



UNILAB
UNIVERSIDADE DA INTEGRAÇÃO INTERNACIONAL
DA LUSOFONIA AFRO-BRASILEIRA
INSTITUTO DE HUMANIDADES E LETRAS DOS MALÊS
LICENCIATURA EM HISTÓRIA

ANA CAROLINA DE OLIVEIRA VELOSO

**A HEMEROTECA DIGITAL BRASILEIRA NA PESQUISA HISTÓRICA: UMA
ANÁLISE DOS SIMPÓSIOS NACIONAIS DE HISTÓRIA DA ANPUH (2013-2019)**

SÃO FRANCISCO DO CONDE

2022

ANA CAROLINA DE OLIVEIRA VELOSO

**A HEMEROTECA DIGITAL BRASILEIRA NA PESQUISA HISTÓRICA: UMA
ANÁLISE DOS SIMPÓSIOS NACIONAIS DE HISTÓRIA DA ANPUH (2013-2019)**

Monografia apresentada como Trabalho de Conclusão de
Curso da Licenciatura em História da Universidade da
Integração Internacional da Lusofonia Afro-Brasileira
(UNILAB).

Orientador: Prof. Dr. Eric Brasil Nepomuceno.

SÃO FRANCISCO DO CONDE

2022

Universidade da Integração Internacional da Lusofonia Afro-Brasileira
Sistema de Bibliotecas da Unilab
Catalogação de Publicação na Fonte

V555h

Veloso, Ana Carolina de Oliveira.

A Hemeroteca Digital Brasileira na pesquisa histórica : uma análise dos simpósios nacionais de História da ANPUH (2013-2019) / Ana Carolina de Oliveira Veloso. - 2022. 28 f. : il. color.

Trabalho de Conclusão de Curso (Graduação) - Instituto de Humanidades e Letras dos Malês, Universidade da Integração Internacional da Lusofonia Afro-Brasileira, 2022.

Orientador: Prof. Dr. Eric Brasil Nepomuceno.

1. Pesquisa histórica - Brasil. I. Associação Nacional de História (Brasil) - Congressos - 2013-2019. II. Biblioteca Nacional (Brasil). Hemeroteca Digital - Pesquisa. III. Título.

BA/UF/BSCM

CDD 981.0063

AGRADECIMENTOS

Primeiramente agradeço a Deus pelo dom da vida. Em segundo as pessoas por quem vivo, Ana Beatriz da Cruz Veloso e Ana Ariel de Oliveira Carvalho, respectivamente irmã caçula e filha. Ao Vinicius Carvalho (pai da minha filha) por todo apoio, carinho e paciência. A todos os amigos e colegas que estiveram comigo durante esses últimos 5 anos.

Aos meus professores pelo conhecimento construído em sala (presencialmente e no remoto), nos corredores do Malês, palestras, rodas de conversa, grupos de pesquisa, etc. Em especial ao meu orientador Eric Brasil por escrever o projeto de pesquisa que tem como um de seus frutos essa monografia, por confiar no meu potencial mesmo quando eu o neguei, por me permitir ser sua bolsista para desenvolver essa pesquisa maravilhosa. E a minha amiga Priscila Valverde Silveira, bolsista Pibic/Unilab projeto “Pós-Abolição: sentidos usos e produção acadêmica (2004-2019)” também orientada pelo prof. Eric Brasil, pela parceria, momentos de discussões sobre o tema e incentivo.

Estendo meus agradecimentos a Fundação de Amparo à Pesquisa do Estado da Bahia (FAPESB) pela bolsa de financiamento, ao PIBIC-Unilab, a Pró-Reitoria de Pesquisa e Pós-Graduação da Unilab, também agradeço ao Laboratório de Humanidades Digitais da UFBA pelo suporte, pelos debates, que contribuam diretamente no desenvolvimento da pesquisa.

Agradeço a Anita Lucchesi e ao Paulo Alves Junior por aceitarem o convite de compor a banca avaliadora desta defesa.

Não posso deixar de agradecer aos artistas que foram minha trilha sonora nesses últimos 2 anos, Min Yoon-gi (Suga-BTS), Jay Park e Sam Kim (샘김), a arte deles foi fundamental nesse momento de pandemia da Covid-19.

RESUMO

Esta monografia consiste no Trabalho de Conclusão do Curso de História, pela Universidade da Integração Internacional da Lusofonia Afro-Brasileira. Este trabalho tem como objetivo apresentar uma análise do uso da Hemeroteca Digital Brasileira da Biblioteca Nacional (doravante HDB) nos anais dos Simpósios Nacionais de História da ANPUH (Associação Nacional de História). Para tal utilizamos uma metodologia que combina técnicas de raspagem de dados, através de scripts desenvolvidos em linguagem Python, e análise qualitativa através do software de apoio a análise de dados qualitativos (CAQDAs) - ATLAS.ti 7. Com os resultados deste trabalho esperamos trazer um panorama de como a HDB vem sendo utilizada e apresentada nos anais da ANPUH, assim como a importância da metodologia de pesquisa ao utilizar uma hemeroteca digital.

Palavras-chaves: Associação Nacional de História (Brasil) - Congressos - 2013-2019. Biblioteca Nacional (Brasil). Hemeroteca Digital - Pesquisa. Pesquisa histórica - Brasil.

ABSTRACT

This monograph consists of the Completion Work of the History Course, by the University of International Integration of Afro-Brazilian Lusophony. This work aims to present an analysis of the use of the Brazilian Digital Scientific Journal Library of the National Library (henceforth HDB) in the proceedings of the National History Symposia of the ANPUH (National Brazilian Association of History). For this we use a methodology that combines data scraping techniques, through scripts developed in Python language, and qualitative analysis through the software to support qualitative data analysis (CAQDAs) - ATLAS.ti 7. With the results of this work we hope to bring an overview of how HDB has been used and presented in the annals of ANPUH, as well as the importance of research methodology when using a digital scientific journal library.

Keywords: National Association of History (Brazil) - Congresses - 2013-2019. National Library (Brazil). Digital Scientific Journal Library - Research. Historical research - Brazil.

SUMÁRIO

1	INTRODUÇÃO	8
2	A HEMEROTECA DIGITAL BRASILEIRA E A PESQUISA EM HISTÓRIA	9
2.1	BASE DE DADOS ANPUH 2013-2019	10
2.2	A HDB	10
3	A CONSTRUÇÃO DE UMA METODOLOGIA DIGITAL PARA PESQUISA HISTÓRICA	12
4	ENCONTRANDO A HDB E ENTENDO SEUS USOS: ANÁLISE DOS DADOS	19
4.1	FAMÍLIA DE CÓDIGOS ELEMENTOS_TEXTUAIS	19
4.2	RECORTE_TEMPORAL	20
4.3	RECORTE_ESPACIAL	21
4.4	TEMÁTICA_GERAL	22
4.5	HDB_UTILIZAÇÃO	23
4.5.1	HDB_UTILIZAÇÃO: metodologia	24
5	CONCLUSÃO	26
	Referências	27
	Glossário	28

1 INTRODUÇÃO

A base de dados analisada é um dos resultados gerados pelo Projeto de Pesquisa *História Digital: acervos e ferramentas digitais para pesquisa e ensino*, edital Proppg 01/2020 - Programa Institucional de Bolsas de Iniciação Científica 2019-2020 (PIBIC/FAPESB). O projeto tinha como objetivo principal analisar os usos e impactos da Hemeroteca Digital Brasileira nas pesquisas em História desde seu lançamento em 2012 e listar e caracterizar as Bibliotecas Nacionais com Hemerotecas digitais de acesso público e gratuito existentes na América Latina. Os resultados alcançados pelo projeto serão apresentados no capítulo 1. Esta monografia se insere na área de História Digital, uma área em franca expansão dentro do campo mais amplo das Humanidades Digitais. Esta última pode ser definida como “uma área acadêmica interdisciplinar que fornece metodologias específicas da área das tecnologias digitais para serem incorporadas na investigação nas Humanidades como um todo” (ALVES, 2016). Encontramos perspectivas de História Digital que se limitam a fontes que passaram por uma rematerialização digital e em como isso facilita o acesso as mesma, sem necessidade de deslocamento físico dos pesquisadores ou de quem as queira consultar. Outras abrangem fontes que já nascem no meio digital e as formas como podem ser acessadas, sua veracidade entre outros pontos (BRASIL; NASCIMENTO, 2020; ROMEIN et al, 2020; SILVEIRA, 2018). Entendemos que o conceito de História digital ainda está em disputa e debate e que tal processo ainda está em curso, especialmente no Brasil, onde a área possui recente engajamento de pesquisadores e pesquisadoras (LUCCHESI; SILVEIRA; NICODEMO, 2020).

