of the disk space required to persist the collected files on a four months period and annually, it's possible to conclude visually that the median values are good values for the estimate of growth: 10,85% on four months period and 31,30% yearly. The yearly growth rate is three times greater than the value assumed by [1], because not only the web growth with more resources but also each collected resources is on average greater. So in reality an estimate of growth of the required disk space of a web archive needs to account this two dimensions at the same time.

Figure 1: Minimum collection size in TB, its average and standard deviation of all population.



Using the estimate growth described previously it's possible to estimate the evolution of the occupied disk space to preserve the collected files. The figure 3 shows a graph with the number of accumulated disk space in Terabytes on each four-month time period, an estimate until May 2022 and the available useful disk in Terabytes to persist them. The conclusion is that the disks should be full during the first four months of 2022.

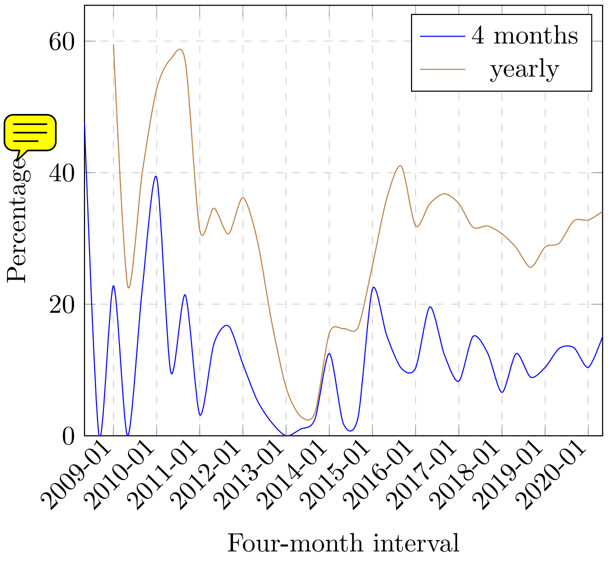
Another important consideration is that on the current architecture isn't possible to use all the available disk on the Document servers. Because the Arquivo.pt sub systems requires that every ARC or WARC file of the same collection should be put on the same server (there is another copy of collection on its mirror Document server). Nevertheless, it's possible to move the smallest collections to fit on blank spaces. This would require some time to copy, manage and apply the required configuration changes. Also, the indexing processes that run on the Document servers also need some available disk space, if one server doesn't have any available disk, the indexing processes could run unbalanced without using all the available resources.

To compensate the difficulty of use all the entire Document server disk, it's possible to disable the copy to the Document servers and persist temporary only on the Crawler server, this measure of last resource could only be used for the current crawled collection, taking the risk of only have one copy of the data.

7 Data storage capacity index CDXJ

The information collected by Arquivo.pt is indexed, one of those indexes are the CDXJ indexes. Those indexes are used by the Pywb service to replay the archived pages. The read performance of those indexes is crucial in order to have a good replay experience of the archived content. Currently, those indexes are stored on Solid State Drive disks on the Broker servers. The CDXJ indexes grows with the number of collected files, because each index file contains a line for each collected resource. Given the increase of the web in general a forecast of the CDXJ files is very important. The forecast is based on the generic Arquivo.pt crawling configuration of three AWP and four FAWP per year with a budget of 10 000 resources per host.

Figure 2: Evolution of the percentage of growth of the disk space required to store the collected files on 4 months period and annually.



The Arquivo.pt currently has one year embargo period, this means the minimum time to put a crawling collection in production is one year. Another important consideration is the current working method of Arquivo.pt to put the majority of the collections in production only in January. So in practice the embargo period has an average of 1,5 years - minimum value of 1 year and a maximum value of 2 years.

The current backlog of CDXJ indexing is none for collections started to crawl after on the 1 January, 2020. Some small crawling jobs are still active, but given its small size, they won't put a stress on the storage capacity.

Another important consideration is the rest of the software that is run on the Broker servers, particularly those that its data could growth rapidly. The Patching and the Save Page Now features, given its natures of fix an archived page with missing resources and to preserve an entire page and its resources, could require many disk to save information. A page with big resources like a couple of videos could demand many disk so save them, this on multiple pages could put a stress on the available space on the Broker servers. Currently this information is copied automatically on every business day from the two Broker servers to a pair of Document servers, so normally the Arquivo.pt doesn't store too much backlog related to this information. Given the novelty of those features and its particularities is very difficult to predict the minimum reserved space for those functionalities. Nevertheless, from 19 September of 2019 until 21 December of 2019 the patching saved 39 GB and from 1 January 2019 to 6 July of 2020 389 GB.

