O PROCESSAMENTO DE LINGUAGEM NATURAL APLICADO À ARQUIVOLOGIA: UMA ANÁLISE DAS PUBLICAÇÕES CIENTÍFICAS NACIONAIS E INTERNACIONAIS

NATURAL LANGUAGE PROCESSING APPLIED TO ARCHIVOLOGY: AN ANALYSIS OF NATIONAL AND INTERNATIONAL SCIENTIFIC PUBLICATIONS

Altair Lucas Packeiser Junior[1]
Camila Schwinden Lehmkuhl[2]

RESUMO

Com os avanços tecnológicos, e uso massivo de novas ferramentas para automatização de processos, distintas áreas do conhecimento tem feito uso do processamento de linguagem natural como base para desenvolvimento de sistemas, visando otimizar suas tarefas. Com a Arquivologia não é diferente, os fazeres arquivísticos demandam conhecimentos e práticas específicas, seja em contato com o papel ou em meios digitais, respeitados seus princípios. Nesse sentido, o uso dessas ferramentas tecnológicas pode auxiliar o arquivista em suas atividades, dentre elas, aqui elencadas, as funções arquivísticas: diagnóstico, criação/produção, classificação, avaliação, descrição/indexação, entrada de documentos, preservação/conservação, e difusão/acesso/acessibilidade. O objetivo geral deste trabalho é analisar produções bibliográficas nacionais e internacionais sobre o processamento de linguagem natural (PLN) relacionados à Arquivologia, tendo por base as funções arquivísticas. Enquanto que os objetivos específicos são: a) levantar as produções nacionais e internacionais que abordam o tema PLN e Arquivologia; b) comparar o que foi recuperado no levantamento e identificação entre as bases nacionais e internacionais, tendo como referência as funções arquivísticas; c) identificar possibilidades de relação entre as funções arquivísticas e o PLN. Para responder aos objetivos, utilizou-se como metodologia a pesquisa bibliográfica. Como resultados, observou-se menor número de publicações nacionais em relação às publicações internacionais, principalmente aquelas relacionadas aos interesses da Arquivologia e Ciência da Informação, bem como maior diversidade nas publicações internacionais, o que permite assinalar que a temática tem sido mais abordada em âmbito internacional, do que no Brasil. Por fim, considera-se que a pesquisa em tela é inicial, que poderá servir de base para pesquisas futuras sobre o tema, que cada vez mais estará presente na vida dos arquivistas.

Palavras-chave: Arquivologia; Processamento de linguagem natural; Funções arquivísticas; Ciência da Informação.

ABSTRACT

With technological advances and the massive use of new tools to automate processes, different areas of knowledge have made use of natural language processing as a basis for developing systems to optimize their tasks. Archivology is no different, archival activities require specific knowledge and practices, whether in contact with paper or digital media, while respecting its principles. In this sense, the use of these technological tools can help archivists in their activities, including the archival functions listed here: diagnosis, creation/production, classification, appraisal, description/indexing, document input, preservation/conservation, and dissemination/access/accessibility. The general objective of

[1] Graduando de Arquivologia - Centro de Ciências da Educação, Centro de Ciências da Informação. Universidade Federal de Santa Catarina. E-mail: packeiser.jr@gmail.com
[2] Profª. Dra, Orientadora no curso de Arquivologia. Universidade Federal de Santa Catarina. Email: camila.lehmkul@ufsc.br

this work is to analyze national and international bibliographic productions on natural language processing (NLP) related to Archivology, based on archival functions. The specific objectives are: a) to collect national and international productions on the subject of NLP and Archivology; b) to compare what was retrieved in the survey and identification between national and international databases, with reference to archival functions; c) to identify possible relationships between archival functions and NLP. In order to meet the objectives, bibliographical research was used as the methodology. The results showed that there were fewer national publications than international publications, especially those related to the interests of Archivology and Information Science, as well as a greater diversity of international publications, which shows that the subject has been covered more internationally than in Brazil. Finally, it is considered that this research is initial and could serve as a basis for future research on the subject, which will be increasingly present in the lives of archivists.
Keywords: Archivology; Natural Language Processing; Archival Functions; Information Science.

1 INTRODUÇÃO

A constante produção documental, seja em papel ou em meio digital, demanda fazeres arquivísticos ininterruptos para sustentar o acúmulo dos arquivos. Para gerir os documentos arquivísticos de forma mais facilitada, a informática proporciona ferramentas de automação que podem auxiliar nos processos, possibilitando realizar atividades com um tempo de resposta menor, dependendo de como foram programados os seus processos e seus objetivos.

Consequentemente, na era da informação mais acessível, a comunicação se torna maior, criando necessidades de respostas mais rápidas, mais intuitivas, menos robotizadas. As experiências de usuários torna-se mais simplificada com o uso de *chatbots*, por exemplo, que utilizam de processamento de linguagem natural (PLN) para compreender solicitações dos usuários, e então entregar uma resposta compreensível e de interesse do usuário.

Os fazeres da Arquivologia, aqui caracterizados a partir das oito funções arquivísticas (LEHMKUHL, 2021), podem também contar com o auxílio do processamento de linguagem natural, proporcionando melhor organização e recuperação dos documentos de arquivo. Por exemplo, a partir da elaboração de diagnósticos/reconhecimento de documentos para criar indexadores ou na automação de partes desses processos, como descrição, classificação, avaliação, preservação, dentre outras. A partir desse pressuposto, as questões que permeiam a pesquisa em tela são as seguintes: como o PLN está sendo relacionado à Arquivologia nas publicações acadêmicas brasileiras e internacionais?

Partindo desses questionamentos, o objetivo geral deste trabalho é analisar as produções bibliográficas nacionais e internacionais sobre o processamento de linguagem natural relacionado à Arquivologia, tendo por base as funções arquivísticas. Os específicos são: a) levantar as produções nacionais e internacionais que abordam o tema PLN e Arquivologia; b) comparar o que foi recuperado no levantamento e identificação entre as bases nacionais e internacionais, tendo como referência as funções arquivísticas; c) identificar possibilidades de relação entre as funções arquivísticas e o PLN.

Como procedimentos metodológicos, se utiliza da pesquisa bibliográfica em bases de dados acadêmicas, com o intuito de identificar as publicações que abordam o PLN, tendo como base as funções arquivísticas.

O trabalho justifica-se por se tratar de uma ferramenta que está em constante evolução tecnológica e computacional, que visa facilitar a atuação de gestores e atender melhor os usuários de todas as áreas de conhecimento, com muita ou pouca prática com as tecnologias da informação, podendo utilizar a escrita ou a fala (gravação de áudio) para produzir ou recuperar informações. Outro fator importante é o interesse do autor deste trabalho nas constantes atualizações da tecnologia, necessitando de aprendizagem do profissional com a utilização das máquinas em prol de uma gestão de documentos de qualidade.

Com o presente trabalho, pretende-se observar quantitativamente os interesses e preocupações da Arquivologia e as funções arquivísticas, utilizando do PLN para resoluções no meio digital, priorizando os usuários e a qualidade da informação. Sabe-se que com a tecnologia, a entrega de informações pode ser facilitada, de acordo com os interesses dos usuários, mas ainda se faz necessário o acompanhamento do profissional que detém conhecimento sobre o acervo, baseando-se nos ciclos de atividades arquivísticas, ou pelas funções que abordaremos a seguir.

Por fim, a pesquisa está estruturada inicialmente com o referencial teórico, que abordará o PLN e suas possibilidades de automação. Na sequência, a Arquivologia, abordando a evolução dos fazeres da área, com destaque para as funções arquivísticas. Os processos metodológicos aplicados na pesquisa são apresentados no ponto seguinte, com os detalhes dos processos para a realização da coleta e análise dos dados. E ao final, apresenta-se as conclusões e sugestões para futuros estudos.

2 PROCESSAMENTO DE LINGUAGEM NATURAL

No final dos anos de 1940 (RINO, 1987), a preocupação dos cientistas no uso das máquinas (que tinham objetivos primordiais de realizar cálculos no menor tempo possível) era de as transformar em uma ferramenta mais "inteligente", semelhante ao comportamento humano, como a capacidade de conversar (de forma escrita ou sonora) e se comportar de forma aleatória.

As primeiras ideias de utilização da linguagem para automatizar ações humanas foram para a criação de tradutores, decifrando códigos para escrever em outra língua (SILVA, 2019). Segundo Lucia Rino (1987, p. 1), esses trabalhos automatizados serviriam não apenas para a pesquisa, "mas também em tarefas diárias e diversões, permitindo que leigos também se utilizassem dos recursos da informática". A comunicação como experiência do usuário se origina na correlação dos estudos do tratamento de linguagem natural na área de Inteligência Artificial (IA).