Nessa monografia buscamos aplicar, em uma pesquisa prática, técnicas de coleta e tratamento digitais de fontes primárias, ao mesmo tempo em que analisamos os usos de um repositório digital de fontes digitalizadas, a HDB. Por conseguinte, dividimos a monografia em três capítulos. No primeiro, apresentaremos o projeto que deu origem à pesquisa e caracterizamos a HDB. Buscamos assim, localizar a pesquisa e ao mesmo tempo justificar a escolha da HDB como elemento central do trabalho.

No segundo capítulo, “A construção de uma metodologia digital para pesquisa histórica”, apresentamos os aspectos metodológicos da construção da pesquisa, detalhando as etapas de coleta, limpeza, organização e codificação. Por fim, no capítulo três, “Encontrando a HDB e entendendo seus usos: Análise dos dados”, apresentamos um conjunto de análises sobre a base de dados e as conclusões atingidas. Mostramos como a HDB tem crescido em importância e recorrência nas pesquisas recentes, entretanto, esse uso não vem acompanhado de reflexão teórica e cuidados metodológicos apropriados.

2 A HEMEROTECA DIGITAL BRASILEIRA E A PESQUISA EM HISTÓRIA

Ao longo de um ano de trabalho no projeto “História Digital: acervos e ferramentas digitais para pesquisa e ensino”, entre 2020 e 2021, desenvolvi o plano de trabalho “História Digital: hemerotecas digitais e ferramentas digitais para pesquisa e ensino”. Apesar das dificuldades enormes geradas pela pandemia da Covid-19, atingimos os objetivos gerais e listamos abaixo os resultados alcançados.

Construímos uma Biblioteca de Referências no Zotero, que pode ser acessada nesse [link](#), e que continua sendo alimentada. A mesma foi divulgada no site da UNILAB e as visitas já feitas são um dos principais resultados. A biblioteca foi intitulada História Digital e foi desenvolvida em conjunto com a discente Priscila Valverde Silveira do curso de Bacharelado Interdisciplinar em Humanidades (IHLM/UNILAB), bolsista Pibic/Unilab projeto “Pós-Abolição: sentidos usos e produção acadêmica (2004-2019)”, também sob a orientação do professor Eric Brasil no âmbito do curso de História e em diálogo com o Laboratório de Humanidades Digitais da UFBA (LABHDUFBA).

Buscando compreender o contexto atual da difusão de hemerotecas digitais na América Latina e Caribe, realizamos um levantamento de sites de Bibliotecas Nacionais latino-americanas e caribenhas. Após o levantamento, selecionamos doze bibliotecas nacionais que possuem repositórios digitais de jornais e revistas acessíveis de forma remota. Em seguida criamos um roteiro para análise e o compilamos em cada interface, chegando a algumas conclusões importantes.

Sete delas não possuem em sua página dados sobre seu lançamento, temos lançamentos nos anos: 2004, 2006, duas em 2013, 2015. Todas as bibliotecas possuem acesso livre a sua plataforma; sete bibliotecas possuem em seu site um resumo do que se pode encontrar na plataforma do acervo, umas tendo um resumo bem simples enquanto outras possuem seção destinada para este fim.

Em todas é possível fazer o download dos pdf sem reconhecimento ótico de caracteres (OCR) e apenas uma permite o download com OCR. Apenas duas bibliotecas contam com a ferramenta de leitura da página online com OCR. Dos filtros, é possível fazer busca por palavras-chave em todas as bibliotecas, 10 aceitam expressões exatas e múltiplos termos.

Além da avaliação das características gerais das interfaces, produzimos considerações gerais para cada biblioteca. Todos os dados levantados nesta etapa da análise podem ser consultados em planilha publicada na web através deste [link](#).

Buscando publicizar os resultados e gerar um produto mais acessível, criamos o mapa interativo na ferramenta *StoryMapsJs*, inserindo os dados coletados na análise online das bibliotecas. Em cada item do mapa, além dos dados individuais, está disponível o [link](#) para acessar a planilha da análise que está disponível na web. Podemos acessar o mapa nesse [link](#).

Apesar do número crescente de repositórios on-line com jornais e revistas digitalizadas, ainda percebemos que suas interfaces são muito variadas e os recursos, filtros e possibilidades de uso se resumem majoritariamente em busca por palavras-chave e acesso a imagens das páginas digitalizadas, seja em formatos de imagens ou PDF.

A HDB não é diferente, porém seu volume e precisão na busca a colocam em destaque nas hemerotecas latinoamericanas. Por isso e por sua relevância nas pesquisas nacionais, buscamos uma base de dados e método capaz de analisar seu uso nas pesquisas em História no Brasil.

2.1 BASE DE DADOS ANPUH 2013-2019

Composta pelos trabalhos completos publicados nos Anais dos Simpósios Nacionais de História (SNH) da Anpuh entre 2013, primeiro ano com SNH após o lançamento da HDB, e 2019, ano do último Simpósio até o momento do início da pesquisa. Esse conjunto de dados foi coletado e organizado a partir da raspagem de dados realizada por *scripts* escritos com a linguagem de programação Python, desenvolvidos pelo professor Eric Brasil em parceria com o Laboratório de Humanidades Digitais da UFBA. Os resultados dessa etapa da pesquisa estão disponíveis [nesse link](#).

Essa base é um recorte da base completa gerada pelo programa Anais-Anpuh, cujo código está disponível nesse [repositório do GitHub](#), e que reúne todos os anais dos SNH desde 1963.

2.2 A HDB

A HDB foi lançada em 2012, hoje contando com mais de 7,542 periódicos e 20 milhões páginas digitalizadas do acervo de periódicos - jornais, revistas, anuários, boletins etc - da Fundação Biblioteca Nacional (BETTENCOURT; PINTO, 2013)¹. Seu [portal](#) nos possibilita acessar o acervo por diversas formas, combinando palavras-chave com os filtros datas, locais e

¹ O acervo digital está em constante ampliação e mais informações podem ser acessadas no link: <http://bndigital.bn.gov.br/hemeroteca-digital/> Acessado em 30 de maio de 2022.

jornais (BRASIL; NASCIMENTO, 2020). Escolhemos utilizar a HDB por ser um dos acervos digitais de acesso livre nacional mais consolidado e conhecido por pesquisadores e professores.

O repositório da HDB nos dá a possibilidade de busca a partir de três filtros:

a) por periódico, que envolvem jornais, revistas e outros documentos como mensagens governamentais; b) por período, em intervalos de dez anos iniciando em 1740 a 1749 com o “Folheto de Lisboa: entre Douro e Minho” de 1741 de Lisboa, Portugal até o ano atual; e c) por local, que além de todos os estados brasileiros, aponta que o acervo contém documentos pontuais publicados em outros países Paraguai, Argentina, México, Portugal, Inglaterra, Estados Unidos e França. (Azevedo; Pessoa; Neta, 2019, p. 49).

Ainda segundo os autores, pode-se refinar os resultados a partir de palavras ou frases. Tendo em vista o tamanho do acervo, essas ferramentas possibilitam um resultado mais preciso das ocorrências dentro das relações feitas pelos filtros. O que otimiza bastante o tempo de pesquisa ao fazer o levantamento de dados, ao mesmo tempo que nos demanda atenção e cuidados principalmente como a possibilidade de leituras fragmentadas, problemas de reconhecimento de caracteres, escolhas pouco eficientes dos parâmetros de busca (BRASIL, no prelo) se encontrar temas correlatos com o recorte da pesquisa numa busca com menos filtros. O que nos leva a concordar com Telles (2017, p. 90) ao dizer que os dados “não são uma representação direta do registro histórico, mas resultam do uso de ferramentas específicas para a sua obtenção”.

Que leva à questão de como as hemerotecas digitais vêm sendo utilizadas para pesquisa e o quão importante é se discutir a metodologia imposta para o levantamento dos dados. Pode parecer para muitos que só informar os filtros aplicados seja o suficiente, mas consideramos que discutir o porquê da escolha desses filtros e como eles são apresentados nas interfaces gráficas dos repositórios são pontos tão relevantes quanto à metodologia da análise dos dados, já que eles são resultado de um processo que levará a um determinado resultado. Encontramos essa questão presente no artigo de Brasil e Nascimento (2020) quando os autores indagam quais resultados a interface da HDB retornaria caso não se tivesse a opção de um dos filtros disponíveis.

O uso de ferramentas digitais, sejam elas para a construção da base de dados quanto para sua análise, levanta questionamentos novos em relação aos procedimentos metodológicos, sejam elas em fontes nascidas digitalmente ou digitalizadas. Ou seja, é necessário pensar em uma “metodologia particular, porém fundamentada nos princípios básicos já consagrados da pesquisa historiográfica, apenas adaptados ao formato digital” (Almeida, 2011, p. 25).

