



**UNIVERSIDADE ESTADUAL DO NORTE DO PARANÁ**  
**CAMPUS LUIZ MENEGHEL - CENTRO DE CIÊNCIAS TECNOLÓGICAS**  
**SISTEMAS DE INFORMAÇÃO**

**ALEXIA GUILHERME BIANQUE**

**Um Estudo de Caso sobre a Indexação automática de documentos oficiais da UENP baseado em *layouts*.**

Bandeirantes

2015

**ALEXIA GUILHERME BIANQUE**

**Um Estudo de Caso sobre a Indexação automática de documentos oficiais da UENP baseado em *layouts*.**

Trabalho de Conclusão de Curso submetido à Universidade Estadual do Norte do Paraná, como requisito parcial para obtenção do grau de Bacharel em Sistemas de Informação.

Orientador: Prof. Dr. Ederson Marcos Sgarbi

Co-Orientador: Prof. Esp. Wellington Della Mura

Bandeirantes

2015

**ALEXIA GUILHERME BIANQUE**

**Um Estudo de Caso sobre a Indexação automática de documentos oficiais da UENP baseado em *layouts*.**

Trabalho de Conclusão de Curso submetido à Universidade Estadual do Norte do Paraná, como requisito parcial para a obtenção do grau de Bacharel em Sistemas de Informação.

**COMISSÃO EXAMINADORA**

---

Prof. Dr. Ederson Marcos Sgarbi  
UENP – *Campus* Luiz Meneghel

---

Prof. Me Glauco Carlos Silva  
UENP – *Campus* Luiz Meneghel

---

Prof. Esp. Wellington Della Mura  
UENP – *Campus* Luiz Meneghel

Bandeirantes, 19 de agosto de 2015

## RESUMO

O trabalho visa à solução de um problema comum na área empresarial e pública: a necessidade das organizações em gerenciar as informações contidas em documentos que se encontram arquivadas de forma desestruturada. A quantidade de documentos impressos armazenados tem crescido exponencialmente e esse aumento faz surgir à necessidade de encontrar uma maneira eficiente para que o processo de localização dos documentos seja preciso e rápido. Por meio dos metadados desses documentos, da indexação e da recuperação da informação isso torna-se possível. O trabalho tem como objetivo a criação de uma ferramenta que reconheça os metadados, faça a indexação automática do documento eletrônico e que proporcione ao usuário a recuperação da informação rapidamente e ofereça a busca por dados de forma precisa.

**Palavras-chave:** Recuperação da informação, metadados, busca, documentos eletrônicos.

## ABSTRACT

*The work is aimed at solving a common problem in business and public area: the need for organizations to manage the information that are unstructured. The amount of stored paper documents has grown exponentially and this increase raises the need to find an efficient way for the process of searching for and locating documents is accurate and fast. Through the metadata of these documents, indexing and information retrieval that becomes possible. The work aims to create a tool that understands the metadata, make automatic indexing of electronic documents and provides the user with the retrieval of information quickly and offer the search for precise data.*

**Keywords:** *information retrieval, metadata, search, electronic documents.*

## LISTA DE TABELAS

Tabela 1 - Elementos do Dublin Core. ....	19
Tabela 2 - Metadados. ....	34
Tabela 3 - Tabela de Demonstração do Grau de Relevância dos Metadados e suas respectivas justificativas. ....	36
Tabela 4 - Comparação de Ferramentas de Indexação e Pesquisa.....	39
Tabela 5 - Descrição e Demonstração dos elementos de processo usados no diagrama. ....	42

## LISTA DE FIGURAS

Figura 1 - Blocos do documento rotulados por números. Fonte: OLIVEIRA, 2014. ....	16
Figura 2 - Áreas demarcadas no documento. Fonte: OLIVEIRA, 2014. ....	17
Figura 3 - Processos de um Sistema de Recuperação de Informação. Fonte: CARDOSO, 2000. .....	23
Figura 4 - Etapas para o processo de indexação automática e consulta. Fonte: O Autor, 2015. .....	30
Figura 5 - Documento Oficial da UENP. Fonte: Site da UENP, 2014. ....	32
Figura 6 – Sub Processo – Obtenção dos blocos do Documento. Fonte: O autor, 2015. ....	33
Figura 7 - Estrutura do Documento do tipo Ato Executivo e seus respectivos metadados definidos. Fonte: O autor, 2015.....	35
Figura 8 - Processo de Indexação. Fonte: O autor, 2015. ....	42
Figura 9 - Processo de Recuperação da Informação. Fonte: O autor, 2015.....	43
Figura 10 - Diagrama de caso de uso. Fonte: O autor, 2015.....	45
Figura 11 - Painel de Controle acessado via navegador Web. Fonte: O autor, 2015.....	46
Figura 12 - Campos (fields) definidos. Fonte: O autor, 2015. ....	47
Figura 13 - Configuração do Full Text. Fonte: O autor, 2015.....	48
Figura 14 - Configuração da Relevância dos campos. Fonte: O autor, 2015.....	49
Figura 15 - Detecção automática de Idioma. Fonte: O autor, 2015. ....	49
Figura 16 - Certificação de Idioma. Fonte: O autor, 2015. ....	50
Figura 17 - Configuração da <i>HighLighter</i> . Fonte: O autor, 2015. ....	50
Figura 18 - Configuração das <i>Stop Words</i> . Fonte: O autor, 2015. ....	51
Figura 19 - Interface de Indexação. Fonte: O autor, 2015. ....	52
Figura 20 – Índice - Documento XML. Fonte: O autor, 2015. ....	52
Figura 21 - Interface de Busca. Fonte: O autor, 2015.....	53
Figura 22 - Teste de Tempo e Integridade, utilizando Metadados. Fonte: O autor, 2015 .....	56
Figura 23 - Teste de Tempo e Integridade, utilizando <i>Full Text</i> . Fonte: O autor, 2015 .....	57

## LISTA DE SIGLAS

GED	Gestão Eletrônica de Documentos
RI	Recuperação da Informação
SRI	Sistema de Recuperação da Informação
UENP	Universidade Estadual do Norte do Paraná
SGBD	Sistemas Gerenciadores de Bancos de Dados
OCR	<i>Optical Character Recognition</i>
PDF	<i>Portable Document Format</i>
XML	<i>Extensible Markup Language</i>



# SUMÁRIO

1. INTRODUÇÃO.....	11
1.1. CONTEXTO E DELIMITAÇÃO DO TRABALHO .....	12
1.2. FORMULAÇÃO DO PROBLEMA.....	12
1.3. OBJETIVOS.....	13
1.3.1. Objetivo Geral.....	13
1.3.2. Objetivos Específicos.....	13
1.4. JUSTIFICATIVA.....	13
1.5. METODOLOGIA DA PESQUISA .....	14
1.5.1. CLASSIFICAÇÃO DA PESQUISA .....	14
1.6. ORGANIZAÇÃO DO TRABALHO .....	14
2. FUNDAMENTAÇÃO TEÓRICA.....	15
2.1. DOCUMENTOS E SUAS TRANSFORMAÇÕES .....	15
2.1.1. DOCUMENTO DIGITALIZADO.....	15
2.1.2. PROCESSAMENTO DE IMAGEM.....	15
2.1.3. DOCUMENTO ELETRÔNICO .....	17
2.2. METADADOS.....	18
2.2.1. <i>DUBLIN CORE METADARA ELEMENT SET (DCMES)</i> .....	19
2.3. INDEXAÇÃO.....	21
2.3.1. INDEXAÇÃO AUTOMÁTICA .....	21
2.4. RECUPERAÇÃO DA INFORMAÇÃO .....	22
2.4.1. SISTEMAS DE RECUPERAÇÃO DA INFORMAÇÃO.....	22
2.4.2. MODELOS DE RECUPERAÇÃO DA INFORMAÇÃO.....	24
2.5. FERRAMENTAS.....	25
2.5.1. <i>APACHE LUCENE</i> .....	26
2.5.2. <i>APACHE SOLR (SEARCHING ON LUCENE REPLICATION)</i> .....	26
2.5.3. <i>htDig</i> .....	27
2.5.4. <i>Datapark Search Enigne</i> .....	27
2.5.5. <i>mnoGoSearch</i> .....	28
2.5.6. <i>Xapian</i> .....	28
2.5.7. <i>ElasticSearch</i> .....	29
2.5.8. <i>Tsearch2</i> .....	29
3. METODOLOGIA PROPOSTA.....	30
3.1. NECESSIDADES DO USUÁRIO .....	31
3.2. DEFINIÇÃO DAS ÁREAS .....	32
3.3. METADADOS.....	34

3.4.	MECANISMO DE INDEXAÇÃO .....	37
3.5.	INDEXAÇÃO E RECUPERAÇÃO .....	41
3.5.1.	PROCESSO INDEXAÇÃO.....	42
3.5.2.	PROCESSO DE RECUPERAÇÃO .....	43
4.1.	MODELO DE CASO DE USO.....	45
4.2.	FERRAMENTA SOLR .....	45
4.3.	CONFIGURAÇÕES PARA INDEXAÇÃO .....	46
4.3.1.	CAMPOS ( <i>FIELDS</i> ) DO DOCUMENTO .....	47
4.3.2.	<i>FULL TEXT</i> .....	48
4.3.3.	GRAU DE RELEVÂNCIA DOS METADADOS .....	48
4.3.4.	DETECÇÃO DO IDIOMA AUTOMATICAMENTE .....	49
4.3.5.	BUSCA <i>HIGHLIGHTER</i> .....	50
4.3.6.	<i>STOPWORDS</i> .....	50
4.4.	INDEXAÇÃO AUTOMÁTICA .....	51
4.4.1.	ÍNDICE GERADO .....	52
4.5.	CONSULTA .....	53
5.	CONSIDERAÇÕES FINAIS .....	55
6.	CONCLUSÃO .....	58
6.1.	TRABALHOS FUTUROS.....	58
	Referências Bibliográficas.....	59
	APÊNDICE A – Imagens de diferentes tipos de Documentos da UENP .....	64
	APÊNDICE B – Descrição dos Casos de uso .....	66

# 1. INTRODUÇÃO

Com o número crescente de documentos produzidos pelas organizações, o armazenamento e a busca por informações importantes pode se tornar uma tarefa demorada e árdua. Quanto maior a quantidade de informações, maior a necessidade de um gerenciamento eficiente a fim de transformá-las em conhecimento. Neste contexto, a Gestão Eletrônica de Documentos (GED) tem se destacado como uma ferramenta estratégica para garantir agilidade na recuperação de documentos (BALDAM; VALLE; CAVALCANTI, 2002). O GED vem sendo utilizado para designar a “utilização de técnicas automatizadas para gerenciar documentos de arquivo, independentemente de seu *layout*” (THOMAZ; SANTOS, 2003).

O GED surgiu a partir da necessidade das organizações em gerenciar as informações que se encontravam desestruturadas, visando facilitar o acesso ao conhecimento explícito da corporação. Pode ser considerado como um conjunto de soluções utilizadas para assegurar a produção, administração, manutenção e destinação aos documentos (SANTOS, 2002).

Os documentos necessitam ser armazenados por meio de um processo de indexação, facilitando assim, a busca por informações. Para que a indexação e a busca tornem-se possíveis existe a necessidade de reconhecer os metadados dos documentos eletrônicos. Segundo Fanning (2006), "os metadados são a chave para se ter acesso à informação que precisamos, quando precisamos". De nada adianta possuir uma solução para gerenciar os documentos se não houver a preocupação com a interpretação dos dados contidos nos mesmos (CHESTER, 2006).

A informação passa a ser insumo para qualquer atividade e para que ela seja útil e relevante tem que estar disponível no momento certo, além disso, “De nada adianta a informação existir, se quem dela necessita não sabe a sua existência ou se ela não puder ser encontrada”. Por isso existe a necessidade das informações serem recuperadas rapidamente e de forma correta quando requeridas pelo usuário por meio de busca por palavras-chave (MARCONDES e SAYÃO, 2001, p. 26).

O acesso a informação que há a necessidade de ser recuperada se dá por meio dos Sistemas de Recuperação da Informação (SRI) que são sistemas que identificam

entre um grande conjunto de informações, aquelas que são realmente úteis, isto é, que estão de acordo com o que o usuário solicitou (ARAÚJO JUNIOR, 2007).

## 1.1. CONTEXTO E DELIMITAÇÃO DO TRABALHO

Com o crescimento do número de documentos nas organizações surge um grande problema, encontrar informações que o usuário necessite. Para que essa dificuldade seja resolvida é preciso o uso de tecnologias que facilitem a busca, trazendo resultados corretos ou que se relacionem diretamente com o termo da pesquisa.

O resultado da consulta está diretamente ligado à maneira que o documento foi indexado. Para que a busca seja rápida e retorne documentos relevantes, será avaliada a necessidade do usuário e a partir daí, será feita a escolha dos metadados contidos nos documentos para que as informações sejam indexadas da melhor forma visando facilitar a busca para o usuário.

Os documentos que serão base do acervo de arquivos são da Universidade Estadual do Norte do Paraná (UENP). A UENP gera uma diversidade de documentos diariamente, como: Editais, Portarias, Licitações, etc. O trabalho abordará a indexação de três tipos, sendo esses: Ato Executivo, Portaria e Ordem de Serviço. O motivo de serem utilizados estes três tipos de documentos é que os três são semelhantes em relação ao *layout* seguindo o mesmo padrão de elaboração.

## 1.2. FORMULAÇÃO DO PROBLEMA

As áreas empresariais e públicas trabalham com uma grande carga de documentos diariamente, existem vários fatores que podem gerar problemas com este grande fluxo, tais como: espaço e segurança de armazenamento, preservação, lentidão na busca por documentos, etc.

Por meio das pesquisas realizadas até o presente momento, nenhum trabalho implementou a indexação de forma automática através da detecção de blocos em documentos eletrônicos oficiais, ou seja, baseado em *layout* de documentos. Outro fator importante é a preservação dos documentos por meio da digitalização e armazenamento.

## 1.3. OBJETIVOS

### 1.3.1. Objetivo Geral

Este projeto visa o desenvolvimento de uma aplicação de indexação automática de documentos digitalizados da UENP com base em seus metadados.