O processamento de linguagem natural (PLN) é uma área da linguagem computacional que desenvolve aplicações com base na compreensão da linguagem humana. Um dos resultados comumente utilizados na área da informática é a Inteligência Artificial, que utiliza bancos de dados para responder a determinadas demandas. As pesquisas em PLN podem ser baseadas em três aspectos da comunicação, que segundo Silvio Pereira (2011) são: som (fonologia); estrutura (morfologia e sintaxe); significado (semântica e pragmática).

> O PLN é genuinamente multi-disciplinar, congregando, principalmente, estudos nas áreas de Ciência da Computação, Linguística e Ciências Cognitivas. A pesquisa em PLN divide-se em duas sub-áreas de trabalho: interpretação e geração. (BARROS; ROBIN, 2001, p. 1)

Como exemplificado por Barros e Robin (2001), a "interpretação" é uma tradução da pergunta do usuário, essa em linguagem natural, transformando então em uma linguagem de consulta, integrando os bancos de dados e suas ferramentas de busca adaptadas para a linguagem necessária ("*query language*"). Já a "geração" é uma tradução de um conteúdo pré-definido para uma linguagem natural, produzindo textos mais próximos daqueles produzidos por pessoas, como resumos.

Além das áreas de trabalho citadas por Barros e Robin, Lucia Rino (1987, p. 39) adiciona os "tradutores de textos e os interpretadores de estórias" como tipos de sistemas de PLN. Focando na área de interpretação, Rino utiliza o termo "sistemas de consulta a bases de dados", em que há uma relação de sistemas de perguntas e respostas, podendo ser divididos em módulos que realizarão determinadas tarefas, como análise de sentenças; acesso ao conhecimento específico; geração de respostas; interação com o usuário.

Para uma boa recuperação da informação, inicialmente necessitam de linguagem documentária (CAMPOS, 2001), em que informações são representadas nos documentos, como a utilização de tesauros e uma classificação que auxiliam na produção de bancos de dados estruturados, principalmente para os dados que os usuários buscam. Os documentos podem ser estruturados com os metadados, utilizando softwares para interpretar e disponibilizar documentos relevantes para quem está pesquisando. Mas, há problemas com relação aos documentos em ambiente digital, como a falta de padrões e critérios em preenchimento de campos indexadores em sistemas informatizados, e uma linguagem documentária bem estabelecida, que integre, por exemplo, instrumentos de classificação documental (SOUSA; ARAÚJO JÚNIOR, 2013), assim possibilitando utilizar o PLN para automatizar as indexações de acordo com os critérios selecionados pelo profissional arquivista.

A indexação, como um dos elementos principais para o PLN, necessita de módulos que auxiliam na arquitetura de sistemas de PLN, entre eles, aquele voltado para o contexto:

> O **Modelo de Domínio** fornece o contexto *enciclopédico*, armazenando conhecimento a respeito das entidades, relações, eventos, lugares, e datas do domínio, em algum formalismo de IA - *e.g.*, Lógica de Predicados, Redes Semânticas, Frames, Scripts ou Hierarquias de Tipos. (WINSTON, 1992 apud BARROS; ROBIN, 2001, p. 4, grifo do autor)

Como abordado por Barros e Robin (2001), o "Modelo de Domínio" é uma base de conhecimento que está relacionada com os "contextos", elemento primordial para a atuação dos profissionais arquivistas, associando documentos e arquivos de acordo com suas tramitações, comunicações, linhas temporais, entre outros elementos circunstanciais e que será abordado em específico na seção a seguir.

3 ARQUIVOLOGIA

Além de ser uma ciência que estuda a gestão, organização, preservação e tratamento de arquivos, a Arquivologia se atenta também ao papel do gestor de arquivos, que será o responsável pela governança arquivística. Para tais atividades, a história relata que a humanidade passou por mudanças relacionadas ao modo de difundir a informação, passando por várias gerações nas formas de registros e suportes, da escrita cuneiforme e o alfabeto em tabletes de argila, até os registros binários com acesso por meio de computadores.

O arquivo, historicamente, confunde-se com a história da escrita nas civilizações pré-clássicas (REIS, 2006), sendo um elemento fundamental para o desenvolvimento da administração e consequentemente para o fortalecimento do conceito de Arquivo. No século XVI a mudança nas ações com os arquivos começam a influenciar as concepções jurídicas, aumentando a criação de manuais que abordam a arquivística, até que durante a Revolução Francesa, como parte de seus pressupostos, o arquivo passa a ser disseminado e de livre acesso.

Segundo Luís Reis (2006), há vários marcos que formam a arquivística atual, como a publicação do Manual dos Arquivistas Holandeses no século XIX e o nascimento do Conselho Internacional de Arquivos em 1950, proporcionando o debate sobre os conceitos, fundamentos, princípios, procedimentos, assim promovendo a ciência.

> Um movimento macro e social de preservação da memória, de novas formas de registro de informação e demais recursos da tecnologia tem feito com que nas últimas décadas a Arquivologia busque novos olhares, se remodele, se recrie, se atualize. (LEHMKUHL, 2021, p. 31)

Com a constante atualização das tecnologias, a Arquivologia necessita estar adaptando-se aos novos modos de produção de informação, conforme apresentado acima por Lehmkuhl (2021), a necessidade de integração da Arquivologia com as tecnologias da informação, demonstrando as preocupações que o profissional arquivista deverá observar, preparado para realizar as funções arquivísticas, independente do suporte.

O arquivista, enquanto profissional que estará lidando com o meio digital, necessitará de conhecimentos que poderão constituir subsídios para as máquinas, e em conjunto com profissionais de Tecnologia da Informação (TI), poderão ser constituídos sistemas para que

sejam automatizadas as execuções de suas atividades. Como resultado, espera-se uma entrega de informações de qualidade aos usuários, e para tal, é necessário que esse profissional detenha conhecimentos que apenas uma graduação pode oferecer, remetendo à Arquivologia como ciência que o formará, preparado para lidar com a informação não estruturada, independente do meio em que está, papel ou digital, distintamente do profissional de TI que não terá tais conhecimentos em sua formação geral. Dentre esses conhecimentos arquivísticos, parte está contemplado nas funções arquivísticas.

3.1 FUNÇÕES ARQUIVÍSTICAS

As funções arquivísticas dão sustentação à Arquivologia, enquanto que os princípios que regem a área constituem sua espinha dorsal (COUTURE, 2016). Dentre as funções que serão utilizadas para essa pesquisa, estão aquelas relacionadas à releitura de Lehmkuhl (2021): diagnóstico, criação/produção, avaliação, classificação, descrição/indexação, entrada de documentos, preservação/conservação, difusão/acesso/acessibilidade. Elas contemplam grande parte do fazer arquivístico, regido por princípios, normas e técnicas (ROUSSEAU; COUTURE, 1998) e serão descritas a seguir.

O diagnóstico arquivístico pode ser considerado como uma "análise das necessidades" (COUTURE et al. 2003), originado da Administração, tem como objetivo, o conhecimento da organização em que o arquivista estará atuando ou desenvolvendo certa atividade. Observar e identificar como estão sendo geridos os documentos, propondo soluções para as necessidades identificadas para então se tomar a decisão a respeito do que poderá ser feito.

Criação/produção de documentos remete à qualidade da informação, evitando a "criação ou manutenção de informações ou documentos desnecessários" (LEHMKUHL, 2021, p. 69). Utiliza-se das técnicas de identificação arquivística, importantes para conhecer a produção e organização documental, e também para o estabelecimento de padrões para os tipos documentais desde o seu nascimento. Com essa identificação estabelecida serão criados documentos de forma padronizada, evitando se produzir documentos com nomes distintos, o que pode acarretar em dificuldades para a classificação e avaliação desses documentos.

Classificação como função arquivística remete aos movimentos da informação orgânica, possibilitando contextualizar os documentos produzidos, para então dar acesso ao

acervo documental de forma clara. Segundo o Dicionário Brasileiro de Terminologia Arquivística (ARQUIVO NACIONAL, 2005, p. 49), a classificação aborda a "organização dos documentos de um arquivo ou coleção" e "análise e identificação do conteúdo de documentos, [...] podendo-se-lhes atribuir códigos". Como resultado dessa ação, estão os planos de classificação de documentos (PCD), um dos instrumentos base para o desenvolvimento das atividades de gestão documental.