3 A CONSTRUÇÃO DE UMA METODOLOGIA DIGITAL PARA PESQUISA HISTÓRICA

Buscando analisar os usos e impactos da Hemeroteca Digital Brasileira nas pesquisas em História, utilizamos uma metodologia que combina técnicas de raspagem de dados, através de scripts desenvolvidos em linguagem Python, e análise qualitativa através do software de apoio a análise de dados qualitativos (CAQDAs) - ATLAS.ti 7. Essa análise foi aplicada sobre uma base de dados composta por trabalhos completos publicados nos Anais dos Simpósios Nacionais de História da Anpuh entre 2013 (primeiro SNH após o lançamento da HDB) e 2019, ano do último Simpósio até o momento do início da pesquisa.

A Associação Nacional de História (Anpuh) surge a partir do *I Simpósio de Professores de História do Ensino Superior* que ocorreu em 1961, tendo como objetivo discutir o currículo dos cursos de História defronte aos problemas da reforma universitária que ocorria naquele período, onde muitas perguntas não obtiveram respostas, reforçando a necessidade de um espaço para se discutir não só o currículo do Curso mas também as produções ligadas à ele (FALCON, 2011). Segundo o site da ANPUH, o Simpósio Nacional de História ocorre bianualmente tendo suas edições estaduais organizadas nos anos de seu intervalo.

Com essa estrutura, com eventos estaduais em um ano e o SNH no outro, e com sua longevidade, os eventos da Anpuh estão consolidados como espaços fundamentais para os debates, embates e apresentações de reflexões historiográficas brasileiras.

O conjunto de dados utilizados nesta monografia foi criado a partir da raspagem de dados realizada por dois *scripts* escritos em Python, desenvolvidos pelo professor Eric Brasil em parceria com o Laboratório de Humanidades Digitais da UFBA.

O primeiro deles, *script-anais-anpuh.py*, realiza a raspagem dos trabalhos em pdf de todos os Simpósios Nacionais da Anpuh entre 1963 até 2017, disponíveis atualmente no site da associação, que podem ser acessados a partir do [link](#) . O *script-anais-anpuh-2019.py* foi desenvolvido como desdobramento do primeiro para raspar os trabalhos publicados nos anais do 30º Simpósio Nacional de História, realizado no ano de 2019 em Recife, PE. Isso foi necessário pois esses anais constavam, até o momento da pesquisa, apenas no site específico do evento de 2019, e não no repositório do site da Anpuh. Ambos os scripts podem ser acessados no repositório Anais-Anpuh no GitHub do Laboratório de Humanidades Digitais da UFBA, através deste [link](#) .

Ambos os scripts foram escritos em [Python 3.8](#), sendo esta uma linguagem de programação com uma curva de aprendizagem menor, que te permite trabalhar rapidamente, e

integrar diferentes sistemas com maior eficiência. O código utilizou bibliotecas e módulos do Python 3 voltadas para manipulação de funções do sistema operacional (os), para operar com expressões regulares (re) e acessar urls (urllib), realizar downloads da web (wget), extrair dados de arquivos HTML e XML ([Beautiful Soup](#)) e manipulação e análise de dados (Pandas).

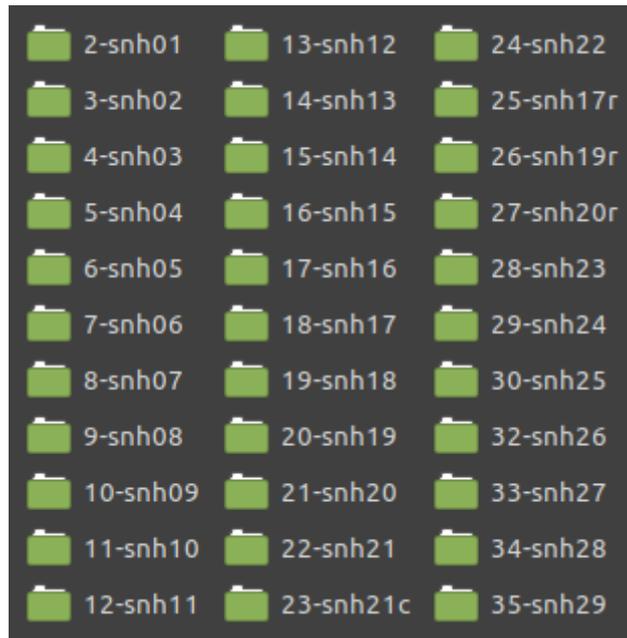
O *script-anais-anpuh.py* tem o seguinte funcionamento quando executado: Cria pasta para salvar os PDFs, após verificar se a mesma não existe no local, com o módulo 'os'; Acessa a URL dos Anais com a biblioteca 'urllib' e realiza a análise do HTML da mesma com a biblioteca 'BeautifulSoup'; Cria uma lista de eventos a partir da página principal; Acessa as páginas de cada evento contidas na lista criada anteriormente através de uma iteração; Em cada item da lista de eventos, o script busca todos os papers da primeira página e cria uma nova lista.

Nessa lista de papers de uma dada página, o script realizará as ações de encontrar as informações de cada paper, inclui essas informações em uma lista (que depois gerará um CSV com os dados), busca se há pdf disponível e se ele não é repetido e faz download do PDF. Após realizar essas ações para todos os itens de uma página, busca a próxima página de papers do evento, se não houver, passa para o próximo evento e repete as ações em um loop até o último evento disponível.

O *script-anais-anpuh-2019.py* realiza a mesma lógica de ação a partir do site do 30º SNH da Anpuh, sendo que nesse site só constam os papers de um evento.

Segundo documentação do programa, “o script retorna para o usuário todos os pdfs disponíveis em todas as páginas de todos os Simpósios Nacionais da Anpuh desde 1963 até 2017. São criadas pastas com o número de cada evento para o armazenamento dos arquivos em PDF”. (BRASIL, ANDRADE, NASCIMENTO, 2020)

Figura 1



Fonte: <https://github.com/LABHDUFBA/Anais-Anpuh/blob/master/images/pastas.png>

É importante notar que muitos papers não estão com pdf disponível no site, assim como nas edições mais antigas encontramos arquivos que contém vários papers num único PDF. O script também gera um arquivo CSV (*comma-separated values*) contendo os seguintes valores para cada paper: Autor(es)/Instituições, Título, Tipo, Evento, Ano, Link do Arquivo. Esse arquivo pode ser aberto como uma planilha e trabalhado em banco de dados. (BRASIL, ANDRADE, NASCIMENTO, 2020)