Assuming the working method described before, the median of growth of the disk space required to store on Broker servers the CDXJ indexes of a four months period is 9,78% and with an average of 10,48%, annually the values increases respectively to 31,56% and 28,18%. The figure 4 displays the accumulated disk space in Gigabytes required to store the CDXJ indexes, its growth (using the ratio factor of 9,78%), the accumulated occupied disk space in Gigabytes using the embargo time difference and its growth (using an annual ratio factor of 31,56%).

It's forecast that on September 2021 the crawled CDXJ indexes shouldn't fit on the current brokers servers, but because the embargo period, only on January 2023 the Broker servers won't have any space left to store those indexes. Nevertheless, if we account the required disk space for other services, like the Patching and Save Page Now, or if the Arquivo.pt needs to put a collection in production before the embargo methodology described before, the January 2022 is a better limit date.

8 Page full text search

For each new collection that Arquivo.pt puts into production, extra resources are needed to make the collections searchable. The Page Search Full Text System is the most demanding component in terms of resources. In addition to storage for indexes, it takes a lot of memory RAM and CPU power to get the search results in an acceptable time for the billions of documents available to search. It is necessary to estimate the annual consumption of these resources, as well as the quality degradation of the Search System as more content becomes available.

The indexing system generates a different Page Search Full Text index for each Arquivo.pt collection. Each index is placed on the Query servers and run a Java $^{\text{TM}}$ process that include the Nutchwax software.

Figure 3: Prediction of data storage needed to save collections on each branch excluding RAID.