A avaliação faz parte da gestão documental, abordando a "intervenção no *ciclo de vida dos documentos* desde sua produção até serem eliminados ou recolhidos para guarda definitiva" (BERNARDES, 1998, p. 12, grifos do autor). Bernardes (1998, p. 14) também afirma que a avaliação consiste em "identificar valores e definir prazos de guarda para os documentos de arquivo". A ferramenta utilizada para o processo de avaliação é a tabela de temporalidade e destinação de documentos (TTDD), que é baseada no PCD, e será responsável pela redução da massa documental, consequentemente auxiliará na preservação dos documentos históricos, na melhor relação de custo-benefício dos recursos humanos e materiais e na recuperação da informação.

Descrição/indexação está intimamente ligada à classificação e ao acesso, abordando elementos intrínsecos e extrínsecos que representam os documentos, independente dos suportes. Com a descrição pode-se criar vários instrumentos, como catálogos, guias, repertórios e inventários (ROUSSEAU; COUTURE, 1998), o que exigirá do arquivista conhecer os seus usuários e suas demandas.

Entrada de documentos, como função arquivística, pode ser associada à "aquisição", abordando a chegada dos documentos ao arquivo, necessitando de gestão e conhecimento do espaço disponível para a guarda do acervo. De acordo com o Dicionário Brasileiro de Terminologia Arquivística (ARQUIVO NACIONAL, 2005, p. 85), a entrada de documentos pode ser realizada por "comodato, compra, custódia, dação, depósito, doação, empréstimo, legado, permuta, recolhimento, reintegração ou transferência". O que demandará do arquivista planejamento de seu espaço, seja físico ou digital, para recebimento desse acervo.

Preservação/conservação, trata de ações preventivas realizadas para maior durabilidade do suporte em que o documento encontra-se registrado, digital ou em papel, e para que estes não necessitem de reparações. O objetivo é que não se perca a informação, independente do meio de acondicionado (físico e digital), possibilitando o manuseio dos

documentos (acesso e difusão). O Arquivo Nacional (2005, p. 135) acrescenta que a preservação é a "Prevenção da deterioração e danos em documentos, por meio de adequado controle ambiental e/ou tratamento físico e/ou químico".

Difusão, acesso e acessibilidade, se relacionam com a disseminação da informação contida nos arquivos para os usuários (LEHMKUHL, 2021). O Arquivo Nacional (2005, p. 19), em seu dicionário de terminologia arquivística, descreve "acesso" como "possibilidade de consulta a documentos e informações" e "tornar acessível os documentos e a promover sua utilização", que podemos relacionar a "acessibilidade" descrito como "condição ou possibilidade de acesso a serviços de referência, informação, documentação e comunicação".

Em várias funções descritas, há possibilidades de utilizar o PLN para auxiliar no fazer arquivístico, possibilitando a interação dos profissionais arquivistas para inserção de novas tecnologias da informação em prol da organização da informação com qualidade, ponto que será mais explorado a partir da seção 5.

4 PROCEDIMENTOS METODOLÓGICOS

Para alcançar os objetivos propostos na pesquisa em tela, será utilizada a pesquisa bibliográfica, que segundo Gil (2017, p. 33), "é elaborado com base em material já publicado", incluindo materiais impressos (livros, revistas, teses, anais de eventos, entre outros) e outros tipos de fontes no meio digital (discos, fitas magnéticas, internet). A pesquisa bibliográfica pode ser utilizada tanto para a busca de dados, como também embasar a teoria dos trabalhos acadêmicos com a produção do capítulo dedicado à revisão bibliográfica apresentado anteriormente.

O método de pesquisa utilizado para a proposta de trabalho será quali-quantitativo, ou seja, utilizará da pesquisa quantitativa para a mensurar as informações em números, com a pesquisa qualitativa para valorizar a complexidade e dinâmica dos fenômenos humanos e sociais (SCHNEIDER; FUJII; CORAZZA, 2017). A união dos dois métodos possibilita análises e discussões mais ricas, gerando um quadro mais geral, por também "aumentar o conhecimento sobre determinado tema, alcançar os objetivos traçados, observar e compreender a realidade estudada" (BRÜGGEMANN; PARPINELLI, 2008, p. 564).

A pesquisa bibliográfica foi realizada entre os dias 16 e 22 de abril de 2023. Para a coleta de dados, foram utilizadas as bases de dados de produções acadêmicas nacionais: Base de dados em Ciências da Informação (BRAPCI), voltada para as áreas da Ciência da Informação, Biblioteconomia e Arquivologia; Base de Dados em Arquivística (BDA), voltada especificamente para a produções científicas e técnicas da Arquivologia ou Arquivística; Biblioteca Digital Brasileira de Teses e Dissertações (BDTD), abordando trabalhos de mestrado e doutorado, voltados para todas as áreas do conhecimento. Como base de dados para publicações internacionais, foi utilizado a SCOPUS, base de dados da Elsevier, que possui publicações acadêmicas internacionais voltadas também para todas as áreas de conhecimento, além de se representar como uma base mundial com quase 20% de artigos pelo mundo todo[3].

A pesquisa foi realizada utilizando as seguintes palavras-chave nos campos de busca, “linguagem natural” AND “arquiv*”, respeitando os critérios dos buscadores para poder ampliar as buscas, com operador lógico “AND”, e com as variações da palavra “arquivo”, como, “arquivologia”, “arquivística”, “arquivar”, entre outras possibilidades. Na pesquisa internacional, foi realizada a pesquisa em inglês e espanhol, substituindo “linguagem natural” por “natural language” e “lenguaje natural”. O buscador internacional não reconhece a utilização de variações (uso do asterisco), por isso foi utilizado a palavra “arquivo” em inglês e espanhol, substituindo por “archive” e “archivo”.

Quadro 1 - Relação de produções acadêmicas recuperadas nas bases de dados

Base de dados	Linguagem natural	AND arquiv*	Utilizados na pesquisa
BRAPCI	113	5	4
BDTD	823	79	14
BDA	9	1	8
Scopus (ing)	126566	92	52
Scopus (esp)	255	0	0
Total	127766	177	78

Fonte: Elaborada pelo autor (2023)

[3] Fonte: https://www.elsevier.com/pt-br/about

Dentre os recuperados e selecionados, não houve critérios de seleção de tipos documentais, sendo possível a análise de artigos, livros, publicações e materiais didáticos. Conforme a tabela acima exposta, foram selecionados os trabalhos que expressam nos títulos e resumos a abordagem do tema deste trabalho, ou seja, o uso de PLN em arquivos ou voltados para a Arquivologia. O resultado a partir dessa pesquisa levantou 177 publicações, somando todas as bases de dados utilizadas. Dentre os artigos, dissertações e teses buscados, foram selecionadas 78 publicações que estão relacionadas com o tema da pesquisa, e 74 se excluirmos aqueles que estão duplicados em mais de uma base de dados.

No caso do Banco de Dados da Arquivística, o termo "linguagem natural" já está inserido no campo da arquivística, não sendo necessário excluir os dados de produções que não abordam "arquiv*". Na busca por publicações acadêmicas na língua espanhola, não há nenhum trabalho científico abordando a linguagem natural e arquivos. Essas produções consideradas relevantes serão analisadas de acordo com seus objetivos de integração do PLN com arquivos, e suas preocupações com as funções arquivísticas.

Quantitativamente, os dados adquiridos por coleta em base de dados serão distribuídos de acordo com a leitura dos títulos e resumos, mensurando os estudos de acordo com relação do PLN com a Arquivologia, e categorização das funções arquivísticas empregadas no conteúdo das publicações acadêmicas.

Após a identificação das categorias que melhor se enquadram na pesquisa, foram selecionados alguns termos dos títulos e resumos que remetem às funções arquivísticas, exibindo em um quadro que possa comparar as menções e abordagens nacionais e internacionais. Ressalta-se que as pesquisas analisadas podem estar categorizadas dentro de mais de uma função arquivística de acordo com seus objetivos e propostas abordadas.

5 RESULTADO E ANÁLISE DOS DADOS

Para a análise dos dados, a coleta nas bases nacionais (BRAPCI, BDA e BDTD) e internacional (Scopus) foi baseada no conteúdo dos títulos e resumos apresentados nas bases, para categorizar em "quantidade de publicações recuperadas", "quantidade de publicações utilizadas", e as oito funções arquivísticas (diagnóstico, criação/produção, classificação, avaliação, descrição/indexação, entrada, acesso/difusão, preservação/conservação). Como critério, os conteúdos informacionais que interessam para a quantificação foram relacionados

ao tema da pesquisa, aprofundando apenas o teor dos resumos e seus objetivos explanados, relacionando o processamento da linguagem natural aplicada às funções arquivísticas.

