Universidade de Brasília – UnB

Faculdade de Ciência da Informação – FCI

A PRÁTICA DA INDEXAÇÃO AUTOMÁTICA NO DSPACE PELAS BIBLIOTECAS DIGITAIS E REPOSITÓRIOS INSTITUCIONAS DE BRASÍLIA

Juliana Araujo Gomes de Sousa

Orientadora: Profa. Dra. Fernanda Souza Monteiro

Brasília

2015

JULIANA ARAUJO GOMES DE SOUSA

A PRÁTICA DA INDEXAÇÃO AUTOMÁTICA NO DSPACE PELAS BIBLIOTECAS DIGITAIS E REPOSITÓRIOS INSTITUCIONAS DE BRASÍLIA

Trabalho de Conclusão de Curso apresentado ao Curso de Biblioteconomia da UnB como requisito parcial para a obtenção do título de Bacharel em Biblioteconomia.

Orientadora: Prof. Dra. Fernanda Souza Monteiro

Brasília

2015

Titulo: A prática da indexação automática no Dspace pelas bibliotecas digitais e repositórios institucionais de Brasília.

Aluna: Juliana Araújo Gomes de Sousa.

Monografia apresentada à Faculdade de Ciência da Informação da Universidade de Brasília, como parte dos requisitos para obtenção do grau de Bacharel em Biblioteconomia.

Brasília, 15 de dezembro de 2015.

Fernanda de Souza Monteiro - Orientadora
Professora da Faculdade de Ciência da Informação (UnB)
Doutora em Ciência da Informação

Dulce Maria Baptista – Membro
Professora da Faculdade de Ciência da Informação (UnB)
Doutora em Ciência da Informação

Milton Shintaku – Membro externo
Coordenador de Articulação, Geração e Aplicação de Tecnologia (IBICT)
Doutor em Ciência da Informação

S725p

Sousa, Juliana Araujo Gomes de

A prática da indexação automática no DSpace pelas bibliotecas digitais e repositórios institucionais de Brasília / Juliana Araujo Gomes de Sousa.- 2015.

82 f. : il.

Orientadora: Fernanda Souza Monteiro

Monografia (Graduação) – Universidade de Brasília, Faculdade de Ciência da Informação, Curso de Graduação em Biblioteconomia, 2015.

1. Indexação automática. 2. DSpace I. Título.

CDU 025

AGRADECIMENTOS

A meus pais, pela dedicação, compreensão, paciência e por acreditarem que um dia eu iria terminar este trabalho.

A minha irmã, pelas conversas e pelo apoio incondicional.

Aos amigos, que me acompanharam durante toda a minha vida acadêmica e que tornaram a vida na UnB mais feliz.

Ao Alfredo´s, lugar que proporcionou muitas conversas, desabafos, ideias, bons drinks e muita pizza.

A professora Fernanda, pela dedicação, pela orientação, pela paciência e incentivo durante toda a elaboração deste trabalho.

Finalmente, agradeço a todas as pessoas que, de alguma forma contribui para a realização deste trabalho.

“Eu estava deitado na cama a noite e disse: “Eu vou desistir, pro inferno com isso!”. E outra voz em mim dizia: “Não desista! Salve uma pequena brasa, uma faísca. E nunca dê essa faísca, pois enquanto você a tiver, sempre poderá começar uma chama maior.”

Charles Bukowski

RESUMO

Tendo em vista o aumento exponencial das publicações em meio digital e a dificuldade de indexar grandes quantidades de documentos com assuntos diferentes, foi realizada uma pesquisa com o intuito de compreender qual o motivo que os repositórios e as bibliotecas digitais de Brasília não utilizam a indexação automática de documentos textuais em seu acervo digital. Baseando-se nisso, foi realizado um estudo de caso sobre os repositórios institucionais e as bibliotecas digitais de Brasília que utilizam o software DSpace, que por sua vez, possibilita a indexação automática de documentos textuais que compõem a sua base de dados. Para tal, realizou-se uma pesquisa mista, com base na aplicação de questionário via e-mail e entrevista. Por meio da análise dos dados que foram coletados, concluiu-se que a não utilização da indexação automática está vinculada a falta de conhecimento específico dos profissionais bibliotecários. Entretanto, todos que não possuem conhecimento mostraram-se interessado em saber como a indexação automática de texto completo funciona no DSpace, com o intuito de implementa-la futuramente.

Palavras-chave: indexação, indexação automática, indexação automática no DSpace, DSpace.

ABSTRACT

In reason of the exponential growth of publications in digital media and the difficulty to index a lot of documents with different subjects a research was conducted in order to understand for what reason Brasilia's repositories and digital libraries do not use automatic indexing for textual documents in its digital collection. Based on that, a case study was conducted about the institutional repositories and digital libraries in Brasília that use Dspace software, which enables automatic indexing of textual documents that makes its database. To this end, a mixed survey was carried out, based on the questionnarie via e-mail and interview. By analyzing the data that was collected, it was concluded that the non-use of automatic indexation is linked to lack of expertise of librarians. However, all of those who lack knowledge were interested in understading how full-text automatic indexing works in DSpace, in order to implement it in the future.

Palavras-chave: indexing, automatic indexing, Dspace automatic indexing, DSpace.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE SIGLAS

BD	- Biblioteca Digital
BDjur	- Biblioteca Digital Jurídica
BDMPF	- Biblioteca Digital Ministério Público Federal
BDSF	- Biblioteca Digital do Senado Federal
BCE	- Biblioteca Central
CD	- Compact Disc
CNJ	- Conselho Nacional de Justiça
CNMP	- Conselho Nacional do Ministério Público
DFL	- Digital Library Federation
GID	- Gerenciamento da Informação Digital
HP	- Hewllet Packard
IBICT	- Instituto Brasileiro de Informação em Ciência da Informação
KWIC	- Key Word in Context
KWOC	- Key Word out of Context
MIT	- Massachusetts Institute of Technology
MPF	- Ministério Público Federal
OAI	- Open Archives Initiative
PLN	- Processamento de Linguagem Natural

R.I	- Recuperação da Informação
STJ	- Superior Tribunal de Justiça
TST	- Tribunal Superior do Trabalho
UCB	- Universidade Católica de Brasília
UFBA	- Universidade Federal da Bahia
UnB	- Universidade de Brasília
UniCEUB	- Centro Universitário de Brasília
UNISIST	- Sistema Mundial de Informação Científica
URL	- Uniform Resource Locator

Sumário

1 Introdução

Há algum tempo é notório uma queda na produção da mídia impressa, seja jornais, revistas, livros e periódicos científicos. Em contraponto a isso surge o aumento das publicações eletrônicas.

As publicações eletrônicas requerem um baixo custo de produção, já que o próprio autor pode editar, revisar e depositar, reduzindo os custos com editorial e o tempo para se publicar é extremamente menor do que o de uma publicação impressa.

São inúmeras as vantagens que a publicação eletrônica carrega, porém uma das desvantagens é que não há profissionais o suficiente para organizar toda essa informação na mesma proporção em que são publicadas. Devido a isso, muita informação está perdida em meio digital por falta de organização, no sentido em que a representação da informação do que está sendo publicado não tem sido satisfatória, o que compromete diretamente na recuperação da informação (RI). De acordo com Araújo Júnior (2007), a indexação tem papel fundamental na recuperação da informação.

Dessa maneira, Pinto (2000), definiu o processo de indexação como:

> A indexação é uma atividade que desmonta o discurso montado pelo autor do documento, à medida que ela faz recortes neste discurso. Assim, ela permite passar de um documento constituído (um documento primário) à sua reconstituição em um novo documento-índice (um documento secundário), o qual é formado não pela representação do conteúdo do documento inicial, mas pela representação dos elementos indicadores do seu conteúdo e que vão se constituir na chave de acesso a recuperação da informação.

Existem três formas de se indexar: a indexação manual, que é feita por humanos; a indexação automática, que é feita por um software, que pode utilizar diferentes métodos; a indexação que combina os dois tipo, a indexação manual e a indexação automática, que é conhecida como indexação semiautomática.

A indexação automática é opção interessante quando se trata de um acervo digital que contém documentos textuais heterogêneos e que o objetivo é que a informação chegue mais rápido até os usuários.

A fim de auxiliar na indexação automática um dos softwares para implementação de BD e repositórios institucionais mais utilizado que é o DSpace, que disponibiliza essa função.

O DSpace é um software livre de código aberto e totalmente personalizável, mantido pela DuraSpace.É capaz de atender as demandas de qualquer instituição e qualquer tipo de material que se deseja preservar.

O foco deste trabalho se concentra na análise da utilização da indexação automática de documentos textuais, mais especificamente na indexação automática realizada pelo software DSpace.

1.1 Definição do problema

A sociedade tem caminhado para uma realidade que a busca por informação cientifica não é mais feita inicialmente e primordialmente em uma biblioteca tradicional (física), mas sim na web, em que se pode encontrar bibliotecas digitais e periódicos científicos de acesso aberto e repositórios institucionais.

O número de artigos científicos publicados em meio digital é muito alto e o processo de indexação manual é bastante moroso, com isso surgiram vários softwares que utilizam diferentes critérios para realizarem automaticamente a indexação desses documentos. O problema relacionado a esses softwares é que o processo utilizado para fazer a indexação possivelmente não terá o mesmo índice de precisão que uma indexação feita manualmente.

Para ter acesso a informação de forma satisfatória é necessário que os documentos tenham sido indexados de maneira eficiente. Segundo Vieira (1984), a indexação é uma das operações mais significativas que compõe o ciclo documentário.

O processo de indexação manual faz com que o usuário tenha que esperar mais tempo para ter acesso a um documento. Segundo Bertrand (1994), a indexação manual de um documento leva em média 30 minutos para ser feita. Para Pinto (2000), um dos fatores que afetam a qualidade da indexação manual que é a falta de coerência entre indexadores e a dificuldade de se escolher uma grande quantidade de conceitos.

Na indexação manual existem problemas, como citado anteriormente, a morosidade do processo, o tempo que o indexador tem é bastante limitado devido ao aumento exponencial de publicações, a complexidade do assunto, etc. Já em relação a indexação automática tem-se a necessidade de atender os critérios utilizados pelo software e que possivelmente não vão atender aos mesmos critérios estabelecidos para a indexação que é feita manualmente.

Para Neves (2009 apud, RODRÍGUEZ; GONZÁLEZ 1999) são quatro os fatores que fazem pensar em adotar a indexação automática, são eles:

- O alto custo da indexação humana, em termos de tempo, suscitou a ideia de explorar de maneira eficaz, a um custo e tempo reduzidos, o volume constantemente crescente de informação. Essa questão motivou estudos que para comparar a indexação humana e a indexação automática;
- Aumento exponencial da informação eletrônica e a proliferação de textos completos;
- A Gestão Eletrônica de Documentos (GED) e a informatização dos processos documentais;
- A automatização de processos cognitivos e a pesquisa crescente e os avanços em Processamento de Linguagem Natural (PLN). A automatização de processos cognitivos permite o surgimento de sistemas inteligentes, que somados ao PLN, podem lidar com a atividade de indexação. Porém, os autores alertam para complexidade da linguagem e afirmam que um sistema não pode lidar globalmente com ela, sendo capaz apenas de reconhecer cadeias de caracteres.

Apesar do que foi citado acima a indexação manual ainda é prática comum em alguns repositórios nacionais e quiçá internacionais. Ainda com base no pensamento de Pinto (2000), um dos motivos pelos quais a indexação automática ainda não passou a vigorar 100% é que tanto a indexação automática quanto a semiautomática ainda não apresentam resultados satisfatórios na recuperação da informação.

Com base no que foi encontrado na literatura sobre a indexação automática, suas vantagens e adequação ao contexto das bibliotecas e repositórios digitais, é importante investigar se esta opção do software DSpace é utilizada e, em caso negativo, quais são os motivos da não utilização.

A não utilização pode estar relacionada com a falta de conhecimento e/ou treinamento adequado para utilizar a ferramenta e a falta de suporte da equipe de informática das instituições no auxilio da customização e personalização do software, pois para que o software realize a indexação automática é necessário um conhecimento especifico que pode estar além da formação do bibliotecário.

2 Objetivos e justificativa

Este trabalho tem como objetivos:

2.1 Objetivo geral

- Identificar quais são os motivos que levam a não utilização da indexação automática do DSpace nas biblioteca digitais e repositórios institucionais de Brasília.

2.2 Objetivos específicos

- Identificar quais bibliotecas digitais de Brasília utilizam a plataforma de software DSpace;
- Definir o nível de conhecimento dos profissionais sobre indexação automática e a tecnologia disponibilizada pela ferramenta para realizar esse processo;
- Identificar quais as características da indexação automática feita pelo DSpace que não atendem as necessidades de indexação das bibliotecas digitais de Brasília.

2.3 Justificativa

Em repositórios institucionais ou bibliotecas digitais de órgãos públicos os documentos que compõem a base de dados são produzidos pelos seus servidores e nem sempre essas produções estão relacionadas com a área fim da instituição. Essa variedade de assuntos aumenta o grau de dificuldade na indexação manual, pois o profissional levará um pouco mais de tempo para indexar um documento em que o assunto lhe é completamente novo e foge dos padrões a que está acostumado. Por exemplo, em uma biblioteca jurídica os profissionais estão acostumados com assuntos jurídicos, mas quando vão lidar com documentos que foram produzidos por servidores da instituição que acolhe pessoas de várias formações acadêmicas, o profissional pode se deparar com um artigo que trate sobre obras arquitetônicas ou sobre saúde mental.

Atualmente muitas instituições trabalham com vocabulário controlado, mas esse vocabulário não consegue compreender documentos que estão fora da área fim da instituição e que por políticas internas precisam ser disponibilizados para os usuários. Segundo Lancaster (2004) um vocabulário controlado deve melhorar a coerência da indexação quanto aos termos a serem usados para indexar o documento em relação a um grupo de documentos, mas é bem possível que a diminua no nível de um único documento. (LANCASTER, 2004, p. 74).

Visto que o vocabulário controlado é utilizado para manter uma coerência na indexação, pode-se dizer que a indexação automática também é capaz de manter uma coerência, pois o método utilizado para indexar um item será o mesmo para indexar um acervo inteiro.

Com base nesse pensamento e também de acordo com o objetivo das BD e dos repositórios que resumidamente convergem para o mesmo objetivo, que é, disponibilizar tudo que é produzido pela instituição em formato digital que identificou-se a necessidade de averiguar porque a indexação automática não é utilizada, já que poderia diminuir o tempo em que a informação não estaria disponível para os usuários.

Para delimitar o campo de pesquisa, optou-se por estudar os repositórios e as BD que utilizam o software DSpace, porque sabe-se que este software possibilita a indexação automática de texto completo para documentos textuais e também por um dos softwares de implementação de acervo digital mais utilizados na atualidade.

3 Revisão de literatura

Descreve-se nesta revisão de literatura, pesquisas, conceitos e características acerca do tema abordado.

O conteúdo coletado será apresentado nas próximas seções que serão divididas respectivamente em quatro partes, a saber: indexação, indexação automática, DSpace, repositório institucional e bibliotecas digitais.

Não pretendeu-se realizar uma revisão de literatura exaustiva, mas apresentar os principais pontos (de acordo com a finalidade do trabalho) em cada seção.

3.1 Indexação

Historicamente o homem sempre buscou criar meios para organizar o que era produzido. Com a popularização dos livros surgiram vários métodos, como a criação de cabeçalhos descritivos para cada capítulo. Nos mosteiros eram feitas listas que indicavam a localização de cada exemplar.

Collinson (1971) afirma que a indexação em grande escala, no sentido de gerar índices complexos, surgiu após a Bíblia inglesa, em que os homens não iriam conseguir consultar ou citar determinada passagem, então Alexandre Cruden fez a compilação da primeira concordância da Bíblia no ano de 1737.

Para Silva e Fujita (2004) a indexação surge somente a partir da geração de índices, de acordo com as autoras:

> "(...) a atividade de indexação, como processo, é realizada mais intensamente desde o aumento das publicações periódicas e da literatura técnico-científica, surgindo a necessidade de criação de mecanismos de controle bibliográfico em centros de documentação especializados."

Para os autores Silva e Fujita (2004) e Collison (1972) a indexação teve seu auge quando começaram as publicações de periódicos. Em 1901, H. W Wilson lançou o *Reader1s Guideto Periodical Literature*, em que cada artigo era indexado pelo seu autor e pelo assunto específico e havia várias remissivas que ligavam um assunto a outros correlatos (COLLISON ,1972, p. 11).

Inicialmente a atividade de indexar estava totalmente ligada a descrições dos documentos de uma base de dados. Pois a partir dessa função que surgiu os principais conceitos que estão relacionados com a indexação, por exemplo, quando um usuário faz buscas em uma base de dados ele vai ter como resultado uma finidade de documentos, entretanto, a quantidade de itens relevantes ou pertinentes que serão recuperados está associada a política de indexação da base de dados, ou a definição do nível de especificidade e exaustividade, a qualidade do vocabulário e a escolha dos termos na estratégia de busca. Essas características vão influenciar na revocação e na precisão.

Não tem tanta importância a maneira que a indexação é feita, seja ela feita de forma rudimentar ou utilizando índices complexos, o seu objetivo geral será a representação temática de documentos com o intuito de recuperar a informação. Pode-se notar que a indexação está relacionada com a representação da informação dos documentos, mas que podem ser definida pelos autores abaixo:

Indexação é a representação do conteúdo temático de um documento por meio dos elementos de uma linguagem documentária ou de termos extraídos do próprio documento (CUNHA E CAVALCANTI, 2008).

Para Naves e Kuramoto (2006) indexação é o processo intelectual que envolve atividades cognitivas na compreensão do texto e a composição da representação do documento. Para Wellish (1995 apud LIMA 2006), indexação é como o ato de indicar ou apontar o conteúdo intelectual de uma coleção. Segundo Vieira(1988) a indexação é uma técnica de análise de conteúdo que condensa a informação significativa de um documento, através da atribuição de termos, criando uma linguagem intermediária entre o usuário e o documento.

Pinto (2000), define indexação como uma atividade que desmonta o discurso montado pelo autor do documento, à medida que ela faz recortes neste discurso.

Cintra (1983, apud HOLANDA, 2012, p. 46) a indexação é definida como a tradução de um documento em termos documentários, isto é, em descritores-termos, cabeçalhos de assunto, termos-chave, que tem por função expressar o conteúdo do documento. Enquanto que para Holanda (2012),a indexação é definida como uma "tradução lexical" das unidades lexicais da língua em que está escrito o documento, para unidades lexicais de uma linguagem documentária.

Lima (2003), define indexação sendo o procedimento intelectual que envolve atividades cognitivas a compreensão e a composição da representação da informação.