### 1.3.2. Objetivos Específicos

- Definir o conjunto de metadados relevantes para os documentos da UENP;
- Definir uma ferramenta para indexação de texto;
- Implementar o protótipo de indexação automática baseado em metadados;
- Implementar o protótipo para pesquisa baseada no índice gerado;

## 1.4. JUSTIFICATIVA

Atualmente os órgãos públicos estão enfrentando um grave problema: a elevada quantidade de documentos impressos. Devido a esse grande número de papéis, necessita-se de um amplo espaço físico para arquivar tais documentos, com isso as buscas por arquivos específicos tornam-se mais demoradas, e ainda não há a garantia de durabilidade com esse meio de armazenamento.

O uso de sistemas de gestão eletrônica de documentos, GED, visa tornar as consultas mais ágeis. Soluções GED utilizam aplicações de reconhecimento ótico de caracteres para extrair caracteres de textos dos documentos digitalizados. Por meio de um reconhecimento automático de *layout* é possível realizar a identificação de blocos que auxiliaram na indexação dos documentos, havendo assim, a capacidade de pesquisa nos documentos eletrônicos.

O desenvolvimento de uma aplicação de indexação automática de documentos eletrônicos solucionará diversos problemas e trará vantagens, tais como:

- A conservação de documentos eletrônicos é mais fácil de ser gerenciada, pois o armazenamento é feito com segurança e a probabilidade da perda dos documentos eletrônicos é mínima. Se os documentos físicos estiverem guardados em uma sala e

acontecer algum desastre por causa natural como: incêndio, alagamento, os documentos serão perdidos para sempre, causando um imenso dano.

- Rapidez na busca por um documento indexado automaticamente, permitindo a busca por palavras, guardando os documentos em pastas conforme seu título.
- Preservação dos documentos, pois documentos eletrônicos não envelhecem e nem desgastam com o tempo.

## 1.5. METODOLOGIA DA PESQUISA

O projeto desenvolvido possui caráter qualitativo, pois ambiente natural é a fonte direta para coleta de dados e o pesquisador é o instrumento chave. O principal foco de abordagem é o estudo de técnicas de GED com o objetivo de implementar uma solução automática para documentos oficiais por meio da definição de metadados para a realização da indexação automática e a recuperação de informações de forma rápida.

### 1.5.1. CLASSIFICAÇÃO DA PESQUISA

A pesquisa realizada visa gerar conhecimento para a solução de um problema específico: a indexação automática e a consulta em documentos eletrônicos, sendo assim uma pesquisa do tipo qualitativa e aplicada.

## 1.6. ORGANIZAÇÃO DO TRABALHO

O trabalho está organizado da seguinte maneira: O capítulo 2 apresenta a metodologia proposta e a classificação de pesquisa, o capítulo 3 a fundamentação teórica. No capítulo 4 é apresentado a proposta, no 5 o desenvolvimento do projeto, no 6 os resultados obtidos e por fim no capítulo 7 encontra-se a Conclusão do trabalho.

## 2. FUNDAMENTAÇÃO TEÓRICA

Neste capítulo serão apresentados os conceitos teóricos necessários para a realização deste trabalho. O capítulo está organizado da seguinte forma: na Seção 3.1 são apresentados os conceitos de Documentos e seus tipos, na Seção 3.2 estão descritos os Metadados e a ferramenta *Dublin Core*; na Seção 3.3 a Indexação e na Seção 3.4 estão expostos os conceitos de Recuperação da Informação, com a técnica de Busca, a ferramenta *Lucene* e a ferramenta *SOLR*.

### 2.1. DOCUMENTOS E SUAS TRANSFORMAÇÕES

O documento é o conjunto de informações que agrega dados estruturados, semiestruturados e não-estruturados e que representam o conhecimento produzido ao longo de um processo da organização (CENADEM, 2015). O documento é construído a partir da necessidade de transmissão de informações e não como a maneira que o processo se dá.

Os documentos são a base do suporte de informação para a tomada de decisão, o gerenciamento e o controle, possibilitando um melhor rendimento para as empresas e/ou organizações em seus processos otimizados. (SPRAGUE JR, 1995).

#### 2.1.1. DOCUMENTO DIGITALIZADO

O documento digitalizado é uma cópia digital de um documento original existente. Os documentos digitalizados poupam os originais do manuseio e consequente degradação. (WEBB, 2000). O documento digitalizado é geralmente uma imagem, sendo assim, os caracteres contidos no mesmo não são reconhecidos pelo sistema de computador, tornando-o inválido para uma pesquisa de palavras-chave.

#### 2.1.2. PROCESSAMENTO DE IMAGEM

O documento digitalizado passa por um processamento de imagem que é composto por diversos filtros com o intuito de remover ruídos advindos da aquisição ou até mesmo da própria imagem original. Após aplicados os filtros o documento está pronto para aplicação das técnicas de reconhecimento automático de *layout* (OLIVEIRA, 2014).

O reconhecimento de *layout* é realizado da seguinte forma: para rotulação dos blocos foi utilizada a técnica de vizinhança que junta as linhas mais próximas do documento, para que seja possível a formação dos blocos. Após isso são delimitadas as áreas por todo documento (OLIVEIRA, 2014).

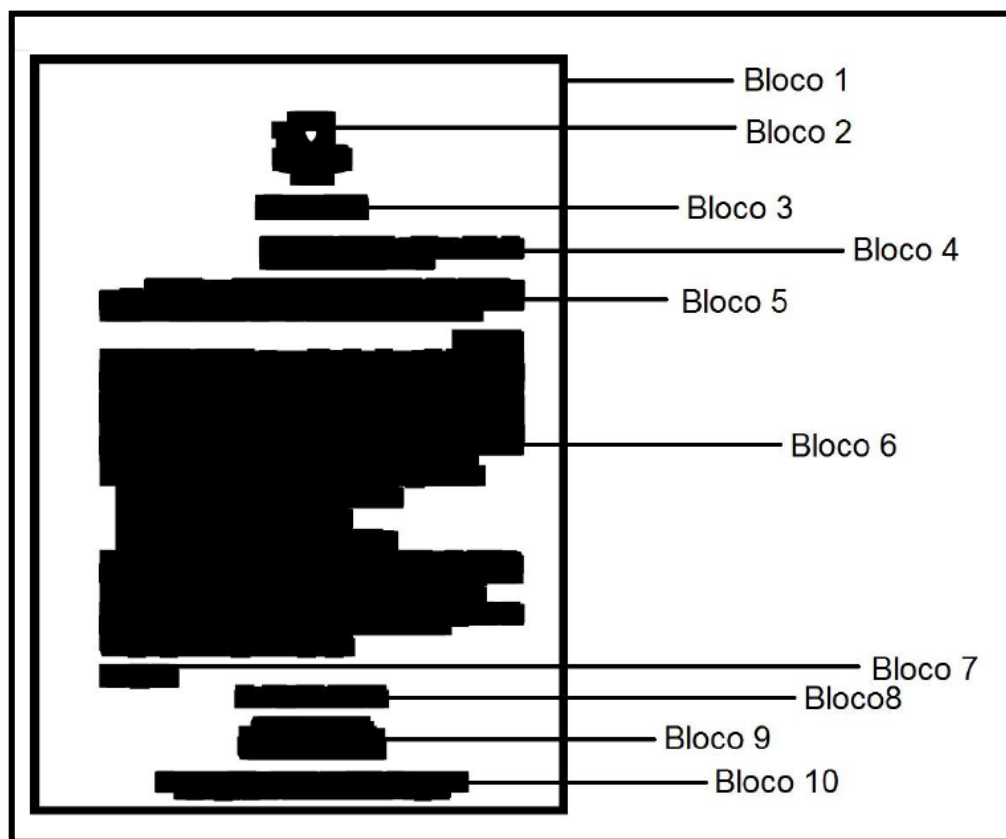


Figura 1 - Blocos do documento rotulados por números. Fonte: OLIVEIRA, 2014.

A partir da Figura 1 são obtidos os blocos de texto existentes no *layout* do documento. Após a análise dos documentos estudados, observou-se que para todos os documentos estudados os blocos serão os mesmos.

As áreas delimitadas consideradas relevantes podem ser observadas na Figura 2.



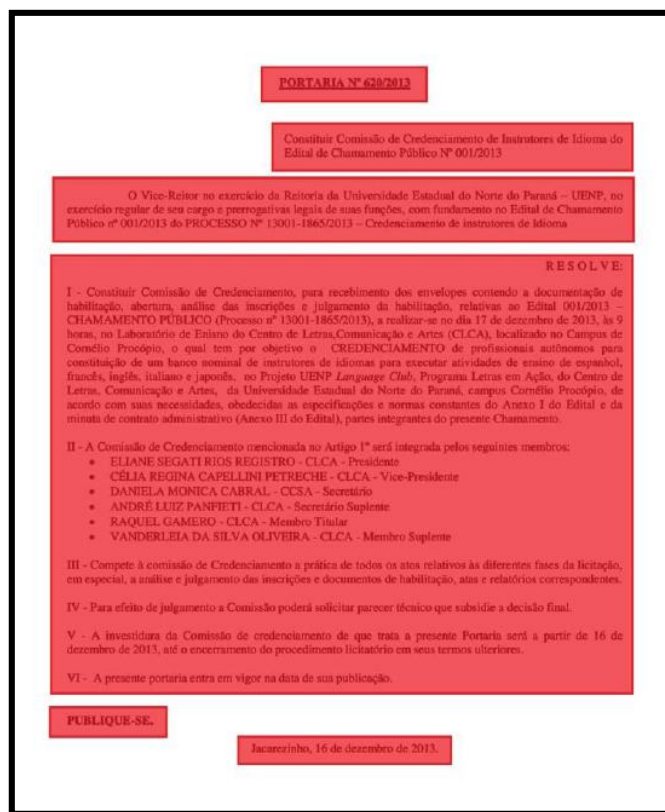


Figura 2 - Áreas demarcadas no documento. Fonte: OLIVEIRA, 2014.

Depois que os blocos estão rotulados é aplicado o OCR (*Optical Character Recognition*) que extrai os caracteres do documento digitalizado. Ao fim de todo este processo, são gerados documentos *.txt* que correspondem as áreas que foram demarcadas no documento e deverão ser utilizados no processo de indexação para que se possa ser gerado um documento eletrônico (OLIVEIRA, 2014).

### 2.1.3. DOCUMENTO ELETRÔNICO

Documento eletrônico é todo registro gerado ou recebido por uma entidade pública ou privada, no desempenho suas atividades, armazenado e disponibilizado ou não, através de sistemas de computação (WEBB, 2000).

Os documentos eletrônicos geralmente são armazenados em microcomputadores e podem ser acessados por diversos usuários.

## 2.2. METADADOS

Metadado é um termo genérico para dados que caracterizam qualquer tipo de informação” (GILS, 2001). São conjuntos de dados estruturados que identificam os dados de um determinado documento e que podem fornecer informação sobre o modo de descrição, administração, requisitos legais de utilização, funcionalidade técnica, uso e preservação (DCMI, 2004).

Informação estruturada que descreve, explica, localiza, ou ainda possibilita que um recurso informacional seja fácil de recuperar, usar ou gerenciar. O termo metadados designa dados sobre dados, ou informação sobre informação (NISO, 2015).

Os metadados tem por funcionalidade o fornecimento de contexto para entender os dados. É a ferramenta para transformar dados brutos em conhecimento (IKEMATU,2001).

Metadados devem ser legíveis tanto por pessoas como também por programas computacionais. Programas podem ajudar a organizar a tornar mais fácil a atividade de encontrar a informação relevante.

Os metadados podem ser classificados em três tipos, sendo esses:

- Metadados descritivos que descrevem uma fonte de informação para fins de identificação e recuperação utilizando elementos como: título, autor, resumo e palavras-chave; (DOS SANTOS, 2011).
- Metadados estruturados que descrevem a organização interna dos objetos e das relações entre eles; (DOS SANTOS, 2011).
- Metadados administrativos que apoiam as atividades de gerenciamento do acervo de recursos de informação como controle de permissões de acesso, localização de arquivos e critérios de avaliação da qualidade. (DOS SANTOS, 2011)

Existem vários esquemas que auxiliam na escolha de metadados, a seguir será exposto o método usado pelo *Dublin Core Metadata Element Set*.

### 2.2.1. DUBLIN CORE METADARA ELEMENT SET (DCMES)

A iniciativa *Dublin Core Metadata Initiative* teve origem em 1994, em Chicago, Estados Unidos, durante a segunda Conferência Internacional sobre a Web. A DCMI voltou a se reunir em 1995, dessa vez em Ohio, também nos Estados Unidos. Dessa reunião surgiu o padrão de metadados que vem sendo amplamente utilizado no mundo todo (PEREIRA; RIBEIRA JUNIOR; NEVES, 2005).

O padrão de metadados Dublin Core pode ser definido como um conjunto de elementos descritivos com a finalidade de facilitar a descrição de recursos digitais disseminados na Internet (SOUZA, 2000). É composto de somente quinze elementos descritivos simples e auto-explicativos.

Suas principais características são:

- simplicidade na descrição de recursos;
- interoperabilidade semântica – promove o entendimento comum dos descritores; ajuda a unificar padrões de descrição de conteúdos, aumentando a possibilidade de interoperabilidade semântica entre disciplinas;
- consenso internacional – padrão de descrição de reconhecimento e aceitação internacional no tocante à cobertura e escopo dos recursos;
- extensibilidade – permite agregar outros metadados e constitui-se em alternativa aos modelos de descrição mais elaborados, demorados e caros

O esquema de descrição de metadados *Dublin Core* é simples e facilita o uso para os criadores e mantenedores de documentos eletrônicos, auxiliando na RI de metadados (LAGOZE, 1996).

Apresenta-se a abaixo uma breve descrição dos quinze elementos utilizados pelo *Dublin Core*, de acordo com Weibel (1997):

Tabela 1 - Elementos do Dublin Core.