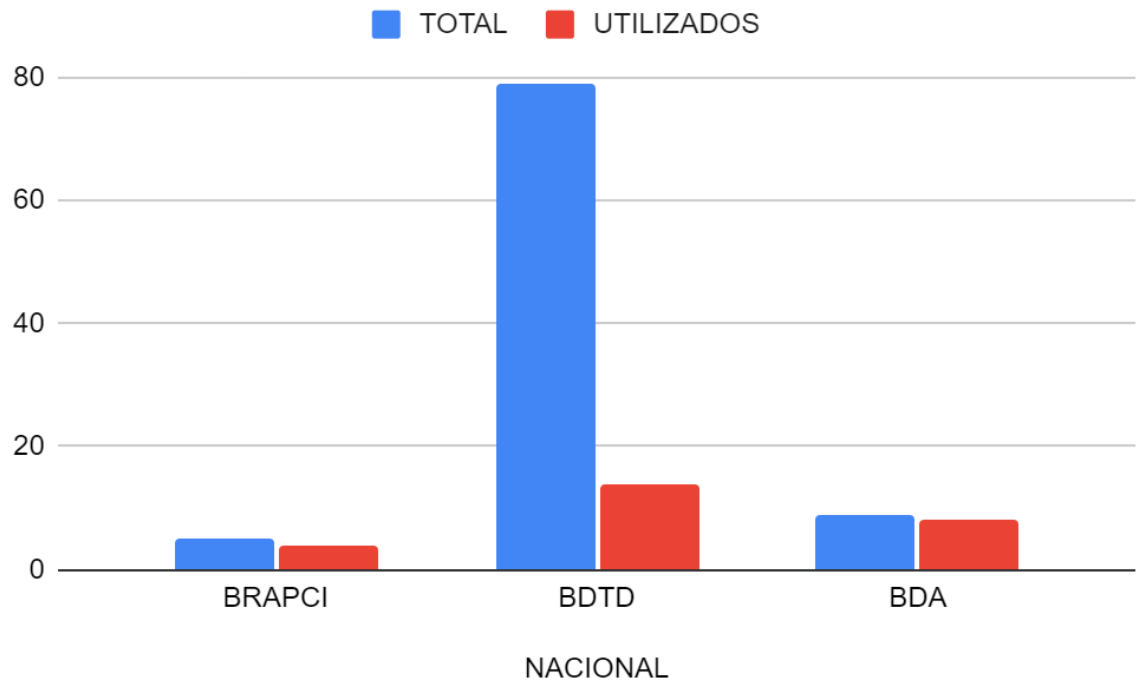
Como abordado nos procedimentos metodológicos, a relação entre o total de publicações recuperadas e o total de trabalhos selecionados de acordo com os critérios, representam 44% de aproveitamento. Observando as publicações nacionais, foi recuperado um total de 93 publicações que integram "linguagem natural" e "arquiv*" (e suas variações). Dentre as publicações de interesse para a pesquisa, foram selecionadas 26, o que representa 28% do total de trabalhos recuperados (93 publicações).

A BRAPCI, voltada para o campo da Ciência da Informação, retornou cinco trabalhos publicados, e foram selecionados quatro desses trabalhos para as análises. Já a BDA, focada para a arquivística, recuperou 9 publicações de acordo com o critério de busca, e apenas uma delas também ficou inutilizada para essa pesquisa. A BDTD, que abrange várias áreas do conhecimento, além de divulgar os trabalhos de graus especializados, como mestrado e doutorado, recuperou 79 publicações ao total, e utilizou-se 14 deles, havendo assim uma maior variação entre os recuperados e utilizados.

No entanto, dentre os recuperados e utilizados, três publicações estão nas bases da BRAPCI e da BDA. Uma delas é intitulada "A indexação na recuperação da informação em arquivos: uma abordagem inicial", produzida por Mariane Costa Pinto (2016), que aborda a teoria da indexação para a recuperação da informação (ou a descrição para a difusão/acesso) em arquivos. Outro trabalho publicado em ambas as bases foi produzida por Maria Almeida, Walter Moreira, Luciana Davanzo e Marcia Vitoriano, com o título "Identificação de elementos para construção do vocabulário controlado: contribuição do diagnóstico de arquivo" (2021), também abordando a teoria da indexação, mas voltado para o diagnóstico.

Já o terceiro trabalho publicado, dos autores Renato Tarciso Barbosa Sousa e Rogério Henrique de Araújo Júnior, de 2013, nomeado "A classificação e a taxonomia como instrumentos efetivos para a recuperação da informação arquivística", é uma proposta de metodologia para auxiliar a classificação de documentos de arquivo. Como diferencial, essa publicação está além das bases nacionais, sendo possível o seu acesso também por bases internacionais, como é o caso da base integrante do estudo aqui apresentado, a Scopus, que possui uma versão em inglês para leitura na base, e está inserida para a quantificação da base internacional.

Figura 1 - Publicações recuperadas em âmbito nacional



Fonte: Elaborada pelo autor (2023)

Observa-se pela figura 1 acima que os trabalhos voltados para a área de Ciências da Informação e Arquivologia possuem menos publicações recuperadas, em comparação com a BDTD, mas seu aproveitamento é maior por abordar os temas principais para o presente trabalho.

Considerando as publicações internacionais, a base Scopus recuperou 92 publicações que abordam "linguagem natural" e "arquivo". Dentre elas, foram selecionados 52 trabalhos publicados, considerando os critérios das informações, havendo uma diferença de 56,5% entre o total e os utilizados para compor as análises.

Como destaques, podemos apontar alguns trabalhos publicados na base internacional, como o de Carlos Eduardo de Lima Joaquim e de Thiago de Paulo Faleiros, intitulado "BERT Abordagem de autoaprendizagem com rótulos limitados para classificação de documentos"[4], que em 2022 abordaram a automação da classificação de documentos para administração

[4] Original: BERT Self-Learning Approach with Limited Labels for Document Classification;

pública brasileira, de modo prático e preocupado com os modelos de requisitos do e-ARQ Brasil.

Outra publicação de destaque é a de Vicenç Ruiz Gómez e Aniol Maria Vallès, que publicaram em 2020 o artigo "#Conte: o caminho entre o ativismo arquivístico e o(s) arquivo(s) social(is)"[5], um estudo prático com a participação de membros da Society of Catalan Archivists and Records Managers (AAC), para utilizar o PLN em postagens de redes sociais, observando a relação das atividades arquivísticas no arquivamento social na web.

Dos dados aqui expostos, observou-se e quantificou-se as publicações com relação às atividades arquivísticas. Dentre as oito funções, os estudos em âmbito nacional abordaram mais a descrição e indexação, seguido do acesso e difusão. Como pode ser observado no quadro 2 e na figura 2, há funções arquivísticas não abordadas ou mencionadas em seus títulos e resumos.

Quadro 2 - Relação das funções arquivísticas abordadas nas produções selecionadas

Funções/Bases	**BRAPCI**	**BDTD**	**BDA**	**Scopus**
Diagnóstico	1	5	1	35
Criação	0	3	0	7
Classificação	1	6	1	15
Avaliação	0	0	0	4
Descrição/Indexação	4	14	8	45
Entrada/Aquisição	0	0	0	3
Acesso/Difusão	2	12	6	44
Preservação/Conservação	0	3	0	19