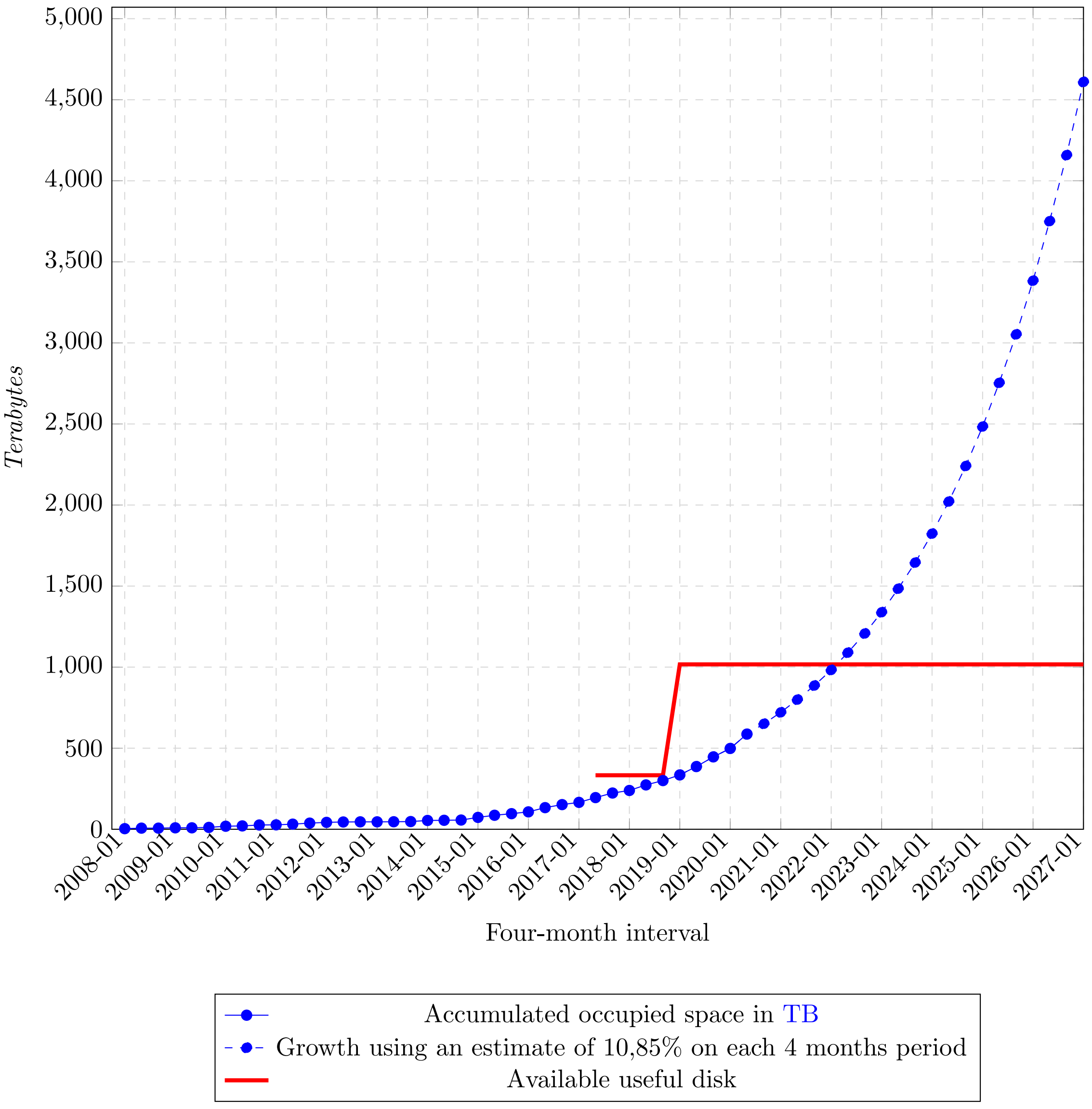


Figure 4: Prediction of data storage needed to save the indexes CDXJ on each Broker server excluding RAID.

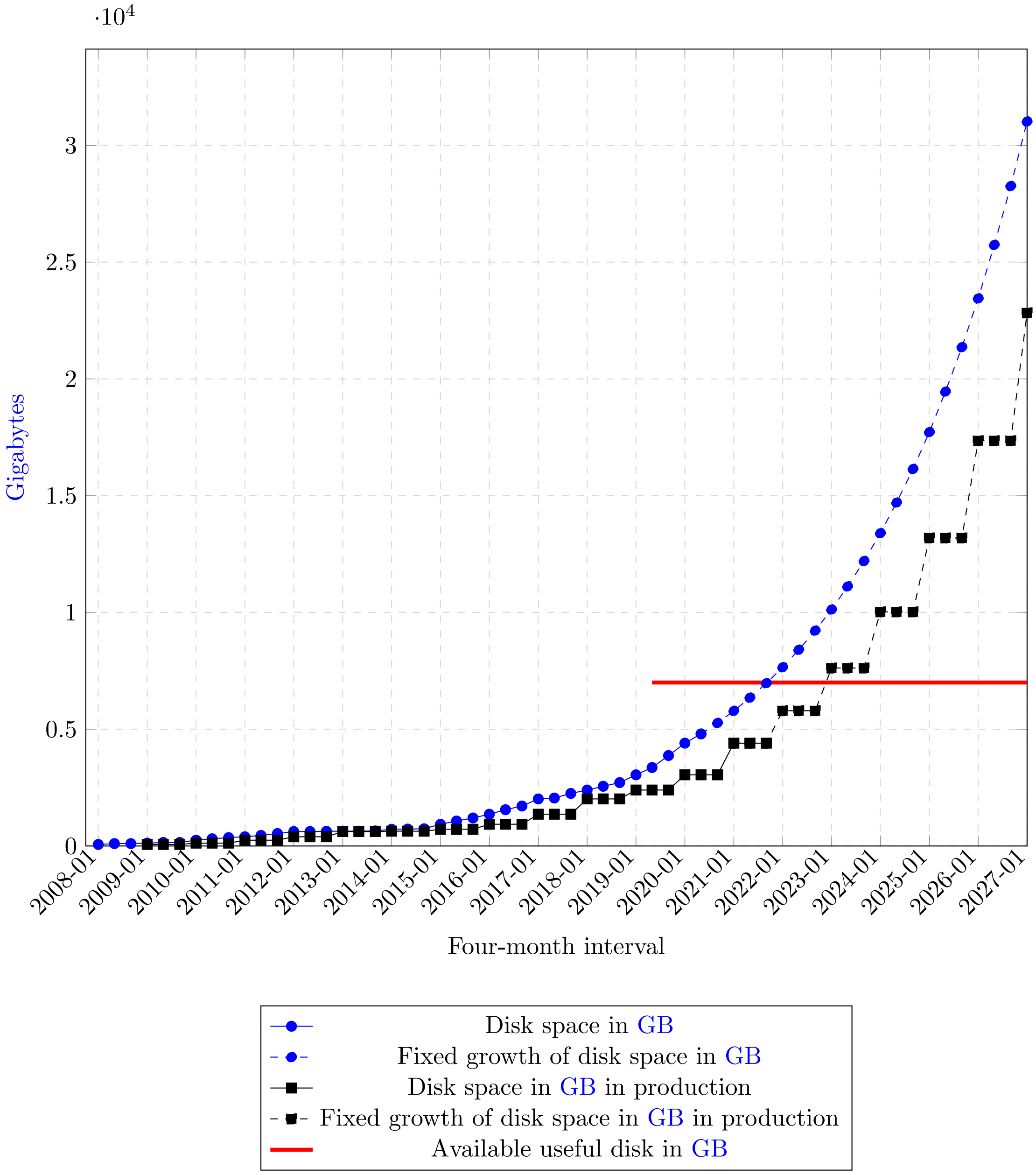
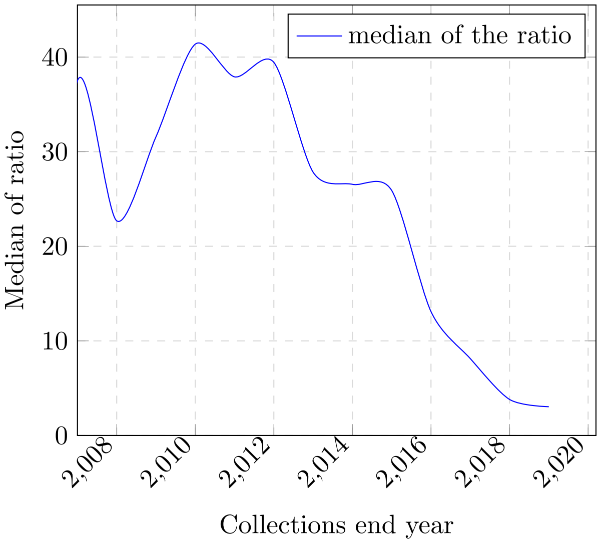