Item	Metadado	Descrição
1	<i>Title</i> (Título)	O nome dado ao documento eletrônico pelo autor ou editor.
2	<i>Author or Creator</i> (Autor)	Pessoas ou organizações responsáveis pelo conteúdo intelectual do objeto. (Ex.: autores no caso de documentos escritos; artistas, fotógrafos ou ilustrador no caso de recursos visuais).

3	<i>Subject and Keywords</i> (Assunto)	Representa o assunto do documento eletrônico, podendo ser definido a partir de sistemas de classificação (CDD, CDU, LCSH) ou Thesaurus, ou simplesmente por uma palavra ou conjunto de palavras.
4	<i>Description</i> (Descrição)	Descrição do conteúdo, podendo ser resumo para DLO ou descrição no caso de recursos visuais.
5	Publisher (Editor)	Entidades responsáveis por tornar o documento disponível na presente forma, tais como editor, universidades ou entidades corporativas.
6	<i>Other Contributors</i> (Outros Colaboradores)	Outras pessoas que contribuíram para a realização da obra (editores, tradutores, ilustrador etc.)
7	<i>Date</i> (Data)	A data quando o documento foi disponibilizado na presente forma.
8	<i>Resource Type</i> (Tipo de recurso)	Gênero do recurso, tais como: <i>home page</i> , novela, poema, dicionário, <i>software</i> aplicativo, arquivo de dados etc.
9	<i>Format</i> (Formato)	A manifestação física do documento eletrônico, tais como: Postscript, HTML ou WordPerfect 6.1.
10	<i>Resource Identifier</i> (Identificação)	Série ou número usado para identificar o documento (URL, ISBN etc.).
11	<i>Source</i> (Fonte)	O documento (impresso ou eletrônico) do qual se originou o recurso eletrônico.
12	<i>Language</i> (Idioma)	Idioma do conteúdo intelectual do documento.
13	<i>Relation</i> (Relação)	Relacionamento com outros documentos impressos ou eletrônicos (por exemplo, imagens em um documento, capítulos em um livro ou itens em uma coleção).
14	<i>Coverage</i> (Cobertura)	Localização espacial ou duração temporal característica do documento.
15	<i>Rights Management</i> (Direito Autoral)	Informação sobre direitos autorais.

Fonte: Weibel, 1996 – Tradução: O autor, 2015.

Cada componente do Dublin Core apresentados na Tabela 1 é opcional. Metadados bem escolhidos geram uma indexação organizada e posteriormente uma ágil recuperação da informação.

## 2.3. INDEXAÇÃO

Para que as informações sobre determinado documento possam ser encontradas e recuperadas, é necessário que haja a indexação dos mesmos. A indexação é um processo que visa obter o acesso à informação dos documentos, por intermédio de termos ou códigos. O índice é o mais importante instrumento para recuperar a informação, tendo em vista que o mesmo é como uma “chave” que dá acesso à informação contida nos documentos, ou como uma “ponte” entre o conteúdo de um acervo de documentos e os usuários (ROBREDO, 2005).

Segundo Gomes (2012), “é na análise que consiste a definição do assunto de um documento, para o atendimento às necessidades de recuperação de informação por determinado perfil de usuário”.

Durante a indexação são obtidas as informações do documento através da análise do seu conteúdo e traduzidos para uma linguagem de indexação. Esta representação identifica o documento, definindo seus pontos de acesso para a busca (FERNEDA, 2003).

### 2.3.1. INDEXAÇÃO AUTOMÁTICA

A automação do processo de indexação só é possível por meio de uma simplificação na qual se considera que os assuntos de um documento podem ser derivados de sua estrutura textual através de métodos algorítmicos. A principal vantagem da automação está no seu baixo custo, considerando o crescente barateamento dos computadores e dos softwares (FERNEDA, 2003).

Os métodos automáticos de indexação geralmente utilizam “filtros” para eliminar palavras de pouca significação (*stop words*). Essa forma de indexação seleciona formas significantes (termos ou frases) dos documentos, desconsiderando os significados que os mesmos podem possuir de acordo com os contextos (FERNEDA, 2003).

O processo de indexação automática é similar ao processo de leitura-memorização humano, sendo seu princípio geral baseado na comparação de cada palavra do texto com uma relação de palavras vazias de significado. Essa relação deve ser previamente estabelecida e o resultado dessa comparação conduz, por eliminação, a considerar que as palavras restantes do texto são palavras significativas.

## 2.4. RECUPERAÇÃO DA INFORMAÇÃO

O processo de RI consiste em identificar, no conjunto de documentos de um sistema, quais documentos atendem à necessidade de informação do usuário. O usuário de um sistema de RI está interessado em recuperar a informação sobre um determinado assunto e não em recuperar dados que satisfazem sua expressão de busca. Essa característica é o que diferencia os sistemas de recuperação de informação dos Sistemas Gerenciadores de Bancos de Dados (SGBD) (FERNEDA, 2013).

No centro do processo de RI está a função de busca, que compara as representações dos documentos com a expressão de busca dos usuários e recupera os itens que supostamente fornecem a informação que o usuário procura (FERNEDA, 2013).

A disseminação da internet vem fazendo com que a disponibilidade destes dados cresça vertiginosamente, demandando soluções e sistemas que permitam a organização e manipulação destes dados, considerando todas as suas características. Estes sistemas são conhecidos como sistemas de recuperação de dados (ELMASRI, NAVATHE;2004).

### 2.4.1. SISTEMAS DE RECUPERAÇÃO DA INFORMAÇÃO

Baeza-Yates e Ribeiro-Neto (1999, p. 1), definem Sistemas de Recuperação de Informações (SRI) como sistemas que lidam com as tarefas de representação, armazenamento, organização e acesso aos itens de informação.

Para Lancaster & Warner (1993 p. 4-5), os SRI são a interface entre uma coleção de recursos de informação e uma população de usuários que desempenham as seguintes tarefas: aquisição e armazenamento de documentos; organização e controle desses; e distribuição e disseminação aos usuários.

O usuário digita o que deseja numa forma de consulta que possa ser processada por um SRI. Esta consulta é feita utilizando palavras-chaves que se resumem na necessidade de informação do usuário (CARDOSO, 2002).

O SRI deve de alguma forma “interpretar” o conteúdo das informações encontradas nos documentos de uma coleção e ordena-los de acordo com um grau de relevância. Relevância é a palavra central de um SRI. O SRI deve recuperar todos os

documentos que são relevantes a uma consulta de um usuário e o menor número possível de documentos não relevantes (CARDOSO, 2002).

A Figura 2 descreve o processo de recuperação de informação em sistemas proposto por (Cardoso 2000 apud Gey 1992) apresentando a maneira a qual se dá a recuperação de informação em sistemas automatizados.

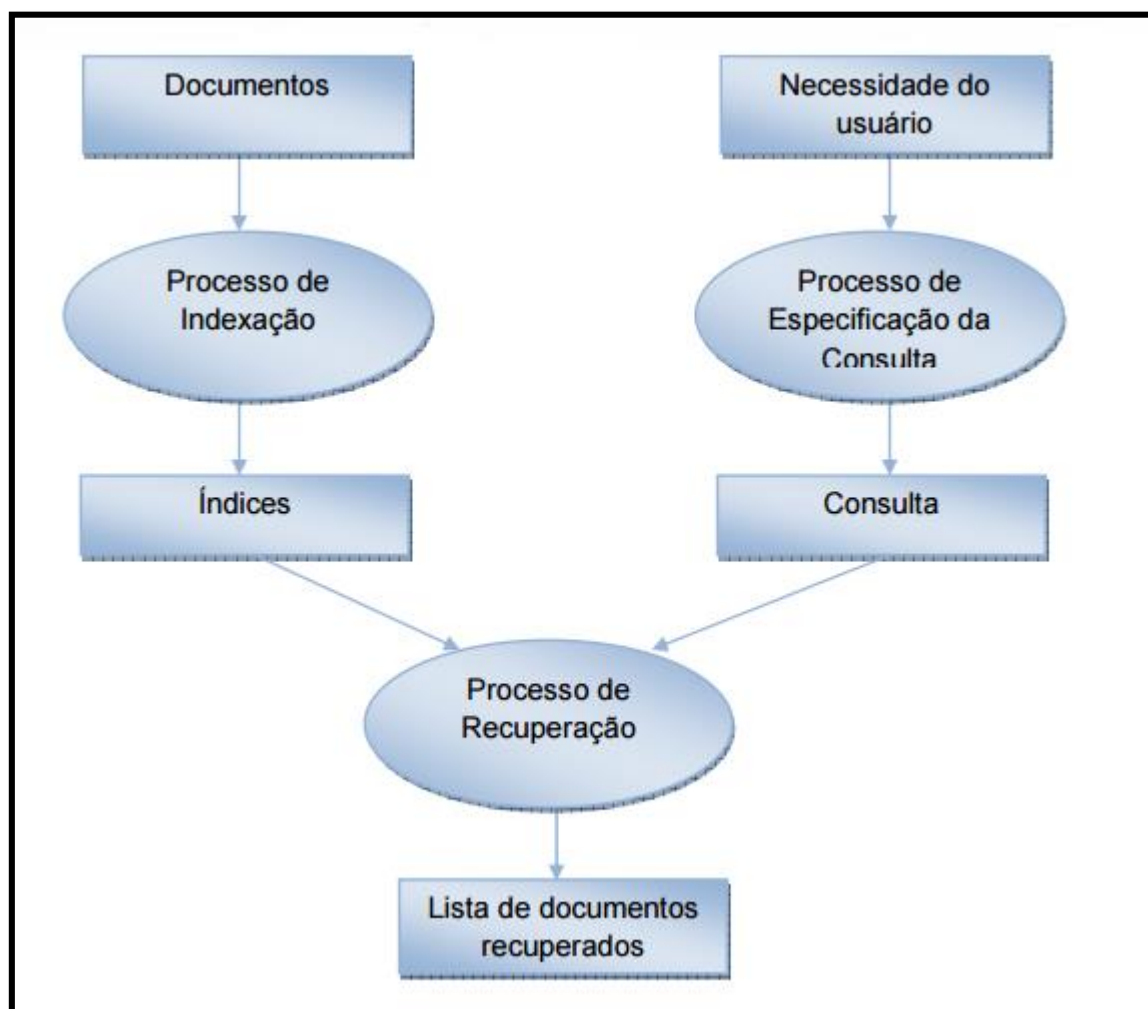


Figura 3 - Processos de um Sistema de Recuperação de Informação. Fonte: CARDOSO, 2000.

De acordo com a Figura 3, os processos básicos da recuperação da informação são: indexação, especificação de consulta e recuperação. A partir de uma coleção de documentos executa-se o processo de indexação para que seja gerado um índice. Este índice é composto pelos termos que representam o documento, a presença do índice facilita a recuperação dos documentos nos futuros processos de busca. O usuário interage com o sistema conforme sua necessidade, especificando o termo a ser

pesquisado no índice. Feita a pesquisa nos índices são apresentados os documentos que apresentam em seu índice o termo pesquisado.

## 2.4.2. MODELOS DE RECUPERAÇÃO DA INFORMAÇÃO

Existem três modelos clássicos na RI. O modelo booleano, vetorial e probabilístico. Além deles, existem também modelos alternativos que tentam refinar ou introduzir novos conceitos. Abaixo serão apresentados os três modelos citados acima.

### BOOLEANO

Um documento é considerado relevante ou não relevante a uma consulta, não existe resultado parcial e não há informação que permita a ordenação do resultado da consulta. Mais utilizado para recuperação de dados do que para recuperação de informação (CARDOSO, 2002).

Vantagens do modelo booleano:

- Expressividade completa se o usuário souber exatamente o que quer;
- É facilmente programável e exato.

Desvantagens do modelo booleano:

- Pessoas lidam com conhecimento parcial;
- Saída não é ordenada.

### VETORIAL

Cada documento é representado como um vetor de termos e cada termo possui um valor associado que indica o grau de importância (peso) deste no documento. As distâncias entre um documento e outro indicam seu grau de similaridade, ou seja, documentos que possuem os mesmos termos acabam sendo colocados em uma mesma região do espaço e, em teoria, tratam de assuntos similares (CARDOSO, 2002).

Vantagens do modelo vetorial:

- Atribuir pesos aos termos melhora o desempenho;
- É uma estratégia de encontro parcial (função de similaridade), que é melhor que a exatidão do modelo booleano;



- Os documentos são ordenados de acordo seu grau de similaridade com a consulta.

Desvantagens do modelo vetorial:

- Ausência de ortogonalidade entre os termos, isto poderia encontrar relações entre termos que aparentemente não têm nada em comum;
- É um modelo generalizado;
- Um documento relevante pode não conter termos da consulta.

### PROBABILISTICO

Os termos indexados dos documentos e das consultas não possuem pesos pré-definidos. A ordenação dos documentos é calculada pesando dinamicamente os termos da consulta relativamente aos documentos. Busca-se saber a probabilidade de um documento D ser ou não relevante para uma consulta Q (CARDOSO, 2002).

Vantagens do Modelo Probabilístico:

- Por obrigar o Princípio da Ordenação Probabilística, o modelo comporta-se otimamente (os documentos são ordenados de forma decrescente por suas probabilidades de serem relevantes);
- Algumas evidências parecem indicar que este modelo tem um desempenho melhor que o do modelo vetorial (por experimentos realizados).

Desvantagens do Modelo Probabilístico:

- Assume a independência entre os termos;
- O modelo não faz uso da frequência dos termos no documento.

Na seção seguinte serão apresentadas as ferramentas utilizadas e será citado quais modelos de RI dão base à ferramenta.

## 2.5. FERRAMENTAS

Nesta seção serão descritas as ferramentas estudadas para o trabalho.

### 2.5.1. APACHE LUCENE

*Lucene* é uma biblioteca de RI, escrita originalmente por *Doug Cutting*, que permite a indexação e busca em aplicações JAVA. Em setembro de 2001, a biblioteca juntou-se a família de soluções JAVA de código aberto da *Apache Software Foundation's Jakarta*, atraindo assim mais desenvolvedores que a tornaram mais robusta e, conseqüentemente, a biblioteca gratuita mais popular de recuperação de informação e busca (Gospodnetic et al., 2009).