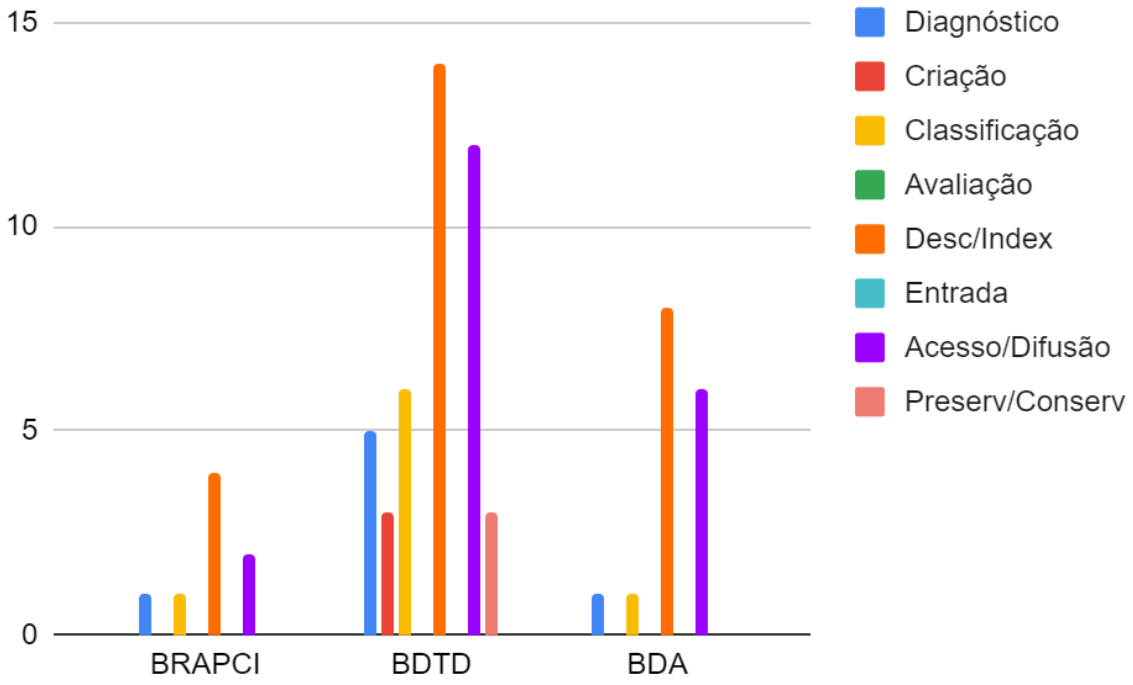
Fonte: Elaborada pelo autor (2023)

Os números exibidos acima representam quantitativamente as abordagens das funções arquivísticas com o PLN, com algumas repetições de trabalhos em mais de uma base de dados, conforme abordado anteriormente. Para articularmos cada função, relacionamos os números comparando as bases nacionais e internacionais. A figura abaixo representa

[5] Original: #Cuéntalo: the path between archival activism and the social archive(s)

graficamente a relação das publicações recuperadas nas bases de dados nacionais, que abordam as funções arquivísticas em seus títulos e resumos.

Figura 2 - Quantidade de menções às funções arquivísticas em âmbito nacional



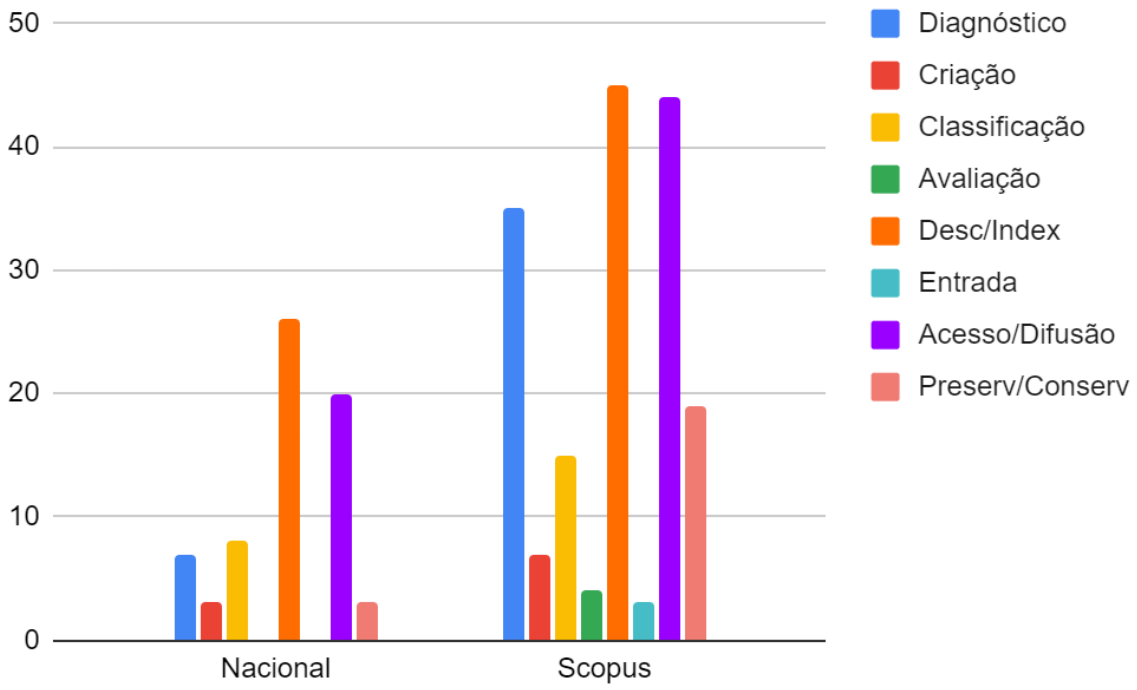
Fonte: Elaborada pelo autor (2023)

Todos os trabalhos publicados nas bases nacionais abordam a descrição/indexação com o PLN em arquivos, demonstrando o principal objetivo de automação para a interação com documentos. Percebe-se também a falta de abordagem em outras funções, como a avaliação documental e a entrada de documentos, e apenas na BDTD a função de criação/produção e de preservação/conservação são mencionados.

Em âmbito nacional, a função de diagnóstico possui apenas 7 menções, enquanto na base da Scopus foram 35 trabalhos que abordaram o tema. Similar disparidade está na atenção à preservação/conservação, sendo 3 trabalhos nacionais abordando a função, enquanto 19 trabalhos internacionais preocupam-se com a automação da função.

Comparando as publicações, ainda podemos observar as menções das funções avaliação (4 menções) e entrada (3 menções) na Scopus, não mencionadas nas bases nacionais, mas podendo ser possível integrar às atividades arquivísticas.

Figura 3 - Comparação das funções arquivísticas mencionadas em âmbito nacional e internacional



Fonte: Elaborada pelo autor (2023)

Somando os números de possibilidades que o PLN pode proporcionar aos arquivos, as publicações nacionais comparadas às internacionais são discrepantes em alguns aspectos, como pode ser observado na figura 3.

Para uma melhor análise das abordagens entre as publicações nacionais e internacionais, exemplificamos alguns trabalhos selecionados que mencionam explicitamente ou implicitamente as funções arquivísticas, comentando as similaridades e diferenças entre as publicações, assim possibilitando elaborar um quadro que será apresentado posteriormente com os termos mencionados nos resumos expostos nas bases de dados em âmbito nacional e internacional.

Observando as menções da função de diagnóstico nos trabalhos selecionados em âmbito nacional, estudos da área da Ciência da Informação e Arquivologia abordam mais explicitamente o tema, enquanto os trabalhos de teses de mestrado e doutorado de várias áreas do conhecimento estão preocupados nas resoluções de problemas, apontando os possíveis pontos adversos e suas resoluções. Explicitamente, podemos exemplificar o trabalho já mencionado anteriormente, de Maria Almeida, Walter Moreira, Luciana Davanzo e Marcia Vitoriano (2021), que mencionam o "diagnóstico" como um elemento importante para a resolução do problema exposto pelos autores no título e no resumo. Implicitamente, Rinaldo José de Lima (2009) trabalha, em sua tese de mestrado, "Extração de informações adaptativas de páginas web por indução supervisionada por extratores"[6], a função de diagnóstico preocupado com a "identificação de informações", abordando a linguagem natural para extração de informações de documentos estruturados e não-estruturados.

Em âmbito internacional, as menções de diagnóstico também são similares às teses nacionais, de forma implícita, partindo principalmente dos problemas relacionados às linguagens documentais para a linguagem técnica (de interesse da área de conhecimento). Em 2022, um coletivo de autores italianos publicaram o artigo "Novas perspectivas para a gestão de patrimônios multilíngues e multialfabéticos por meio da extração automática de conhecimento: a abordagem DigitalMaktaba"[7], mencionando "*establish procedures*" como elemento de resolução de diagnóstico, para então possibilitar "procedimentos para criação, gerenciamento e catalogação de patrimônio arquivístico em alfabeto não-latino" (BERGAMASCHI et al., 2022, tradução nossa).

A função de criação/produção de arquivos é mencionada apenas em três trabalhos nacionais, que focam na automação da tarefa, mas integrada à função do acesso e da recuperação da informação. Como exemplos podemos citar a formação de repositórios, elaboração de "perguntas frequentes" a partir da interação com os usuários. Como destaque, elencamos a automatização de contratos, apresentado por Marina Vieira em sua tese de mestrado no ano de 2022, intitulado "Representação de contratos inteligentes como diagramas

[6] Original: Extraction d information adaptative de pages web par induction supervisée d extracteurs

[7] Original: Novel Perspectives for the Management of Multilingual and Multialphabetic Heritages through Automatic Knowledge Extraction: The DigitalMaktaba Approach

de estado"[8], que visa a "criação" utilizando a linguagem natural para facilitar o entendimento e contribuição dos usuários.