UNISIST (1981), define a indexação como a ação de descrever e identificar um documento de acordo com seu assunto. Dias e Naves (2007), apresentam a seguinte definição para o termo indexação:

> No contexto do tratamento da informação, o termo indexação possui dois sentidos: um, mais amplo, quando se refere à atividade de criar índices, seja de autor, título, assunto, tanto de publicações (livros, periódicos) quanto de catálogos ou banco de dados, em bibliotecas ou centros de informação. O outro sentido, mais restrito, se refere apenas à indexação ou à catalogação de assuntos das informações contidas em documento.

A partir das definições do termo indexação que foram citadas, pode-se concluir que a atividade de indexar concentra-se em representar o conteúdo através de termos, entretanto nenhum autor exemplifica qual a melhor maneira de se fazer/selecionar os melhores termos.

Com base no referencial teórico é notável que a indexação tenha a função de representar a informação por meio de expressões ou termos, sejam eles selecionados de forma livre ou utilizando uma linguagem documentária, por exemplo, um tesauro. Câmara Júnior (2007), a indexação parte da ideia que a seleção do documento tem como ponto de partida o acesso a informação documentária.

Com base no exposto acima e visando manter a coerência e a qualidade na indexação em 1981 UNISIST publicou um documento "princípios da indexação".

> Este documento tem por objetivo o estabelecimento de princípios válidos e consistentes a serem seguidos quando se determina o assunto de um documento. Com o propósito de indexação e recuperação, os conceitos contidos no documento podem ser representados por termos selecionados da linguagem natural (ex: palavras-chaves) ou por símbolos (ex: número de classificação).

Entre as recomendações propostas pela UNISIST, constavam as etapas que são necessárias para fazer uma indexação, que serão apresentadas a seguir.

3.1.1 Etapas na indexação

As etapas da indexação que foram citadas tanto pela UNISIT e por Lancaster (2004), são orientações que servem tanto para a indexação manual quanto para a indexação automática (que será abordada na próxima seção), porém, na indexação automática quem realiza essas etapas são os softwares e não o indexador.

Robredo (2005), apresenta as etapas do processo indexação:

- análise conceitual do conteúdo do documento;
- expressão dessa análise, por meio de um conjunto de frases ou palavras;
- tradução da descrição dos assuntos relevantes para a linguagem de indexação;
- organização das descrições dos assuntos de acordo com a sintaxe da linguagem de indexação.

Vieira (1984), elenca três fases semelhantes as descritas por Robredo para a realização da indexação feita manualmente.

- compreensão do conteúdo do documento por meio da leitura completa do texto, título, resumo, entre outras partes que compõem o documento;
- identificação de conceitos, de modo a estabelecer o ambiente lógico;
- seleção dos conceitos, observando a exaustividade, especificidade e consistência

Pinto (2000), também afirma que a indexação passa por três fases, que de acordo com a autora, são:

- análise conceptual;
- tradução;
- controle de qualidade.

A indexação manual pode ser dividida em duas etapas: análise conceitual e a tradução. A análise conceitual é que faz o trabalho se tornar moroso, pois é nessa etapa que o indexador vai determinar do que se trata o documento. Também é nessa etapa em que a capacidade cognitiva; os conhecimentos relativos ao tema e a capacidade de compreensão do indexador serão de suma importância para que se tenha uma boa indexação. Entretanto o indexador não tem tempo o suficiente para ler um documento inteiro, por isso Lancaster (2004 apud *METHODS FOR EXAMINING DOCUMENTS*), elencou quais as partes dos textos devem ser examinadas cuidadosamente: título, resumo, sumário, introdução, ilustrações e palavras grafadas.

A tradução é a segunda etapa da indexação. Nessa etapa o indexador vai analisar os conceitos que ele selecionou durante a análise conceitual. Nessa segunda etapa Lancaster

(2004), elenca que há dois métodos há serem escolhidos, que é o método de indexação por atribuição e a indexação por extração (indexação derivada). Na indexação por atribuição são atribuídos termos à partir de outras fontes e quanto a atribuição por extração os termos escolhidos para representar o tema do documento são extraídos do próprio documento.

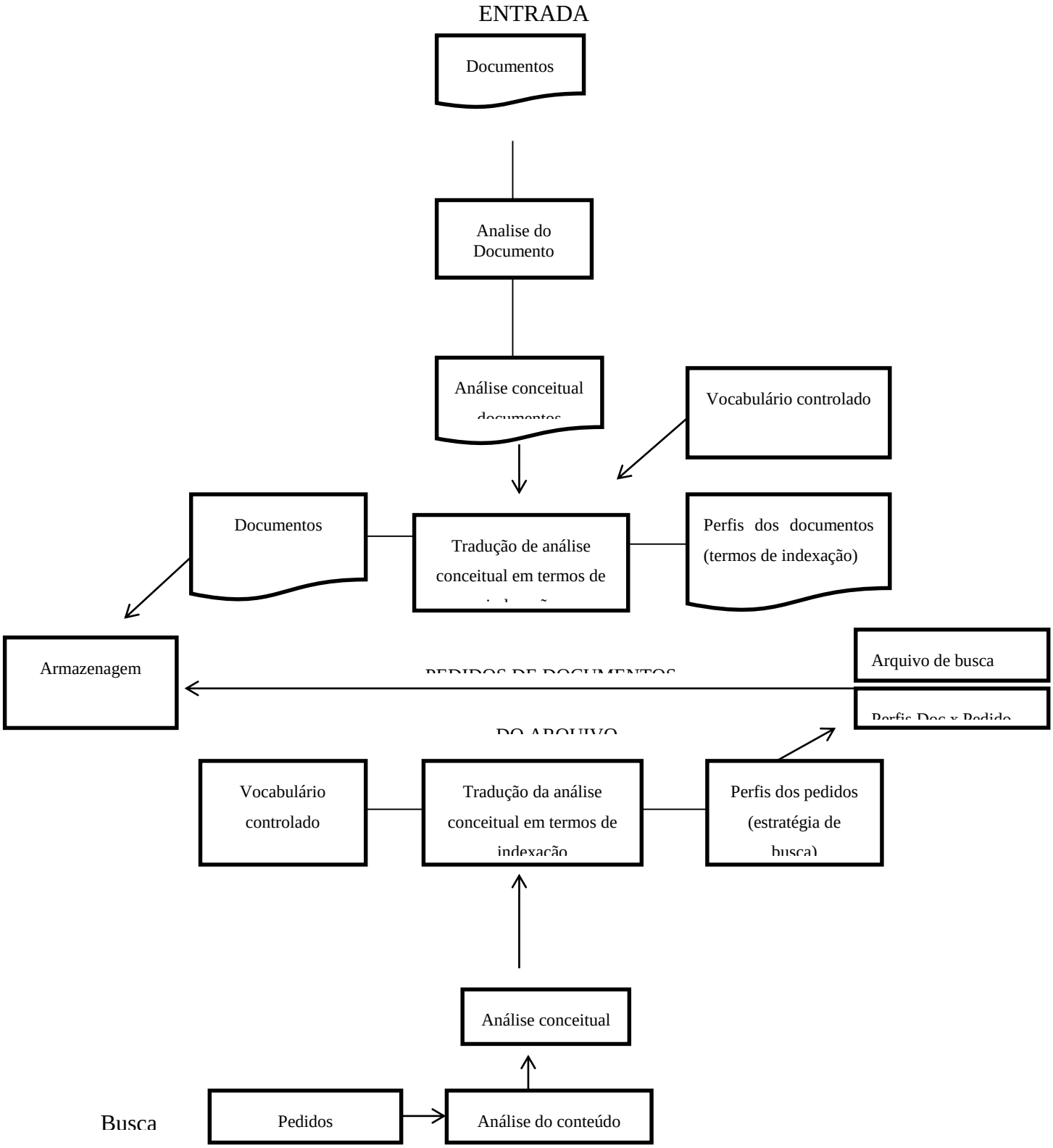
Para Pinto (2000), a etapa de tradução é considerada a mais complexa pois ela exige que o indexador siga algumas regras que são determinadas pelo controle de autoridades e as linguagens documentárias (tesauro). Isso é responsável por causar ruído ou silêncio durante a recuperação da informação. Para Chaumier (1980), ruídos são os documento não pertinentes à questão; os documentos pertinentes existentes no acervo , não recuperados durante a pesquisa, denomina-se silêncio (ausência de resposta). Para tentar amenizar o ruído e/ou silêncio, Fidel (1994, apud PINTO, 2000), cita algumas considerações que devem ser levadas em conta durante o processo de indexação manual, que são:

1. as fontes dos termos de indexação: em quais fontes de vocabulários de indexação os indexadores podem se apoiar para escolher os termos que vão compor os índices ?Existem regras que limitam o indexador aos termos dos tesauros utilizados pelo sistema, e outras permitem que sejam utilizados os termos da língua natural;
2. a precisão: que grau de precisão o indexador pode utilizar para traduzir os conceitos em termos de indexação? Os termos selecionados para o índice devem ser tão precisos que substituam o conceito ou eles devem ter um sentido mais geral ?
3. o peso: o peso relativo dos conceitos de um documento pode ser definido pelo indexador ? No exemplo anterior qual conceito terá o peso maior?
4. a fidelidade: em que medida a tradução deve ser fiel ? Como ser fiel em uma tradução quando o conceito não tem um descritor correspondente? O indexador poderá usar os termos aproximados ?
5. a linguagem do usuário: o indexador pode designar os termos de um índice em uma linguagem mais próxima da do usuário ? Por exemplo, através dos seus perfis é possível estabelecer regras que poderão lhes guiar na escolha dos termos de indexação mais adequados aos seus ?

Pinto (2000), coloca como resposta para as questões de Fidel, que os itens 1,2 e 5, são fáceis de resolver porque estão ligados a questões operacionais, entretanto os itens 3 e 4 não são fáceis pois a definição do peso dos conceitos pertencentes a um documento implica em um processo subjetivo.

Para representar essas três fases da indexação manual, Lancaster apud VIEIRA (1984, p. 9) fez um fluxograma que pode ser visualizado abaixo.

Figura 1: Processo de indexação manual



Fonte: Vieira, 1984

Como foi visto na literatura, os autores tem uma preocupação em ditar ou enumerar passos de como se fazer a indexação, desta forma faz parecer com que o processo de indexação é algo mecânico, entretanto, em contraponto a esta ideia, Collinson (1971), diz que a indexação, com efeito, não é um processo mecânico: para ter utilidade, requer reflexão e ponderação em todas as fases de seu desenvolvimento.

A seguir será mencionado os tipos de indexação e os problemas que estão relacionados a indexação de modo geral.

3.1.2 Tipos de indexação

No que se refere a classificação pelo processo, existem três maneiras de indexar: indexação manual, indexação semiautomática e indexação automática.

Naves e Kuramoto (2006) mencionam que o termo indexação manual não é justo, pois não valoriza o processo de indexação e tampouco o principal ator desse processo, o indexador, de acordo com os mesmos a indexação manual deveria ser chamada de indexação não automática ou de indexação realizada por seres humanos.

Para Pinto (2000), a indexação semiautomática é aquela que combina a indexação manual e automática sendo realizada da seguinte maneira: inicialmente o sistema faz uma indexação automática dos documentos levando em conta as ocorrências das palavras mais frequentes no texto. Em um segundo momento, o indexador humano refina a lista dos termos propostos pelo sistema fazendo os ajustes e/ou complementações necessárias.

A indexação automática visa à mecanização das atividades descritas no tópico anterior, com o objetivo de agilizar e auxiliar o processo intelectual realizado pelos profissionais da área (BORGES, 2009).

No contexto da pertinência. para Lancaster (2004), existem dois tipos de indexação: a indexação seletiva e a indexação exaustiva. A indexação seletiva leva em consideração os conceitos específicos em função dos temas tratados no documento; enquanto na indexação exaustiva, procura extrair do documento o maior número de conceitos de forma a cobrir seu conteúdo de maneira mais completa (PINTO, 2000).

Tanto na indexação seletiva quanto na indexação exaustiva existem alguns problemas, por exemplo, a indexação exaustiva corrobora com a revocação, ou seja, aumenta a probabilidade de recuperar itens inúteis. Enquanto que a indexação seletiva faz com que aumente a precisão, ou seja, aumenta a possibilidade de recuperar documentos uteis.

De modo mais abrangente existem outros problemas que estão vinculados à atividade de indexação que podem afetar diretamente na qualidade desta. Isto será mais explorado na subseção seguinte.

3.1.3 Problemáticas na indexação

O resultado da indexação é o índice, de acordo com Feitosa (2006), tem a seguinte função:

> ...elaboração desses índices, que são instrumento utilizados para a representação do conteúdo de documentos primários, é facilitar a recuperação de informações relativas ao documento indexado ou resumido.

Segundo Collinson (1971) mesmo quando o índice se aproxima de proporções adequadas podem ocorrer numerosos defeitos, e o principais estão relacionados a partes do documentos que não são indexadas, por exemplo: ilustrações, prefácios, prólogos, introduções, notas de rodapé, bibliografias e até mesmo os cabeçalhos. Para o mesmo autor, quando um leitor não consegue encontrar uma informação no índice, mas que está contem no documento, quer dizer que o indexador falhou.

Como já foi dito anteriormente, a indexação manual é um processo moroso e caro. Robredo (2005), explana que a indexação requer tempo e exige conhecimentos adequados do indexador, o que torna uma atividade cara. Borges (2008), afirma que o processo de indexação manual ainda é caro.

Os problemas relacionados a indexação manual vão além da falta de indexadores comparados a quantidade de documentos que são publicados diariamente. O problema também está relacionado a capacitação do indexador, que necessita ter um conhecimento muito grande sobre tudo e também dominar outros idiomas.

A indexação é uma atividade que requer muito do conhecimento prévio do indexador. Para Neves, Dias e Pinheiro (2006, p.142),

> Para compreender um texto, os indivíduos lançam mão de todo o conhecimento prévio armazenado na memória de longo prazo, demandando, inclusive, possíveis esquemas de procedimento existentes na memória semântica. O conhecimento anterior facilita o processamento do texto e a compreensão, por oferecer uma estrutura na qual o conteúdo do material lido possa ser relacionado. A integração do conhecimento passado com o texto que está sendo lido permite aos leitores formar o que é chamado por Van Dijk e Kintsch (1983) e Kintsch (1998) de “modelo situacional”. Este consiste na combinação das informações (ou proposições – unidades abstratas de significado) retiradas do texto com as proposições formadas a partir de

> conhecimentos gerais preestabelecidos e da experiência pessoal dos leitores.

De acordo com os autores pode-se inferir que o conhecimento do indexador afeta diretamente na qualidade da indexação e também na consistência da indexação.

Por ter um tempo limitado para realizar a atividade de indexação, pode ser que o resultado não seja tão satisfatório tanto para a representação da informação quanto para a recuperação da informação. Isso pode gerar revocação na busca, e segundo Collison (1971),

> A vida é muito breve para que se perca tempo em busca de informações que podem ou não existir no livro que está sendo examinado. Todo leitor sério sabe que é possível pesquisar repetidas vezes em busca de uma informação perdida.

Visando a consistência na indexação, acelerar o processo e diminuir o custo, por volta da década de 50 iniciou-se as pesquisas relacionadas a indexação automática.

3.2 Indexação automática

A grande quantidade de documentos publicados diariamente se intensificou não só no Brasil como no mundo, entretanto não havia mão de obra o suficiente para indexar a passos tão largos, então surgiu a necessidade de iniciar estudos para acelerar o processo de indexação.

Com o intuito de resolver essa problemática em 1950 iniciou-se os estudos sobre a indexação automática.

Na literatura são encontrados os termos indexação automatizada e indexação mecanizada. Geralmente esses termos não abarcam a indexação semiautomática. Silva e Fujita (2004), consideram que a indexação semiautomática necessita da validação do documentalista, enquanto Pinto (2000), afirma que o indexador também tem uma participação no processo de decisão dos termos.

O processo de indexação automática baseia-se, segundo Robredo, "na comparação de cada palavra do texto com uma relação de palavras vazias de significado, previamente estabelecidas, que conduz, por eliminação, a considerar as palavras restantes do texto como palavras significativas".

Para Robredo (1982), Gil Leiva(1999) e Lancaster (2004) o pioneiro na indexação automática foi Hans Peter Luhn. Foi no final da década de 50 que Luhn desenvolveu o método KWIC (*Keyword in context)* [palavra-chave no contexto].

O método KWIC é um índice rotado, derivado, em sua forma mais comum, dos títulos de publicações. O índice KWIC é um método barato de obter certo nível de acesso temático

ao conteúdo de uma coleção e é útil na medida em que os títulos são bons indicadores de conteúdo (LANCASTER, 2004).

Além do índice KWIC tem o índice KWOC que é bem semelhante ao KWIC, porém a palavra-chave usada como ponto de entrada não se repete no título, mas é substituída por um (*) ou outro símbolo (LANCASTER, 2004).

De acordo com Santos (2009), Luhn desenvolveu seu método baseando-se nos estudos desenvolvidos por Zipf, que formulou duas leis sobre a distribuição das palavras em um texto.

No ano de 1948, o professor da Universidade de Harvard, George Kingsley Zipf desenvolveu duas leis sobre a frequência das palavras em um texto. Sua primeira lei (frequência de ocorrência das palavras) está relacionada com as palavras de alta frequência, em que, em um texto suficientemente longo forem colocadas em ordem decrescente de frequência, pode-se verificar que a ordem de série de palavras (R), multiplicada por sua frequência (F) produz uma constante (K), ou seja:

R.F=K

A segunda lei de Zipf esta relacionada a baixa frequência das palavras em um texto, ou seja, em um texto, várias palavras de baixa frequência de ocorrência aparecem o mesmo número de vezes. A lei é enunciada da seguinte maneira:

$$\frac{ln}{l1} = \frac{2}{n(n+1)}$$

Onde:

- In é o numero de palavras que ocorreram N vezes para n < 5 ou n < 6;
- I 1 é o número de palavras que ocorreram uma única vez
- 2 é uma constante atribuída a língua inglesa.

As leis apresentadas acima foram constatadas empiricamente, ou seja, por meio de testes, contudo, não se aplicam em sistemas de informação (FERNANDES, 2013).

De acordo com Mamfrim (1991), Goffman sugeriu a criação de um ponto T, em que ele representa a transição das palavras de alta frequência para as palavras de baixa frequência, ou seja, nesse ponto estão as palavras que são representativas do conteúdo. O ponto T é representado matematicamente como:

$$T= \frac{-1+\sqrt{1+8.\ l\ 1}}{2}$$, onde:

- I1 é o número de palavras que ocorreram uma única vez;
- 8 é uma constante derivada da língua inglesa;
- 2 é uma constante matemática da fórmula de Baskara, para resolução de equações de 2º grau.