O *Apache Lucene* utiliza uma combinação dos modelos vetorial e booleano de recuperação da informação para obter resultados precisos.

A busca consiste em recuperar os documentos que contém um termo informado pelo usuário. No *Lucene* essa operação é extremamente rápida. Mesmo uma consulta complexa, feita em um índice com milhões de documentos, dura menos de um segundo. O resultado da busca pode vir ordenado ou classificado (melhores resultados aparecem primeiro). Opções de consulta fornecidas pelo *Lucene*: (REIS, 2013)

- Busca por palavra-chave ou frase;
- Busca em campos específicos;
- Busca aproximada

Dois etapas principais: indexação e pesquisa:

- A indexação processa os dados originais gerando uma estrutura de dados inter-relacionada eficiente para a pesquisa baseada em palavras-chave.
- A pesquisa, por sua vez, consulta o índice pelas palavras digitadas em uma consulta e organiza os resultados pela similaridade do texto com a consulta.

### 2.5.2. APACHE SOLR (SEARCHING ON LUCENE REPLICATION)

O *Solr* é um mecanismo de indexação e busca textual de código aberto, mantido pela *Apache Software Foundation*, baseado no motor de pesquisa de texto da biblioteca *Lucene* que oferece recursos sofisticados de indexação e busca textual, como: busca com operadores booleanos, busca específica por campo, *highlighting* sobre o resultado da busca, paginação do resultado da busca, *facets* sobre o resultado da busca (recurso presente em *sítes web* de comércio eletrônico), *caching* de busca, integração com banco

de dados relacionais, replicação de bases de dados, interface de administração web, entre outros recursos. (SMILEY; PUGH, 2009).

O *Solr* veio para complementar e agregar diversas funcionalidades do *Lucene*, como com buscas distribuídas, replicações dos índices, clusterização, integração com banco de dados etc. Essa quantidade de novos recursos faz com que o *Solr* seja altamente escalável (BURCHLER, 2010).

As coleções de objetos utilizados pelo *SOLR* são indexadas por meio de definições fortemente estruturadas sobre os campos (*fields*). Cada documento é representado por um conjunto de um ou mais campos, onde cada um corresponde a um conteúdo ou metadados. Os campos podem tomar o valor de *strings*, números, booleanos, datas, entre outros tipos, que podem ser adicionados ao índice. (TEIXEIRA, 2010)

### 2.5.3. *htDig*

O *htDig* é um software livre sendo um sistema de de indexação e busca criado em 1995 por Andrew Scherpbier. Inclui três grupos de arquivos: um conjunto de ferramentas para indexação, um conjunto de ferramentas de busca, e um conjunto de arquivos HTML para a construção da interface de usuário para acessar o motor de busca (HTDIG, 2015).

O *htDig* funciona de forma diferente da maioria motores de busca como a maioria dos motores usam um processo de duas etapas, a construção de um índice e busca-lo. O *htDig* indexa páginas na íntegra, em seguida, processa as páginas em uma forma pesquisável. O último lançamento oficial é *Dig 3.2.0b6* anunciada em 16 de Junho de 2004 (HTDIG, 2015).

### 2.5.4. *Datapark Search Enigne*

*DataparkSearch* é um motor de pesquisa na web. *DataparkSearch* consiste em duas partes. A primeira parte é um mecanismo de indexação (o indexador). O indexador anda sobre referências de hipertexto e encontrados palavras e novas referências no banco de dados. A segunda parte é uma interface para fornecer a pesquisa serviço utilizando os dados recolhidos pelo indexador (DATAPARK SEARCH, 2015).

*DataparkSearch* foi clonado a partir da versão 3.2.16 do *mnoGoSearch* em 27 de Novembro de 2003. O primeiro lançamento da *mnoGoSearch* teve lugar em Novembro

de 1998. O motor de busca tinha o nome de *UDMSearch* até outubro de 2000, quando o projeto foi adquirido pela *Lavtech.Com Corp.* e mudou seu nome para *mnoGoSearch* (DATAPARK SEARCH, 2015).

DataparkSearch é software livre baseado em um motor de busca projetado para organizar pesquisas dentro de um site, grupo de sites, intranet ou sistema local. Escrito em C. distribuído sob os termos da GNU (General Public License) (DATAPARK SEARCH, 2015).

#### 2.5.5. *mnoGoSearch*

O *mnoGoSearch* é um motor de busca de código aberto para sistemas de computadores *Unix* escritos em C, ele pode indexar texto, planilhas, *html*, dados de texto, *xml* e muitos outros tipos de dados usando analisadores externos (MNOGOSEARCH, 2015).

Este motor está pronto para indexar sites multilíngues: uma grande variedade de conjuntos de caracteres e idiomas são suportados e pode ser detectado automaticamente, ele usa a tecnologia de negociação de conteúdo para buscar versões de mesma página em diferentes línguas, ele pode executar sotaque de busca e segmento insensível frases em chinês, japonês e tailandês. É possível usar sinônimos e *fuzzing* baseado no *spell* para estender os resultados da pesquisa. Os resultados podem ser classificados por relevância, horário da última modificação e por título (MNOGOSEARCH, 2015).

A versão para Windows do *MnoGoSearch* apresenta uma interface gráfica do usuário e é vendido sob uma licença comercial. Um exemplo de grande site usando *mnoGoSearch* é *MySQL.com*, o site do sistema de gerenciamento de banco de dados MySQL (MNOGOSEARCH, 2015).

#### 2.5.6. *Xapian*

*Xapian* é uma biblioteca open source de recuperação de informação probabilística. É uma biblioteca motor de pesquisa de texto completo para os programadores. Projetado para ser uma ferramenta altamente adaptável para permitir que desenvolvedores

adicionem facilmente facilidades para indexação e pesquisa para seus próprios aplicativos (XAPIAN, 2015).

Ele é escrito em C ++, com ligações para permitir o uso de Perl, Python, PHP, Java, Tcl, C #, Ruby e Lua. Altamente portátil e funciona em Linux, Mac OS X, FreeBSD, NetBSD, OpenBSD, Solaris, HP-UX, Tru64, IRIX, Microsoft Windows, GNU Hurd, e OS / 2 (XAPIAN, 2015).

### 2.5.7. *ElasticSearch*

*ElasticSearch* é um servidor de buscas distribuído baseado no Apache Lucene. Foi desenvolvido por *Shay Banon* e disponibilizado sobre os termos *Apache License*. *ElasticSearch* foi desenvolvido em Java e possui código aberto liberado como sob os termos da Licença *Apache* (ELASTICSEARCH, 2015).

*Shay Banon* criou o projeto *Compass* em 2004 e ao planejar sua terceira versão, percebeu que seria necessário reescrever grande parte do código para criar uma solução de pesquisa escalável. Assim ele criou uma solução construída a partir do zero, baseada no *Lucene* e usou uma interface comum, *JSON* sobre *HTTP*, adequado para sistemas que utilizam Java como linguagem de programação. A primeira versão do *ElasticSearch* foi lançada em fevereiro de 2010 (ELASTICSEARCH, 2015).

### 2.5.8. *Tsearch2*

O *Tsearch2* é a segunda geração das ferramentas de indexação de texto para o *PostgreSQL*. Composto por uma série de funções, operadores e tipos adicionais que podem ser instalados em uma base de dados a partir do módulo presente no diretório *contrib*. Existem várias ferramentas livres para esse mesmo fim disponíveis na internet, mas nem todas são integradas a um SGBD relacional. Ele parte do princípio que podemos dividir um texto qualquer em palavras-chave e dispor essas unidades em vetores que serão indexados em árvores de busca genéricas. O *Tsearch2* é desenvolvido por *Oleg Bartunov* e *Teodor Sigaev* e faz parte do pacote oficial do *PostgreSQL* desde a versão 7.4, podendo ser instalado também na 7.3 (TSEARCH2, 2015).

### 3. METODOLOGIA PROPOSTA

O trabalho tem como objetivo a implementação de um protótipo que reconheça os metadados, realize a indexação automática de documentos oficiais da UENP e que ofereça a busca por dados, proporcionando ao usuário a recuperação da informação de forma rápida e precisa.

A Figura 4 representa as etapas que devem ser seguidas para que ocorra o processo de indexação e recuperação da informação.

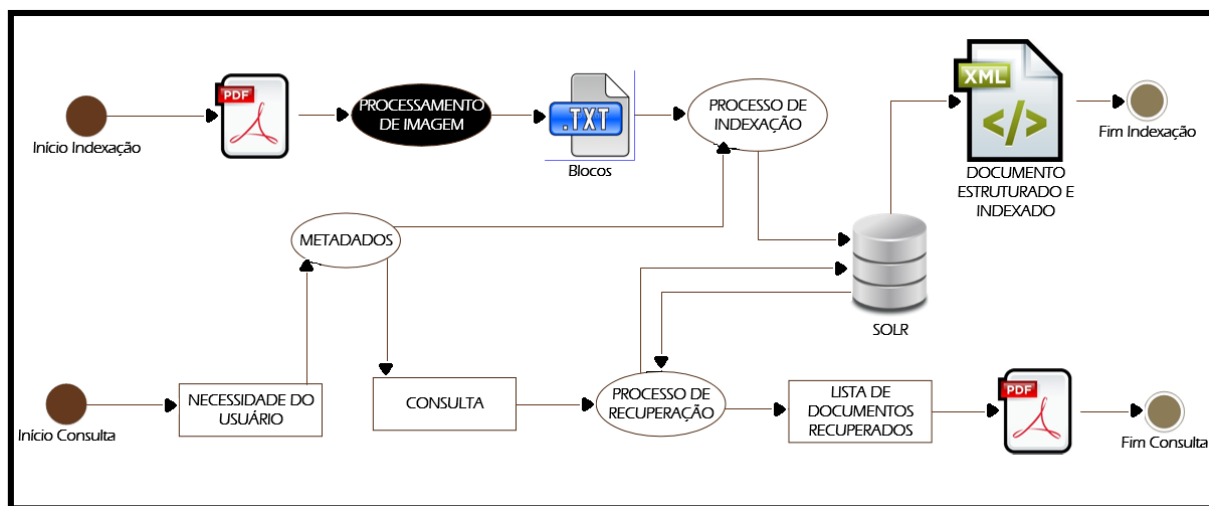


Figura 4 - Etapas para o processo de indexação automática e consulta. Fonte: O Autor, 2015.

O processo de indexação tem como entrada um documento digitalizado do tipo PDF. Aplicando o processamento de imagem no documento *Portable Document Format* (PDF) obtém-se os blocos de texto do tipo texto simples, que serão utilizados para o processo de indexação.

Os blocos serão indexados da seguinte forma: cada bloco de texto corresponderá ao conteúdo de um metadado. Os blocos são enviados para o processo de indexação e por meio de uma configuração o mecanismo já reconhece qual bloco pertence a seu respectivo metadado, estruturando o documento como o original, para que possa ser retornado posteriormente ao usuário.

O processamento de imagem está demonstrado na Figura 3, por uma caixa preta pois o processo é externo, sendo herdado de trabalho Reconhecimento Automático de Blocos para Auxiliar a Indexação em Soluções GED (OLIVEIRA, 2014).

A seguir será realizado a explicação de alguns itens considerados importantes para o desenvolvimento deste trabalho considerando a Figura 3.

### 3.1. NECESSIDADES DO USUÁRIO

O projeto tem como foco a necessidade do usuário, pois esta etapa consiste na definição de quais informações são relevantes para a indexação e a busca baseando-se nos documentos da organização. Definir as reais necessidades e demandas do usuário é um dos mais importantes aspectos do desenvolvimento e implementação de qualquer inovação tecnológica.

As quantidades de documentos impressos armazenados têm crescido mais a cada dia e com isso, surge a necessidade das organizações de gerenciar as informações contidas nos documentos que se encontram desestruturadas.

É importante ter um processo eficiente de busca e localização dos documentos. Por meio deste processo economiza-se tempo, pela rapidez das buscas, espaço físico, pois não há necessidade de armazenar documentos impressos já que o gerenciamento é eletrônico; não há preocupação com a conservação dos documentos, pois dados eletrônicos não se desgastam e caso o armazenamento seja realizado com segurança a probabilidade de perda dos documentos é mínima.

Por meio dos metadados, da indexação e da RI isso torna-se possível. Os documentos oficiais da UENP como: Portarias, Licitações e Resoluções possuem suas particularidades, sendo compostas de acordo com as normas oficiais de elaboração de documentos. As portarias, ato executivo e ordem de serviço, por exemplo seguem um padrão de elaboração, demonstrado pela Figura 5.