Em trabalhos internacionais, há menções implícitas e explícitas relacionadas à criação/produção, com exemplos de criação de arquivos (blocos) a partir de coletas (menção implícita), e produção de catálogos de patrimônios arquivísticos em alfabetos não latinos (menção explícita). Destacamos o trabalho intitulado "Uma visão computacional em arquivos de história oral"[9], produzido por Pessanha e Salah em 2022, que cita o termo "*creation*" para as possibilidades de análises que integram o reconhecimento de voz com PLN.

Menções à função de classificação são mais explícitas em âmbito nacional, relacionando as palavras "categorias", "diversidades" e "seleção" para exemplificar as necessidades e resoluções dos trabalhos publicados. Destacamos o trabalho de Kamilla Cardoso, intitulado "FINDOS – Uma ferramenta para identificação automática de unidades de rastreamento", publicado em 2016, que aborda os problemas de padronizações, e cita o "contexto organizacional" como um elemento importante para a solução.

Em trabalhos internacionais, há trabalhos explícitos que mencionam "*classification*", e implícitos que utilizam as palavras como "*similar*", "*links*", "*structure*" e "*workflow*", que traduzidos podem se relacionar com trabalhos onde as relações ou ligações entre arquivos auxiliam para a automatização. Em 2022, os franceses Bechet, Antoine, Auguste e Damnati produziram o artigo "Geração e resposta de perguntas para explorar coleções de Humanidades Digitais"[10], que aborda a exploração dos documentos digitais que tenham relação com as coleções documentais.

A avaliação é uma das funções com menos menções em trabalhos publicados, não havendo publicações nacionais que abordam o tema, e apenas quatro publicações internacionais que relatam as preocupações com a avaliação documental. Entre as menções, todos relacionam a guarda dos arquivos, podendo exemplificar os cuidados dos trabalhos que mencionam os sistemas legais de acúmulo e guarda de informações, agregando apenas aqueles de interesse dos trabalhos propostos, observando a sustentabilidade digital.

---

[8] Original: Representation of smart contracts as state diagrams
[9] Original: A Computational Look at Oral History Archives
[10] Original: Question Generation and Answering for exploring Digital Humanities collections

Como exemplo de menção, o trabalho apresentado por Christopher A. Lee em 2018 , "Avaliação e seleção assistida por computador de materiais de arquivo"[11], abordam os termos "appraisal" e "selection", como elementos do problema, necessitando abordar a linguagem natural com a linguagem computacional para auxiliar na automatização da avaliação e seleção de documentos e arquivos que objetivem a descrição e preservação desses materiais.

Descrição/Indexação é a função arquivística mais mencionada nos trabalhos publicados, sendo a principal ferramenta para o sucesso do PLN, que necessita de comparações com outras linguagens, como as linguagens técnicas (medicina, biologia), linguagem computacional (algoritmos), e até linguagem comportamental (emoções e reações). Nas publicações nacionais, principalmente aquelas voltadas para a Ciência da Informação e Arquivologia, todas as selecionadas mencionam ou relacionam a descrição e indexação como elemento principal integrante da linguagem natural, como o trabalho de Xavier, Silva e Gomes, de 2015, intitulado "Uma arquitetura híbrida para a indexação de documentos do Diário Oficial do Município de Cachoeiro de Itapemirim", abordando a mineração de textos para estruturar as informações.

Já em publicações internacionais, podemos destacar o trabalho de Jane Greenberg, que em 2018 apresentou "A aplicabilidade do processamento de linguagem natural (PLN) às propriedades e objetivos arquivísticos"[12], abordando as possibilidades que o PLN pode proporcionar em questão de velocidade e consistência de resultados operacionais para arquivos em meio eletrônico.

Outro tema abordado neste trabalho com menos menções recuperadas é a função de entrada, não abordada nas publicações nacionais, mas exposto em três trabalhos internacionais que trabalham a coleta e acúmulo de documentos, com duas publicações voltadas para opiniões e entrevistas, e uma voltada para a submissão de artigos acadêmicos. Citando "*submit*" como elemento de interesse, o trabalho apresentado pelos singapurenses Tian, Kashyap e Kan, "ServiceMarq: Extraindo contribuições de serviço da chamada de artigos"[13], os autores abordam a necessidade de organização dos trabalhos acadêmicos enviados de forma

[11] Original: Computer-Assisted Appraisal and Selection of Archival Materials

[12] Original: The Applicability of Natural Language Processing (NLP) to Archival Properties and Objectives

[13] Original: ServiceMarq: Extracting Service Contributions from Call for Papers

voluntária por acadêmicos, assim preocupados com a conferência dos arquivos adquiridos em meio digital.

Parte integrante do PLN, o acesso e difusão de arquivos é uma das funções importantes para o sucesso da recuperação da informação, que é mencionada explicitamente em grande parte dos trabalhos nacionais selecionados. Implicitamente, as publicações que mencionam a função abordam a relação dos usuários com os arquivos disponíveis. Mariane Costa Pinto, publicou em 2016 o trabalho "Proposta para a criação de um vocabulário controlado a partir do Sistema de Informações do Arquivo Nacional do Brasil (SIAN)", abordando a "recuperação da informação" como um dos objetivos, beneficiando a relação entre os especialistas e os usuários.

Em publicações internacionais, a preocupação dos estudos não é voltada exclusivamente para a recuperação, ampliando os interesses aos compartilhamentos de pesquisas, feedback de usuários, inclusão social e diversidade. Em setembro de 2022, Giannini e Bowen abordaram no artigo "Cultura computacional: transformando a prática e a educação em arquivos para um mundo pós-Covid"[14], a preocupação dos sistemas de arquivos que são complexos, mas podem ser culturalmente "*inclusive and diverse*".

A função de preservação e conservação é pouco mencionada em âmbito nacional, com três teses selecionadas que abordam o futuro das informações, preocupados com o repositório que irão submeter as informações e os suportes que poderão possibilitar o acesso futuro. Katiuscia Andrade em sua dissertação publicada em 2013, "Astrolábio: um corpus de redações escolares do Ceará anotado multidimensionalmente conforme a TEI P5", menciona a preservação das características estruturais e dos conteúdos transcritos a partir das codificações.

Em publicações internacionais, as preocupações estão na interoperabilidade dos dados e na longa duração do acesso das informações, com a transformação das informações não estruturadas em dados (para acesso digital). Além da informação do futuro, os trabalhos internacionais também abordam os registros históricos, relacionando pessoas e atividades com as coleções (livros, fotos, áudios, vídeos digitais e analógicos), abordando o social e não apenas o digital. Preocupados com acontecimentos históricos, Clough, Tang, Hall e Warner abordaram em 2011, no artigo intitulado "Vinculando dados de arquivo à localização: um

[14] Original: Computational Culture: Transforming Archives Practice and Education for a Post-Covid World

estudo de caso nos Arquivos Nacionais do Reino Unido”[15], a relação de acontecimentos históricos registrados com a localização geográfica, baseando-se na necessidade de guarda e manutenção de arquivos com mais de mil anos, em formato físico e digital.

Quadro 3: termos mencionados em publicações que referenciam as funções arquivísticas

	Nacional	**Internacional**
Diagnóstico	"diagnóstico"; "verificar"; "mapeamento"; "análise exploratória"; "identificação de informações"; "identificar, armazenar e rastrear"; "identificação";	"diagnoses"; "analysis"; "general issues"; "establish procedures"; "problem"; "requires"; "benchmark valuations"; "recognized"; "challenges"; "identify"; "selection and appraisal"; "management";
Criação Produção	"criação"; "montar"; "geração";	"creating"; "incorporating"; "documenting";
Classificação	"classificação"; "segmentos"; "estrutura"; "coleções"; “contexto organizacional”;	"classification"; "classifying"; "collections"; “links”; "relationship"; "similarity"; "functional requirements and workflow"; "framework"; "selection"; "structured";
Avaliação	*	"appraisal"; "analysis"; "value"; "law - storing";
Descrição Indexação	"descrição"; "indexação"; "texto"; "termos"; "semântica"; "léxico"; "sintático"; "padrões"; "dados"; "expressão"; "vocabulário controlado"; "linguagem documentária"; "tags - etiquetas";	"description"; "index"; "data"; "metadata"; "textual"; "language"; "extract"; "semantic"; "ISAD(G)"; "datification"; "pattern"; "representation"; "hashtags"; "summarization";
Entrada	*	"submit"; "collect";
Acesso Difusão	"acesso"; "recuperação"; "resgatar"; "diálogo"; "localizar"; "pesquisa"; "retornar"; "prover"; "rastrear"; "disponibilização";	"retrieval"; "research"; "share"; "engagement"; "inclusive"; "adaptable"; "access"; "transcription"; "visualize"; "interaction"; "precision"; "accuracy"; "finding"; "result"; "provide"; "query"; "browsing";

[15]Original: Linking archival data to location: a case study at the UK National Archives

Preservação Conservação	"preservação"; "futuro";	"historical"; "long-term archival"; "transform"; "storing"; "preserve"; "convert"; "digitalized"; "maintain"; "memory";

Fonte: Elaborada pelo autor (2023)

Como observado nas análise de funções arquivísticas e comparado dos termos utilizados nos trabalhos selecionados, há diferentes focos de utilização do PLN para a mesma função arquivística, como pode ser observado na função de acesso e difusão, com trabalhos nacionais mais voltados para a recuperação da informação, enquanto os trabalhos internacionais ampliam os estudos para a adaptação e inclusividade, voltado para o usuário. Outro ponto importante para observarmos são as abordagens das funções de avaliação e entrada, sendo ausente nas publicações nacionais e pouco explorada em âmbito internacional.