A criação do ponto T possibilitou aplicar as leis bibliométricas que trabalham com frequência de palavras como instrumento de indexação em sistemas de informação. (MAMFRIM, 1991)

O método de Luhn foi pioneiro na realização de índices e deu o "ponta pé" inicial para surgirem novos modelos de indexação automática.

Na década de 60 surge um método bem diferente do método estatístico utilizado por Luhn no método KWIC, que é o chamado processamento de linguagem natural (PLN).

O processamento da linguagem natural pode ser definido como qualquer utilização do computador para a manipulação da linguagem natural (FERNANDES, 2013). O PLN é abordado do ponto de vista da análise do conhecimento morfológico, sintático, semântico e pragmático.

Para melhor compreensão sobre os processos e métodos de indexação automática, serão abordados nas próximas seções os conceitos de indexação automática sob a óptica de vários autores, posteriormente encontra-se os modelos de indexação automática e suas características.

3.2.1 Conceituações

A partir da leitura do tópico anterior pode-se notar o quão a indexação automática é uma atividade que perpassa por várias áreas do conhecimento, por exemplo, a informática e a linguística, isso causa uma certa discordância na definição do termo e do que é a indexação automática.

Afim de ter mais consistência e semelhança entre os conceitos optou-se por utilizar apenas documentos que tratam da indexação automática do ponto de vista da biblioteconomia e ciência da informação.

Robredo (1982), relata que o processo de indexação automática se desenvolve seguindo um esquema bastante semelhante ao processo de leitura-memorização, em que o processo de leitura, em que não interessam as letras, mas a ideia que elas representam, quando organizadas em palavras ou conjuntos de palavras.

O processo de memorização pode ser dividido em duas etapas: uma memorização temporária e inconsciente, na qual só serão memorizadas as palavras significativas. E a segunda etapa é uma memorização mais permanente dos conceitos em que atribui-se o nome de memória. A indexação automática é um processo que pode utilizar diferentes métodos desenvolvidos para programas de computador (VIEIRA, 1988).

Mamfrim (1991), afirma que a indexação automática consiste na mecanização desse processo no todo ou em parte, visando a estabelecer rotinas que reduzam a interferência da subjetividade do indexador, tanto na análise do documento, quanto na seleção dos termos significativos.

Anderson e Perez-Carballo (2001, apud SANTOS, 2009) definem indexação automática como a "análise do texto por meio de algoritmos de computador".

Silva e Fujita (2004), utilizam a seguinte definição: " a indexação automatizada seria, portanto, aquela resultante do trabalho intelectual de um profissional para checagem do valor dos termos atribuídos a um documento por um programa de computador". Os mesmos autores definem indexação automática como que realiza a indexação por meio de programas de computador sem nenhum tipo de validação posterior por profissionais.

Cabe ressaltar que diante da diferença entre os termos indexação automática e indexação automatizada, neste trabalho consideram-se as definições pertinentes ao termo indexação automática, já que pode-se considerar a indexação automatizada como semelhante ao método de indexação semiautomática.

Neves (2009, apud FERREIRA, 2013), reafirma que a indexação automática seria a execução de um processo de representação de documentos, porém, realizada por meio de programas ou algoritmos de computador que "varrem" o documento e realizam a representação do conteúdo sem a intervenção direta do indexador.

Também chamada de indexação assistida por computador e de indexação semiautomática, esse tipo de indexação é considerada um modelo de extração com características estatísticas e probabilísticas (BORGES, 2008).

A partir do referencial teórico conclui-se que a indexação automática é um processo realizado por um software que apresentará como resultado um índice composto por palavras que representam a informação contida naquele documento e esse índice pode ser gerado de maneira derivativa ou atribuitiva.

Na próxima seção será abordado alguns modelos de indexação automática.

3.2.2 Métodos de indexação automática

A indexação automática utiliza diferentes métodos computacionais para identificar mecanicamente quais são as palavras significativas de um documento eletrônico. A forma com que essas palavras serão extraídas vai depender do método de indexação automática que é utilizado.

Os modelos de indexação divide-se em dois tipos: indexação automática por extração e por atribuição.

3.2.2.1 Indexação por extração automática

Um dos métodos mais simples de indexação é o método de indexação por extração que pode ser utilizado tanto na indexação manual quanto na indexação automática.

Na indexação automática o método de extração surgiu em 1950, ela era baseada na frequência em que as palavras apareciam no texto e a partir disso era gerada uma lista de termos, ou seja, os termos que mais apareciam no texto eram considerados os termos mais significativos e que poderiam representar a informação contida no documento. Porém nem sempre as palavras com maior frequência são bons termos.

O pioneiro e já citado anteriormente foi o método KWIC, que se baseava na frequência das palavras no texto. Abaixo encontra-se o algoritmo do método desenvolvido por Luhn.

Figura 2: Algoritmo simplificado para gerar o índice KWIC

Fonte: Robredo, 2005, p. 170

Borges (2008), descreve brevemente quais são as tarefas que envolvem a indexação automática por extração, que são as seguintes:

- Contar palavras num texto;
- Cotejá-las com uma lista de palavras proibidas;
- Eliminar palavras não significativas (artigos, preposições, conjunções, etc.);
- Ordenar as palavras de acordo com sua frequência.

A maioria dos métodos de indexação automática estão relacionados ao método de extração de termos, pode-se citar o método da frequência ou análise estatística, ou seja,

quanto maior for a frequência de uma palavra no texto – excluindo-se as stop words[1] – maior é a capacidade dessa palavra ser descritor. Em 1958 Luhn demonstrou em seus trabalhos que a frequência com que uma palavra se repete em um texto está diretamente ligada a capacidade dessa palavra representar o conteúdo do documento, esse método de extração de palavras ou termos através da frequência com essa palavra é encontrada é conhecido como método da frequência. SPARCK JONES, (apud VIEIRA, 1986), diz que a frequência pode ser estabelecida através da:

a. ocorrência total da palavra no documento: a palavra é contada todas as vezes que aparece, fazendo-se o somatório das vezes em que co-ocorre, posteriormente;
b. ocorrência única da palavra no documento: conta-se somente uma vez a palavra, independentemente do número de vezes que ela aparece;
c. ocorrência da palavra na coleção: a contagem é realizada somando-se seu aparecimento na coleção.

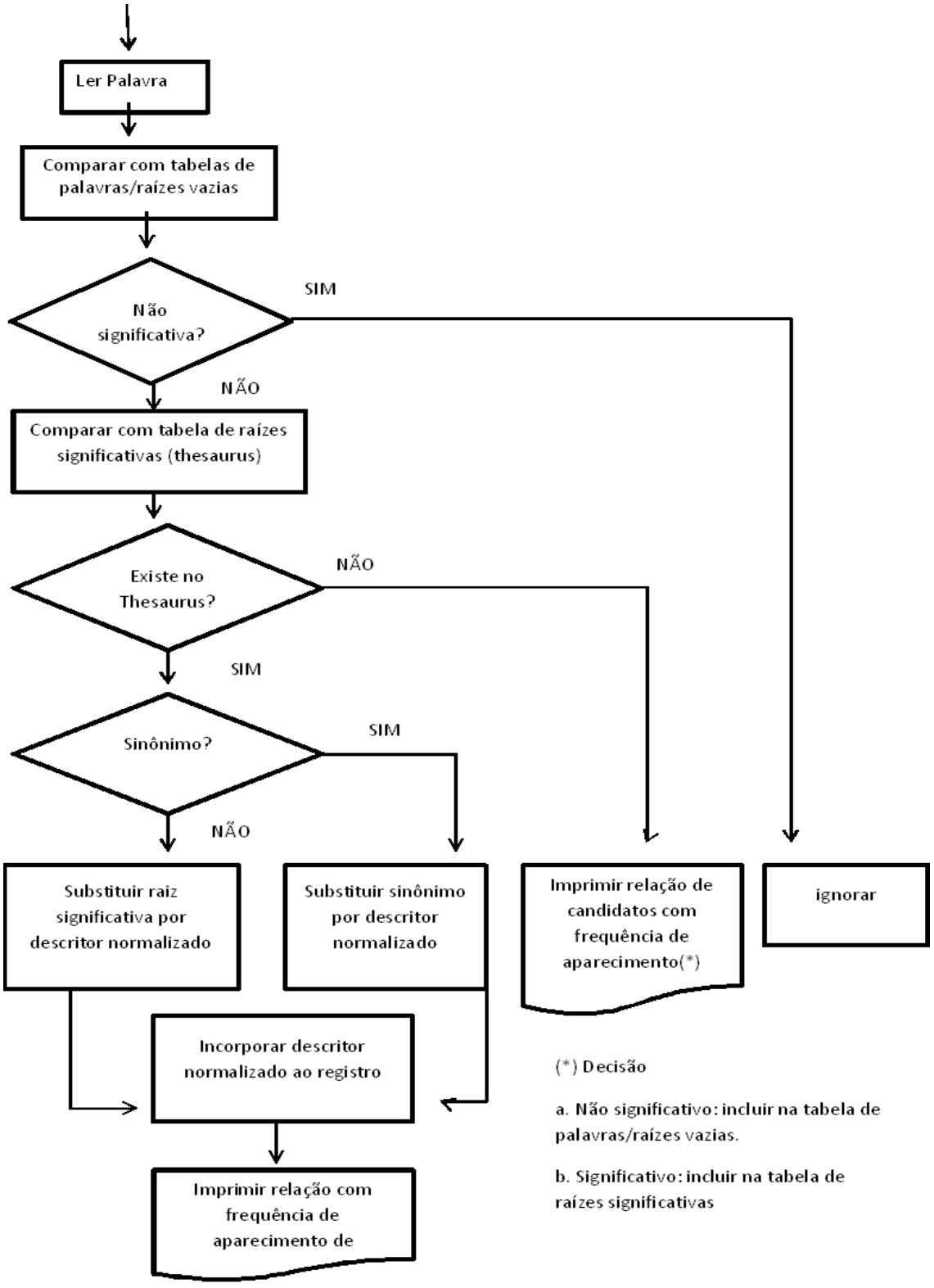
O método da frequência também pode ser utilizado na indexação manual, pois quando o indexador percebe a repetição de determinada palavra ele pode levar em consideração no momento da análise conceitual.

Tem-se também o método da frequência inversa em que examina-se a relação inversa, ou seja, quanto maior a frequência de um termo menor será a sua capacidade representativa.

Na década de 80, Robredo, ilustrou um algoritmo de indexação automática que baseia-se na comparação de cada palavra do texto com uma relação de palavras vazias de significado, previamente estabelecidas que conduz, por eliminação, a considerar as palavras restantes do texto como palavras significativas, (Vieira 1984, apud ROBREDO 1982).

[1] Stopwords sinônimo de palavras vazias

Figura 3: Algoritmo de indexação automática



Fonte: Vieira, 1984

Acima foram apresentados dois métodos possíveis para gerar índices automáticos. Ambos utilizam métodos estatísticos probabilísticos para tal.

Entretanto com o avanço dos estudos sobre linguística e de linguagem documentária no campo da indexação automática surgiram métodos que são capazes de combinar os métodos estatísticos, com a linguística e a linguagem documentária (vocabulário controlado) sem a necessidade de interferência humana no processo.

3.2.2.2 Indexação por atribuição automática

O método de indexação por atribuição automática é bastante complexo, pois para cada expressão significativa do documento é necessário ter um perfil de palavras sinônimos para aquela expressão. Entretanto a escolha da expressão vai depender da frequência com uma expressão semelhante aparece no documento. Lancaster (2004), utiliza o termo chuva ácida para exemplificar o método. O perfil de palavras para chuva ácida pode ser: chuva ácida, precipitação ácida, poluição atmosférica, dióxido de enxofre, etc. Porém se o termo chuva ácida aparecer com uma frequência alta, então o termo de indexação será chuva ácida.

Um exemplo de indexação por atribuição é o método de atribuição de peso, que de acordo com Salton, (1973, apud VIEIRA 1984) é uma forma de atribuir-lhes valores semânticos para torná-los mais precisos, sem no entanto diminuir sua capacidade de revocacão. É baseado na frequência de cada descritor.

O peso pode ser atribuído, de acordo com Sparck Jones (1976 apud, VIEIRA 1984):

a) pela frequência total ou frequência única — a palavra recebe o mesmo valor do número de sua frequência;

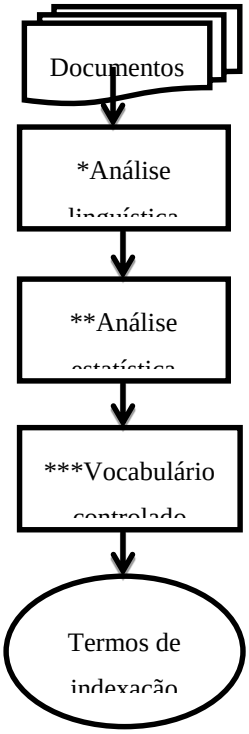
b) pela fonte — se a palavra se encontrar em um documento reconhecido como relevante, receberá um peso maior do que outra existente em um documento menos relevante;

c) pela fonte e usuário — o usuário é quem julgará se o documento recuperado é relevante ou não. Se for, os termos utilizados na estratégia de busca terão, posteriormente, seu valor aumentado;

d) pela frequência na coleção.

Um pouco diferente do método de atribuição de peso, Gil Leiva (2008), apresenta um modelo de arquitetura de um sistema de indexação automática misto

Figura 4: Modelo de arquitetura de um sistema de indexação automática



(*) Decomposição do texto em palavras e frases;
Obtenção da categoria gramatical das palavras;
Desambiguador morfológico;
Desambiguador sintático.

(**) Cálculos de frequência de ocorrência de palavras no texto;
Cálculos de frequência de ocorrência de palavras na coleção;
Ranking das melhores palavras que representam o texto

Fonte: Narukawa, 2011, p. 57

Cada modelo exemplificado acima tem seus pontos positivos e também suas limitações, o que é importante notar nos diferentes modelos é a evolução de cada método.

Essas são apenas umas das tantas formas que se pode fazer a indexação automática com sistemas de computador.

A partir dos métodos de indexação que foram citados brevemente é possível notar que **a indexação é uma área do conhecimento** interdisciplinar, porque ela une técnicas de diferentes áreas, por exemplo: linguística e a informática. Gil Leiva (apud, SANTOS 2009), lista as áreas que contribuem para a evolução da indexação automática:

- Linguística: contribui com os aspectos semânticos e sintáticos;
- Terminologia: utilização de linguagens documentárias na indexação automática;
- Informática: área responsável por realizar sistematicamente as etapas da indexação automática;
- Linguística computacional: tratamento computacional da linguagem e línguas naturais;
- Estatística: processos matemáticos para cálculo da maior ou menor frequência dos termos do documento;

- Inteligência artificial: desenvolvimento de sistemas que são capazes de realizar uma tarefa de forma similar a maneira que um humano a executa.

3.2.3 Características

As características da indexação automática estão relacionadas com os modelos que foram apresentados acima, pode-se retirar alguns itens que foram citados acima para exemplificar.

A indexação automática é caracterizada por uma série de fatores, por exemplo, é realizada mecanicamente; o método de extração de termos não é tendencioso quanto a indexação manual; a maioria dos métodos utilizam a linguagem natural e proporciona maior coerência na indexação.

Por utilizar mais a linguagem natural, enquanto que na indexação manual é comum utilizar vocabulário controlado para a representação temática dos documentos, verificou-se que a indexação por linguagem natural por vezes pode ser melhor, porque o indexador mantém a forma com que o autor quis expressar determinado assunto.

Essa questão remete ao problema em se construir algoritmos que levam em consideração as questões de semântica e sintaxe do conteúdo desses documentos (BORGES, 2009).

Foskett (1973, apud SILVA; FUJITA, 2004), afirma que a indexação automática diferentemente do processo de indexação manual não apresenta nenhum esforço intelectual e um dos problemas causados por essa diferença é que os índices automáticos não representam de forma satisfatória o assunto dos documentos da mesma forma que a indexação humana faz (SILVA;FUJITA, 2004).

Em oposição a afirmativa de Silva e Fujita, Vieira (1986) em sua dissertação de mestrado concluiu através de testes comparativos feitos entre a indexação manual e a indexação automática, que não havia grandes diferenças entre uma e outra, porém a indexação automática era mais interessante pelo baixo custo e pelo aumento da produtividade.

Santos (2009), produziu um quadro que sintetiza os tipos de sistemas de indexação automática e suas características. O quadro pode ser conferido abaixo:

Quadro 1: Critérios para Classificação dos métodos de Indexação Automática

Modelo de Indexação Automática	Descrição
Sistemas não linguísticos	Inclui as linhas que seguem modelos estatísticos,, bibliométricos e

	infométricos.
Sistemas linguísticos (PLN)	São as linhas que já consideram um processamento de linguagem natural nos níveis morfológico, sintático e semântico. Por exemplo, com a utilização de vocabulários controlados ou o uso dos sintagmas nominais para representação; e sistemas baseados em regras (MachineAided-Indexing)
Sistemas Inteligentes	Sistemas de indexação automática que se baseiam em algoritmos de Aprendizado de Máquina, permitindo a inferência automática das regras para a classificação dos documentos, podendo incluir o uso de um conjunto de documentos pré-classificados manualmente.
PLN + Sistemas Inteligentes	Trata-se da última geração de sistemas de indexação que une todos os modelos existentes, com a utilização de técnicas e instrumentos de Processamento de Linguagem Natural (incluindo os instrumentos de processamento morfológico, sintático, semântico, pragmático para a composição de uma base de conhecimentos).

Fonte: Santos, 2009, p. 56

Após descrever sobre as características da indexação automática e sobre como os vários sistemas se comportam para gerar um índice automaticamente, na próxima seção será mencionado a evolução da indexação automática, de como ela começou e quais são os principais desafios atuais.

3.2.4 Evolução da indexação automática

Como foi apresentado em subseções anteriores, já está claro que o inicio da indexação automática ocorre com a criação das leis de Zipf e que foi anterior a criação do método KWIC de Luhn. Esses dois estudos foram apresentados na década de 50.

Na década de 70 surgiram outros sistemas de geração automática de índices, como por exemplo o SMART e o MEDIars. De acordo com Ferreira (2013), o sistema SMART funciona sem análise manual do conteúdo. Trechos do documento são introduzidos no computador e uma variedade de procedimentos automáticos de análise de texto é utilizada para produzir para cada item um 'conceito vetor'.

O MEDIars, faz uso de um vocabulário controlado, ou seja ele compara as palavras que foram utilizadas para fazer a busca com uma lista de palavras chaves determinada para os documentos (FERREIRA, 2013).