Figura 2

Y	Autor(es)/Instituições	Título	Tipo	Evento	Ano	Link do Arquivo
0	Maria Helena Capelato	os desafios da anpuh frente à crise brasileira: a luta pela preservação da demo	Conferência	XXIX Simpós	2017	https://anpuh.org.br/
1	Abner Neemias da Cruz (Univ	percursos políticos independentistas:a primeira geração diplomática, o impéri	Comunicação	XXIX Simpós	2017	https://anpuh.org.br/
2	A. Ricardo Abdalla (Pontificia	comensalidade e memória árabe na área central da cidade de são paulo	Comunicação	XXIX Simpós	2017	https://anpuh.org.br/
3	Acácia Regina Pereira (UERJ	currículo de história como responsável pela formação da identidade e memóri	Comunicação	XXIX Simpós	2017	https://anpuh.org.br/
4	Adalberto Paranhos (UNIVER	à flor da pele: pulsações do desejo feminino na música popular brasileira dos a	Comunicação	XXIX Simpós	2017	https://anpuh.org.br/
5	Adalberto Coutinho de Araújo	democracia e socialismo: propostas de políticas econômicas do operariado em	Comunicação	XXIX Simpós	2017	https://anpuh.org.br/
6	Adna Gomes Oliveira (UERJ)	o itamaraty, o corpo diplomático e a onu no início da guerra fria: a atuação bra	Comunicação	XXIX Simpós	2017	https://anpuh.org.br/
7	Adriana Aparecida Pinto (Uni	a imprensa e discurso sobre mulheres na imprensa rondonopolitana nos anos f	Comunicação	XXIX Simpós	2017	https://anpuh.org.br/
8	Adriana Maria de Souza Ziery	elementos religiosos da ascensão de d.joão i ao poder: o messias, o povo e a ci	Comunicação	XXIX Simpós	2017	https://anpuh.org.br/
9	Adriana Barreto de Souza (UI	osé narcizo de magalhães e menezes, a profissão militar e as milícias de home	Comunicação	XXIX Simpós	2017	https://anpuh.org.br/
10	Aécio Lessa Macedo (Univers	as potencialidades do estudo da história local em sala de aula: análise de uma	Comunicação	XXIX Simpós	2017	https://anpuh.org.br/
11	Adriane Piovezan (FIES Facul	rituais fúnebres militares: o túmulo dos fuzileiros navais mortos na intentona i	Comunicação	XXIX Simpós	2017	https://anpuh.org.br/
12	Afonso Henrique Sant Ana B.	labaredas do rio - abordagens histórico-sociais do rio e do corpo de bombeiros	Comunicação	XXIX Simpós	2017	https://anpuh.org.br/
13	Agnaldo Kupper (Universidac	futebol e contextos	Comunicação	XXIX Simpós	2017	https://anpuh.org.br/
14	Agenor Sarraf Pacheco (UNIV	a folia da cidade-floresta: patrimônio do afeto e lutas pela memória em melga	Comunicação	XXIX Simpós	2017	https://anpuh.org.br/
15	Agostinho Júnior Holanda Cc	a santa casa da misericórdia do maranhão e a intervenção dos presidentes-pro	Comunicação	XXIX Simpós	2017	https://anpuh.org.br/
16	Alan Christian de Souza Sant	lauro sodré em revista: textos, traços e retratos do senador paraense no perio	Comunicação	XXIX Simpós	2017	https://anpuh.org.br/
17	Airton José Cavenaghi (Anhe	mapas turísticos e sua representação cultural: marcel mauss e seu "ensaio sobi	Comunicação	XXIX Simpós	2017	https://anpuh.org.br/
18	Aíman Jorge Henrique Franc	ivan ribeiro e a "via prussiana" no mundo rural brasileiro	Comunicação	XXIX Simpós	2017	https://anpuh.org.br/
19	Ailton José Morelli (Universit	a importância dos espaços nas memórias de infância	Comunicação	XXIX Simpós	2017	https://anpuh.org.br/
20	Alana Cavalcanti Cruz (UNIVE	memórias de moradores da rua duque de caxias, joão pessoa-pb, da primeira r	Comunicação	XXIX Simpós	2017	https://anpuh.org.br/
21	Alan Ricardo Duarte Pereira i	trajetórias militares no império português: notas introdutórias sobre a família	Comunicação	XXIX Simpós	2017	https://anpuh.org.br/
22	Alberto Dias Mendes (UERJ)	ânio quadros e as influências de bandung	Comunicação	XXIX Simpós	2017	https://anpuh.org.br/
23	Alanna de Jesus Teixeira (UF	a representação do passado em anatole france: literatura e história	Comunicação	XXIX Simpós	2017	https://anpuh.org.br/
24	Alec Ichiro Ito (USP)	a sucessão episcopal no bispado e diocese de congo e angola entre 1607 e 161	Comunicação	XXIX Simpós	2017	https://anpuh.org.br/
25	Aldina da Silva Melo (UNIVER	a África no ensino de história no brasil	Comunicação	XXIX Simpós	2017	https://anpuh.org.br/
26	Alessandra Gonzalez de Carv	da "zona de contato" ao "limite antagônico": disputas entre criollos e indígena:	Comunicação	XXIX Simpós	2017	https://anpuh.org.br/

Fonte: <https://github.com/LABHDUFBA/Anais-Anpuh/blob/master/images/exemplo-csv.png>

Após a raspagem dos documentos, temos um dataset composto de trinta e quatro pastas, do primeiro SNH até o trigésimo (quatro eventos possuem pastas de resumos além dos eventos completos). Um total de 15.157 arquivos em PDF. Para os objetivos dessa pesquisa, filtramos esse conjunto de dados cronologicamente, a partir do 27º SNH no ano de 2013. Essa seleção se justifica pelo fato da HDB ter sido lançada ao público em 2012. Esse novo dataset é formado por 6.434 arquivos em PDF. É importante lembrar que esse número corresponde aos trabalhos completos que foram publicados nos anais dos eventos dos anos de 2013, 2015, 2017 e 2019, pois o número de trabalhos submetidos, aprovados e apresentados no evento daria um número muito superior a esse .

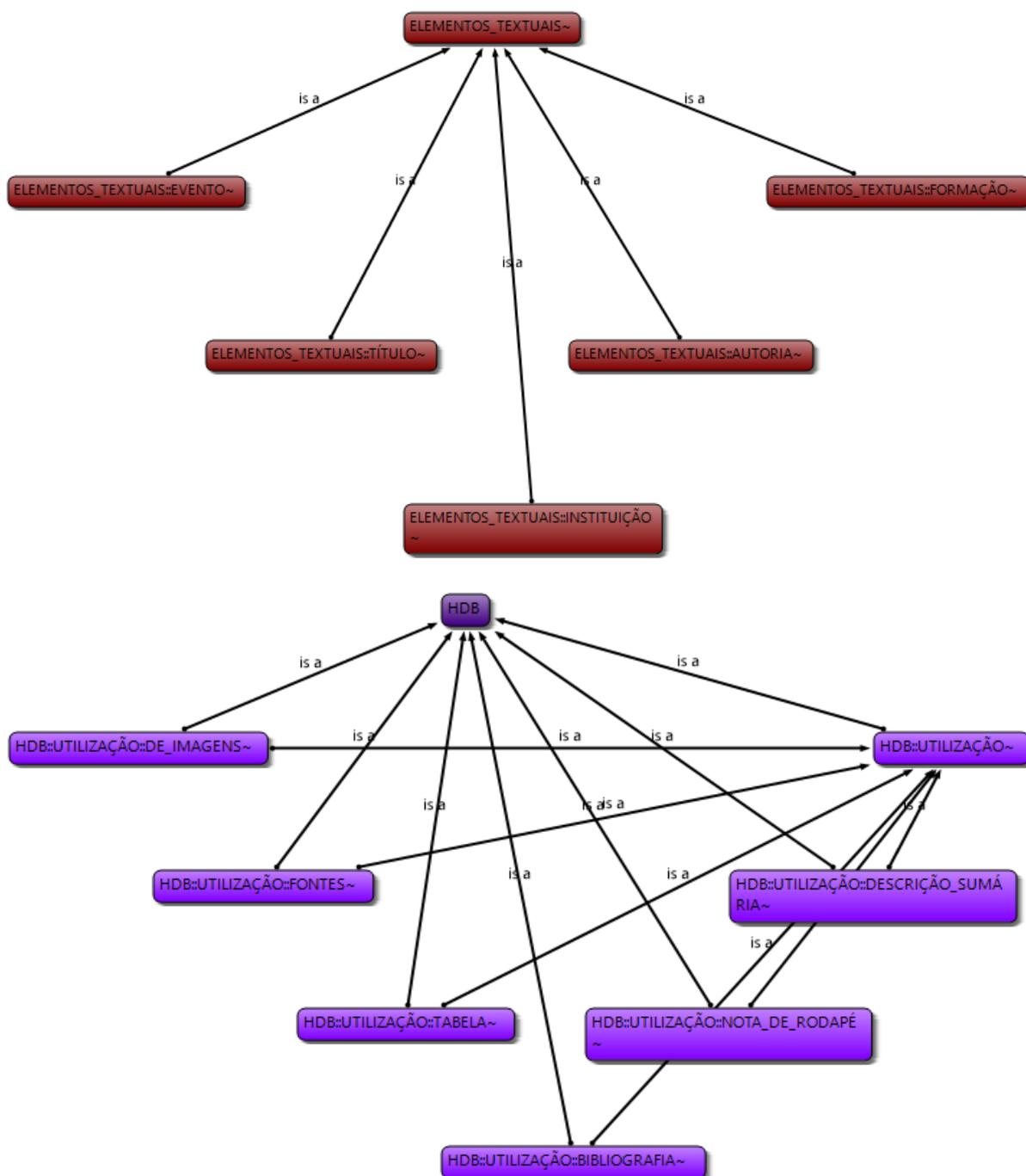
Para a análise dos dados, criamos um projeto no programa ATLAS.ti 7, definido como Unidade Hermenêutica, e inserimos todos os 6.434 PDFs coletados. Cada PDF nesta unidade é denominado de *Document* ou *Primary Document*. Para selecionar quais continham o termo Hemeroteca Digital Brasileira, utilizamos uma função denominada *auto-coding*, que realiza uma seleção automatizada de trechos (*quotation* ou citação) a partir de um trecho de texto definido pelo pesquisador. Utilizamos expressões regulares para que a codificação automatizada fosse mais precisa, fizemos uso do operador | (que significa *ou*) entre um conjunto de expressões: `hemeroteca|memoria.bn.br`. Não utilizamos a opção *Case Sensitive*, assim o resultado desconsidera diferença entre maiúsculas e minúsculas. Como nesse momento se tratava de uma estratégia para selecionar todos os arquivos que contivessem o termo, preferimos que todo o corpo do PDF fosse selecionado ao invés de parágrafos ou apenas a palavra, resultando na identificação de 113 PDFs.