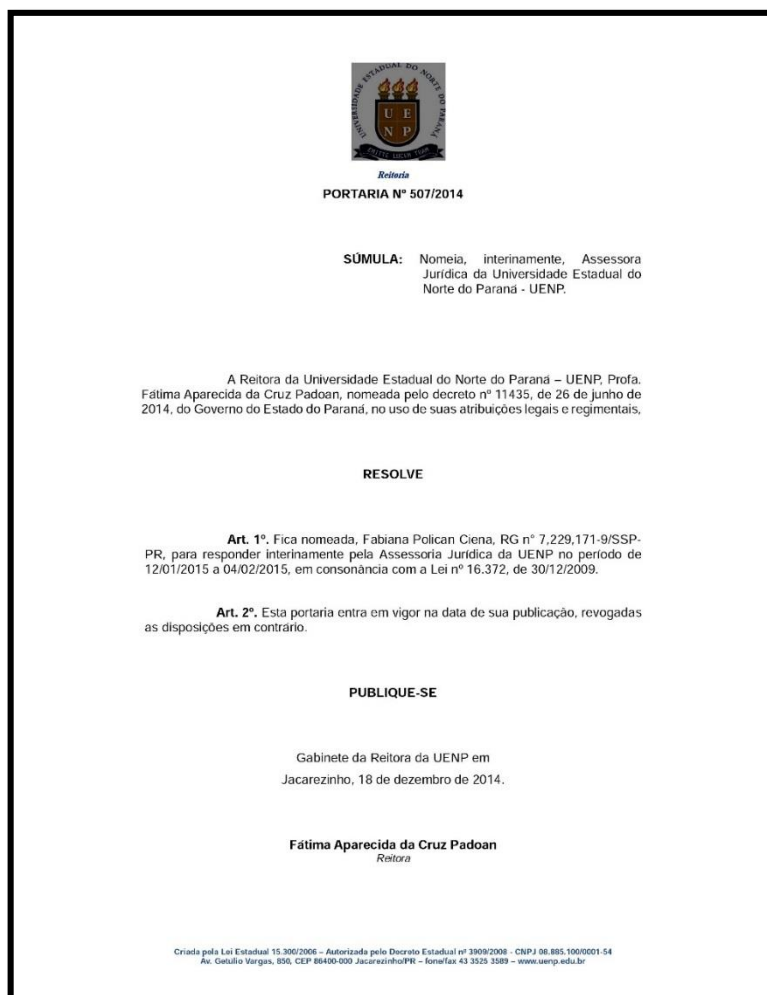


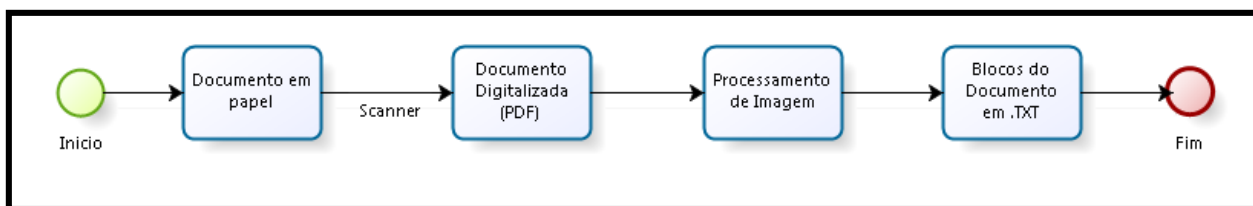
Figura 5 - Documento Oficial da UENP. Fonte: Site da UENP, 2014.

Como pode ser visto na Figura 5 a Portaria possui título, súmula, cidade, data, assinatura e através destas informações é possível determinar quais dados são relevantes para realizar a indexação e a pesquisa dos documentos.

### 3.2. DEFINIÇÃO DAS ÁREAS

O documento passa por um processamento de imagem que divide o documento em áreas. Cada área do documento PDF, por meio da utilização de filtros, se transforma em um bloco do tipo texto simples. Através do diagrama apresentado na Figura 6, pode-se observar as etapas necessárias para que essa transformação ocorra.





**Figura 6 – Sub Processo – Obtenção dos blocos do Documento. Fonte: O autor, 2015.**

O diagrama apresentado pela Figura 6 é um sub processo de uma tarefa presente no processo de indexação, pois os blocos são gerados a partir do trabalho Reconhecimento Automático de Blocos para Auxiliar a Indexação em Soluções GED (OLIVEIRA, 2014).

A partir de um documento da UENP impresso é gerado um documento digitalizado por meio de um *scanner*, sendo este a cópia digital do documento existente em papel. O documento digitalizado passa por um processamento de imagem que é composto por diversos filtros. Após aplicados os filtros são aplicadas as técnicas de reconhecimento automático de *layout*, e a partir são gerados os blocos do tipo texto simples (OLIVEIRA, 2014).

São gerados dez a partir de cada documento, sendo eles: (OLIVEIRA, 2014)

- O bloco 1 é uma borda criada;
- O bloco 2 é o timbre da instituição;
- O bloco 3 é o título do documento;
- Os blocos 4 e 5 fazem parte da súmula;
- Os blocos 6 e 7 são o corpo do texto, ou seja, a informação mais relevante do documento;
- O bloco 8 contém a data e local de expedição do documento;
- O bloco 9 o nome do autor do documento;
- Por fim o bloco 10, o rodapé do documento.

Após a análise dos documentos estudados, observou-se que para todos os documentos estudados os blocos serão os mesmos. Os blocos relevantes são os blocos do 3 ao 9, que serão explicados na próxima seção, os restantes serão descartados.

O próximo passo é definir os metadados dos documentos.

### 3.3. METADADOS

Metadado é a possibilidade de transformar um dado em informação e poder compara-la e combina-la com outra informação. É a técnica de documentar, recuperar, manipular e organizar de maneira estruturada os dados contidos em um documento digital.

Esta etapa consiste na determinação da importância dos dados que servirão como base para a indexação.

Por meio de um estudo realizado na estrutura dos documentos que a UENP gera diariamente foram definidos seus respectivos metadados. Os metadados extraídos são do tipo descritivos, pois facilitam a identificação do assunto e o conteúdo do objeto, tendo o intuito de recuperar documentos, utilizando elementos como: título, autor, data e palavras-chave.

A escolha de metadados foi realizada baseada na necessidade do usuário, pois este manipulará e utilizará as informações dos documentos conforme a precisão. A definição dos metadados dos documentos eletrônicos consiste na combinação de alguns elementos existentes no esquema de metadados *Dublin Core*. Utilizou-se cinco elementos existentes no sendo esses: Título, Assunto, Descrição, Data e Autor. O metadado “Introdução” foi escolhida a partir das informações contidas no *layout* dos documentos da UENP. Ao todo seis metadados que serão apresentados na Tabela 2.

Tabela 2 - Metadados.


<b>Elemento</b>	<b>Descrição</b>
Título	O nome dado ao documento eletrônico pelo autor ou editor.
Assunto	Representa o assunto do documento eletrônico.
Introdução	Informações sobre o autor do documento.
Descrição	Descrição do conteúdo.
Data	A data que o documento foi criado.
Autor	Pessoas ou organizações responsáveis pelo conteúdo do documento.

Fonte: O autor, 2015

A partir da necessidade do usuário foi possível definir a maneira com que os dados irão se comportar: como serão indexados e armazenados e como a informação será

retornada, visando sempre que o utilizador final obtenha um documento relevante à sua pesquisa.

Abaixo está exposta a Figura 7 que tem por finalidade demonstrar a estrutura de um documento da UENP e exibir seus metadados no documento estruturado.



**Titulo**

ATO EXECUTIVO Nº 003/2014 – GR/UENP

**Assunto**

**Súmula:** Estabelece período de recesso administrativo na Universidade Estadual do Norte do Paraná – UENP.

**Introdução**

A Reitora da Universidade Estadual do Norte do Paraná – UENP, Profa. Fátima Aparecida da Cruz Padoan, nomeada pelo decreto nº 11435, de 26 de junho de 2014, do Governo do Estado do Paraná, no uso de suas atribuições legais e regimentais,

**Descrição**

**RESOLVE**

**Art. 1º.** Estabelecer que no período de 22 de dezembro de 2014 a 02 de janeiro de 2015, não haverá atividades administrativas na Universidade Estadual do Norte do Paraná – UENP, exceto nos setores que, a critério de seus responsáveis, não possam sofrer solução de continuidade.

**Art. 2º.** Este Ato Executivo entra em vigor na data de sua publicação, revogadas as disposições em contrário.

**Data**

Jacarezinho, 16 de dezembro de 2014.

**Autor**

*Original assinado*  
**Fátima Aparecida da Cruz Padoan**  
*Reitora da UENP*

Criada pela Lei Estadual 15.300/2006 - Autorizada pelo Decreto Estadual nº 3909/2008 - CNPJ 08.885.100/0001-54  
Av. Getúlio Vargas, 850 - CEP 86400-000 Jacarezinho/PR - fone/fax 43 3525 3589 - www.uemp.edu.br

Figura 7 - Estrutura do Documento do tipo Ato Executivo e seus respectivos metadados definidos. Fonte: O autor, 2015.

Como já citado acima os metadados foram definidos conforme a necessidade do usuário. A partir disso, será aplicado um grau de relevância para cada metadado.

O grau de relevância será definido para que haja um ranqueamento no retorno das consultas, exibindo nas primeiras posições os resultados que são mais significativas para o usuário. Abaixo será apresentado na Tabela 3 o grau de relevância de cada metadado escolhido e a justificativa.

**Tabela 3 - Tabela de Demonstração do Grau de Relevância dos Metadados e suas respectivas justificativas.**

<b>Metadados</b>	<b>Grau de Relevância dos Metadados para Ranquear a Pesquisa</b>	<b>Justificativa</b>
Titulo	3.0	Os títulos dos documentos da UENP são como uma chave identificadora pois cada título é único, já que este contém o tipo de documento, um número identificador e o ano de publicação. Exemplo: Portaria Nº 206/2015
Súmula	1.5	A súmula é uma breve descrição (verbete) do assunto tratado no documento
Introdução	1.0	Geralmente contém informações sobre o autor do documento, como por exemplo: decreto que o nomeou, a data de nomeação
Corpo do Texto	2.5	No corpo do texto contém o conjunto de informações que dá contexto ao documento. Geralmente contém nomes de pessoas, número de documentos pessoais.
Local/Data	1.0	Cidade e a data, as quais o documento foi publicado
Autor	1.0	Pessoa responsável pelo conteúdo intelectual do documento

Fonte: O autor, 2015

Após escolhido o grau de relevância para os metadados, o próximo passo é indexar. Para ocorrer a indexação é necessário eleger a ferramenta que fará essa tarefa. Na seção seguinte foram avaliadas diversas ferramentas e suas respectivas funcionalidades.

### 3.4. MECANISMO DE INDEXAÇÃO

Uma das principais tarefas foi a escolha da ferramenta na qual é a base para o desenvolvimento deste trabalho. Há muitos aspectos a considerar para se ter certeza de que o sistema escolhido atenderá às necessidades da organização.

Após estudar as ferramentas, *Apache SOLR*, *Indexing htDig*, *Datapark Search Enigne*, *Mno Go Search*, *Xapian*, *Elastic Search*, *TSearch 2*, foi reproduzida uma tabela de comparações entre suas funcionalidades.

As funcionalidades comparadas na Tabela 4 serão explicadas e listadas abaixo.

**Open Source:** É o *software*, ferramenta ou aplicação que é duplamente livre e de código aberto. É livremente licenciado para conceder a usuários o direito de uso, cópia, estudo, mudança e melhoria em seu *design* através da disponibilidade de seu código fonte. É livre de taxas e preços.

**API para diferentes linguagens de programação:** Bibliotecas de várias linguagens de programação que se comunique com o Indexador.

**Indexa diferentes tipos de Documento:** A ferramenta suporta a indexação de diversas extensões de documentos. Por exemplo: .doc, .ppt, .xls, .pdf.

**Texto completo (*full text*):** Disponibilidade de indexar e buscar o todo conteúdo do texto, sem divisão de campos.

**Dividir um documento em campos:** Cada documento é representado pelo conjunto de um ou mais campos

**Replicação de dados:** Possibilidade de replicar os dados em outros servidores tornando a aplicação escalável e acessível por meio de outras máquinas.

**Caching de busca:** Quando uma pesquisa é feita, a mesma será guardada no cache e estará disponível para reutilização. Os objetos salvo em cache serão válidos enquanto o Índice existir.

**Sinônimos:** Dicionário de sinônimos para o conteúdo de indexação.

**Diferentes idiomas:** Classifica documentos automaticamente por idioma.

**Busca *Highlighter*:** Destaque no termo pesquisado.

**Relevância por Termo:** Dar peso aos campos escolhidos, para que os resultados sejam precisos e listados por ordem de importância (ranqueados).

**Descartar acento:** Descarta os acentos contidos nas palavras.

**Operadores Booleanos:** São palavras que têm o objetivo de definir para o sistema de busca como deve ser feita a combinação entre os termos ou expressões de uma pesquisa. São eles: *AND* (e), *OR* (ou), *NOT* (não).

**Analisadores:** Quebra o conteúdo de entrada em palavras individuais e faz a conversão de textos para letras minúsculas.

***Spell Checker*:** É o verificador ortográfico. Caso o usuário digite uma palavra de forma errada o sistema oferece uma correção ortográfica automaticamente.

***Stop Words*:** são palavras que podem ser consideradas irrelevantes para o conjunto de resultados a ser exibido em uma busca. Exemplos: as, e, os, de, para, com.

Tabela 4 - Comparação de Ferramentas de Indexação e Pesquisa

<b>Ferramentas X Funcionalidades</b>	<b>Apache SOLR</b>	<b>htDig</b>	<b>Datapark Search Engine</b>	<b>MnoGoSearch</b>	<b>Xapian</b>	<b>Elastic Search</b>	<b>Tsearch2 (Módulo PostgreSQL)</b>
<i>Open Source</i>	X	X	X		X	X	X
<i>API para diferentes linguagens de programação</i>	X				X	X	
<i>Indexa diferentes tipos de Documento</i>	X		X	X	X		
<i>Texto completo (full-text)</i>	X	X	X	X	X	X	X
<i>Dividir um documento em campos</i>	X			X			X
<i>Replicação de dados</i>	X		X	X		X	
<i>Caching de buscas</i>	X		X				
<i>Sinônimos</i>	X	X	X		X	X	
<i>Diferentes idiomas</i>	X	X	X	X	X		X
<i>Busca Highlighter</i>	X					X	X
<i>Relevância por Termo</i>	X		X		X	X	
<i>Descartar acento</i>	X	X	X				
<i>Operadores Booleanos</i>	X	X	X	X	X	X	X
<i>Analísadores</i>	X		X	X		X	
<i>Spell Checker</i>	X		X		X	X	X
<i>Stop Words</i>	X		X	X	X	X	X

Fonte: O autor, 2015

A ideia é que a ferramenta escolhida seja do tipo *open source* pois idealizamos que este projeto tenha uma continuidade por meio de uma integração com outros trabalhos e torne-se um GED de software livre para que seja usado não somente na UENP, mas também em outros ambientes públicos.

Há a necessidade de uma ferramenta que indexe diferentes tipos de documentos pois se tratando de instituições, pode haver a necessidade de indexar documentos de diferentes extensões, ou até documentos que não são semelhantes ao *layout* abordado neste trabalho, sendo assim, gera a necessidade da ferramenta trabalhar da maneira *full text* para que possa ser indexado documentos que se diferem no seu tipo e *layout*.