Na década de 80 e 90 os principais programas de indexação automática já se preocupavam com o processamento de linguagem natural, apesar que, segundo Gil Leiva (1999, apud NARUKAWA;LEIVA; FUJITA, 2009), os estudos sobre PLN iniciaram-se ainda na década de 60, mas de acordo com a pesquisa realizada por Ferreira (2013), só em 1983 com o sistema SPIRIT, seguidamente surgiram o Hirst em 1987, Automindex e Analisador morfossintático em 1991; SRIAC em 1997; IILICO em 1998; KanaCustomerMessaging System, Brightware e SISA em 1999; NPLwin e Zstation em 2000; Semantic Agent e ThoughtTreasure em 2003; Atenea e SiRILiCO em 2005; DocMir em 2007, Coh – Metri x e LIWC em 2009.

Atualmente os principais estudos sobre sistemas de indexação automática se relacionam com programas que possam fazer uma leitura sintática e semântica dos documentos, ou seja, pretende-se que os sistemas de indexação automática sejam capazes de realizar um trabalho bem semelhante ao que é realizado por indexadores humanos. Mas a semântica e a sintaxe desempenham uma função bastante importante na indexação automática que é, identificar a estrutura lexical das frases e o significado dos termos que representam o conteúdo do documento (BORGES; MACULAN; LIMA, 2008).

Outro tipo de indexação que surgiu em meio a essa explosão informacional na Web, foi a folksonomia. Para Wal (2006, apud CATARINO; BAPTISTA, 2007), Folksonomia é o resultado da atribuição livre e pessoal de etiquetas (*tagging*) a informações ou objetos (qualquer coisa com *URL*), visando à sua recuperação. A atribuição de etiquetas é feita num ambiente social (compartilhado e aberto a outros).

A folksonomia tem sido bastante utilizada em redes sociais como, FaceBook, Instagram e o Youtube também utiliza uma forma de etiquetagem para categorizar os vídeos.

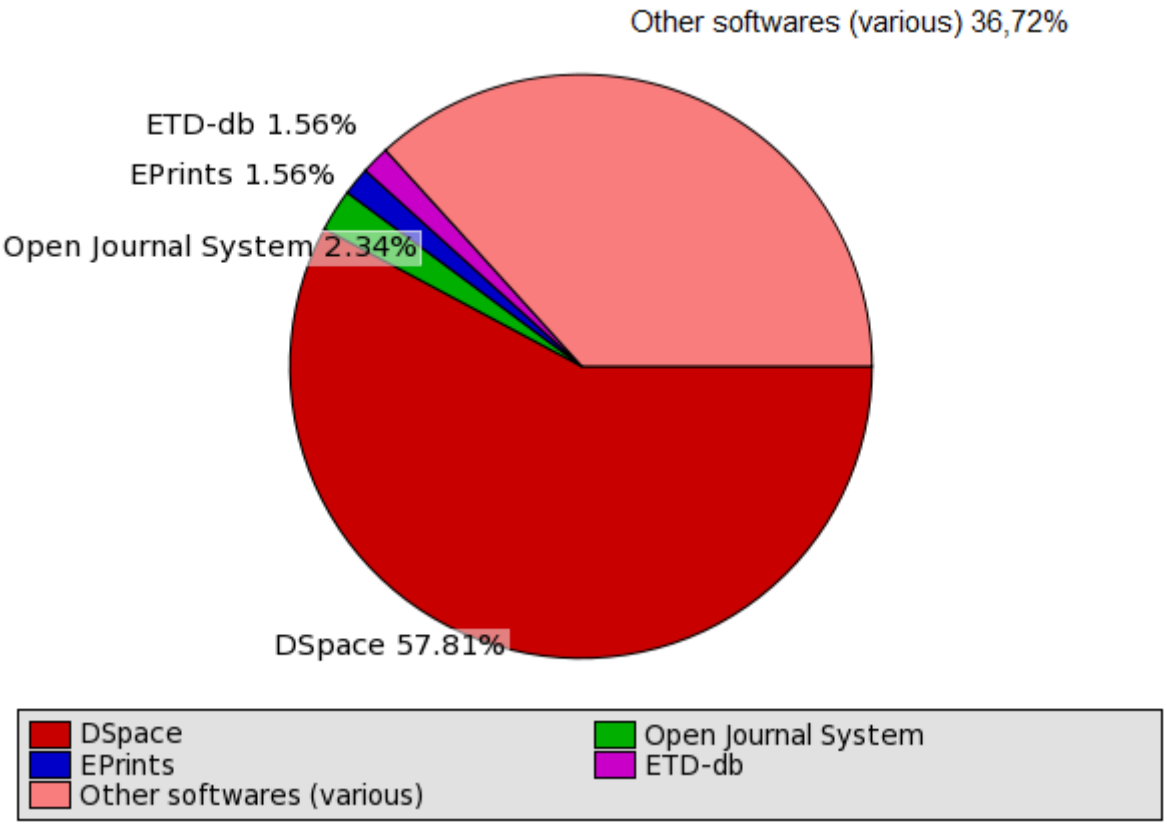
3.3 DSpace

O DSpace é um software livre resultante do projeto colaborativo entre o MIT Libraries e a Hewlett-Packard Company. É essencialmente utilizado para a implementação de repositórios/bibliotecas digitais, pois ambos têm funções bem semelhantes, como por exemplo: armazenar e gerenciar informações em meio digital; preservar e disponibilizar a produção intelectual.

O DSpace além de documentos textuais também pode gerenciar documentos imagéticos; arquivos de áudio e vídeo; publicações multimídia e páginas da web.

Entre os tantos softwares existentes para implementação de repositórios institucionais e bibliotecas digitais, o DSpace é o mais utilizado. De acordo com o registry of open access repositories (ROAR) mais da metade dos repositórios institucionais e/ou bibliotecas digitais no mundo utilizam o DSpace.

Figura 5: quais softwares são mais utilizados no mundo



Fonte: ROAR, 2015.

Através de uma rápida pesquisa no site oficial do software é possível encontrar a maioria das instituições no mundo que utilizam a ferramenta e pode-se fazer uma estimativa de que mais de 1000 instituições hoje utilizam o DSpace.

Entre essas 1000 instituições destacam-se 13 categorias completamente diferentes que utilizam o software, entretanto a maioria das instituições que o adotaram são de cunho acadêmico, governamental e centros de pesquisa.

Já foram lançadas 10 versões: 1.0 (2002), 1.1 (2003), 1.2 (2005), 1.3 (2005), 1.4 (2007), 1.5 (2009), 1.6 (2010), 1.7 (2011), 1.8 (2012), 3.x (2013), 4x (2014) e 5.1 (2015).

Desde a primeira versão o DSpace recebe atualizações e melhorias na sua funcionalidade que caminha de acordo com as necessidades dos usuários gestores.

Shintaku e Meirelles (2009), citam a evolução da indexação automática de texto completo da versão 1.3, porém as melhorias no software não param por aí. Pode-se citar a possibilidade do uso de vocabulário controlado; nova interface OAI-PMH; estatísticas baseadas em Solr; importação de registros com base em referências bibliográficas; formulários baseados nos tipos de documentos; tagcloud; embargo; adaptação a plataforma de acesso, tanto para JSPUI a partir da versão 3.0 quanto para XMLUI a partir da versão 5.1; open search; importação de lista de controle de autoridade; tecnologia shibboleth; implementação de RDF; definição de metadados para usuários; internacionalização, permite mudar o idioma da página inicial do repositório institucional/ biblioteca digital; etc

3.3.1 Metadados

O padrão de metadados utilizado no DSpace é o Dublin Core, mas o DSpace permite que se escolha outro padrão desde que sejam definidos todos os campos.(SHINTAKU; MEIRELLES, 2009, p. 23).

O Dublin Core baseia-se no princípio de que a descrição do documento deve ser elaborada pelo seu produtor ou criador (CAMPELLO, 2006, p.62).

A versão original do Dublin Core possui 15 elementos que são: título, criador, assunto, descrição produtor, colaborador data, tipo, formato, identificador, fonte, idioma, relação, cobertura e direitos. O catalogador não necessita utilizar todos os 15 campos, isso vai depender da necessidade de descrição do material a ser catalogado.

Comumente as instituições optam por utilizar o padrão de metadados Dublin Core qualificado ou Qualified Dublin Core (QDC). No quadro abaixo pode-se encontrar os campos mais comuns.

3.3.2 Requisitos do sistema

Todo software possui pré-requisitos específicos para que funcionem como o esperado. Para que se execute corretamente o DSpace é necessário notar quais são os sistemas e ferramentas que são requisitos para sua instalação.

Abaixo está listado quais são as ferramentas necessárias para que o funcionamento do software seja pleno. As recomendações vão desde os sistemas operacionais, incluindo ferramenta java, banco de dados e web server. Também é importante salientar que pode-se utilizar outras ferramentas, entretanto não há instruções no manual de como utiliza-las.

Quadro 2: Requisitos do sistema

	• OS UNIX-like ou Microsoft Windows • Java JDK • Apache Maven • Apache Ant • Banco de dados relacional: (PostgreSQL ou Oracle). • Mecanismo de Servlet: (Jakarta Tomcat 4.x, Jetty, Caucho Resina ou equivalente).

Pra qualquer versão do DSpace será necessário a instalação dos softwares acima de acordo com as versões necessárias.

Mesmo utilizando as mesmas ferramentas é possível notar algumas diferenças entre versões mais antigas e versões mais recentes. O impacto disso no software é que cada versão traz alguma mudança na maneira de executar a ferramenta. A ferramenta Java que nas primeiras versões era utilizada a versão Oracle Java JDK 5 e na versão 1.7.x começou a utilizar a versão JDK 6, ou seja, a versão superior utiliza novos recursos de linguagem que tornam o software mais “limpo”.

O manual indica duas ferramentas de banco de dados, entretanto cada uma opera de maneira distinta, cabe ao gestor decidir qual seria a melhor ferramenta, pois de acordo com o manual do DSpace não tem nenhuma ferramenta capaz de exportar Postgres para Oracle de maneira automática.

Cada elemento apresentado no quadro está relacionado com algum tipo de mudança no software, desde a instalação até a submissão e importação/exportação de itens e também relaciona-se com as funções que o software desempenha. Na próxima seção será listadas algumas das principais funções do DSpace.

3.3.3 Funções

O DSpace é uma das ferramentas mais utilizadas para a construção de repositórios institucionais e bibliotecas digitais. Os repositórios institucionais têm como principal característica representar a produção intelectual de uma instituição, aumentando a visibilidade tanto da instituição quanto dos autores (COSTA; LEITE, 2006). Para que isso aconteça o software deve ter funções que possibilitam o fluxo de informação científica. Sayão (2009), listou as seguintes funções do DSpace:

- Facilitar a captura e depósito de materiais, incluindo os metadados sobre esses materiais;
- Facilitar o acesso fácil aos materiais, tanto pela listagem quanto pela busca;
- Facilitar a preservação em longo prazo dos materiais;
- Armazenamento e recuperação de objetos digitais;
- Identificação via metadados;
- Ferramentas de busca simples e avançada;
- Fluxo de submissão que pode ser adequado às necessidades de cada instituição;
- Preservação digital.

Sayão cita a R.I como uma das funções do DSpace, função essa que está diretamente ligada ao processo de indexação. Por esse motivo, na próxima seção será abordada como a indexação pode ser feita no software.

3.3.4 Indexação no DSpace

A indexação realizada no DSpace pode ser feita de duas formas, semiautomática ou automática. A indexação semiautomática ocorre quando indica-se os metadados que devem ser utilizados para gerar o índice. Essa indexação é semiautomática, porque o preenchimento dos campos que serão utilizados para fazer a indexação do documento foi realizado por humanos.

A indexação automática no software é a indexação de texto completo, que segundo Shintaku e Meirelles (2009) consiste na criação de índices textuais em que todas as palavras se tornam pontos de recuperação para o documento.

A indexação automática de texto completo do DSpace é semelhante a indexação que é feita em banco de dados

A indexação de texto completo só é possível para documentos textuais que se encontram nos formatos PDF, html e Word e outros que podem ser extraídos texto. Quando se opta por utilizar a indexação de texto completo o software consegue identificar e ignorar as palavras que não são relevantes, ou seja, as stopwords.

No software já tem uma lista de stopwords, porém o gestor pode incluir e/ou retirar palavras dessa listagem que ele acredita que pode interferir na recuperação de documentos daquele repositório.

É possível que o administrador configure a quantidade de palavras que serão extraídas do documento para gerar o índice textual.

No manual do DSpace, recomenda-se que o índice seja periodicamente atualizado. Essa periodicidade vai depender da quantidade de submissões que são feitas em um período de tempo.

A maioria dos repositórios contém documentos textuais, ativar o parâmetro de indexação automática de texto completo tornaria a recuperação de documentos mais eficiente.

3.3.4.1 Indexação automática no DSpace

A indexação automática de texto completo no DSpace é feita utilizando o Apache Lucene ou o Solr, a diferença é que nas versões anteriores a versão 3.0 utilizava-se o Lucene e nas versões posteriores a 3.0 utiliza-se o Solr.

Ambos são um software de código aberto desenvolvido pela Apache Software Foundation com o objetivo de fazer buscas e indexar dados que podem ser convertidos para texto.

A indexação com o Lucene/Solr é feita com a utilização de um analisador. Esse analisador é que define quais são as regras de extração dos termos do documento.

Durante a implementação de um analisador para gerar o índice ocorre algumas etapas que são definidas por Sonawane (2009), extração das palavras, remoção das palavras comuns, ignorar pontuação, redução de palavras para o formato de raiz, alteração das palavras para minúsculas, etc. Somente depois dessas etapas executadas é que o índice será gerado.

As etapas acima citadas podem ser diferentes dependendo do tipo de analisador que será utilizado. No quadro abaixo encontra-se os quatro tipo de analisadores disponível no Lucene/Solr.

Quadro 4: Analisadores integrados do Lucene

Analisador	**Operações realizadas nos dados do texto**
WhitespaceAnalyzer	Divide os tokens[2] em espaço em branco
SimpleAnalyzer	Divide o texto em caracteres que não sejam letras (números, caracteres japonês, acrônimos etc.) e coloca o texto em minúsculo
StopAnalyzer	Remove as palavras irrelevantes (desnecessárias para procura) e coloca o texto em minúscula
StandardAnalyzer	É capaz de "tonkenizar" endereços de e-mail; acrônimos, caracteres chinês, japonês e coreanos; números.

Fonte: usando o Apache Lucene para a procura de texto: https://www.ibm.com/developerworks/br/java/library/os-apache-lucenesearch/

Além desses analyser que foram citados no quadro acima, tem-se o BrazilianAnalyzer, a diferença do BrazilianAnalyzer para o StopAnalyzer é que no primeiro contém as stopwords da língua portuguesa.

O processo de análise é feito antes da indexação. Após o processo de indexação o Lucene gera um índice, semelhante ao que encontramos ao final de um livro.

Atualmente o Solr no DSpace só conta a opção do StandardAnalyzer e do BrazilianAnalyzer.

Uma das facilidades do Lucene/Solr no DSpace é recuperar as palavras no singular e no plural; masculino e feminino; palavras com grafias diferenciadas, mas com som semelhantes. Por exemplo: "Sousa" e Souza" (SHINTAKU E MEIRELLES, 2009).

A partir do exposto pode-se notar que o termo indexação também está relacionado com a ciência da computação. Nesta seção foi abordada a indexação da perspectiva da ciência da computação que tem a pretensão de apenas recuperar uma informação, não sendo importante que os documentos recuperados vão ser de fato úteis para o usuário.

Para a elaboração deste trabalho preocupou-se em utilizar repositórios e bibliotecas digitais que trabalhassem majoritariamente com documentos textuais, que é o tipo de documento que o Lucene/Solr consegue indexar.

[2] A análise converte os dados do texto em tokens e esses tokens são incluídos como termos no índice Lucene

3.4 Repositório institucional

Os repositórios institucionais começam a surgir em meados dos anos 90, como uma alternativa que as instituições acadêmicas tiveram para manter o fluxo da comunicação científica, já que as assinaturas dos periódicos estavam cada vez mais caras.

Em consequência do difícil acesso aos periódicos também surgiu o movimento de acesso aberto ou acesso livre. Segundo Kuramoto (2006), não há um consenso sobre a tradução correta do termo open acess.

O movimento do acesso aberto teve ínicio na Declaração de Budapeste no ano de 2001. O movimento defende a ideia de que os artigos científicos devem ter acesso aberto, principalmente por meio digital.

Em 11 de abril de 2003 ocorreu uma reunião que uniu vários representantes de diversas áreas do conhecimento a fim de discutir sobre a implementação do acesso aberto.

Em consequência disso no mesmo ano foi publicado quais as características que uma publicação deve possuir para ser considerada de acesso aberto, que são: responsáveis pelos direitos de copyright devem conceder acesso aberto e perpetuo ao trabalho; disponibilizar uma versão integral do trabalho incluindo-se todo o material suplementar e uma cópia da permissão em formato digital.

Segundo Cunha e Cavalcante (2008), repositório institucional está diretamente ligado com a memória intelectual de uma comunidade ou organização. Enquanto que para os mesmos autores, a biblioteca digital armazena documentos e informações em forma digital em sistema automatizado, geralmente em rede, que pode ser consultado a partir de terminais remotos.

A partir das definições acima pode-se inferir que quando se trata de repositório institucional os itens que constituem aquela base de dados tem que necessariamente está ligado com a instituição que o mantem, enquanto que nas bibliotecas digitais não possui restrição para o que deve compor a coleção.

Em um glossário do IBICT traz a seguinte definição de repositório institucional:

> São sistemas de informação que armazenam, preservam, divulgam e dão acesso à produção intelectual de comunidades universitárias. Ao fazê-lo, intervêm em duas questões estratégicas: - contribuem para o aumento da visibilidade e o "valor" público das instituições, servindo como indicador tangível da sua qualidade; - permitem a reforma do sistema de comunicação científica, expandindo o acesso aos resultados da investigação e reassumindo o controle acadêmico sobre a publicação científica (INSTITUTO, 2007, s/p).

Para Lynch (2003, apud MARTINS, 2009), repositórios institucionais são “um conjunto de serviços que uma universidade oferece aos membros da sua comunidade, para a gestão e disseminação de materiais digitais, criados pela instituição e pelos seus membros”.

De acordo com os autores citados acima, pode-se concluir que os repositórios institucionais estão vinculados a memória de instituições acadêmicas, contribuindo assim com a comunicação científica e o movimento de “livre acesso” a informação científica.

Os repositórios institucionais também podem ser temáticos, ou seja, os repositórios temáticos possuem documentos relacionados a uma área específica do conhecimento, enquanto que os repositórios institucionais não temáticos possuem documentos em várias áreas do conhecimento.