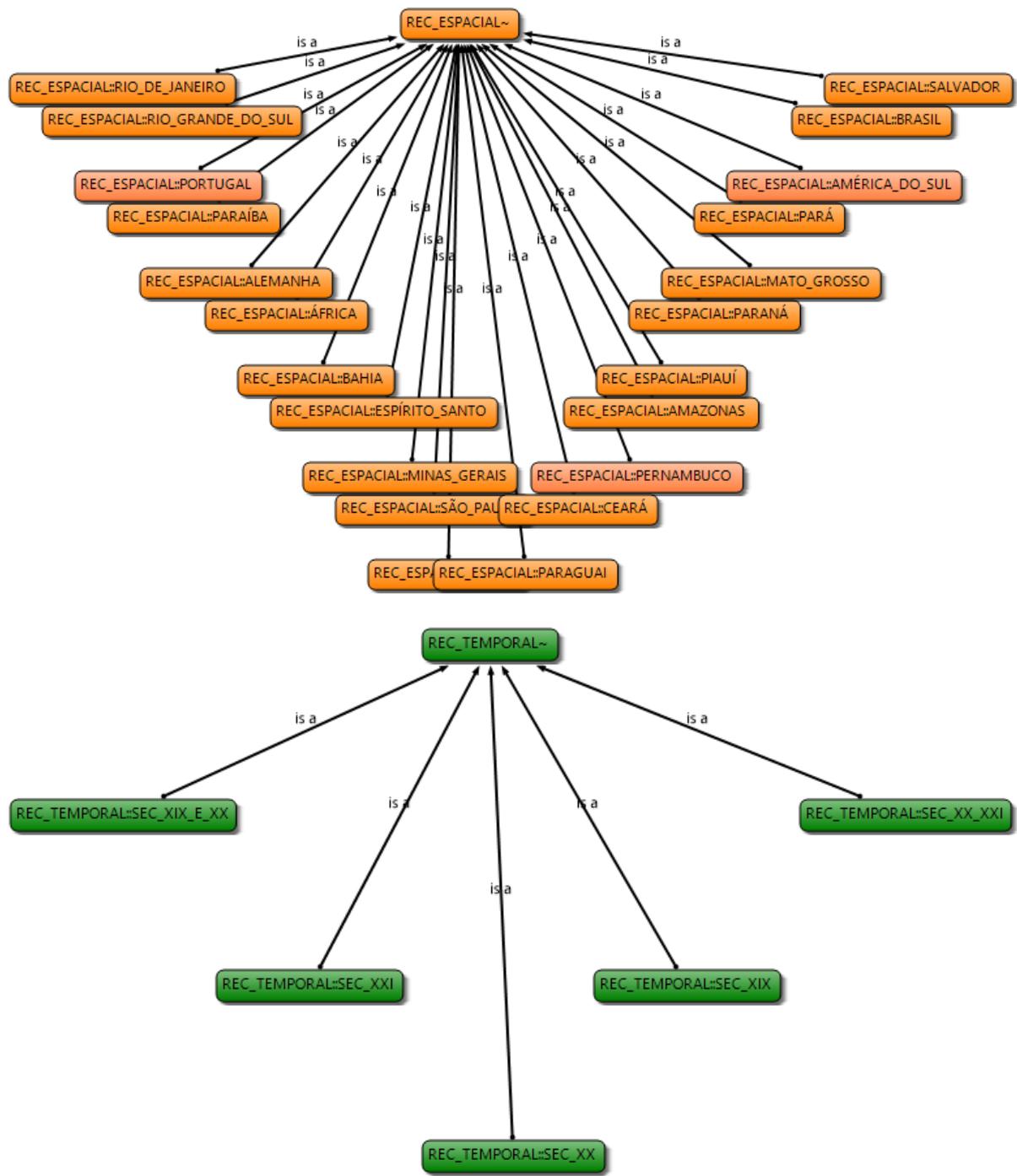
O ATLAS.ti (Archive of Technology, Life word and Language for Text Interpretation) é um aplicativo categorizado como CAQDAS (*Computer Assisted Qualitative Data Analysis*) para apoio a pesquisa qualitativa de dados através da leitura computadorizada de diversas fontes digitais. É um software de código fechado, com direitos exclusivos, sendo necessário licença para uso. Criado a partir de um projeto de 1986 para lidar com entrevistas com sobreviventes de Chernobyl, a versão inicial do Atlas.ti buscava auxiliar a seleção e categorização de fragmentos das entrevistas.(NASCIMENTO, 2017, p. 6)

Com o passar dos anos o programa foi aprimorado, incluindo hoje funcionalidades variadas, com a criação de códigos (codes) - normalmente uma palavra ou frase que sintetiza alguma compreensão teórica - para um conjunto de trechos selecionados (quotations). É possível criar famílias de códigos, códigos filhos e ainda estabelecer links entre eles; o programa também gera redes de relações variadas; execução de buscas e pesquisas e criação e

memos, anotações reflexivas que pode ser adicionadas a qualquer outro elemento.(BRASIL e NASCIMENTO, 2022)

Com a base de dados pronta, discutimos sobre a maneira de formatar a ferramenta com as compreensões teóricas sobre Hemerotecas Digitais transformando-as em códigos de análise, os quais ficaram pré-definidos em famílias de códigos e códigos filhos. Gerados as seguintes relações:







A pesquisa se propôs a construir sua base e a análise a partir de ferramentas digitais, o que não excluiu a necessidade de leitura próxima dos papers, ou seja, se fez necessário a leitura dos 113 arquivos. Como tinha-se por objetivo identificar no corpo do paper a frase/parágrafo que justificasse a vinculação de um determinado código, imaginamos que seria possível identificar no *resumo* ou na *introdução*, sendo um equívoco de nossa parte, já que em muitos papers só foi possível fazermos essa identificação no desenvolvimento do texto ou até mesmo na conclusão. Essa etapa tomou quase o dobro do tempo estipulado inicialmente, mas ao mesmo tempo rendeu dados e resultados que possibilitam desdobramentos não pensados a priori.

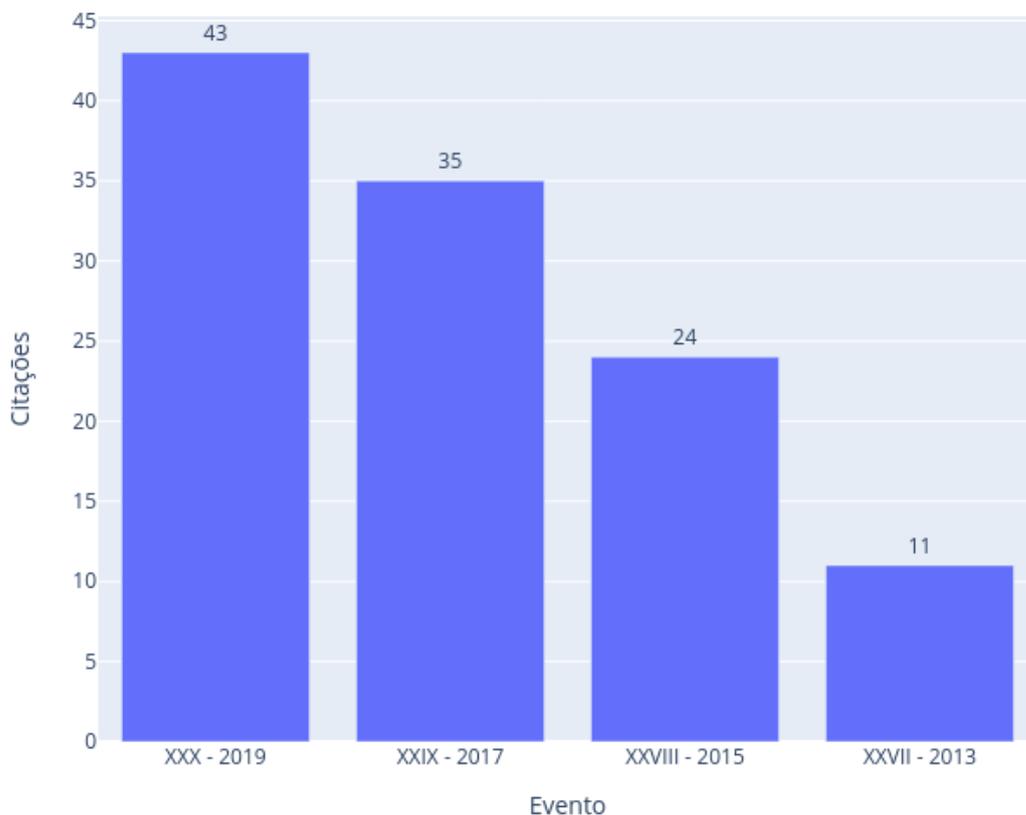
4 ENCONTRANDO A HDB E ENTENDO SEUS USOS: ANÁLISE DOS DADOS

Neste capítulo pretendemos analisar como a HDB é utilizada nos anais da ANPUH, análise esta feita pela base de dados de 113 PDF, onde descrevemos no capítulo anterior o processo de sua construção. Organizamos o capítulo e a análise pelos cinco códigos-família: ELEMENOS_TEXTUAIS; RECORTE_TEMPORALI; RECORTE_ESPACIAL; TEMÁTICA_GERAL; HDB_UTILIZAÇÃO. Cada um consiste em um tópico, contando com um subtópico do código-filho *Metodologia*, que é vinculado ao código-família HDB_UTILIZAÇÃO. .