No presente trabalho há o desejo de desenvolver uma busca otimizada para o usuário, sendo assim, será necessário o uso de funcionalidades que melhoram o retorno da pesquisa, assim como: sinônimos, *caching*, descarte de acentos, operadores booleanos, resultado ranqueado, analisadores, descarte de palavras irrelevantes, um verificador ortográfico, caso o usuário digite a palavra pesquisada de maneira errada e por fim, destaque no termo que foi pesquisado.

Como observado na Tabela 4, a ferramenta que mais se aproxima das necessidades da aplicação foi o *Apache SOLR* que utiliza a biblioteca de busca do *Apache Lucene*.

São ferramentas que se agregam otimizando a busca e a indexação. O *Lucene* foi desenvolvido com base na linguagem Java, o que oferece uma maior flexibilidade na implementação do protótipo da solução, que será desenvolvido também nesta linguagem. O *SOLR* é baseado no motor de pesquisa de texto da biblioteca *Lucene* que oferece os seguintes recursos sofisticados de indexação e busca textual:

- Pesquisa ranqueada - os melhores resultados são retornados primeiro.
- Tipos poderosos de consulta: consultas por frase, consultas de proximidade, consultas de intervalos.
- Altamente escalável através da replicação de dados para outros servidores;
- Busca por campos (título, autor, etc)
- Triagem por qualquer campo
- Múltiplo índice busca com resultados mesclados
- Permite a atualização simultânea e pesquisa



- Mecanismo de armazenamento configurável (*codecs*).








No projeto o SOLR tem como função indexar e recuperar documentos em seu servidor por meio de uma interface desenvolvida pelo autor do projeto.

### 3.5. INDEXAÇÃO E RECUPERAÇÃO

Há a necessidade do usuário de indexar e posteriormente recuperar os documentos do acervo. A indexação e a recuperação serão feitas por meio de uma interface simples e intuitiva, afim de facilitar o acesso do usuário aos documentos.

A Tabela 5 serve como explicação dos elementos de processos utilizados nos diagramas. Serão apresentados diagramas nas etapas de indexação e consulta.

Tabela 5 - Descrição e Demonstração dos elementos de processo usados no diagrama.

Ícone	Elementos de Processo	Descrição
	DataStore	É usado para simbolizar um depósito de Dados que oferece as atividades um mecanismo para resgatar ou atualizar informações
	Início	É usado para simbolizar o início da atividade.
	Fim	É usado para simbolizar o fim da atividade.
	Objeto de Dados	Fornecem informações sobre como documentos, dados e outros objetos são usados.
	Tarefa	É usado para simbolizar uma Tarefa que é uma atividade dentro do processo
	Tarefa de Usuário	É usado para simbolizar uma Tarefa típica de um fluxo de trabalho onde o humano executa a tarefa com o auxílio de um aplicativo de software
	Sub Processo	É usa para simbolizar uma atividade que contém outras atividades

Fonte: O autor, 2015

As seções abaixo demonstrarão e descreverão os processos necessários para obter-se a indexação automática e a busca baseada em metadados.

### 3.5.1. PROCESSO INDEXAÇÃO

A partir de cada documento são gerados seis arquivos do tipo texto simples, cada arquivo corresponde ao conteúdo de um metadado. A Figura 8 representa as etapas para que seja feita a indexação dos blocos para que seja gerado um documento.

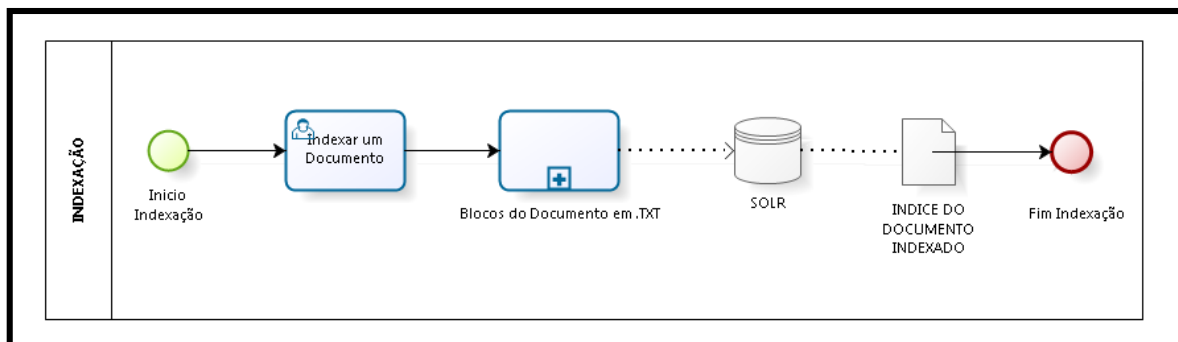


Figura 8 - Processo de Indexação. Fonte: O autor, 2015.

O usuário acessa a interface do sistema para indexar um documento no acervo. A aplicação dará a opção ao usuário de selecionar a pasta em que se encontra os seis

blocos de tipo texto simples, que estão aptos a serem utilizados no processo de indexação. Por meio da Figura 8, podemos observar que Blocos do Documento em .TXT está representado como uma tarefa que possui sub processo.

A Indexação dos blocos será feita por meio do *Solr*, mecanismo de indexação e busca textual baseado no motor de busca da biblioteca *Lucene*. Cada bloco foi definido como um determinado metadado pelo autor e cada documento será representado por um conjunto de seis campos. Todo campo corresponde ao conteúdo de um metadado.

Após o usuário selecionar a pasta do documento que deseja indexar, os seis blocos contidos dentro da pasta serão ligados aos seus respectivos campos. A partir disso, os blocos serão estruturados em um único documento, por meio de uma estrutura *Extensible Markup Language* (XML) gerada pelo SOLR, e indexado, gerando um índice.

Após indexado pode-se utilizar dados contidos no documento para buscas por palavras, permitindo que o usuário encontre qualquer documento que contenha uma determinada palavra ou frase.

### 3.5.2. PROCESSO DE RECUPERAÇÃO

A busca será implementada de acordo com as necessidades do usuário, por meio de uma interface que se comunica com a ferramenta *Apache SOLR*, o qual, se integra com a biblioteca *Apache Lucene* e possibilita uma busca textual com termos destacados. A Figura 9 representa as etapas para que seja feita uma consulta nos documentos indexados.

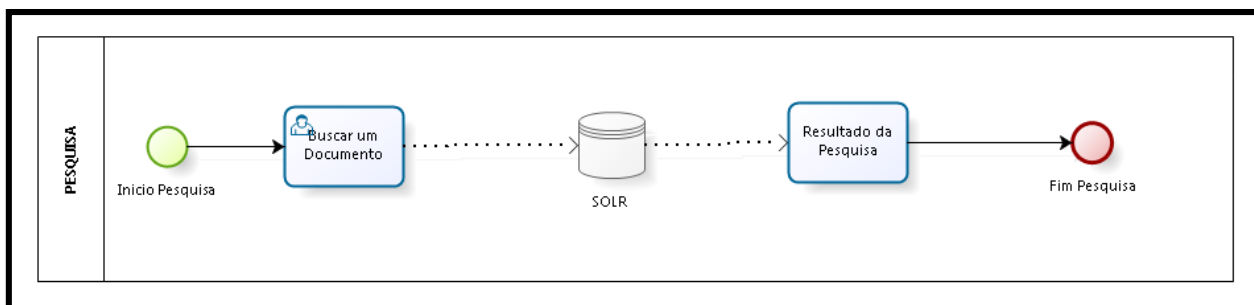


Figura 9 - Processo de Recuperação da Informação. Fonte: O autor, 2015.

O usuário solicita uma pesquisa, é feita a requisição no servidor do *Solr*, onde se encontra o acervo de documentos indexados da UENP. A requisição retorna ao usuário uma lista de documentos que correspondem a pesquisa solicitada. Após o usuário

encontrar o documento que satisfaça sua necessidade, ele tem a possibilidade de acessar o documento digitalizado em PDF.

## 4. DESENVOLVIMENTO

### 4.1. MODELO DE CASO DE USO

Os casos de uso da aplicação estão mostrados na Figura 10.

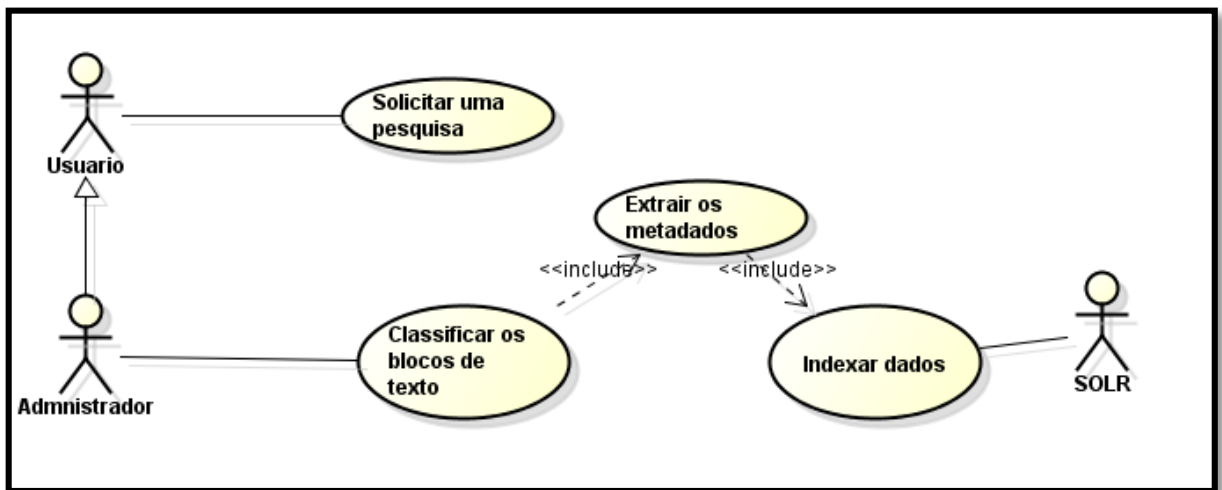


Figura 10 - Diagrama de caso de uso. Fonte: O autor, 2015.

Os casos de uso apresentados na Figura 10 estão descritos e explicados no Apêndice B.

### 4.2. FERRAMENTA SOLR

Para que o *Solr* funcione é necessário ter o servidor *Apache* já rodando em sua máquina. Para obtenção do *Solr* foi realizado o *download* no site do *Apache*, depois de concluído é necessário descompactá-lo para começar o processo de inicialização. A versão usada é a 5.1.0.

A inicialização deve ser feita pelo terminal, dentro da pasta *do Solr* e utilizar o seguinte comando: `bin/solr -p 8983 -e techproducts`

Após iniciado torna-se possível acessar o *Solr* por meio de uma interface de administração *Web* pode-se verificar as configurações e realizar operações administrativas como testes de *queries* e análise de estatísticas, demonstrado pela Figura 11.

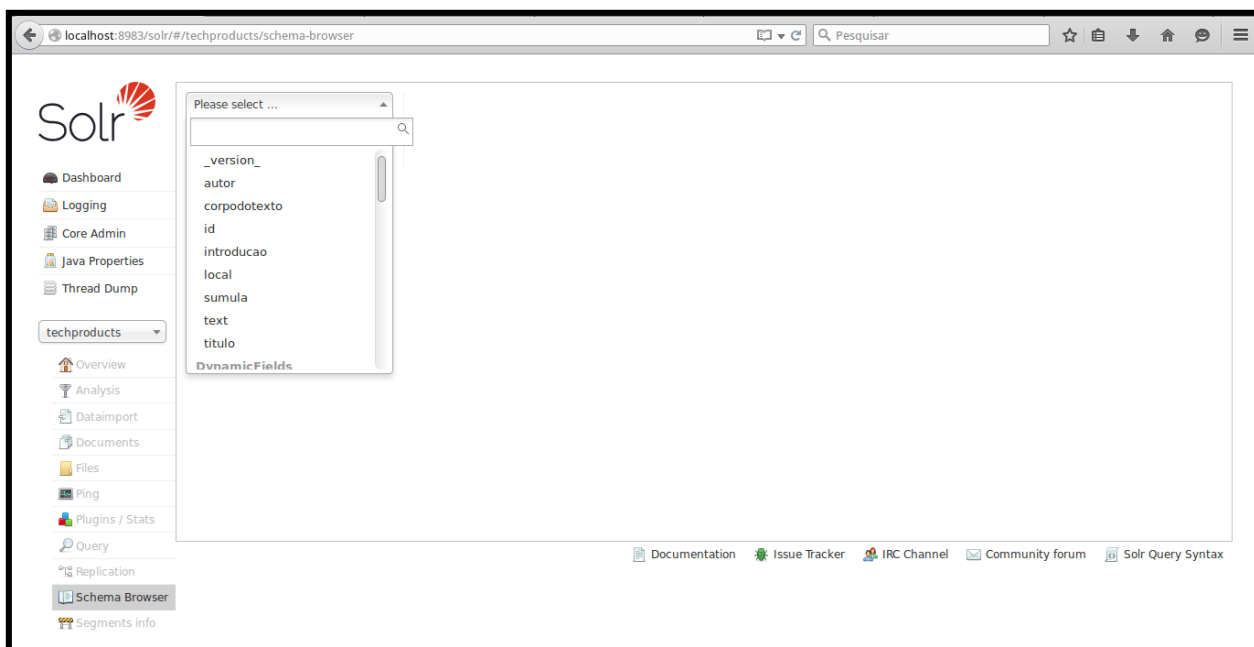


Figura 11 - Painel de Controle acessado via navegador Web. Fonte: O autor, 2015.

Ao acessar esta interface pode-se observar, por meio da Figura 11, um menu lateral que possibilita o administrador observar os arquivos de configurações por meio do item *File*, visualizar o número de documentos indexados no acerto através do item *Overview*, importar uma base de dados por meio do item *DataImport*, executar uma pesquisa acessando o item *Query*, além dessas, há uma grande gama de funcionalidades que a interface administrativa pode oferecer ao gestor.