Os repositórios temáticos / institucionais seguem o movimento do livre acesso a informação, pois os autores começaram a perceber que a visibilidade seria maior com a política de acesso livre. Essa percepção fica visível quando Stevan Hanard (2001 apud, MARCONDES; SAYÃO, 2009), diz que:

> Ao contrário dos autores de livros e artigos de revista, que escrevem para explorarem direitos ou por honorários, os autores de artigos de periódicos revisados por pares escrevem apenas pelo “impacto da pesquisa”. Para ser citados e tomar parte na construção da pesquisa de outros pesquisadores, seus resultados têm de ser acessíveis aos seus usuários potenciais. Do ponto de vista dos autores, o acesso pago aos seus resultados é tão contraproducente como o acesso pago a anúncios comerciais [...]

Com o surgimento do movimento do acesso livre a informação científica, surge também novas tecnologias a fim de fomentar esses ideais de disseminação da informação. Na mesma filosofia de acesso aberto à informação surgem os softwares livres que vão de encontro com o Open Archives Initiative (OAI).

Em 1999, foi realizada uma convenção onde foi criada a OAI, que definiu algumas especificações para que se tivesse um nível mínimo de interoperabilidade entre os repositórios.

Kuramoto (2006) descreve que as características para um arquivo e-prints são os mecanismo de submissão; sistema de armazenamento a longo prazo, política para preservação de documentos e interoperabilidade entre todos os repositórios.

Atualmente, tem-se várias plataformas de repositórios *open source.* Citarei brevemente os softwares mais utilizados atualmente que são: DSpace, E-prints e Fedora.

- **DSpace[3]:** DSpace é um software livre desenvolvido pelo MIT (Massachusetts Instituteof Technology) e pelos Laboratórios Hewlett-Packard para criação de repositórios institucionais, é um sistema de livre acesso destinado ao armazenamento, preservação e a disseminação de conteúdo digital. (
- **E-prints[4]:** foi criado pela SchoolofElectronicsand Computer Science ofUniversityof Southampton, também é um software de acesso livre e diz ser a ferramenta mais fácil e rápida de criar repositórios institucionais.
- **Fedora[5]:** (FlexibleExtensible Digital ObjectandRepositoryArchitecture) é um software open source para repositórios digitais que foi desenvolvido pela a Universityof Virginia e pela Cornell University, tem como principal característica gerenciar qualquer tipo de conteúdo digital, afim de preservar e disseminar os conteúdos digitais.

3.4.1 Características

A principal função de um repositório institucional está relacionada com a memória institucional, com a preservação digital do que foi produzido por uma comunidade específica.

De acordo com Dodebei (2009) um repositório institucional é uma base de dados digital e virtual (web-baseddatabase), de caráter coletivo e cumulativo (memória da instituição), de acesso aberto e interoperável que coleta, armazena, dissemina e preserva digitalmente a produção intelectual da instituição.

Nos repositórios é permitido importar e exportar; armazenar e recuperar objetos digitais, porém essa função não é exclusiva dos repositórios, uma base de dados também tem as mesmas funcionalidades, por isso (Heery& Anderson, 2005, apud PORDEUS, 2013) identificaram quatro características que diferenciam um repositório de qualquer outra coleção digital.:

- Os conteúdos são depositados num repositório, quer pelo autor, proprietário ou por terceiro;
- A arquitetura do repositório gere tanto conteúdo como metadados;
- O repositório oferece um conjunto de serviços básicos mínimos, ex.: colocar, encontrar, pesquisar, controle de acesso.

---

[3]http://dspace.org/)

[4]http://www.eprints.org/uk/

[5]http://www.fedora.info/

Segundo Pordeus (2013) o foco e a motivação para criar repositórios digitais podem também diferir, de acordo com o contexto e as comunidades onde foram construídos, entretanto neste trabalho será considerado como repositório apenas os que estão vinculados a uma instituição de ensino e pesquisa.

3.5 Biblioteca digital

As bibliotecas tradicionais são aquelas em que a informação está contida em um suporte físico.

Foi assim durante muito tempo, mas a partir do momento em que foram criadas as bases de dados para consultar referências bibliográficas com o avanço da internet, as necessidades informacionais dos usuários ficando cada vez mais complexas, pensaram-se então que as bibliotecas tradicionais pudessem oferecer um novo serviço informacional que diminuísse o tempo de espera que o usuário tinha para ter acesso a informação. Com isso criaram-se as bibliotecas digitais.

Os conceitos de biblioteca digital podem ser bem diferentes, isso acontece porque a expressão é utilizada em várias áreas do conhecimento. Para manter uma semelhança nos conceitos que serão citados abaixo, optou-se por escolher textos que estão totalmente ligados a ciência da informação.

De acordo com Tammaro e Salarelli (2008) a expressão biblioteca digital tem duas outras expressões sinônimas, que são biblioteca eletrônica e biblioteca virtual. Para Cunha e Cavalcanti (2008),

> biblioteca eletrônica provê acesso não somente ao seu próprio acervo mas também, por meio de redes eletrônicas, a outros tipos de documentos e serviços providos por outras bibliotecas. É vista como uma biblioteca fisicamente identificável, mas que não possui material impresso e que faz parte de uma biblioteca digital.

Para os mesmos autores a definição de biblioteca virtual é:

> Acervo informacional eletrônico que pode ser acessado, de forma remota, e que está hospedado em diversos computadores. Esse tipo de biblioteca não implica localização física, seja para o usuário final, seja para a fonte. O usuário pode acessar a informação a partir de qualquer ponto e a informação pode estar em qualquer lugar.

Os dois conceitos citados anteriormente surgiram bem antes da expressão biblioteca digital. Para acabar com as discussões, em 1997 nos Estados Unidos, alguns pesquisadores definiram no Workshop on Distributed Knowledge Work Environments o que seria considerado a melhor definição para biblioteca digital.

> [...] o conceito de ‘biblioteca digital’ não é simplesmente o equivalente ao de uma coleção digitalizada dotada de instrumentos de gestão da informação. É, antes, um ambiente que reúne coleções, serviços e pessoas para apoiar todo o

> ciclo vital de criação, disseminação, uso e preservação de dados, informação e conhecimento.

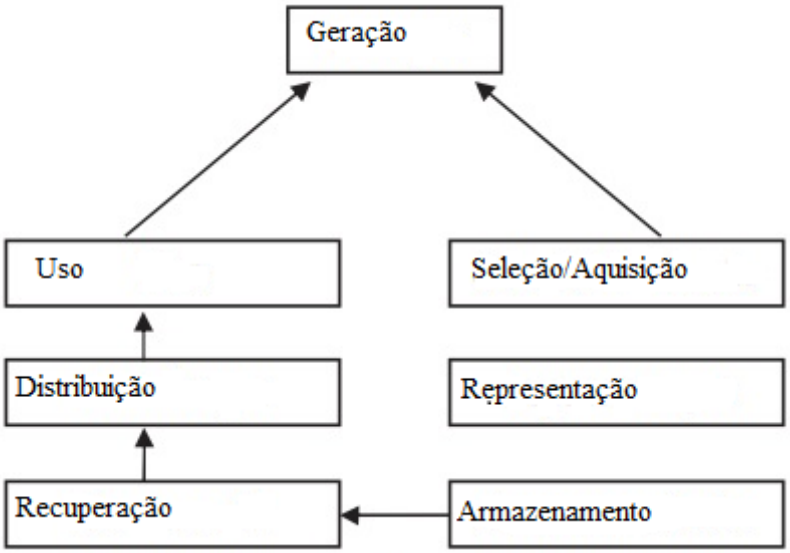
Para este trabalho utilizarei a definição da Digital Library Federation (DFL):

> Bibliotecas digitais são organizações que fornecem os recursos, inclusive o pessoal especializado, para selecionar, estruturar, oferecer acesso intelectual, interpretar, distribuir, preservar a integridade e garantir a permanência no tempo de coleções de obras digitais, de modo que estejam acessíveis, pronta e economicamente, para serem usadas por uma comunidade determinada ou por um conjunto de comunidades.

As bibliotecas digitais são bem parecidas com os repositórios institucionais. Segundo Leite (2012) todo repositório institucional de acesso aberto pode ser considerado um tipo de biblioteca digital. Entretanto, nem toda biblioteca digital pode ser considerada um repositório institucional.

As bibliotecas digitais não são tão diferentes das bibliotecas tradicionais, pois tanto em uma quanto a outra os documentos que serão disponibilizados para o usuário passarão pelo ciclo da informação, que está representado na figura abaixo.

Figura 6 – Ciclo informacional



Fonte: Ponjuán-Dante (1988)

3.5.1 Características

O acervo das bibliotecas digitais estão armazenados em formato digital, isso faz com que o usuário tenha acesso a informação de forma mais cômoda e rápida. Enquanto nas bibliotecas tradicionais o usuário necessita ir até a biblioteca para encontrar a informação desejada correndo o risco do livro estar emprestado, reservado ou até mesmo desaparecido. Já nas bibliotecas digitais isso não ocorre, porque o servidor pode estar há quilômetros de

distância do usuário (clientes) que mesmo assim estes terão acesso ao que se deseja há qualquer momento.

Arms (2000, apud TAMMARO;SALARELLI, 2008) descreve alguns benefícios que a biblioteca digital pode trazer para o usuário:

- Informação entregue diretamente aos usuários: ao invés de ir à biblioteca, os usuários, de qualquer lugar e a qualquer hora, podem ter acesso à biblioteca;
- Melhoramento da pesquisa: as bibliotecas digitais representam um notável aperfeiçoamento dos sistemas de buscas em bases de dados, possibilitando pesquisas integradas e tornando disponíveis serviços em rede, como a possibilidade de navegação entre diversas coleções e a personalização das interfaces;
- Melhor colaboração: as bibliotecas digitais podem favorecer a colaboração entre usuários, por exemplo, compartilhando os mesmos recursos digitais e criando outros de forma cooperativa;
- Atualização das informações: as bibliotecas digitais estão sempre atualizadas. O tempo para publicação é muitas vezes longo, mas a biblioteca digital, em compensação, pode incluir rapidamente os recursos na coleção;
- Melhor uso das informações: ampliação do número de usuários potenciais e também reutilização e personalização dos recursos com relação a diferentes faixas de usuários com diferentes níveis de idade e competência;
- Diminui o fosso digital: as bibliotecas digitais, ao diminuir os limites tradicionais das bibliotecas em matéria de tempo, espaço e cultura, podem ajudar a reduzir a distância que dificulta o acesso à informação.

Além dessas características que estão relacionadas com a comodidade do usuário e o crescimento das bibliotecas digitais. Cunha (1999), citou algumas características que podem ser encontradas nas bibliotecas digitais, que são:

- Acesso remoto pelo usuário, por meio de um computador conectado a uma rede;
- Utilização simultânea do mesmo documento por duas ou mais pessoas;
- Inclusão de produtos e serviços de uma biblioteca ou centro de informação;
- Existência de coleções de documentos correntes onde se pode acessar não somente a referência bibliográfica, mas também o seu texto completo. O percentual de documentos retrospectivos tenderá a aumentar à medida que novos textos forem sendo digitalizados pelos diversos projetos em andamento;

- Provisão de acesso em linha a outras fontes externas de informação (bibliotecas, museus, bancos de dados, instituições públicas e privadas);
- Utilização de maneira que a biblioteca local não necessite ser proprietária do documento solicitado pelo usuário;
- Utilização de diversos suportes de registro da informação tais como texto, som, imagem e números;
- Existência de unidade de gerenciamento do conhecimento, que inclui sistema inteligente ou especialista para ajudar na recuperação de informação mais relevante.

Essas características são resultantes de bibliotecas digitais que foram projetadas pensando-se em cada atividade que pode ser executada pelo usuário e estruturando-as de forma que seja um serviço realmente superior aos proporcionados pelas bibliotecas tradicionais.

3.5.2 Funções

As bibliotecas digitais inicialmente eram bem similares as bibliotecas tradicionais em termos de oferecer acesso a informação porém com o avanço das tecnologias de informação principalmente da internet, faz com que os gestores desses ambientes digitais busquem por mais opções para otimizar o serviço prestado.

Para Arellano (1998) a função da biblioteca digital está muito além de apenas ser a junção de um bom software com uma base de dados. O autor lista as seguintes tarefas que uma biblioteca digital deve desempenhar:

- Criar um ambiente compartilhado que conecte os usuários à coleções de informação pessoal, coleções encontradas em bibliotecas convencionais e coleções de dados usadas por cientistas.
- Desenvolver interfaces de informações gerais ou especializadas relevantes aos seus usuários;
- Prover acesso a um grande número de fontes de informação e coleções de qualidade, ambas em versões *on-line*, integrando-as com os objetos físicos da informação;
- Promover um ambiente que permita a experimentação e a incorporação de novos serviços e produtos;
- Facilitar a provisão , disseminação e uso da informação por instituições, grupos e indivíduos;
- Armazenar e processar informação em múltiplos formatos, incluindo texto, imagem, áudio, vídeo, 3-D,

- Intensificar a comunicação e colaboração entre os sistemas de informação para beneficio da sociedade em geral.

Chen (2004),publicou a pirâmide das funcionalidades. Ela desenvolveu essa pirâmide a partir dos modelos de bibliotecas digital da época.

Níveis de Serviço	**Características**	**Exemplos de realizações**
Transformação	Novos serviços, mudanças na Produtividade, apoio à sociedade de aprendizagem.	Bibliotecas digitais integradas.
Transação (workflow)	Baseada no processo de uma transação.	Governo eletrônico, wiki, repositórios institucionais.
Comunicação	Interatividade, ambientes para a colaboração.	Metadados, criação de conteúdos, comunicação unidirecional.
Informação	Comércio eletrônico, universidades virtuais, ensino eletrônico.	Mecanismos de busca, bibliotecas digitais não-interativas, imagens de textos.

Mesmo utilizando informações digitais a biblioteca digital tem o mesmo propósito de qualquer outra biblioteca, que é, adquirir, organizar, disponibilizar e preservar a informação(ARELLANO, 1998).

4 Metodologia

Durante o embasamento teórico do trabalho notou-se que seria necessário utilizar tanto a pesquisa quantitativa, com aplicação de questionários para coletar dados, quanto a pesquisa qualitativa, que será realizada através de entrevista presencial, porque não é possível obter as informações necessárias somente com o método quantitativo.

Gomes e Araujo (2005), apontam que alguns autores defendem a utilização do método qualitativo no campo das ciências sociais, porque os métodos quantitativos são inapropriados

para esse tipo de ciência, pois não conseguem abarcar a complexidade das questões que envolvem o ser humano.

> A Biblioteconomia é uma área do conhecimento que compõe o campo das ciências sócias, portanto notou-se a necessidade de utilizar tanto o método quantitativo como o qualitativo, ou seja, método misto. Método esse definido por Creswell (2010): Uma abordagem da investigação que combina ou associa as formas qualitativa e quantitativa. Envolve suposições filosóficas, o uso de abordagens qualitativas e quantitativas e a mistura das duas abordagens em um estudo. Por isso é mais do que uma simples análise dos dois tipos de dados; envolve também o uso das duas abordagens em conjunto, de modo que a força geral de um estudo seja maior do que a da pesquisa qualitativa ou quantitativa isolada.

Esse método vem sendo bastante utilizado pois os pesquisadores estão obtendo resultados mais satisfatórios ao utilizarem os métodos de pesquisa mistos. May (2004, apud GOMES; ARAUJO 2005), defende essa corrente da seguinte maneira:

> [...] ao avaliar esses diferentes métodos, deveríamos prestar atenção, [...], não tanto aos métodos relativos a uma divisão quantitativa-qualitativa da pesquisa social – como se uma destas produzisse automaticamente uma verdade melhor do que a outra -, mas aos seus pontos fortes e fragilidades na produção do conhecimento social. Para tanto é necessário um entendimento de seus objetivos e da prática.

De acordo com o Creswell (2010) os métodos mistos divide-se em três tipos, que são: métodos mistos sequenciais, em que o pesquisador utiliza um método para expandir os resultados obtidos com outro método; método misto concomitante, que são aqueles em que o pesquisador mistura os dados qualitativos e quantitativos para fazer a análise; métodos mistos transformativos, o pesquisador utiliza um enfoque teórico para uma perspectiva ampla em um projeto que contém dados qualitativos e quantitativos.

Nesse trabalho foi utilizado o método misto sequencial, pois pretende-se completar os resultados obtidos na pesquisa quantitativa com os obtidos na pesquisa qualitativa.

A estratégia para coleta de dados que será utilizada é a transformativa sequencial. Essa estratégia divide-se em duas fases, onde a primeira fase se caracteriza de forma quantitativa (com aplicação de questionário) e a segunda fase de forma qualitativa (através entrevista presencial).

Creswell (2010) afirma que o campo das ciências sociais e da saúde são complexos, por isso a abordagem apenas do método qualitativo ou quantitativo podem ser insuficientes para lidar com essa complexidade.

Este trabalho apresenta um contexto favorável para a utilização do método de pesquisa misto, porque a Biblioteconomia compõe o campo das ciências sociais e que para alcançar os resultados esperados na pesquisa se faz necessário utilizar o método proposto.

O objeto de estudo deste trabalho é a utilização da indexação automática em repositórios institucionais e bibliotecas digitais de Brasília. O objetivo é compreender o motivo pelo qual as instituições que adotam o software DSpace não utilizam a opção de indexação automática que está disponível desde a versão 1.4.

Para alcançar o objetivo proposto e desenvolver este trabalho faz-se necessário cumprir com as seguintes etapas:

1. definir o tema;
2. delimitar estratégias de busca;
3. pesquisar documentos pertinentes ao tema;
4. organizar os documentos encontrados durante o levantamento bibliográfico por ano e por tipo;
5. definir a amostra para a coleta de dados;
6. definir estratégia de coleta de dados;
7. desenvolver os instrumentos para pesquisa quantitativa;
8. elaborar questionário para coleta de dado;
9. tabular dados da pesquisa quantitativa;
10. analisar os dados obtidos com a pesquisa quantitativa;
11. definir amostra da pesquisa qualitativa;
12. elaborar instrumentos para a pesquisa qualitativa;
13. analisar os dados da pesquisa qualitativa;
14. discorrer sobre os dados obtidos a partir da pesquisa quantitativa e da pesquisa qualitativa.

5 Desenvolvimento

Para encontrar documentos pertinentes ao tema do trabalho e elaborar a revisão de literatura que, dividiu-se em 5 seções, realizou-se a pesquisa em algumas bases de dados, como: Scielo; Biblioteca Digital de Teses e Dissertações da UFMG; Repositório Institucional da Universidade de Brasília; Google Acadêmico; Brapci; DataGramaZero; buscas livres no google.