Figure 5: Yearly evolution of the median of the collections ratio of page search full text search index disk space and the collection preserved disk space



Currently the collections that have been indexed using the Page Search Full Text index have a total disk space size of 453,40 TB and a 64,66 TB for its indexes. The ratio of the size of the indexes with the size of its corresponding documents has a median of about 18,03 % and an average of 21,37 % for all the collections indexed by the Arquivo.pt, this value is substantial lower than the 36,9 % calculated by [1]. As the collections grows for more recent years the ratio of the disk space required to save the page search full text indexes have being decreasing. On figure 5 it is possible to view this consistency decreasing. On 2019 the value is 3,04 %.

The Page Seach Full Text indexes run on the Arquivo.pt Query servers. The group of servers has six Dell PowerEdge R620 and two Dell PowerEdge R730xd. The first ones have 15 TB of usable disk space and 32 vCPU. The second ones have 37 TB of usable disk space and 40 vCPU. Each server have 512 GB of memory RAM. Because of the Arquivo.pt architecture of having two mirror branches, there are only half available per each index copy. In total, per branch there are 2.5 TB of memory RAM and 97 TB of usable disk space to save the indexes.

Table 3: Forecast the growth of the page search full text index in Terabytes.

Until Collection end date	**Forecast page search fts idx size TB**
2019-01	59.67
2020-01	64.63
2021-01	71.41
2022-01	79.35
2023-01	90.17
2024-01	104.91
2025-01	121.80
2026-01	152.35
2027-01	189.62

Table 4: Forecast Query servers memory RAM in Terabytes, using all collections crawled until the date and for production using one year embargo period.

Collection date	**Collection in production**	**Memory RAM in TB**
2020-01	2021-01	1.15
2021-01	2022-01	1.46
2022-01	2023-01	1.85
2023-01	2024-01	2.34
2024-01	2025-01	2.97
2025-01	2026-01	3.76
2026-01	2027-01	4.77

The page search index occupy 59,67 TB of disk space on the Query servers for the collections that have finished the crawl at the end of 2018. Those correspond to 334,64 TB of indexed collected information. Using the 3,04 % for the index relative to the collected information for next years, it's forecast that there will be available disk space for the 2023 crawled collections. But, because of Arquivo.pt one year embargo period, those indexes are put in production only on January 2024. On table 3 it's possible to view a more detailed value per each year and an estimation for further years.

Using the estimation of 1 Gigabytes of memory RAM for 7 millions of indexed documents, for each index, it's calculated that on January 2021 it's needed 1,15 TB of memory RAM, on January 2022 1,46 TB, on January 2023 1,85 TB and on January 2024 2,34 TB. This estimation use one year embargo period and add new indexes only in January for the Page Search Full Text service, more details and a forecast until 2027 on table 4. So the current limit of memory RAM would be reached on January 2024 with collections indexed and crawled until January 2023.