### 4.3. CONFIGURAÇÕES PARA INDEXAÇÃO

Para indexar arquivos por meio da ferramenta *SOLR*, se ainda não tiver sido desenvolvido uma interface de indexação automática, é necessário que seja através do terminal por linhas de comando. Precisa-se entrar na pasta *solr* e utilizar o seguinte comando: `bin/post -c blocos example/blocos/*.txt`.

O comando segue os seguintes passos:

- *Bin* é a pasta que contém o arquivo *post* que faz a indexação;
- *-c blocos* é para especificar o nome do core que será indexado o documento;

- *Example* é a pasta que contém os blocos que estão sendo selecionado para indexação;
- O comando *\*.txt*, serão indexados todos os arquivos que estão na pasta do tipo *.txt*;

Para que a Indexação ocorra de acordo com as necessidades do projeto é preciso adaptar o SOLR. A forma como o índice será construído, depende das especificações do *schema* e do *solrconfig* (documentos de configuração do *Solr*). Através destes arquivos é possível definir todos os dados a serem utilizados e como será a representação dos documentos. Nas seções abaixo serão expostas as adaptações feitas.

#### 4.3.1. CAMPOS (*FIELDS*) DO DOCUMENTO

Cada documento é representado pelo conjunto de um ou mais campos, onde cada um corresponde ao conteúdo de um metadado. Por meio da configuração de campos (*fields*) é possível definir o modelo de dados que será utilizado e determinar o tipo de cada dado (*int, float, char, string, text, etc*) e seus atributos.

Os campos dos documentos indexados foram especificados de acordo com os valores que há a necessidade de ser recuperado. Cada campo definido no arquivo de configuração do *Solr* é um metadado, os metadados foram determinados nas seções anteriores, sendo eles: Título, Súmula, Introdução, Corpo do Texto, Local e Autor.

Os diversos campos são utilizados para representar os documentos durante a fase de indexação e de pesquisa.

A Figura 12 demonstra as configurações já aplicadas no *Solr*.

```
<field name="id" type="string" indexed="true" stored="true" required="true" multiValued="false" />
<field name="titulo" type="text_general" indexed="true" stored="true" multiValued="true"/>
<field name="assunto" type="text_general" indexed="true" stored="true" multiValued="true"/>
<field name="introducao" type="text_general" indexed="true" stored="true" multiValued="true"/>
<field name="descricao" type="text_general" indexed="true" stored="true" multiValued="true"/>
<field name="data" type="text_general" indexed="true" stored="true" multiValued="true"/>
<field name="autor" type="text_general" indexed="true" stored="true" multiValued="true"/>
```

Figura 12 - Campos (fields) definidos. Fonte: O autor, 2015.

### 4.3.2. FULL TEXT

Por meio da criação de uma cópia das *fields* o *Solr* faz a indexação *full text*, além da indexação por campos. A configuração necessária para que seja realizada a indexação de tipo *full text* está apresentada na Figura 13.

```
<copyField source="titulo" dest="text"/>
<copyField source="assunto" dest="text"/>
<copyField source="introducao" dest="text"/>
<copyField source="descricao" dest="text"/>
<copyField source="data" dest="text"/>
<copyField source="autor" dest="text"/>
```

Figura 13 - Configuração do Full Text. Fonte: O autor, 2015

### 4.3.3. GRAU DE RELEVÂNCIA DOS METADADOS

Qualidade de dados é um conceito que representa a visão, os critérios e as métricas para avaliar, interpretar e melhorar as fontes de informação além de permitir a seleção das dimensões relevantes dentro de um contexto de aplicação (SATTLER 2008 apud SANTOS).

Após definir os graus de relevância dos metadados é necessário configura-los no SOLR. Para isso é necessário alterar o documento *solrconfig.xml*. O grau de relevância será utilizado na busca *full text*, ou seja, quando o usuário não selecionar nenhuma metadado para sua pesquisa. É preciso passar os campos e seus valores como parâmetro. A configuração está demonstrada na Figura 14.



```

<!-- Query settings -->
<str name="defType">edismax</str>
<str name="qf">
  titulo^3.0 assunto^2.5 introducao^1.0 descricao^1.0 id^3.0 data^1.0 autor^1.0
</str>
<str name="mm">100%</str>
<str name="q.alt">*:*</str>
<str name="rows">3</str>
<str name="fl">*,score</str>

<str name="mlt.qf">
  titulo^3.0 assunto^2.5 introducao^1.0 descricao^2.5 id^3.0 data^1.0 autor^1.0
</str>
<str name="mlt.fl">titulo,assunto,introducao,descricao,id,data,autor</str>
<int name="mlt.count">3</int>

```

Figura 14 - Configuração da Relevância dos campos. Fonte: O autor, 2015

#### 4.3.4. DETECÇÃO DO IDIOMA AUTOMATICAMENTE

Para que o idioma dos campos (*fields*) sejam detectados automaticamente há a necessidade de alterar o documento de configuração *solrconfig*, e deve ficar como apresentada na Figura 15.

```

<processor class="org.apache.solr.update.processor.LangDetectLanguageIdentifierUpdateProcessorFactory">
<lst name="defaults">
  <str name="langid.fl">titulo,assunto,introducao,descricao,data,autor</str>
  <str name="langid.langField">language_s</str>
</lst>
</processor>

```

Figura 15 - Detecção automática de Idioma. Fonte: O autor, 2015.

É necessário especificar para ferramenta quais campos que ela deve fazer o reconhecimento de idioma. Para certificar-se que o idioma correto foi identificado basta verificar qual linguagem o *Solr* selecionou para as configurações da busca *Highlighter*. Por meio da Figura 16, pode-se observar que a linguagem foi identificada corretamente.

```

1633 <boundaryScanner name="breakIterator"
1634         class="solr.highlight.BreakIteratorBoundaryScanner">
1635     <lst name="defaults">
1636         <!-- type should be one of CHARACTER, WORD(default), LINE and SENTENCE -->
1637         <str name="hl.bs.type">WORD</str>
1638         <!-- language and country are used when constructing Locale object. -->
1639         <!-- And the Locale object will be used when getting instance of BreakIterator -->
1640         <str name="hl.bs.language">pt</str>
1641         <str name="hl.bs.country">BR</str>
1642     </lst>
1643 </boundaryScanner>
1644 </highlighting>
1645 </searchComponent>

```

Figura 16 - Certificação de Idioma. Fonte: O autor, 2015.

#### 4.3.5. BUSCA HIGHLIGHTER

A busca *Highlighter* tem extrema importância no projeto, pois é através dela que o usuário irá identificar, por meio de um destaque, onde estão localizados o termo que foi pesquisado.

Para habilitar a busca *highlighter* no *Solr* foi usada a configuração apresentada na Figura 17.

```

<!-- Highlighting defaults -->
<str name="hl">on</str>
<str name="hl.fl">titulo, assunto, introducao, descricao, data, autor</str>
<str name="hl.preserveMulti">>true</str>
<str name="hl.encoder">html</str>
<str name="hl.simple.pre">&lt;b&gt;</str>
<str name="hl.simple.post">&lt;/b&gt;</str>

```

Figura 17 - Configuração da *HighLighter*. Fonte: O autor, 2015.

Como se pode notar na Figura 17, para habilitar a *Highlither* basta passar como parâmetro os campos (*fields*) onde deseja utiliza-la. E o restante das configurações são para marcação, ou seja, destaque do termo pesquisado.

#### 4.3.6. STOPWORDS

Nem todas as palavras presentes nos documentos relevantes para recuperá-los. Essas palavras irrelevantes que não possuem valor semântico e ocorrem com frequência significativa, como por exemplo, artigos, preposições e conjunções. Sua remoção dos índices gerados em SRIs visam diminuir o tamanho do índice, tornar mais rápidas as consultas por frases que envolvam *StopWords* e melhorar o ranking dos resultados.

As *StopWords* serão utilizadas na busca e na indexação, para que seja descartado os termos irrelevantes, tornando a busca mais rápida.

A configuração da *StopsWords* está contida dentro das definições do *TypeField*, ou seja, tipo de dado que foi escolhido pelo administrador no momento da criação das *fields*. No presente trabalho o tipo escolhido foi *text\_general*. É necessário alterar as *stopwords* para português pois vem em inglês como padrão. A configuração aplicada para as *StopWords* está demonstrada na Figura 18.

```
424 <fieldType name="text_general" class="solr.TextField" positionIncrementGap="100">
425   <analyzer type="index">
426     <tokenizer class="solr.StandardTokenizerFactory"/>
427     <filter class="solr.StopFilterFactory" ignoreCase="true" words="lang/stopwords_pt.txt" format="snowball" />
428     <filter class="solr.PortugueseLightStemFilterFactory"/>
429     <filter class="solr.SynonymFilterFactory" synonyms="index_synonyms.txt" ignoreCase="true" expand="false"/>
430     ->
431     <filter class="solr.LowerCaseFilterFactory"/>
432   </analyzer>
433   <analyzer type="query">
434     <tokenizer class="solr.StandardTokenizerFactory"/>
435     <filter class="solr.StopFilterFactory" ignoreCase="true" words="lang/stopwords_pt.txt" format="snowball" />
436     <filter class="solr.SynonymFilterFactory" synonyms="synonyms.txt" ignoreCase="true" expand="true"/>
437     <filter class="solr.LowerCaseFilterFactory"/>
438     <filter class="solr.PortugueseLightStemFilterFactory"/>
439   </analyzer>
440 </fieldType>
```

Figura 18 - Configuração das *Stop Words*. Fonte: O autor, 2015.

#### 4.4. INDEXAÇÃO AUTOMÁTICA

Para que a indexação automática ocorra de forma eficaz, foi realizada uma pesquisa eficiente a fim de encontrar as áreas ideais para indexação do documento. Por isso a análise conceitual é uma das etapas mais importantes da indexação.

Após implementada as mudanças nas configurações do *Solr*, de acordo com a necessidade do projeto, foi desenvolvida uma interface de indexação automática. O indexador de documentos necessita de um método para organizar as informações para que a busca seja mais rápida.

O processo de organização dos blocos em seus respectivos campos é automático, havendo a necessidade de interação do usuário apenas para selecionar a pasta que estão os blocos que serão indexados, a partir daí o *Solr* faz a construção do índice que poderá ser consultado pelo usuário posteriormente. Na Figura 19 está exposta a interface de indexação.

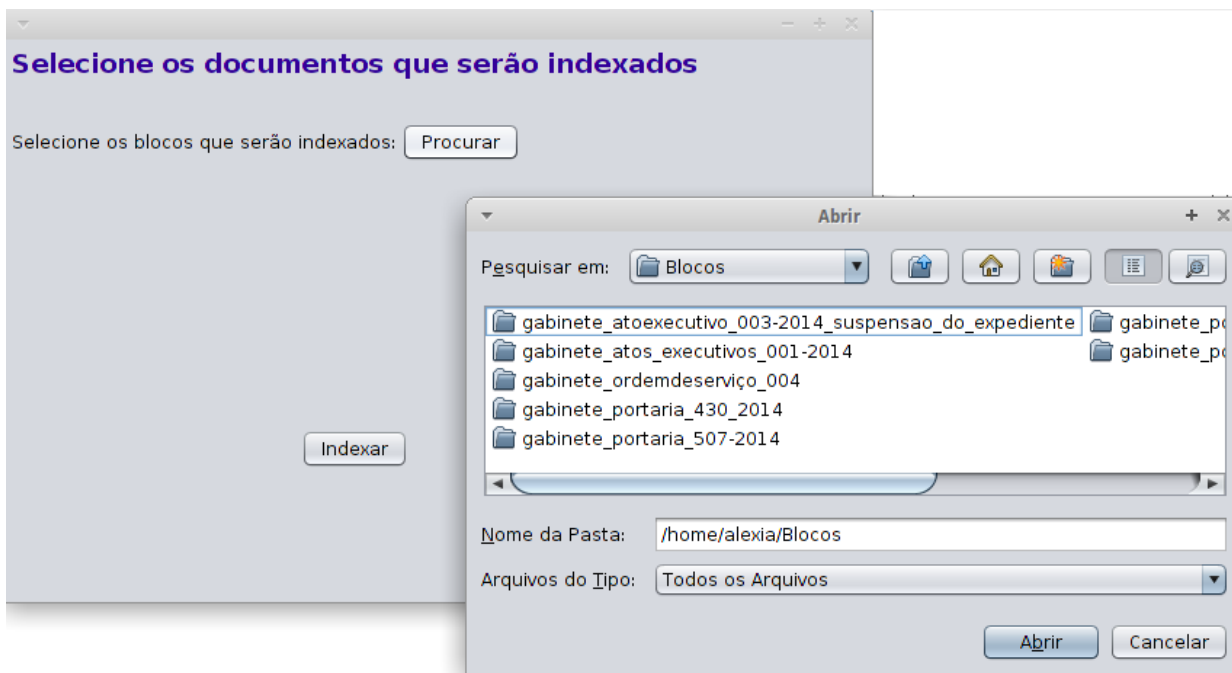


Figura 19 - Interface de Indexação. Fonte: O autor, 2015.

#### 4.4.1. ÍNDICE GERADO

O processo de indexação gera um documento *XML* que funciona como índice, onde são especificados os campos e os conteúdos dos campos que foram extraídos dos documentos. O arquivo *XML* de um documento do tipo Portaria possui a estrutura apresentada na Figura 20.

```
O documento XML não está associado a estilos. A estrutura do documento está representada abaixo.
- <add>
- <doc>
  <field name="id">SP7890</field>
  <field name="titulo">PORTARIA Nw 430/2014</field>
  - <field name="sumula">
    SUMULA: Concede promoção de classe ao Servidor Docente Celso Davi Aoki, de Professor Assistente para Professor Adjunto A.
  </field>
  - <field name="introducao">
    A Reitora da Universidade Estadual do Norte do Paraná - UENP, Profa. Fátima Aparecida da Cruz Padoan, nomeada pelo decreto nº 11435, de 26 de junho de 2014, do Governo do Estado do Paraná, no uso de suas atribuições legais e regimentais, considerando protocolo 13001 -2600/2014
  </field>
  - <field name="corpodotexto">
    RESOLVE Art. 1°. Conceder, a partir de 03 de outubro de 2014, promoção de classe ao Servidor Docente Celso Davi Aoki, RG nº 10.914.138-0-SSP/PR, lotado Campus de Cornélio Procopio, Professor Assistente para Professor Adjunto A, por ter concluído o curso de pós-graduação, em nível de doutorado, em Sociologia, na Universidade Federal do Paraná. Art. 2º. Esta portaria entra em vigor na data de sua publicação, revogadas as disposições em contrário. PUBLIQUE-SE.
  </field>
  - <field name="local">
    Gabinete da Reitora da UENP em Jacarezinho, 20 de outubro de 2014.
  </field>
  <field name="autor">Fatima Padoan</field>
</doc>
</add>
```

Figura 20 – Índice - Documento XML. Fonte: O autor, 2015.