Para tal foram utilizados os seguintes termos de busca: “indexação automática”, “conceitos de indexação”, “história da indexação”, “indexação manual”, “indexação

mecanizada", "sistemas de indexação", "indexação semiautomática"; "manual DSpace", "histórico DSpace", "indexação DSpace", "repositório institucional", "definição biblioteca digital", "acesso aberto".

Após a estruturação e a elaboração da revisão de literatura, iniciou-se o processo de elaboração da metodologia que seria utilizada na pesquisa.

A metodologia foi definida de acordo com os objetivos específicos e geral do trabalho, de forma que fosse possível obter os melhores resultados.

Definida a metodologia, elaborou-se o primeiramente o instrumento para a coleta de dados da pesquisa quantitativa, um questionário composto por questões dicotômicas, de múltipla escolha e questões abertas.

Antes da aplicação efetiva do questionário, foi aplicado um pré-teste na data de 14 de agosto de 2015, com bibliotecários e estudantes de biblioteconomia que possuem contato com a ferramenta (DSpace).

O pré-teste foi realizado com 3 bibliotecários e 3 estudantes de biblioteconomia. Escolheu-se esses participantes no intuito verificar a clareza com que foram redigidas as questões, para que independente do nível de conhecimento sobre a indexação o participante conseguisse responder.

A versão final do questionário[6] foi encaminhada para o e-mail institucional dos participantes da amostra.

Após o recebimento dos questionários respondidos, iniciou-se a tabulação dos dados e análise.

A conclusão dos dados obtidos com a aplicação do questionário determinou quais as instituições deveriam participar da entrevista presencial para esclarecer de maneira mais abrangente os motivos que leva a não utilização da indexação automática.

As instituições aptas a participarem da pesquisa são as que declararem conhecimento sobre a indexação automática no DSpace, porém não a utiliza.

Para realizar a entrevista, identificou-se quais seriam os possíveis problemas que a ferramenta poderia apresentar, que fazem com que prefiram realizar a indexação manual do que considerar a indexação automática de texto completo.

Com a realização da entrevista foi possível obter dados o suficiente para abarcar os objetivos propostos.

---

[6]O questionário completo está disponível no Apêndice A

5.1 Universo da pesquisa e amostra

Atualmente no Brasil tem-se disponível mais de 80 acervos digitais online e que utilizam a plataforma DSpace para disponibilizar conteúdo digital. Essa informação pode ser encontrada na mesma ferramenta que está ilustrada na figura 13 na pesquisa por facetas.

Para fins desta pesquisa, a população investigada resumiu-se aos acervos digitais que estão vinculados à alguma instituição governamental e de ensino e pesquisa que são geograficamente localizadas no Distrito Federal.

A amostra foi definida com o auxílio de uma ferramenta "quem está usando o DSpace[7]" de consulta disponível na internet em que é possível encontrar quais são os repositórios institucionais e as bibliotecas digitais de Brasília que utilizam o software DSpace. Abaixo encontra-se uma figura da ferramenta de pesquisa.

Figura 7: Quem está usando o DSpace

FONTE: http://registry.duraspace.org/registry/dspace

É importante salientar que o universo da pesquisa foi definido através das informações encontradas no site na época em que foi realizada a consulta, mais especificamente em 29/04/2015, portanto, na atualidade pode conter outros repositórios institucionais e bibliotecas digitais de Brasília que não foram contempladas na pesquisa. Segundo os dados coletados no site, há 13 unidades de informação no Distrito Federal que oferecem acesso a conteúdo digital para o público interno e externo que utilizam o DSpace como plataforma.

Como apresentado na revisão de literatura a amostra é composta por repositórios institucionais e bibliotecas digitais de Brasília, que são elencadas nas subseções seguintes.

[7]http://registry.duraspace.org/registry/dspace

5.2.1 Bibliotecas digitais de Brasília

Em Brasília, centro do poder politico nacional, a maioria das bibliotecas está vinculada a órgãos públicos. A maioria dessas bibliotecas é hibrida, ou seja, elas fornecem tanto o acesso a informação em meio físico quanto em formato digital.

A partir do mapeamento que está disponível no site do DSpace e de visita realizada em alguns centros de informação foram encontradas nove bibliotecas digitais em funcionamento no Distrito Federal, que são:

- Conselho Nacional de Justiça (CNJ)

A biblioteca digital do CNJ é composta por artigos, teses, dissertações, monografias, livros e notícias que vão de encontro ao interesse do Poder Judiciário. E diferentemente das outras bibliotecas digitais, aceita-se documentos textuais de autores que não possui vinculo com o órgão. Para submeter um documento o autor necessita enviar uma cópia do documento em formato .pdf ou em um arquivo gravado no CD.

- Ministério da Educação

A coleção da biblioteca digital do MEC é totalmente composta por artigos, legislações e projetos que são de interesse do órgão.

- Superior Tribunal de Justiça

A Biblioteca Digital Jurídica (BDJur) é um repositório, mantido pelo Superior Tribunal de Justiça (STJ) e gerenciado pela Biblioteca Ministro Oscar Saraiva. Os documentos que compõem a coleção da BDJur, são: legislação do STJ, textos doutrinários (artigos de revistas, capítulos de livros, obras raras e trabalhos acadêmicos de caráter jurídico) e documentos produzidos pelas unidades do Tribunal.

- Tribunal Superior do Trabalho

A biblioteca digital do TST é constituída por atos normativos e administrativos, boletim interno do TST, normas jurisprudenciais, produção intelectual de Ministros e servidores e revista do Tribunal Superior do Trabalho.

- Câmara dos Deputados

A Biblioteca Digital da Câmara dos Deputados reúne conteúdos informacionais relevantes para as atividades legislativas. Esses conteúdos são: publicações editadas pela Edições Câmara; trabalhos de órgãos técnicos; obras raras e valiosas; produção acadêmica de servidores; estudos e notas técnicas das consultorias legislativa e orçamento.

- Senado Federal

O acervo digital da Biblioteca Digital do Senado Federal (BDSF) é composta por livros; obras raras; artigos de revista; notícias de jornal; produção intelectual dos senadores e servidores do órgão; legislação em texto e em áudio.

- Ministério Público Federal

A Biblioteca Digital do Ministério Público Federal (BDMPF) é composta pela legislação produzida por todas as unidades do Ministério Público Federal (MPF) e pelo Conselho Nacional do Ministério Público (CNMP); produção bibliográfica da instituição e pela produção intelectual dos membros e servidores. Tribunal de Justiça do Distrito Federal e Territórios

- Biblioteca digital do Tribunal de Justiça do Distrito Federal e Territórios

A Biblioteca Digital do Tribunal de Justiça do Distrito Federal e Territórios é o portal de acesso às coleções digitais dos artigos doutrinários das mais renomadas revistas jurídicas do país, do Caderno Direito & Justiça e da legislação do interesse da Instituição; produção intelectual dos magistrados e servidores dessa Casa de Justiça; sumário dos livros recém-adquiridos.

- Universidade de Brasília – Biblioteca Digital de Monografias

Possui um acervo de monografias de graduação e especialização, enviadas pelo aluno da UnB na conclusão de curso.

5.2.2 Repositórios institucionais de Brasília

Para facilitar as necessidades de informações que serão utilizadas durante a pesquisa optou-se por escolher apenas os repositórios em Brasília.

De acordo com o que foi descrito na seção 6 deste trabalho, conclui-se que repositórios estão relacionados a comunidades de ensino e pesquisa, portanto todos as instituições descritas nessa subseção estão relacionados ao ensino e pesquisa.

Para encontrar quais eram os repositórios institucionais em Brasília, pesquisou-se através da função “quem está usando DSpace” disponível na página oficial do software quais eram as instituições de ensino e pesquisa de Brasília que utilizam o software DSpace em seus repositórios. Através dessa busca encontrou-se a Universidade Católica de Brasília (UCB) e o Centro Universitário de Brasília (UniCeub), porém há o repositório da Universidade de Brasília que não foi encontrado no site do DSpace.

Antes de explicitar sobre cada repositório é importante salientar que algumas das instituições o denominam de biblioteca digital, entretanto neste trabalho será ignorado o que foi definido pela instituição responsável pelo repositório e assumiremos a definição dada pelo IBICT (2012), são sistemas de informação que armazenam, preservam, divulgam e facilitam o acesso à produção intelectual de comunidades universitárias.

Segue-se as instituições que que estão de acordo com a definição dada pelo IBICT:

- Universidade Católica de Brasília

O repositório da Universidade Católica de Brasília UCB é composta por monografias, teses e dissertações de alunos da instituição. Em seu site não é possui informações mais detalhadas sobre a quantidade de documentos depositados e quais são as políticas de submissão.

- Centro Universitário de Brasília

O repositório institucional do UniCEUB, reúne a produção intelectual da sua comunidade universitária, tanto do corpo discente, quanto do corpo docente. Tem como objetivo ser o principal veiculo para a comunicação científica e apoias o processo de ensino e aprendizagem da instituição.

- Universidade de Brasília – Repositório Institucional

Os trabalhos que podem ser acessados no repositório da UnB foram produzidos pela comunidade acadêmica. São depositados monografias, teses, dissertações, livros, capítulos de livros, artigos, trabalhos apresentados em congressos e anais.

5.3 Instrumento de coleta de dados

Para realizar a coleta dos dados será aplicado um questionário que contém questões de múltipla escolha, questões dicotômicas e questões abertas.

Os questionários serão aplicados via internet com a ferramenta de pesquisa SurveyMonkey[8] que está disponível online.

Após a realização do questionário será realizada uma entrevista[9] presencial afim de confirmar as informações fornecidas no questionário e coletar novas informações.

[8]https://www.surveymonkey.com

[9] A entrevista na íntegra está disponível no Apêndice B desta monografia.

5.4 Apresentação dos resultados dos dados da pesquisa

O questionário utilizado nessa pesquisa foi encaminhado para os participantes através do e-mail institucional e específico da seção ou setor responsável por gerenciar a BD e/ou repositório institucional.

A aplicação do questionário foi escolhida por questões referente ao tempo de resposta e a maior contribuição das pessoas envolvidas na pesquisa.

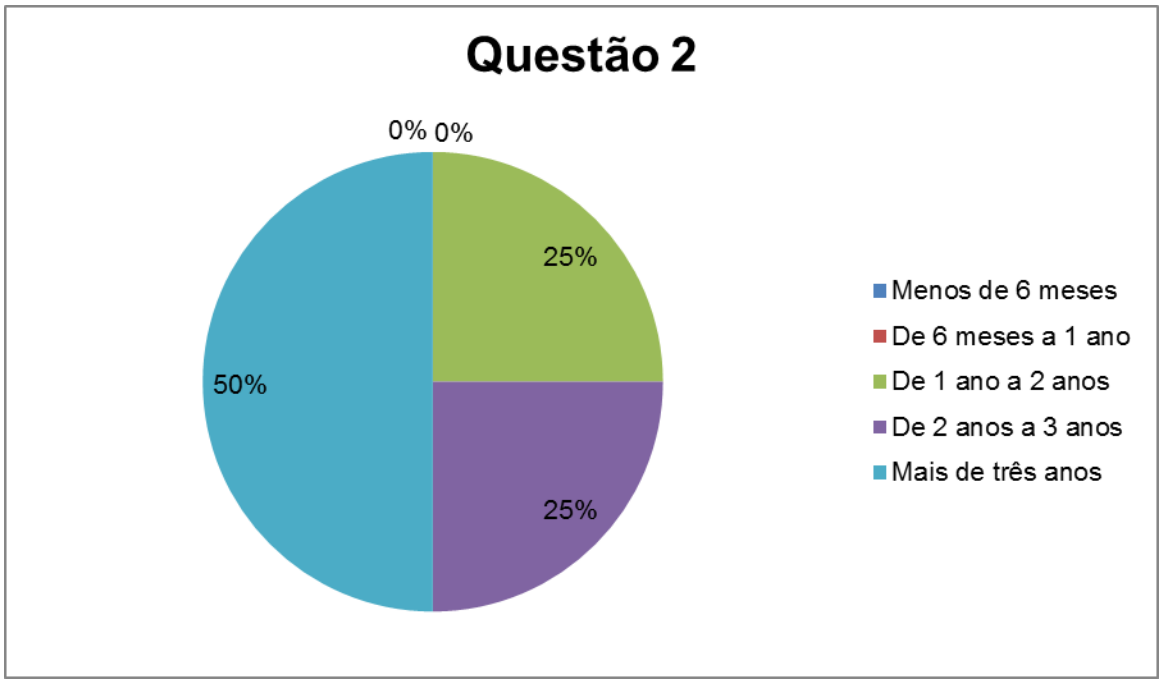
Os dados coletados através do questionário evidenciaram algumas características sobre o contexto dos responsáveis no manejo do software e da indexação realizada por ele.

Todos os dados coletados através do questionário foram tratados de acordo com a estatística descritiva.

5.5 Resultados obtidos por meio da análise do questionário

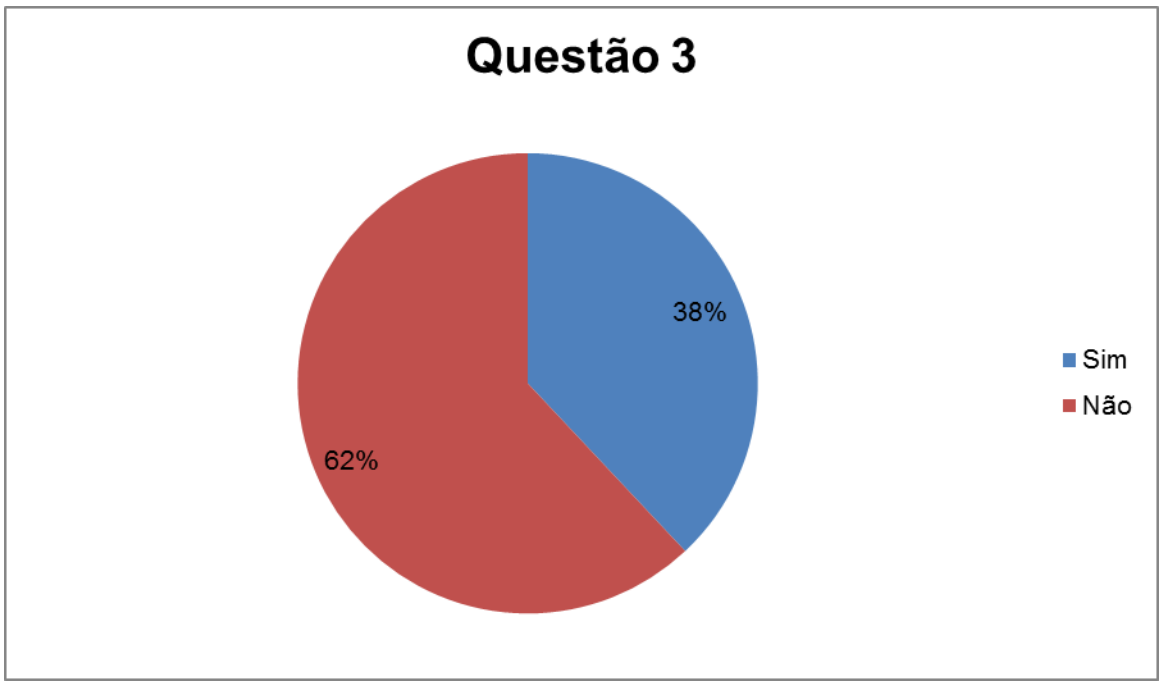
Abaixo encontram-se todos os resultados obtidos por meio da aplicação do questionário. Logo após a tabulação dos dados será apresentado o que pode inferir por meio das respostas dadas pelos participantes da amostra.

Figura 8: Há quanto tempo utiliza o DSpace.



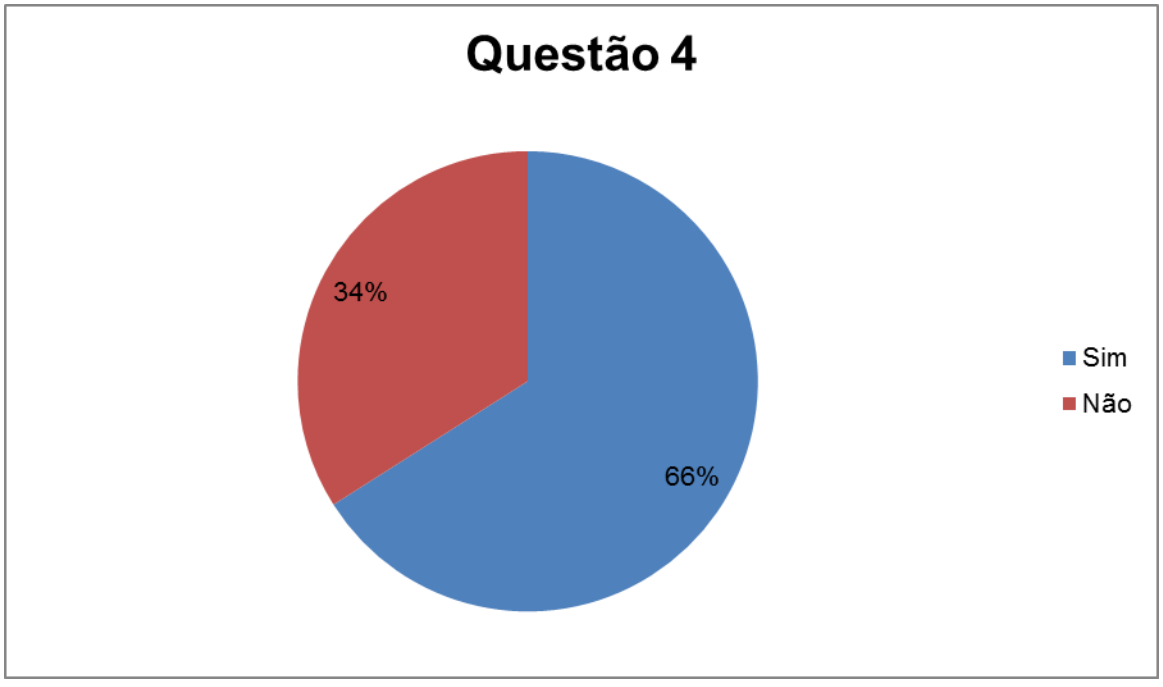
Metade dos entrevistados declararam utilizar o software a mais de três anos.

Figura 9: Recebeu treinamento especializado para utilizar o DSpace?