4.1 FAMÍLIA DE CÓDIGOS ELEMENTOS_TEXTUAIS

Gráfico 1

HDB - Ocorrências por evento (2013-2019)



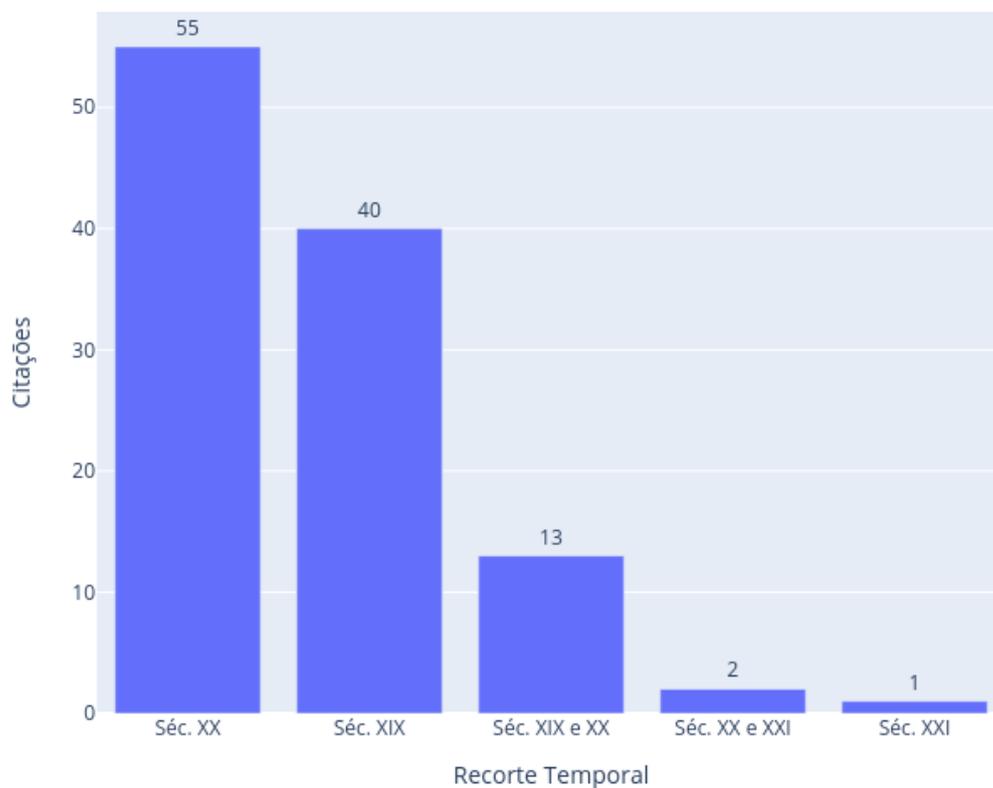
Este código foi criado com o objetivo de quantificar e permitir analisar a frequência de ocorrência de Hemeroteca Digital Brasileira, e sua variação entre os eventos. Tendo como base

os 113 PDFs selecionados, o número de papers que utilizam a HDB tem números crescentes entre os eventos, saindo de 11 papers no Evento XXVII para 43 no Evento XXX, quase quatro vezes mais ocorrências na comparação entre esses dois eventos. Os códigos-filhos foram nomeados conforme a edição do evento: *AUTORIA* com 124 ocorrências, pois há paper com mais de um autor; *INSTITUIÇÃO* com 118 ocorrências, esse número é menor em relação ao código-filho anterior pois nem todos os autores possuíam uma vinculação possível de identificar no corpo do paper; *TÍTULO* possuindo 113 ocorrências em concordância com os número de papers analisados.

4.2 RECORTE_TEMPORAL

Gráfico 2

HDB -Recorte Temporal



O código foi pensado para identificar o período ao qual o paper contextualiza sua pesquisa. Podemos afirmar que as ocorrências estão concentradas nos séculos XIX e XX, com um pouco mais de 48% das ocorrências no código-filho do século XX. Como esse é um código

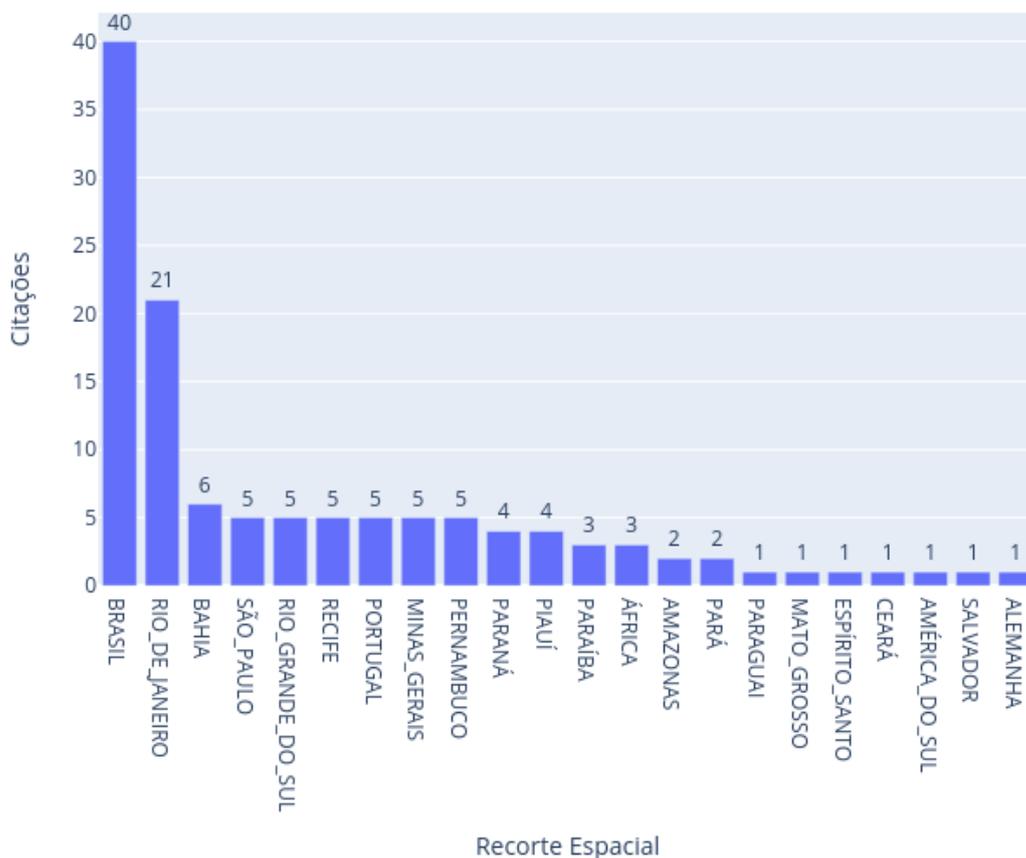
que só se aplica uma vez por paper, tivemos 113 ocorrências. Ao analisarmos o recorte temporal dos papers identificamos cinco temporalidades, que geraram os cinco código-filhos que foram nomeados conforme o/s século/s em qual a temática do paper estava inserida. Inicialmente iremos utilizar etiquetas por século, mas no decorrer da análise próxima percebemos que havia alguns paper que se inseriram num recorte temporal que perpassa por dois séculos, gerando assim os códigos-filho “*Séc.XIX e XX*” e “*Séc.XX e XXI*”.

A predominância das pesquisas sobre os séculos XX e XIX podem ser explicadas pela própria característica do acervo da HDB, cujo escopo acumula periódicos principalmente a partir de meados do século XIX, e com grande incidência de periódicos da primeira metade do século XX.