6 CONSIDERAÇÕES FINAIS

Com o presente trabalho pôde-se conhecer e apresentar uma gama de possibilidades de trabalhar o PLN aplicado às funções arquivísticas, não abordando apenas a descrição e indexação, mais exploradas, mas podendo ser ampliado para outras funções aplicadas à Arquivologia, da teoria à prática, da ação tecnológica à ação social.

O processamento de linguagem natural é apenas uma das áreas de estudo para a inteligência artificial, considerada o futuro da tecnologia, automatizando atividades antes praticadas apenas pela mão humana, mas não necessitando a sua exclusão. O futuro depende da adaptabilidade para a experiência do usuário, obedecendo a regras estabelecidas, assim não havendo possibilidade de extrapolá-las, e objetivando sempre o bem da sociedade.

Acredita-se que os objetivos da pesquisa foram alcançados, tanto o objetivo geral quanto os específicos. Foi possível analisar as produções acadêmicas que abordam o PLN aplicada à Arquivologia, principalmente observando as funções arquivísticas integradas aos propósitos dos estudos. O objetivo específico de levantamento das produções nos revela números que trazem reflexões, principalmente em relação àqueles direcionados para a Arquivologia, mostrando menor número de publicações interessadas em abordar o PLN, e se comparados com os de interesse da Tecnologia da Informação, há uma distância considerada. Outro fator que nos toma a atenção é a ausência de produções em língua espanhola, algo

incomum para uma base de dados como a Scopus, que agrega trabalhos para além da língua inglesa, mas possibilita futuras pesquisas que integrem outras bases de dados para ampliar a gama de publicações de interesses semelhantes.

O objetivo específico de comparar as publicações recuperadas com relação às funções arquivísticas nos mostra que as nacionais abordam, em grande maioria, o diagnóstico, a indexação e a recuperação da informação, objetivo fundamental para o sucesso do PLN. Diferentemente, as publicações internacionais ampliam seus objetivos, mesmo não sendo integralmente voltado para a Arquivologia, mas abordam, nas mesmas funções abordadas em âmbito nacional, outros interesses e objetivos, como o exemplo da preservação e conservação. Mas também, há menções com funções pouco exploradas em seu total, como a entrada de documentos e avaliação, não abordada nacionalmente e pouco explorada internacionalmente.

E no objetivo de identificar as possibilidades de relação das funções arquivísticas com o PLN, observou-se que nos títulos e resumos dos trabalhos selecionados, os termos selecionados no quadro 3 necessitam de contexto com outras palavras e frases, que mesmo assim é possível observar nas menções suas preocupações, em que nas publicações nacionais estão voltadas para a qualidade dos resultados, enquanto as publicações internacionais preocupam-se com os usuários, como pode-se notar nos termos que estão relacionados à adaptação, acessibilidade, e preservação da história.

A Arquivologia também está em constante evolução, necessitando de profissionais capacitados para manusear do físico (papel, pedra, discos, entre outros) ao digital (algoritmos, softwares, banco de dados, entre outros), objetivando a qualidade na entrega aos usuários, do serviço à informação, da melhor maneira no menor tempo. Não é imprescindível o profissional arquivista possuir domínio de programação, mas é necessário que tenha certo conhecimento a respeito, para que sejam elaborados processos que visem a aplicação e prática dos fazeres arquivísticos, hoje muito vinculados ao desenvolvimento de ferramentas, como o uso da inteligência artificial.

A interação das atividades arquivísticas com o processamento de linguagem natural ocorre principalmente na produção de descrições e indexações, por se tratar de linguagens simplificadas para uma resposta (resultado) correta, ou aproximadamente certa para as necessidades dos usuários. Observa-se a interação como um meio para o fim, mas pode ser aplicado do início ao fim, como a partida de criação à preservação. Mas esses estudos

analisados são iniciais, podendo ser aprofundado com mais elementos tecnológicos e atividades arquivísticas.

REFERÊNCIAS

ALMEIDA, Maria F. I. de; MOREIRA, Walter; DAVANZO, Luciana; VITORIANO, Marcia C. de C.. Identificação de elementos para construção do vocabulário controlado: contribuição do diagnóstico de arquivo. **Inf. Inf.**, Londrina, v. 26, n. 1, p. 601-631, jan./mar. 2021. Disponível em: https://ojs.uel.br/revistas/uel/index.php/informacao/article/view/38986/pdf Acesso em: 03 out. 2023.

ANDRADE, Katiuscia de Moraes. **Astrolábio: um corpus de redações escolares do Ceará anotado multidimensionalmente conforme a TEI P5**. Dissertação (Mestrado) – Universidade Federal do Ceará, Departamento de Letras Vernáculas, Programa de Pós-graduação em Linguística, Fortaleza (CE), 2013. 135f. Disponível em: https://repositorio.ufc.br/handle/riufc/8195 Acesso em: 05 out. 2023.

ARQUIVO NACIONAL. **Dicionário brasileiro de terminologia arquivística**. Rio de Janeiro: Arquivo Nacional, 2005. Disponível em: https://simagestao.com.br/wp-content/uploads/2016/01/Dicionario-de-terminologia-arquivistica.pdf Acesso em: 18 jun. 2023.

BARROS, Flávia de Almeida; ROBIN, Jacques. Processamento de Linguagem Natural. **Revista Eletrônica de Iniciação Científica** , v. 1, p. 1-61, 2001. Disponível em: https://www.cin.ufpe.br/~fab/cursos/jai96/ProcessamentoDeLinguagemNatural.pdf Acesso em: 04 jun. 2023.

BECHET, Frederic; ANTOINE, Elie; AUGUSTE, Jérémy; DAMNATI,Géraldine. Question Generation and Answering for exploring Digital Humanities collections. *In.:* **Proceedings of the Thirteenth Language Resources and Evaluation Conference**, Marseille, France. European Language Resources Association, jun. 2022, p. 4561–4568. Disponível em: https://aclanthology.org/2022.lrec-1.486.pdf Acesso em: 04 out. 2023.

BERGAMASCHI, S.; DE NARDIS, S.; MARTOGLIA, R.; RUOZZI, F.; SALA, L.; VANZINI, M.; VIGLIERMO, R.A. Novel Perspectives for the Management of Multilingual and Multialphabetic Heritages through Automatic Knowledge Extraction: The DigitalMaktaba Approach. **Sensors**, 2022, n. 22 (11):3995. Disponível em: https://doi.org/10.3390/s22113995 Acesso em: 03 out. 2023.

BERNARDES, Ieda Pimenta. **Como avaliar documentos de arquivo**. São Paulo: Arquivo do Estado, 1998. Disponível em: https://www.arqsp.org.br/arquivos/oficinas_colecao_como_fazer/cf1.pdf Acesso em: 09 dez. 2023.

BRÜGGEMANN, Odaléa M.; PARPINELLI, Mary Â.. Utilizando as abordagens quantitativa

e qualitativa na produção do conhecimento. **Revista da Escola de Enfermagem da USP**, São Paulo, 2008.

CAMPOS, Maria Luiza de Almeida. **Linguagem documentária**: teorias que fundamentam sua elaboração. Niterói: RJ, EdUFF, 2001. Disponível em: https://bibliotextos.files.wordpress.com/2011/09/livro-linguagem.pdf Acesso em: 06 dez. 2023.