9 Image search

The image search service is consisted by two major software components: the image search indexing and the image search API. The indexes are put on a Apache Solr. The image search indexing transforms the ARC and WARC files of an Arquivo.pt collection into a single JSONL file. The file is formed by list of structures where each row, or grain of information, is a single indexed image. Each row has every information that is relevant to the image search API to expose a search service where it is possible to search the archived images using different search criteria.

A new version of the Image search have been in development during the last couple of months. The current version have 22 millions of indexed images, but on the new version it's forecast to would have about 900 millions for the Arquivo.pt collections. This is increment of 4 000 % in the number of indexed images on the same information. It's estimate that the JSONL files would occupy about 10 - 14 Terabytes and the Apache Solr indexes 2,3 - 3,8 Terabytes. Nevertheless, currently, the information isn't all indexed, so the final values could still change.

The current version of the Image search system would have the capacity for many years. But the new version, because of such prevision of increment of the number of indexed images, would need a fine analysis of the hardware requirements. Currently, because we don't have everything indexed, it's difficult to predict the requirements. Nevertheless, an empirical test have been made. The empirical test was performed on a 256 GB of memory server with 226 millions of indexed images. Comparatively with the current production environment, this is an increment of 10x of the indexed images. The performance was slower than the current production environment, but still usable. This gives a ratio of per million of indexed image per Gigabytes of memory RAM of 0,88. So for the new version of the system, using the 900 millions of indexed images, it is forecast that the system would need 792 Gigabytes of memory RAM per branch.

Because the forecast of 792 Gigabytes of memory RAM for the new version of the image search service is greater than the maximum memory value of any Arquivo.pt server, then a change in architecture would need to be made. Fortunately, the Apache Solr supports a distributed search using the SolrCloud component. This component includes the ability to set up a cluster of Solr servers that combines fault tolerance and high availability with the capabilities to provide distributed indexing and search.

It was calculated that the median of the annual increase of indexed images is 28,46 % for the current system. A new calculation is need after the indexing of all the images using the new version of the image search service. But using the old value has estimation, the new system would need 1 Terabytes of memory RAM after one year, the same value of memory than the current hardware available for this service.

So using the previous described embargo period it's estimate that there won't be any memory for image search Solr servers by January 2022. However it is warned that there is a high uncertainty regarding this date. Further test would be need to have a more finest estimation.

10 Conclusions

The growth ratio of the data storage capacity to preserve the collected web files of the Arquivo.pt collections is 31,30% yearly and on four months period the value is 10,85%. It is forecast that there won't be any available free disk to copy the crawling information to the Document servers during the first four months of 2022.

About the CDXJ indexes, that are used and stored on the two Broker servers, it was calculated that its growth ratio is 31,56% annually and on four months period has a value of 9,78%. It's projected that on January 2022 the Broker servers would require more available disk space to put in production the collections crawled until the end of 2020. This includes a conservative projection for other services like the newly Patching and the Save Page Now services.

Related to the Page full text search system it's estimate that they would reach its limit on January 2024 of memory RAM and disk space.

It is foreseen, and using an empirical analysis, that there won't be any memory on the Solr servers for the new version of the Image search system by January 2022.

References

[1] D. Bicho and D. Gomes. Estudo de gestão de capacidade do arquivo.pt. Technical report, Arquivo.pt, June 2015.

Acronyms

ARC ARC is a lossless data compression and archival format used by web archives to store the collected data prior of the standardization of WARC file format. 4, 6–8, 17

AWP Four-month crawl of the .pt domain and outside of the .pt domain. 7, 8

AWP-ForaPT Four-month crawl outside of the .pt domain. 5, 6

AWP-PT Four-month crawl of the .pt top domain. 5–7

CMS Content management system. 5

CPU Central Processing Unit. 6, 10

DAS Direct Attach Storage. 4

EAWP Special crawls. 5–7

FAWP Daily crawls. 5–8

GB Gigabytes. 3–5, 9, 12, 14, 17

MAWP Monthly crawl. 5, 6

MB Megabytes. 7

OS Operating System. 4, 5

RAID Redundant Array of Independent Disks. 2, 4, 5, 11, 12

RAM Random Access Memory. 3, 4, 6, 10, 14, 16–18

SAN Storage Area Network. 4

SSD Solid State Drive. 8

TB Terabytes. 3–8, 11, 14–18

URL Uniform Resource Locator. 6

vCPU Virtual Central Processing Unit. 14

vCPUs Virtual Central Processing Units. 3–5

WARC The Web ARChive file format. 4, 6–8, 17