4.3 RECORTE_ESPACIAL

Gráfico 3

HDB - Recorte Espacial

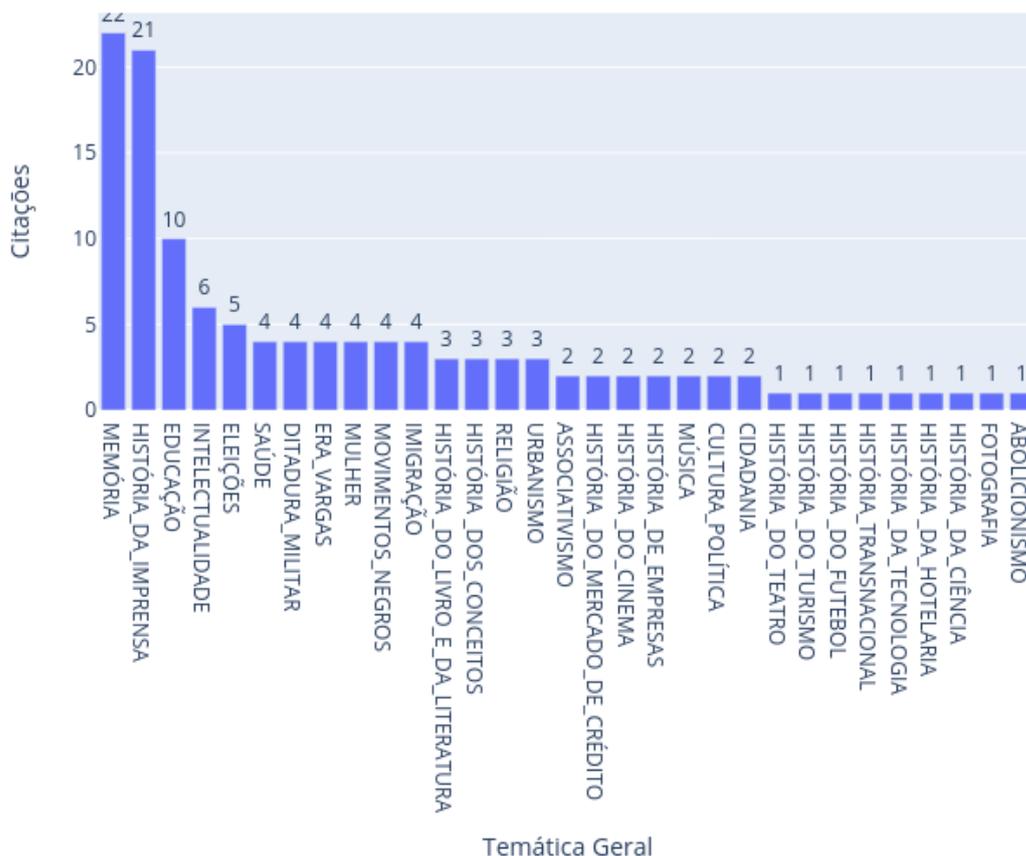


Criado para georreferenciar os papers. O código gerou 122 ocorrências. É importante levar em consideração que alguns papers possuem mais de um recorte geográfico, por isso o número de ocorrências pode ser maior que o de papers. Deste código podemos dizer que um pouco mais de 80% (111) das ocorrências são em território brasileiro, tendo um número mais expressivo nos códigos-filhos: Brasil com 40 ocorrências, quase 33% no total das 122; e o código-filho Rio de Janeiro (Estado) com 21 ocorrências, com quase 18% no total das 122. Tivemos ocorrências que apontavam para territórios fora do Brasil, esses códigos somados geraram 11 ocorrências equivalente a um pouco mais de 9% do total de 122, tendo maior presença no código-filho Portugal gerando cinco ocorrências, um pouco mais que 4% das 122. Essas ocorrências geraram 22 códigos-filhos que foram nomeados conforme lugar/região/território a qual o paper trata.

4.4 TEMÁTICA_GERAL

Gráfico 4

HDB - Temática Geral



Este código foi criado para identificar a principal temática/recorte temático/assunto a qual o paper se insere. Podemos confirmar que as temáticas com maior número de ocorrências são dos códigos-filhos Memória com 22 quase de 18% das 123 ocorrências, e História da Imprensa com 21 um pouco mais de 17% das 123 ocorrências. Um paper pode gerar mais de uma ocorrência dentro desse código já que estamos trabalhando com temáticas mais pontuais.

Esta análise gerou 123 ocorrências, que levaram a criação de 31 códigos-filhos que foram nomeados conforme essa temática/recorte temático/assunto. Inicialmente acreditou-se que o resumo e a introdução do paper seriam suficiente para conseguir definir a qual código-filho o paper iria se inserir, mas em muitos dos casos se fez necessário a leitura do paper por completo para identificar sua temática geral, já que para inserir a marcação era selecionado um parágrafo onde se justificasse esse código.

4.5 HDB_UTILIZAÇÃO

Gráfico 5

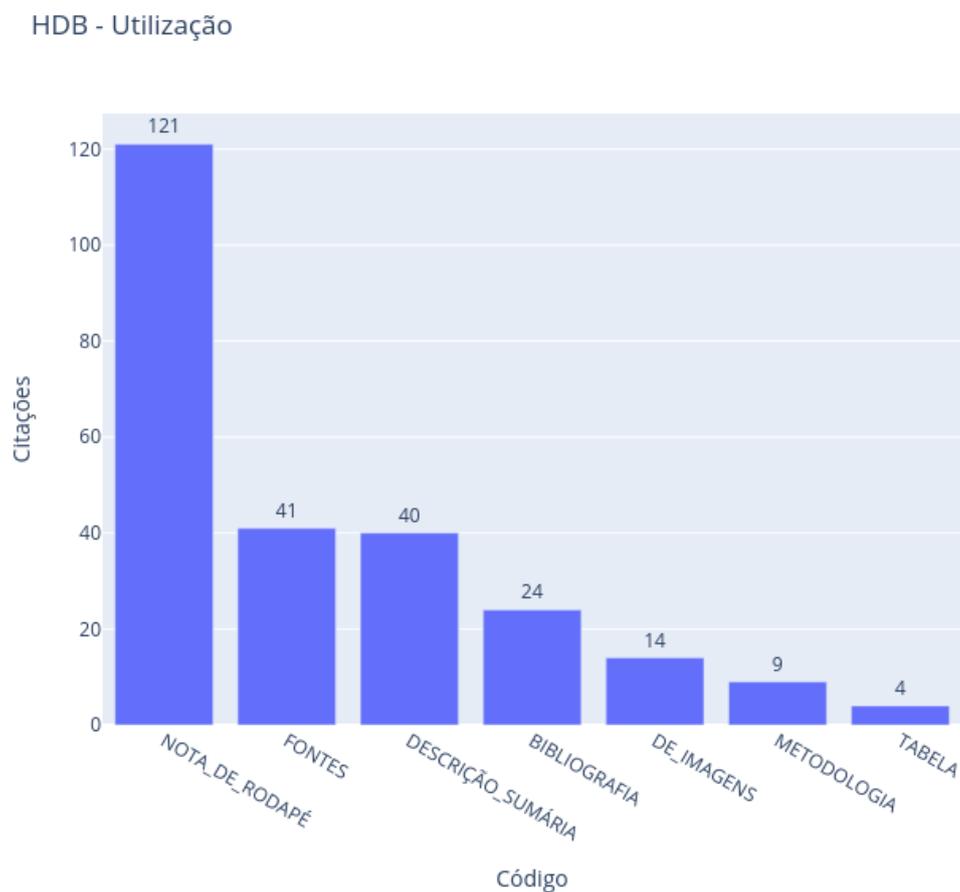


Tabela 5

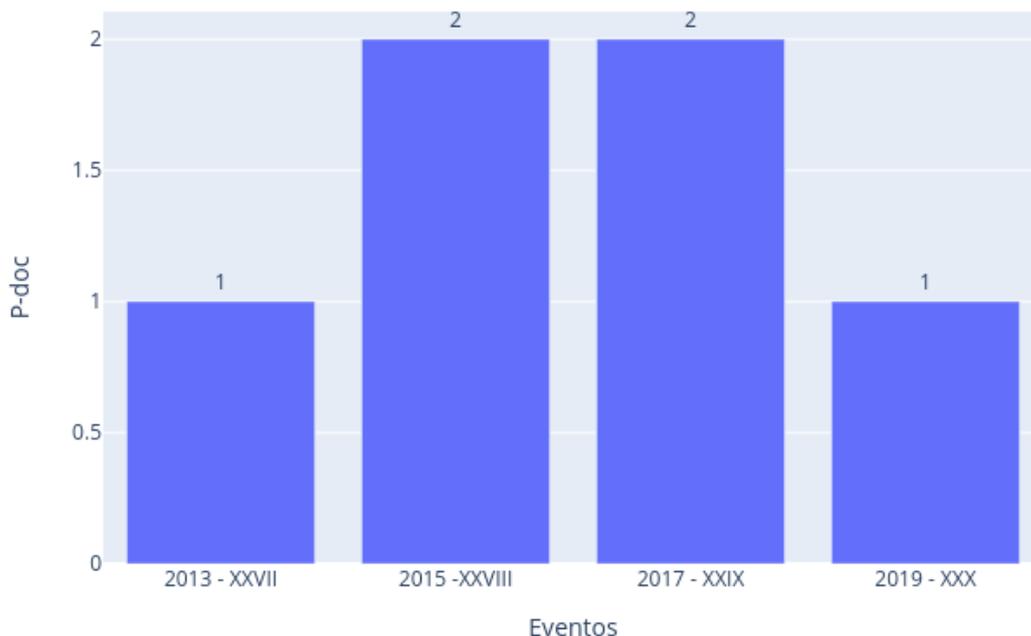
Esse código foi criado com o objetivo de identificar a forma como a Hemeroteca Digital Brasileira (HDB) está presente nos papers e gerou 253 ocorrências nos 113 pdfs.