CARDOSO, Kamilla R. F.. **FINDOS**: uma ferramenta para identificação automática de unidades de rastreamento. Dissertação (mestrado em Ciência da Computação). Universidade Federal de Pernambuco. CIn, Ciência da Computação, Recife, 2016. Disponível em: https://repositorio.ufpe.br/handle/123456789/28086 Acesso em: 04 out. 2023.

Clough, P., Tang, J., Hall, M.M. and Warner, A. Linking archival data to location: a case study at the UK National Archives. **Aslib Proceedings**, 2011, v. 63 n. 2/3, p. 127-147. Disponível em: https://doi.org/10.1108/00012531111135628 Acesso em: 05 out. 2023.

COUTURE, Carol. La discipline archivistique au Canada: état de développement et perspectives d'avenir. **In Situ, Revu dês Patrimoines**, Canadá, v. 30, 2016. Disponível em: https://journals.openedition.org/insitu/13669#quotation. Acesso em: 14 dez. 2020.

COUTURE, Carol et al. **Les fonctions de l`archivistique contemporaine**. Sainte-Foy (Québec) Canadá: Presses de L'Universite du Québec, 2003.

FLICK, Uwe. **Qualidade na pesquisa qualitativa**. Tradução: Roberto Cataldo Costa. Porto Alegre: Editora Artmed, 2009.
GIL, Antonio Carlos. **Como elaborar projetos de pesquisa**. São Paulo: Atlas, 6 ed., 2017.

GEANNINI, Tula; BOWEN, Jonathan P.. Computational Culture: Transforming Archives Practice and Education for a Post-Covid World. **Journal on Computing and Cultural Heritage**, 2022, v. 15 (3) n. 47 p. 1–18. Disponível em: https://doi.org/10.1145/3493342 Acesso em: 05 out. 2023.

GREENBERG, Jane. The Applicability of Natural Language Processing (NLP) to Archival Properties and Objectives. **The American Archivist**, v. 61 (2), p. 400–425, jan. 1998. Disponível em: https://doi.org/10.17723/aarc.61.2.j3p8200745pj34v6 Acesso em: 05 out. 2023.

JOAQUIM, C.E.L.; FALEIROS, T.P. BERT Self-Learning Approach with Limited Labels for Document Classification. **16th International Conference on Learning and Intelligent Optimization**, LION 16 2022; Milos Island; Greece; 5 June 2022 through 10 June 2022. Disponível em: https://www.scopus.com/record/display.uri?eid=2-s2.0-85148488276&doi=10.1007%2f978-3-031-24866-5_21&origin=inward&txGid=60fc0c0fed2cf5b462b7f36454c14b48 Acesso em: 26 nov. 2023.

LEE, Christopher A.. Computer-Assisted Appraisal and Selection of Archival Materials. **IEEE International Conference on Big Data (Big Data)**, 2018. Disponível em: https://ils.unc.edu/callee/p2721-lee.pdf Acesso em: 05 out. 2023.

LIMA, Rinaldo José de; FREITAS, Frederico Luiz Gonçalves de. **Extraction d information adaptative de pages web par induction supervisée d extracteurs**. Dissertação (Mestrado). Programa de Pós-Graduação em Ciência da Computação, Universidade Federal de Pernambuco, Recife, 2009. Disponível em: https://repositorio.ufpe.br/handle/123456789/2000 Acesso em: 03 out. 2023.

LEHMKUHL, Camila Schwinden; SILVA, Eva Cristina Leite da. **Registros civis no Brasil frente às funções arquivísticas**. 2021. 226 p. Tese (Doutorado) - Universidade Federal de Santa Catarina, Centro de Ciências da Educação, Programa Pós-Graduação em Ciência da Informação, Florianópolis, 2021. Disponível em: http://www.bu.ufsc.br/teses/PCIN0256-T.pdf Acesso em: 13 maio 2021.

PEREIRA, Silvio do Lago. Processamento de linguagem natural. São Paulo, USP, 2011. Disponível em: https://www.ime.usp.br/~slago/IA-pln.pdf Acesso em: 04 jun. 2023.

PESSANHA, F.; SALAH, A. A.. A Computational Look at Oral History Archives. **Journal on Computing and Cultural Heritage**, v. 15, Issue 1, Article number 6, February 2022. Disponível em: https://www.scopus.com/record/display.uri?eid=2-s2.0-85123450763&doi=10.1145%2f3477605&origin=inward&txGid=5d0618beda9eade377ec93abffa4fe9c Acesso em: 04 out. 2023.

PINTO, Mariane Costa. Proposta para a criação de um vocabulário controlado a partir do Sistema de Informações do Arquivo Nacional do Brasil (SIAN). CONGRESSO NACIONAL DE ARQUIVOLOGIA - CNA, 7., 2016, Fortaleza. **Anais eletrônicos**... Revista Analisando em Ciência da Informação - RACIn, João Pessoa, v. 4, n. especial, p. 479-490, out. 2016. Disponível em: http://arquivologiauepb.com.br/racin/edicoes/v4_nesp/racin_v4_nesp_artigo_0479-0490.pdf Acesso em: 05 out. 2023.

REIS, Luís. O arquivo e arquivística evolução histórica. **Biblios**, Peru, ano 7, nº 24, abr. - jun. 2006. Disponível em: https://portal.tcu.gov.br/lumis/portal/file//fileDownload.jsp?fileId=8A8182A14D056C05014D061501DB4A94 Acesso em: 18 jun. 2023.

RINO, Lucia Helena Machado. **Uma interface em linguagem natural para recuperação do conhecimento**. Tese (Mestrado em Ciências de Computação) - Faculdade de Ciências da Computação e Matemática Computacional, Universidade de São Paulo, São Paulo, 1987. Disponível em: https://www.teses.usp.br/teses/disponiveis/55/55134/tde-08102019-170319/publico/LuciaHelenaMachadoRino_ME.pdf Acesso em: 04 jun. 2023.

ROUSSEAU, Jean-Yves; COUTURE, Carol. **Os fundamentos da disciplina arquivística**:

glossário. Lisboa, Portugal: Publicações Dom Quixote, 1998.

RUIZ GÓMEZ, V.; MARIA VALLÈS, A. #Cuéntalo: the path between archival activism and the social archive(s). **Archives and Manuscripts**, v. 48, issue 3, 2020. Disponível em: https://link.springer.com/chapter/10.1007/978-3-031-24866-5_21 Acesso em: 26 nov. 2023.

SCHNEIDER, Eduarda M.; FUJII, Rosangela A. X.; CORAZZA, Maria J. Pesquisas quali-quantitativas: contribuições para a pesquisa em ensino de ciências. **Revista Pesquisa Qualitativa**, São Paulo, v. 5, n. 9, p. 569-584, dez. 2017.

SILVA, Wallace. História do processamento de linguagem natural. Medium, 27 jan. 2019. Disponível em: https://medium.com/@wallacehsilva/hist%C3%B3ria-do-processamento-de-linguagem-natural-e2c363a4eac0 Acesso em: 18 jun. 2023.

SOUSA, Renato T. B. de; ARAÚJO JÚNIOR, Rogério H. de. A classificação e a taxonomia como instrumentos efetivos para a recuperação da informação arquivística. **Ciência da Informação**, Brasília, jan./abr. 2013. Disponível em: https://revista.ibict.br/ciinf/article/view/1400/1578 Acesso em: 04 jun. 2023.

TIAN, Shi; KASHYAP, Abhinav Ramesh; KAN, Min-Yen. ServiceMarq: Extracting Service Contributions from Call for Papers. *In.:* **ACM Symposium on Document Engineering 2020 (DocEng '20)**, Virtual Event, CA, USA. ACM, New York, NY, USA, September 29-October 2, 2020. Disponível em: https://doi.org/10.1145/3395027.3419596 Acesso em: 05 out. 2023.

VIEIRA, Marina L. L.. **Representation of smart contracts as state diagrams**. Dissertação (mestrado) - Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós-Graduação em Ciência da Computação, Florianópolis, 2022. Disponível em: https://repositorio.ufsc.br/handle/123456789/241048 Acesso em: 04 out. 2023.

XAVIER, Bruno M.; SILVA, Alcione D. da; GOMES, Geórgia R. R.. Uma arquitetura híbrida para a indexação de documentos do Diário Oficial do Município de Cachoeiro de Itapemirim. **TransInformação**, Campinas, 27(1):83-95, jan./abr., 2015. Disponível em: https://www.scielo.br/j/tinf/a/mJmTKbL94hj89q9p8HfCnLj/?format=pdf&lang=pt Acesso em: 05 out. 2023.