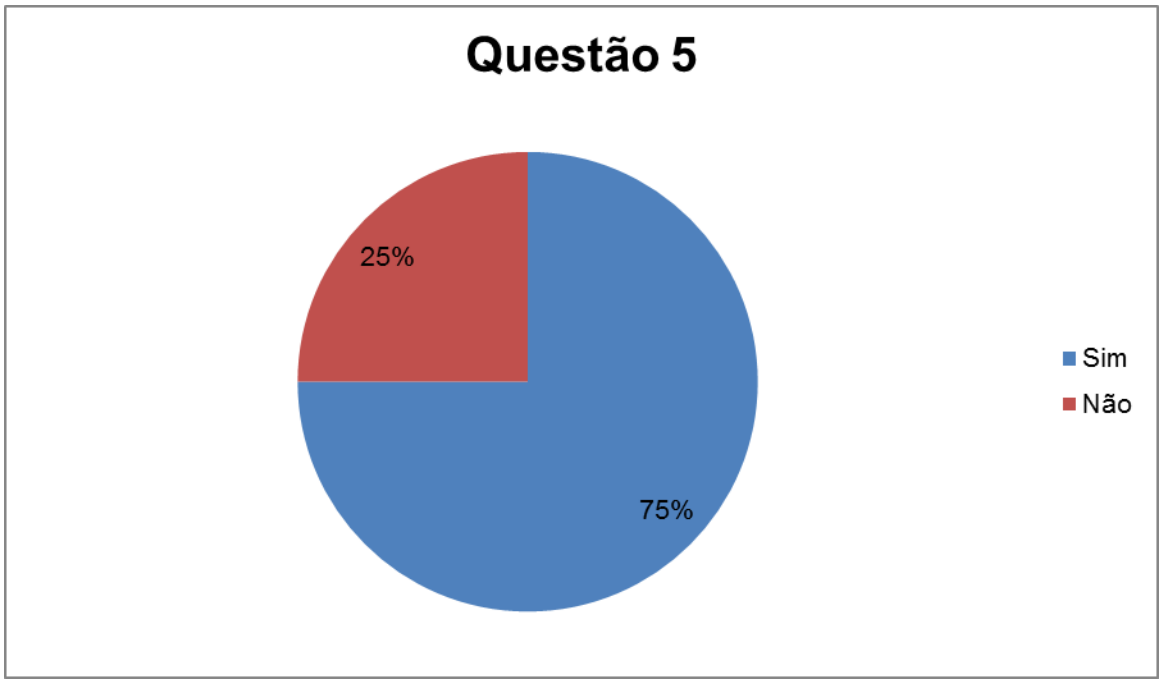
Apenas três instituições declaram ter recebido treinamento especializado para utilizar o software DSpace.

Figura 10: Durante o treinamento, foi abordado que o software possibilita a indexação automática?



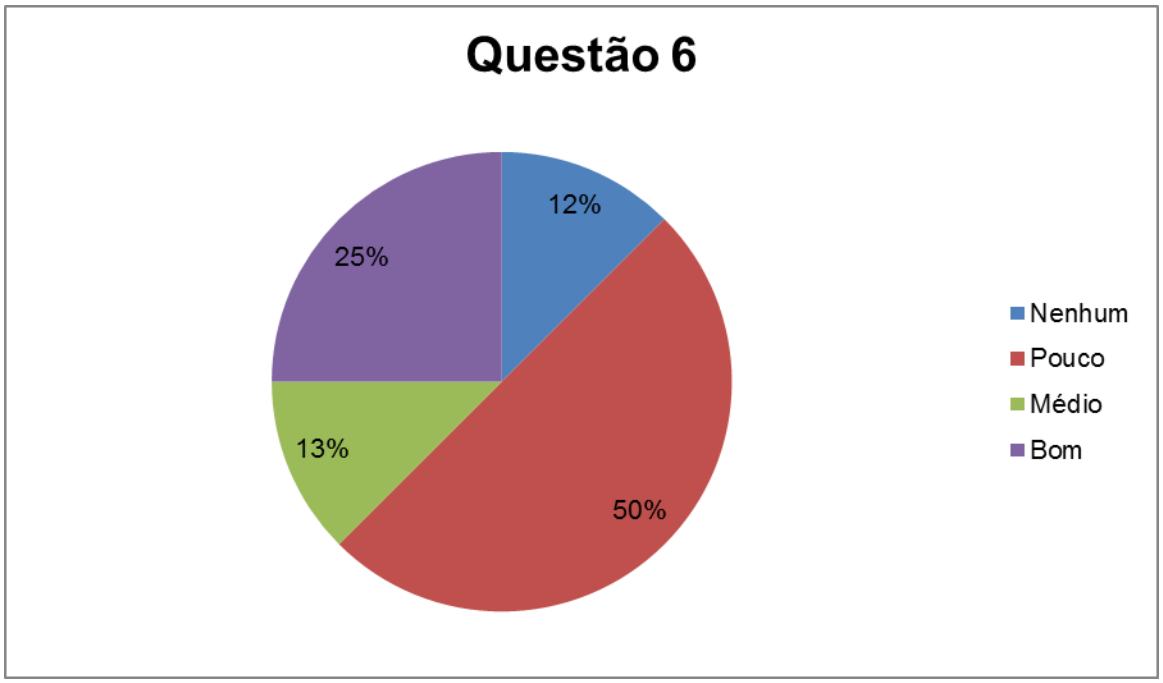
Dos participantes que receberam treinamento especializado, apenas dois afirmaram que durante o treinamento foi abordado a possibilidade de realizar indexação automática no software DSpace.

Figura 11: setor/seção tem apoio da equipe de informática para realizar customização e personalização no software?



Apenas 25% das instituições pesquisadas não possuem apoio da equipe de informática na customização e personalização do software, de acordo com as necessidades da instituição.

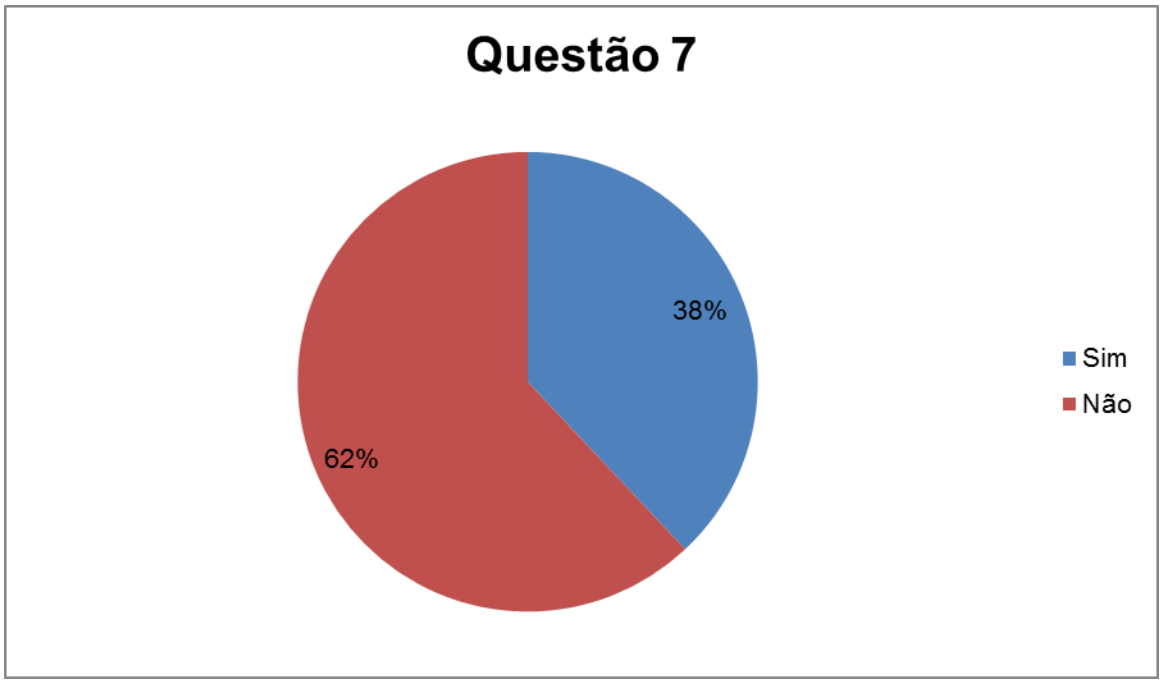
Figura 12: Qual o nível de conhecimento sobre a funcionalidade da indexação automática no DSpace?



Os participantes que declararam ter pouco conhecimento são os que não receberam treinamento especializado e são os que utilizam a ferramenta a um tempo inferior.

Apenas um participante declarou que não tem nenhum conhecimento, apesar de ter declarado utilizar o software a mais de três anos e ter recebido treinamento especializado para utilizar o software.

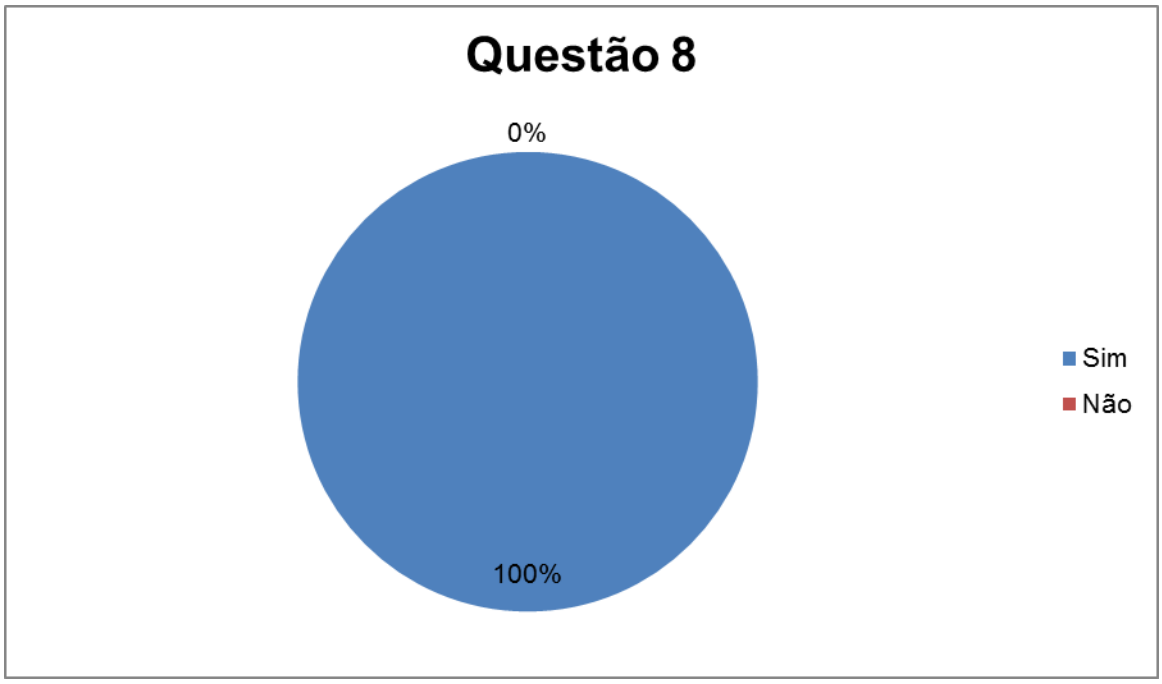
Figura 13: Faz uso da indexação automática no DSpace?



Todos que declararam utilizar a indexação automática, também são os que declararam possuir médio ou bom conhecimento sobre a função e também utilizam a ferramenta a mais de três anos.

Destes apenas um não recebeu treinamento especializado para utilizar a ferramenta.

Figura 14: Se a resposta do item 8 for SIM, responda. A recuperação da informação tem sido satisfatória?



Todos que utilizam a indexação automática no DSpace declaram que a recuperação da informação é bastante satisfatória.

Mas é importante salientar que é do ponto de vista do gestor e não do usuário.

5.6 Interpretação dos resultados

Por meio das respostas obtidas pela aplicação do questionário, conclui-se que todos os respondentes que declararam ter um bom conhecimento sobre o funcionamento da indexação automática no DSpace a utilizam e afirmam que a recuperação da informação é satisfatória.

Apenas um participante manifestou conhecimento sobre a indexação automática no DSpace, mas não a utiliza por ter dificuldade de manejar a ferramenta, por isso prefere utilizar a indexação manual, porque apesar de ser um processo mais moroso ele é mais factível do que utilizar a indexação automática.

Os participantes que não possuem nenhum conhecimento ou pouco conhecimento sobre a indexação automática no software se mostraram interessados em saber mais informações sobre seu funcionamento com o intuito de futuramente poder implementar a indexação automática. Após a análise dos dados obtidos através do questionário, identificou quais eram os candidatos aptos a participarem da entrevista. Os candidatos selecionados para a entrevista foram os que declaram ter algum conhecimento sobre a indexação automática no DSpace, mas não a utilizam.

A entrevista foi realizada no dia 28 de outubro de 2015 com uma servidora do Gerenciamento da Informação Digital (GID), que fica localizado no piso superior da Biblioteca Central (BCE) na Universidade de Brasília (UnB).

Na entrevista questionou-se o porquê de não utilizar a indexação automática e somente ela; e se os motivos eram inerentes as dificuldades advindas do software em questão.

A entrevistada não soube esclarecer as dúvidas referentes às configurações que seriam necessárias realizar para fazer a indexação automática e quais dificuldades poderiam ocasionar a não utilização da indexação automática, porque no momento em que o RIUnB foi implementado ela ainda não trabalhava no setor.

Ainda sobre a indexação automática, a entrevistada declarou que para o contexto do RIUnB não seria possível utilizar apenas a indexação automática, porque a maioria dos acessos ao repositório são provenientes do Google. Ou seja, se utilizasse apenas a indexação automática o acesso de anônimos[10] poderia diminuir circunstancialmente, porque quando o resultado de uma busca é baseada nas palavras do texto a precisão diminui e quando tem-se o auxilio de metadados a precisão aumenta.

Sabendo-se que as estratégias de busca de informação na web são feitas principalmente através dos motores de busca, seria interessante que os termos que estão no documento também fossem pesquisáveis, porém acarretaria problemas ao resultado da busca. Por exemplo, o Google ordena seus resultados de acordo com a frequência e o contexto do usuário, com o intuito de aumentar a precisão, mas quando todas as palavras de todos os documentos em meio digital são recuperáveis por um motor de busca, a precisão pode diminuir.

[10] Usuários que não fazem login

De acordo com o exposto acima e com os dados que foram coletados, pode-se concluir que a não utilização da indexação automática esta relacionada com a carência de conhecimento e treinamento adequado dos profissionais para a utilizarem a ferramenta e também por não oferecer resultados tão satisfatórios fora do site especifico do repositório institucional ou biblioteca digital.

6 Limitações da pesquisa

Toda pesquisa apresenta algumas características que podem dificultar no momento da coleta de dados. Para essa monografia pode-se citar os principais problemas durante a fase da aplicação do questionário e da entrevista presencial.

O problema inicial durante a fase da coleta de dados ocorreu porque na época em que foram aplicado os questionários o país se encontrava em um momento político e econômico complicado, o que ocasionou greve em uma grande parcela das instituições de ensino federais e estaduais e também nos órgãos do poder judiciário.

Esse cenário impactou diretamente no número de abstenção da pesquisa que foi realizada, ou seja, das 12 instituições aptas a participarem da pesquisa, apenas 8 responderam ao questionário.

Um segundo problema relacionado a abstenção é a dificuldade de entrar em contato com algumas instituições, mais especificamente duas instituições onde o principal tipo de contato era através de correio eletrônico.

Também identificou-se uma resistência das pessoas a participarem da pesquisa, muitos se mostraram um pouco resistentes para responder o questionário.

Durante a fase de contato com os candidatos aptos a participarem da entrevista surgiu o único problema da segunda fase. Esse problema era a falta de coerência entre as informações declaradas no questionário e as informações declaradas na entrevista. Por esse motivo obteve-se apenas um participante apto para a entrevista.

Apesar dos problemas relatados acima foi possível encontrar bons resultados afim de cumprir com os objetivos propostos neste trabalho.

7 Considerações Finais

Considerando a proposta desse trabalho, de compreender a não utilização da indexação automática, pode-se afirmar que através da metodologia que foi empregada e das dificuldades que envolveram este estudo os objetivos foram alcançados com êxito.

Com o levantamento bibliográfico e apoiado por uma entrevista realizada com Milton Shintaku, servidor do IBICT, responsável pela coordenação de articulação, geração e aplicação de tecnologia, pôde-se ter uma ideia do porque a indexação automática de documentos textuais não esta sendo amplamente utilizada.

Três os objetivos específicos deste trabalho permitiram responder o objetivo geral, a saber:

- Identificar quais bibliotecas digitais de Brasília utilizam a plataforma de software DSpace;
- Definir o nível de conhecimento dos profissionais sobre indexação automática e a tecnologia disponibilizada pela ferramenta para realizar esse processo;
- Identificar quais as características da indexação automática feita pelo DSpace que não atendem as necessidades de indexação das bibliotecas digitais de Brasília.

O primeiro objetivo específico foi alcançado através do site oficial do DSpace que permite pesquisar quais são as instituições que utilizam o software em uma determinada jurisdição.

O segundo objetivo foi alcançado através da aplicação dos questionários onde possui questões em que o respondente declara seu conhecimento a cerca da indexação automática de texto completo realizada no DSpace.

A resposta para o terceiro objetivo específico obteve-se a partir da entrevista[11] realizada com um dos participantes da amostra, que apontou uma característica do DSpace que não favorece o uso apenas da indexação automática de texto completo. De acordo com o participante entrevistado (informação verbal)[12], a utilização apenas da indexação automática não seria viável porque a maioria dos acessos ao repositório são advindos da pesquisa

[11] A entrevista na íntegra está disponível no Apêndice B desta monografia.

[12] Entrevista concedida ENTREVISTADA. **Entrevista 1.** [out. 2015]. Entrevistador: Juliana Araujo Gomes de Sousa. Brasília, 2015. 1 arquivo .mp3 (30 min). A entrevista na íntegra encontra-se transcrita no Apêndice B desta monografia.

realizada no Google e a versão do DSpace utilizada na instituição não permite a indexação e recuperação direta do PDF por motores de busca.

Com o Google analytics é possível saber como chegaram até o repositório, quanto tempo permanecem na página, os itens visualizados, taxa de rejeição, etc. Através disso é possível saber as estatísticas de acesso advindas de outros caminhos que não o acesso direto da página do repositório, por esse motivo, pelas altas taxas de acesso advindas do Google faz-se necessário manter a indexação manual[13] e a semiautomática.

A fim de observar o comportamento do Google em relação aos resultados de uma pesquisa, realizou-se um teste onde o alvo era encontrar o documento "Manual do DSpace: administração de repositórios". Abaixo encontra-se um exemplo de uma busca simples realizada sem a utilização de filtros ou operadores.

Figura 15: pesquisa pelo nome do autor

Na figura 15 verifica-se o comportamento do Google ao realizar uma busca simples utilizando os nomes dos autores. Nota-se que o documento desejado está entre os primeiros itens a aparecer na primeira página. Tal situação justifica-se, porque na página do repositório da Universidade Federal da Bahia (UFBA), os nomes dos autores estão entre os metadados pesquisáveis.