Para auxiliar na análise do código, criamos sete códigos-filhos. Sendo eles: *BIBLIOGRAFIA* - quando HDB é citada na bibliografia do paper, com 24 ocorrências; *DE_IMAGEM* - quando imagens recolhidas na HDB são usadas no paper, com 14 ocorrências; *DESCRIÇÃO_SÚMARIA* - quando a HDB é citada como parte da pesquisa sem uma análise ou detalhamento da metodologia, 40 ocorrências; *FONTES* - quando a HDB é citada na lista de fontes do paper, com 41 ocorrências; *NOTA_DE_RODAPÉ* - quando a HDB é citada em nota de rodapé ou nota de fim no paper, com 121 ocorrências; *TABELA*: quando a HDB é utilizada como fonte para tabelas e gráficos no paper, com 4 ocorrências; *METODOLOGIA*: quando apresenta alguma reflexão ou categorização da metodologia aplicada na pesquisa utilizando a HDB, com 9 ocorrências

4.5.1 HDB_UTILIZAÇÃO: metodologia

Gráfico 6

HDB - Documentos com código `Metodologia` divididos por eventos.



Este código-filho nos gerou 9 ocorrências em 6 papers, a maior parte deles faz uma descrição simples da HDB, a relacionando ao acervo da Biblioteca Nacional Brasileira,

discorrendo também de forma simples sobre os processos da busca no acervo em relação a sua pesquisa, informando os termos utilizados para a pesquisa dentro das ferramentas de filtro da HDB, descrevendo as ocorrências geradas pela busca.

Um dos papers monta sua base de dados a partir de 3 acervos digitais. Consta a palavra de busca mas não explicita o porquê de utilizar mais de um banco de dados nem se propõe a discutir as semelhanças e diferenças entre eles. Dos seis papers, dois colocam as informações de filtros e ocorrências em nota de rodapé.

Quando consideramos o percentual de papers dentro da base de dados que discute a metodologia utilizada, e ainda assim feita de forma superficial, sem uma discussão detalhada e criteriosa, devemos concordar com Brasil e Nascimento (2020, p. 204) quando afirmam que “[..] o uso da HDB não é acompanhado de uma necessária discussão teórico-metodológica acerca de seus impactos, transformações, possibilidades e limites no fazer do historiador.”

5 CONCLUSÃO

Podemos concluir que a HDB tem sido um recurso cada vez mais utilizado entre pesquisas recentes apresentadas nos Simpósios Nacionais de História. Percebemos que há uma concentração de pesquisas que versam sobre os séculos XIX e XX, e com destaque para trabalhos sobre Rio de Janeiro e São Paulo. Tal configuração pode ser entendida pelo caráter do próprio acervo em questão, visto que a maioria dos periódicos datam de meados do século XIX a meados do XX e que há um número muito elevado de jornais publicados no Rio de Janeiro e São Paulo já digitalizados e disponibilizados pela plataforma.

Mesmo que seja possível perceber uma concentração temática em trabalhos sobre memória, imprensa e educação, percebemos que o acervo da HDB tem possibilitado estudos nas mais diversas áreas da história.

Por fim, e mais importante, percebemos com preocupação que o uso da HDB ainda tem sido restrito à citação de referência aos periódicos em notas de rodapé ou na lista de fontes e bibliografia ao final do texto. Seu emprego como ferramenta de pesquisa não recebe atenção metodológica e muito menos há reflexão sobre seus impactos teóricos e epistemológicos para os resultados apresentados nos papers.

Essa discussão da metodologia se faz cada vez mais necessária tendo em vista o grande número de pesquisadores que fazem uso dos acervos digitais e os impactos das interfaces gráficas na construção do conhecimento histórico. Deixar explícito, de forma detalhada e sistemática, as etapas para o levantamento da base de dados, assim como as ferramentas para a análise, possibilita entender os caminhos trilhados para a produção e relacionar esse caminho com os resultados obtidos, e possibilidades de desdobramentos.

Referências

- ALVES, Daniel. As Humanidades Digitais como uma comunidade de práticas dentro do formalismo acadêmico: dos exemplos internacionais ao caso português. **Ler História**, n. 69, p. 91–103, 2016.
- AZEVEDO, Laís Paula de Medeiros Campos; PESSOA, Lúcia Silvia; NETA, Olívia Moraes de Medeiros. A Hemeroteca Digital Brasileira: Fontes E Possibilidades Para a Pesquisa Em História Da Educação. **Cenas Educacionais**, v. 2, p. 39–55, 2019.
- BETTENCOURT, Angela Maria Monteiro; PINTO, Monica Rizzo Soares. A hemeroteca digital brasileira. **Anais do Congresso Brasileiro de Biblioteconomia, Documentação e Ciência da Informação - FEBAB**, v. 25, n. 0, p. 1028–1038, 2013.
- BRASIL, Eric; ANDRADE, Gabriel; NASCIMENTO, Leonardo F. *Script Anais-Anpuh*. Disponível em: <<http://rgdoi.net/10.13140/RG.2.2.34653.03048>>. Acesso em: 18 jun. 2022. DOI: 10.13140/RG.2.2.34653.03048
- BRASIL, Eric; NASCIMENTO, Leonardo Fernandes. História digital: reflexões a partir da Hemeroteca Digital Brasileira e do uso de CAQDAS na reelaboração da pesquisa histórica. **Revista Estudos Históricos**, v. 33, n. 69, p. 196–219, 2020. Disponível em: <[História digital: reflexões a partir da Hemeroteca Digital Brasileira e do uso de CAQDAS na reelaboração da pesquisa histórica. | Revista Estudos Históricos \(fgv.br\)](http://www.fgv.br/revista-estudos-historicos)> Acesso em: 01 de junho de 2022.
- BRASIL, Eric; NASCIMENTO, Leonardo. Por uma história social digital: o uso de CAQDAS na pesquisa e escrita da História. In: Barros, José D'Assunção (org.) **História Digital: a historiografia diante dos recursos e demandas de um novo tempo**. Rio de Janeiro: Editora Vozes, 2022.
- FALCON, Francisco José Calazans. **Memória e história: a fundação da Anpuh**. SIMPÓSIO NACIONAL DE HISTÓRIA–ANPUH, XXVI, p. 1–18, 2011. Disponível em: <http://www.snh2011.anpuh.org/resources/download/1322758035_ARQUIVO_falcon.pdf> Acesso em: 28 de maio de 2022.
- LUCCHESI, Anita; SILVEIRA, Pedro Telles da; NICODEMO, Thiago Lima. Nunca fomos tão úteis. **Esboços: histórias em contextos globais**, v. 27, n. 45, p. 161–169, 2020.
- NASCIMENTO, Leonardo Fernandes. Combinando webscraping em R e ATLAS.ti na pesquisa em ciências sociais: as possibilidades e desafios da sociologia digital. **18º Congresso Brasileiro de Sociologia**, 2017
- ROMEIN, C. Annemieke; KEMMAN, Max; BIRKHOLZ, Julie M.; *et al.* State of the Field: Digital History. **History**, v. 105, n. 365, p. 291–312, 2020.

Glossário

Códigos filhos: etiqueta que tem por objetivo identificar recorte/filtro específicos dentro dos recortes amplo/área, vinculada diretamente a uma família de códigos.

CSV: Comma Separated Values (em português, valores separados por vírgulas), é um arquivo de texto simples que armazena informações de planilhas e tabelas

Dataset: conjunto de dados; base de dados.

Famílias de códigos: Etiqueta que tem por objetivo identificar um recorte/filtro amplo/área.

OCR: Reconhecimento óptico de caracteres permite reconhecer caracteres em arquivos.

GitHub: plataforma gratuita e de código aberto que permite a criação de repositórios para controle de versões de arquivos com possibilidade de trabalho interativo e remoto entre grupos de desenvolvedores, possibilitando acesso público ou privado dos dados.

Python: linguagem de programação com uma curva de aprendizagem menor, que te permite trabalhar rapidamente, e integrar diferentes sistemas com maior eficiência.

Zotero: software de código livre e aberto, gratuito, que busca, gerencia e armazena referências e citações, criando uma biblioteca de referências onde seus itens podem ser organizados em coleções