[13] Faz uso de um campo específico para elencar os assuntos que o documento compreende.

Quando o termo da busca corresponde a um metadado a relevância e a estratégia de busca utilizada vai influenciar nos resultados da busca, o que ocasiona o comportamento verificado na busca realizada.

Figura 16: pesquisa com um termo específico do texto completo

A figura 16 retrata o resultado da pesquisa utilizando como estratégia de busca um termo específico – estrutura informacional do repositório- que aparece apenas no documento completo.

O documento alvo do exemplo continua a aparecer na primeira página dos resultados. Esse comportamento justifica-se porque o termo utilizado na busca foi mais específico, ou seja, a quantidade de documentos que podem corresponder ao termo pesquisado é limitado.

Figura 17: pesquisa com termo mais genérico do texto completo

Diferentemente dos exemplos apresentados nas figuras 15 e 16, o documento desejado não apareceu na primeira página, verificou-se até a página 15 e o documento especificado não foi encontrado.

A explicação para esse fator podem ser duas, a especificidade de um termo, que ocasiona uma precisão maior nos resultados e a relevância dos documentos associado ao termo pesquisado.

É importante observar que os resultados de uma busca sempre vai depender da especificidade do termo utilizado e das estratégias de busca. Exemplos de estratégia de busca são os operadores booleanos (and, or, not); truncamento ou truncagem ; frase exata (utiliza-se a frase exata entre aspas), etc.

Lamentavelmente o terceiro objetivo específico não foi respondido totalmente, pois durante a pesquisa não havia nenhum participante apto a colaborar com as informações referentes ao comportamento do software em relação ao processo e os resultados da indexação automática realizada pelo servidor de busca de alta performance, Solr.

Mesmo com algumas características negativas do software, a ferramenta de indexação Solr apresenta uma performance muito boa referente a indexação de texto completo e também possui funções que podem aperfeiçoar a pesquisa, como por exemplo, criar uma lista de sinônimos, mas para isso é necessário inseri-la manualmente, termo por termo, no código responsável por executar essa função e habilitar a função.

A ferramenta em si apresenta bons resultados e algumas possibilidades de melhoria na pesquisa, mas para isso é necessário que o gestor conheça muito bem a ferramenta de trabalho, afim de obter sempre os melhores resultados.

Como a amostra é bastante restrita, não é possível generalizar os resultados obtidos e nem elencar outras tantas características da ferramenta que não são do conhecimento da entrevistada e nem do pesquisador. Contudo a pesquisa possibilitou identificar algo que ocorre com uma parcela dos bibliotecários, que é a carência de conhecimento para otimizar os resultados na R.I através do uso correto da indexação automática de documentos textuais realizada pelo DSpace.

Contudo, durante a entrevista ficou claro que a indexação automática de texto completo realizada no DSpace possui suas vantagens, mas ainda possui grandes desvantagens, pode-se citar:

- Funciona melhor quando se faz a busca simples: a indexação automática de texto completo apresenta melhores resultados através da busca simples diretamente no *website* do repositório ou da biblioteca digital, porque na busca facetada só serão pesquisáveis os metadados definidos para tal;
- Não é possível identificar qual a relevância no momento de apresentar os resultados da pesquisa: quando o DSpace apresenta os resultados na busca simples os dados são organizados da seguinte maneira: primeiro aparece os resultados que contem os dados de entrada nos metadados pesquisáveis e posteriormente os que os dados de entrada encontra-se no texto completo Entretanto entre esses dois grupos não há uma organização lógica de apresentação;
- Google não permite acesso ao texto completo: o Google consegue indexar todo documento textual em meio digital, porém em alguns casos ele não consegue fornecer ao usuário o acesso a informação. Isso pode ocorrer principalmente com as bibliotecas digitais, onde para acessar a informação é necessário realizar um login, já que diferentemente dos repositórios institucionais, elas não devem oferecer acesso aberto aos seus documentos.

Essas desvantagens vão de encontro ao que foi verificado na entrevista, ou seja, para o contexto daquele repositório institucional não é recomendável utilizar apenas a indexação automática, ainda faz-se necessário combinar a indexação semiautomática e a automática para obter melhores resultados na recuperação da informação.

Esses problemas podem estar relacionados ao Solr, porque ele é um software que realiza a indexação automática, mas a sua premissa é a de recuperar informações dentro de um banco de dados, ou seja, o usuário fornece uma expressão e o Solr apenas vai mostrar quais são os documentos que possuem aquela expressão que foi fornecida na busca, sem se preocupar com a relevância daquele documento para o usuário.

Conforme foi observado por Pinto (2000), a prática da indexação manual ainda é comum em todos os países, principalmente por dois motivos que são: não oferecer respostas totalmente satisfatórias e/ou porque os sistemas de indexação automática não atingem 100% das unidades de documentação desses países.

Apesar desses problemas, isso não justifica a carência de conhecimento dos bibliotecários que trabalham com o software e se limitam a executar apenas as configurações padrão do software a cada versão.

8 Conclusão

Através do levantamento bibliográfico apresentado nas seções dedicadas a revisão de literatura, foi possível identificar características que corroboram para a pouca utilização da indexação automática realizada no software DSpace. Foram apresentadas peculiaridades de como a ferramenta funciona nas diferentes versões do software para que haja uma melhor compreensão da ferramenta.

Ainda na revisão de literatura observou-se uma finidade de vantagens ao optar por utilizar a indexação automática. Porém também observou-se algumas desvantagens em utilizar apenas a indexação automática. Desta forma alcançou-se o objetivo deste trabalho que é compreender a pouca utilização da indexação automática.

Na revisão de literatura, observou-se que uma porcentagem considerável dos trabalhos científicos que se baseiam na investigação da indexação automática preocupam-se mais em identificar, elencar e explicar quais são os métodos, os softwares e seu funcionamento e pouco se encontra sobre o estudo da prática da indexação automática. Uma explicação para esse fato, é justamente a quantidade limitada e pequena de centros de informação que adotam a prática da indexação automática e também sobre as incertezas da qualidade dessa indexação, apesar de ter estudos que apontam que a indexação automática e a indexação manual produzem um resultado semelhante.

Apesar das vantagens da indexação automática, o DSpace ainda não apresenta um quadro favorável em que se permita suprimir a indexação semiautomática, mas isso não é uma característica específica do software que faz a indexação automática, Solr, mas sim do software e do contexto das bibliotecas digitais e repositórios institucionais e também na maneira em que os usuários procuram a informação na atualidade.

Diferentemente das páginas web, as bibliotecas digitais e os repositórios institucionais exigem que se faça download do documento, e em alguns casos é necessário fazer um login

para ter acesso ao documento completo. Sabendo-se que o Google nem sempre tem autonomia para recuperar esses documentos que estão em uma camada mais profunda da web, ainda faz-se necessário a utilização da indexação manual para que através dos metadados de descrição o Google consiga recuperar a informação de forma mais eficiente.

O DSpace está inserido em um contexto atual de implementação de repositórios institucionais e bibliotecas digitais e a cada versão que é lançada apresenta melhorias e também formas de interoperabilidade com motores de busca na web, é um software que tem uma preocupação com a recuperação da informação, prova disso é que na versão 6[14] possivelmente vão disponibilizar outas opções de analisadores.

Sabendo-se que o termo é polissêmico e que na biblioteconomia ele relaciona-se com a representação da informação e que na computação relaciona-se com a recuperação da informação, faz-se necessário justificar como observou-se um ponto de convergência entre os dois conceitos.

Com o desenvolvimento do trabalho observou-se que nesse caso em particular a maneira que a indexação é utilizada na biblioteconomia e a maneira que o termo é utilizado na computação convergem para o mesmo objetivo, que é além de recuperar a informação também representa-la. Isso foi observado ao perceber que os primeiros softwares de indexação automática funcionam utilizando a linguagem natural dos documentos da mesma forma que os softwares utilizados em sistemas de recuperação da informação operam.

Não perceber uma similaridade entre a maneira que a indexação é utilizada em diferentes áreas do conhecimento também pode ser uma característica que faz com que os profissionais não utilizem a indexação automática, por achar que não funciona de acordo com o que é praticado pela biblioteconomia.

Com base no que foi coletado ao longo do trabalho, pode-se concluir que os problemas apontados sobre a indexação automática realizada no DSpace estão relacionadas principalmente com a carência de conhecimento da ferramenta por parte dos profissionais, do que com a ferramenta em si.

[14] Roadmap versão 6 https://wiki.duraspace.org/display/DSPACE/RoadMap

9 Referências bibliográficas

ANDREEWSKI, A., RUAS, V. Indexação automática baseada em métodos linguísticos e estatísticos e sua aplicabilidade à língua portuguesa. Ci. Inf., Brasília, n. 12, p. 61- 73, 1983.

BORGES, G. S. B. **Indexação automática de documentos textuais:** proposta de critérios essenciais. 2009. 111 f. Dissertação (Mestrado em Ciência da Informação) – Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Minas Gerais. 2009.

CATARINO, Maria Elisabete; BAPTISTA, Ana Alice. Folksonomia: um novo conceito para a organização dos recursos digitais na Web. **Datagramazero: Revista de Ciência da Informação,** [s. L], v. 8, n. 3, jun. 2007. Disponível em: <http://www.dgz.org.br/jun07/Art_04.htm>. Acesso em: 27 jun. 2015.

CÂMARA JÚNIOR, A. T. **Indexação automática de acórdãos por meio de processamento de linguagem natural**. 2007. 141 f. Dissertação (Mestrado em Ciência da Informação) – Departamento de Ciência da Informação e Documentação da Universidade de Brasília, Brasília. 2007.

COLLINSON, R.L**. Índices e indexação**: guia para indexação de livros, e coleções de livros, periódicos, e coleções de livros, periódicos, partituras musicais, com uma seção de referência e sugestões para leitura adicional. Trad. Antônio Agenor Briquet de Lemos. São Paulo: Polígono, 1971.

FERNANDES, Jainne Aragão Carvalho. **Indexação automática**: uma revisão de literatura. 2013. 100 f., il. Monografia (Bacharelado em Biblioteconomia)—Universidade de Brasília, Brasília, 2013.

FERREIRA, Ana Gabriela Clipes. Bibliometria na avaliação de periódicos científicos. **Datagramazero: Revista de Ciência da Informação,** [s. L], v. 11, n. 3, jun. 2010. Disponível em: <http://www.dgz.org.br/jun10/Art_05.htm>. Acesso em: 12 abr. 2015.

FOSKETT, A. C. **A abordagem temática da informação**. São Paulo: Polígono, 1973. 437 p.

FUJITA, M. S. L. A avaliação da eficácia de recuperação do sistema de indexação PRECIS. **Ciência da Informação**, Brasília, DF, v. 18, n. 2, p. 120-134, jul./dez. 1989. Disponível em: < http://revista.ibict.br/ciinf/index.php/ciinf/article/view/1361/987 >

GIL LEIVA, I. **La automatización de la indización de documentos**. Gijón (Astúrias): Eciciones Trea, 1999. 220 p.

GOMES, Hagar Espanha. **Guia prático para a elaboração de índices**. Niterói: Grupo Bibli Inf Doc Cien Soc & Hum, 1983. 68 p

HOLANDA, C. ; BRAZ, M. I. **Indexação automática de conteúdos na web**: análise de sites de museus. Biblionline, João Pessoa, v. 8, n. 1, p. 42-59, 2012.

INSTITUTO BRASILEIRO DE INFORMAÇÃO EM CIêNCIA E TECNOLOGIA. **Boas práticas para a construção de repositórios institucionais da produção científica.** Brasília: Ibict, 2012. 34 p. Disponível em: <http://livroaberto.ibict.br/bitstream/1/703/1/Boas práticas para a construção de repositórios institucionais da produção científica.pdf>. Acesso em: 15 abr. 2015.

LANCASTER, F. Wilfrid. **Indexação e resumos:** teoria e prática. Brasília: Briquet de Lemos/Livros, 1991. 347 p.

MARCONDES, Carlos H. et al (Org.). **Bibliotecas digitais:** saberes e práticas. Salvador: Edufba, 2005. 345 p. Disponível em: <http://livroaberto.ibict.br/bitstream/1/1013/1/Bibliotecas Digitais.pdf>. Acesso em: 10 maio 2015.

MORENO, Fernanda Passini; LEITE, Fernando César Lima; ARELLANO, Miguel Ángel Márdero. Acesso livre a publicações e repositórios digitais em ciência da informação no Brasil. **Perspectiva em Ciência da Informação,** Belo Horizonte, v. 11, n. 1, p.82-94. Disponível em: <http://www.scielo.br/pdf/pci/v11n1/v11n1a07.pdf>. Acesso em: 18 maio 2015.

NARUKAWA, Cristina Miyuki. **Estudo de vocabulário controlado na indexação automática:** aplicação no processo de indexação do Sistema de Indización SemiAutomatica

(SISA). 2011. 224 f. Dissertação (Mestrado) - Curso de Ciência da Informação, Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília, 2011. Disponível em: <http://www.marilia.unesp.br/Home/Pos-Graduacao/CienciadaInformacao/Dissertacoes/narukawa_cm_me_mar.pdf>. Acesso em: 25 abr. 2015.

NAVES, Madalena Martins Lopes; KURAMOTO, Hélio (Coord.). **Organização da informação:** princípios e tendências. Brasília: Briquet de Lemos/Livros, 2006. 142 p.

NEVES, Viviane. **Indexação automática de documentos textuais**: iniciativa dos grupos de pesquisa de universidades públicas brasileiras. 2009. 72 f. Tese (Graduação em Biblioteconomia) - Departamento de Biblioteconomia e Documentação da Escola de Comunicações e Arte, Universidade de São Paulo, São Paulo. 2009.

PINTO, V. B. Indexação documentaria: uma forma de representação do conhecimento registrado. Rev. De Letras, v. 1/2, n. 22, 2000. Disponível em: <http://www.revistadeletras.ufc.br/rl22Art09.pdf>.

ROBREDO, J. **Documentação de hoje e de amanhã**: uma abordagem revisitada e contemporânea da Ciência da Informação e de suas aplicações biblioteconômicas documentárias, arquivísticas e museológicas. 4 ed. Brasília: edição de autor, 2005, 409 p

SHINTAKU, M.; BRÄSCHER, M. Dspace versão 1.4: uma análise das facilidades relacionadas ao assunto. In: SEMINÁRIO INTERNACIONAL DE BIBLIOTECAS DIGITAIS BRASIL, 2007, São Paulo. Disponível em: < www.bibliotecadigital.unicamp.br/document/?down=23471 >. Acesso em: 19 mar. 2015.

UNISIST. Princípios de indexação. R. Esc. Bibliotecon., Belo Horizonte, n. 10, p. 83- 94, 1981.

VIEIRA, S. B. **Análise comparativa entre indexação automática e manual da literatura brasileira de Ciência da Informação**. 1984. 204 f. Dissertação (Mestrado em Ciência da Informação) – Departamento de Ciência da Informação e Documentação da Universidade de Brasília, Brasília, 1984

VIEIRA, S. B. Indexação automática e manual: revisão de literatura. Ci Inf. , Brasília, n. 17, p. 43-57, 1988.

FUJITA, M. S. L. (Org.). **A indexação de livros**: a percepção de catalogadores e usuários de bibliotecas universitárias. Um estudo de observação do contexto sociocognitivo com protocolos verbais. São Paulo: Cultura Acadêmica, 2009.

ROBREDO, J. Indexação automática de textos: uma abordagem otimizada e simples. Ciência da Informação, Brasília, v. 20, n. 2, p. 130-136, jul./dez. 1991.

APÊNDICE A

Questionário realizado para fins acadêmicos.

Nome do respondente:

Instituição em que trabalha:

Formação acadêmica:

1. **Informações pessoais**
   1.1. Sexo: ( ) M ( ) F
   1.2. Idade: ( ) 20 a 25 ( )26 a 30 ( ) 31 a 35 ( ) 36 a 40 ( ) 41 a 50 ( ) acima de 50
2. **Há quanto tempo utiliza o DSpace?**
   ( ) menos de 6 meses
   ( ) de 6 meses a 1 ano
   ( ) de um ano a dois anos
   ( ) de dois anos a três anos
   ( ) mais de três anos
3. **Recebeu treinamento especializado para utilizar o DSpace?**
   ( ) Sim ( ) Não

Se a resposta do item 3 for sim, responda o item 4.

4. **Durante o treinamento, foi abordado que o software possibilita a indexação automática?**
   ( ) Sim ( ) Não
5. **O setor/seção tem apoio do equipe de informática para realizar customização e personalização no software?**
   ( ) Sim ( ) Não
6. **Qual o nível de conhecimento sobre a funcionalidade da indexação automática no DSpace?**
   ( ) Nenhum
   ( ) Pouco
   ( ) Médio
   ( ) Bom

7. **Faz uso da indexação automática no DSpace?**

   ( ) Sim ( ) Não

Se a resposta do item 7 for sim, responda o item 8.

8. **A recuperação da informação tem sido satisfatória?**

   ( ) Sim ( ) Não

9. **Se possui médio ou bom conhecimento sobre a funcionalidade da indexação automática no DSpace e NÃO a utiliza, discorra brevemente sobre os motivos.**

APÊNDICE B

Entrevista realizada em 28/10/2015

Qual versão do DSpace é utilizada?

R: No momento estamos utilizando a versão 1.5 do DSpace, porém pretendemos realizar a atualização para corrigir bugs da versão e obter as melhorias e funcionalidades das versões mais recentes.

Como foi feita a configuração para habilitar a indexação automática de texto completo no DSpace?

R: Não sei te responder, pois quando eu cheguei ao setor o DSpace já tinha sido implementado e a pessoa que foi responsável por essa implementação já não está mais aqui.

A indexação automática de texto completo é utilizada nesse repositório institucional?

R: Acredito que não, mas podemos verificar realizando uma busca simples.

Neste momento realizou-se uma busca simples no repositório institucional onde foi constatado que o software foi configurado para realizar a indexação automática de texto completo.

Verifiquei que vocês utilizam os três tipos de indexação, que são: a automática; a semiautomática, que é realizada por meio de metadados definidos como pesquisáveis e a indexação manual, que consiste na utilização de metadado de assunto utilizando um vocabulário controlado dentro do DSpace, por que não utilizar somente a indexação automática de texto completo?

R: Porque a maioria dos acessos feitos ao repositório são advindos da pesquisa do google, que encontra os documentos a partir dos metadados definidos como pesquisáveis e não consegue pesquisar o termo dentro do texto completo.

Por esse motivo não é possível utilizar apenas a indexação automática de texto completo, faz-se utilizar a indexação de texto completo juntamente como a indexação manual ou a semiautomática ou as três juntas que é o caso do nosso repositório.