A partir do índice gerado o usuário pode realizar consulta nos documentos que estão indexados no acervo.

## 4.5. CONSULTA

A qualidade do processo de recuperação de documentos está ligada ao sistema de indexação utilizado. Para que a busca do sistema seja eficaz, este deve ser capaz de combinar pesquisa por campos com pesquisa em texto. Isso depende do modo como o documento foi indexado na etapa anterior. Como citado na seção anterior foi utilizado para indexação o método *full text* e por campos (*fields*).

A interface de consulta deve interagir com os usuários de uma forma simples e intuitiva, possibilitando a localização de qualquer documento do acervo com base apenas no que o usuário sabe no momento sobre o documento.

O usuário deverá estar consciente de como o documento foi categorizado e quais campos indexados foram associados aos documentos.

A interface desenvolvida oferece ao usuário a possibilidade de escolher o campo que deseja buscar, podendo escolher mais de um campo. Pesquisas por campos permite que o usuário passe por milhões de registros instantaneamente a procura da informação que lhe interessa. A interface desenvolvida está apresentada na Figura 21.

**Busca baseada em MetaDados**

Selecione os MetaDados a serem pesquisados:

Título  Súmula  Introdução

Corpo do Texto  Local  Autor

recesso

Busca realizada em: 0.028 s

[1] Documentos encontrados | 1 Termos encontrados |

Título: [ATO EXECUTIVO No 003/2014 - GR/UENP]  
Sumula: [Sumula: Estabelece período de **recesso** administrativo na Universidade Estadual do Norte do Paraná - UENP.]  
Introdução: [A Reitora da Universidade Estadual do Norte do Paraná - UENP, Profa. Fátima Aparecida da Cruz Padoan, nomeada pelo decreto no 11435, de 26 de junho de 2014, do Governo do Estado do Paraná, no uso de suas atribuições legais e regimentais.]  
Corpo do Texto: [RESOLVE  
Art. 1o. Estabelecer que no período de 22 de dezembro de 2014 a 02 de janeiro de 2015, não haverá atividades administrativas na Universidade Estadual do Norte do Paraná - UENP, exceto nos setores que, a critério de seus responsáveis, não possam sofrer solução de continuidade.  
Art. 2o.  
Este Ato Executivo entra em vigor na data de sua publicação, revogadas as disposições em contrário.]  
Local: [Jacarezinho, 16 de dezembro de 2014.]  
Autor: [Fátima Aparecida da Cruz Padoan]

Figura 21 - Interface de Busca. Fonte: O autor, 2015.

A busca foi implementada de acordo com as necessidades do usuário por meio de uma interface que se comunica com a ferramenta *Apache SOLR*, o qual, se integra com a biblioteca *Apache Lucene* e possibilita uma busca textual com diversas funcionalidades, sendo elas: termos destacados, resultados da pesquisa ranqueados (grau de relevância), dicionários de sinônimos, correções ortográficas (*spell schecker*).

## 5. CONSIDERAÇÕES FINAIS

Neste capítulo, serão apresentadas as reflexões sobre o desenvolvimento e resultados obtidos neste trabalho.

### 5.1. ESTUDO DE CASO

Atualmente há uma elevada produção e busca por informação, se fazendo necessário organizar os dados utilizando tecnologias, para disponibilizá-las ao usuário de maneira satisfatória. A partir disto, existe a necessidade de se criar e melhorar técnicas para que as buscas se tornem mais eficazes. Observa-se que a indexação é a conexão entre os documentos existentes no sistema e quais documentos são recuperados, de acordo com sua necessidade do usuário.

Visando agilizar e automatizar este processo foi desenvolvido uma solução para indexação automática de documentos oficiais da UENP baseando-se em seus *layouts*. O reconhecimento de *layout* melhora a agilidade da pesquisa pois a partir dele pode-se dividir as áreas dos documentos, já que os *layouts* dos documentos estudados nesse trabalho seguem um padrão, ou seja, são similares em sua estrutura e particularidades, e a partir disto encontrar seu conjunto de metadados relevantes e como já visto anteriormente, a pesquisa se torna mais rápida quando se utiliza metadados.

A consulta em documentos baseado em metadados torna a pesquisa mais ágil pois o usuário “diz” ao sistema em qual metadado está a palavra que ele quer encontrar, evitando que o sistema passe por milhões de registros a procura da informação que lhe interessa.

No acervo de documentos do protótipo desenvolvido contém 60 documentos eletrônicos da UENP, os documentos indexados são do tipo: Portaria, Ato Executivo e Ordem de Serviço, que estão exemplificados no Apêndice A

### 5.2. TESTES

Foram feitos testes para comparar o tempo de pesquisa e a integridade dos resultados obtidos.

Por meio da interface quando se seleciona alguma *checkbox* a pesquisa é feita por meio de metadados. Quando não se seleciona nenhuma *checkbox* a pesquisa feita é do tipo *full text*. Para testar foi procurado o mesmo termo nos dois diferentes tipos de pesquisa.

Na Figura 22 está demonstrado o teste utilizando metadados. Foi selecionado para que fosse buscado no metadado Assunto. A busca demorou 1 *ms* e foram encontrados dois termos sendo que, os dois termos estão presentes no metadado Assunto.

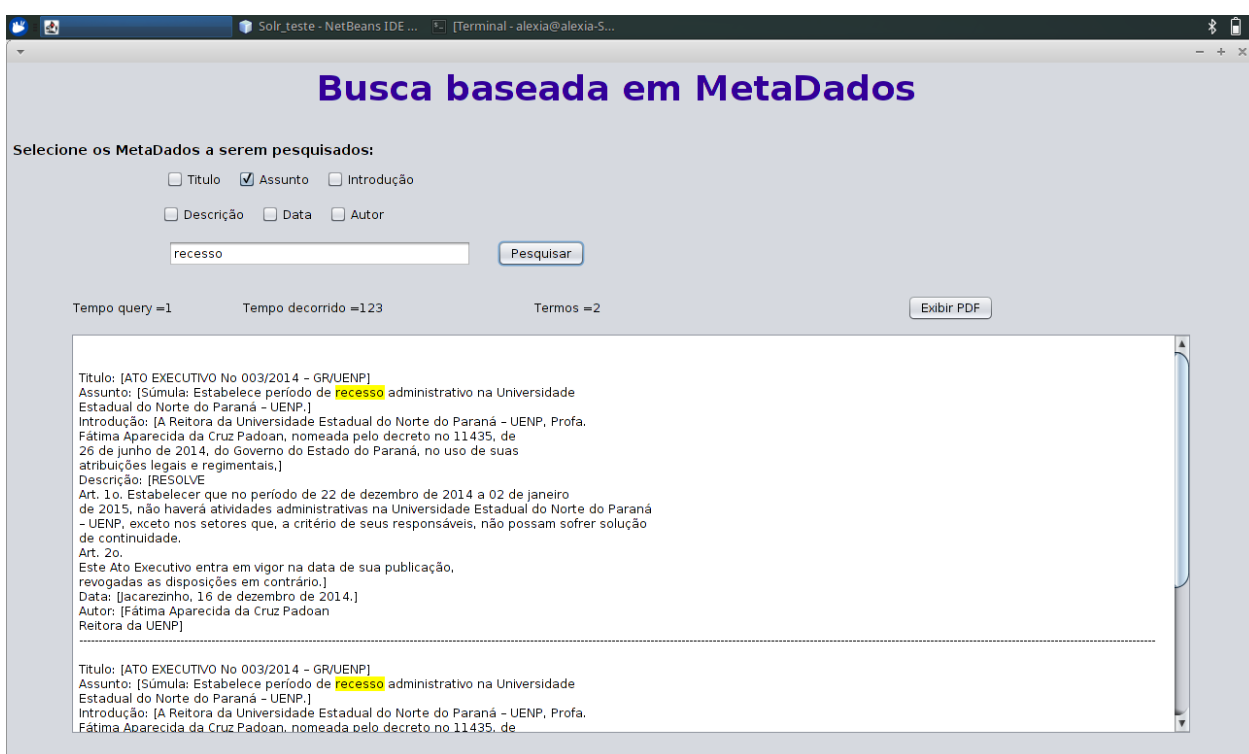
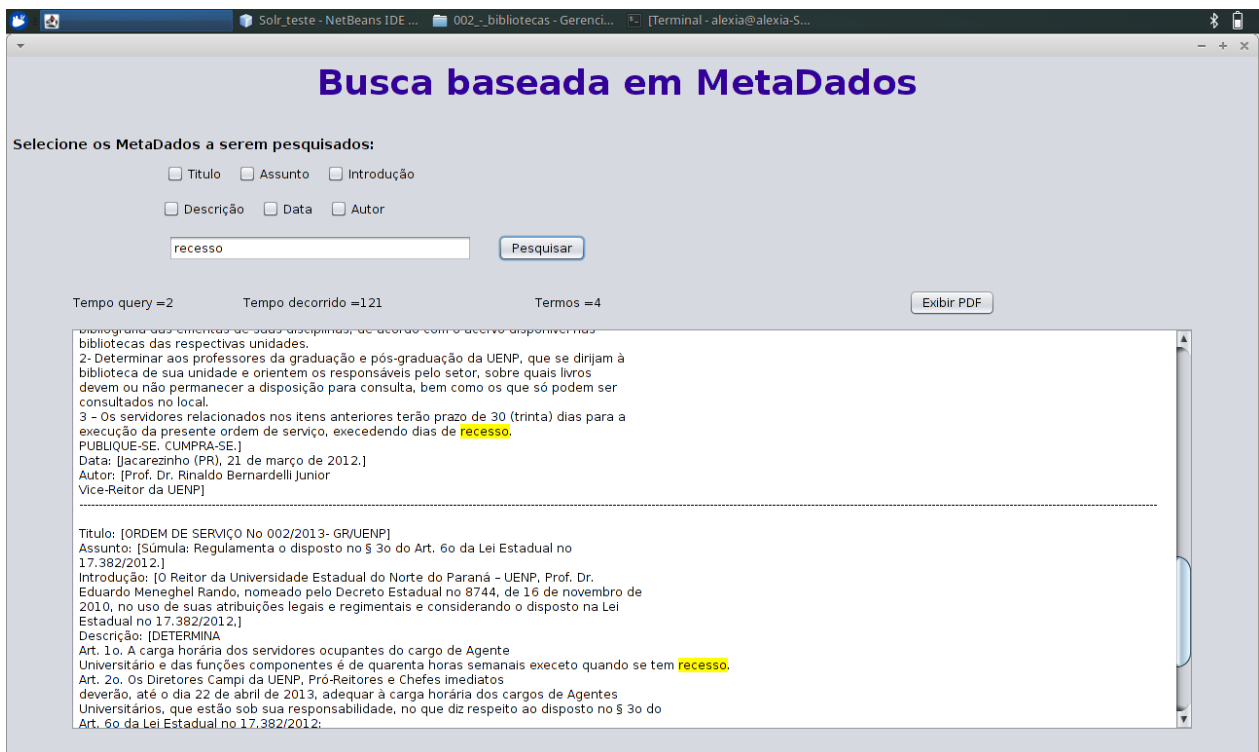


Figura 22 - Teste de Tempo e Integridade, utilizando Metadados. Fonte: O autor, 2015

O teste utilizando *full text*, ou seja, quando não se seleciona nenhuma *checkbox*. O teste demorou 2 *ms* e encontrou quatro termos e está demonstrado abaixo pela Figura 23.





**Figura 23 - Teste de Tempo e Integridade, utilizando *Full Text*. Fonte: O autor, 2015**

Através das Figuras expostas acima pode-se observar que a busca utilizando metadados é mais rápida e mais relevante para o usuário, pois ela retorna apenas dois termos e os dois termos estão no metadado que foi pesquisado.

Já a busca usando *full text* demora 1 ms a mais que a busca por metadado e retorna para o usuário quatro termos, estando, dois deles no Assunto; sendo estes os termos que foram encontrados também na pesquisa por metadados anteriormente; e dois termos na descrição.

## 6. CONCLUSÃO

Neste trabalho foi proposto e desenvolvido uma ferramenta que indexe documentos oficiais da UENP e que posteriormente os recupere baseando-se em seus metadados.

A ferramenta tem como proposta o desenvolvimento de princípios que melhorem a produtividade no ambiente de gestão da UENP, proporcionando melhoras, tais como: economia de espaço físico; a possibilidade de recuperação rápida por meio de categorias e por várias pessoas ao mesmo tempo; e ainda a possibilidade de se fazer cópias de segurança.

Com o resultado deste trabalho confirmamos que seus objetos foram cumpridos por meio da utilização do *Apache Solr* juntamente com o conjunto de metadados.

A combinação do *Apache Solr* e o conjunto de metadados escolhidos resultaram em uma interface de indexação e consulta, sendo esta, uma proposta de aplicação da RI, que facilita ao usuário o acesso as informações contidas nos documentos.

As principais contribuições dessa pesquisa são a extração e classificação automática de metadados dos documentos oficiais da UENP. A extração se faz partir do modo como foi organizada e desenvolvida a indexação. Ambos fatores, resultam diretamente no resultado e no tempo de busca por um termo.

A ferramenta desenvolvida se mostra uma boa solução para os GEDs. Tendo em vista, o GED como um todo, a ferramenta criada auxiliaria nas tarefas de indexação automática e na capacidade de recuperar informações baseando-se nos metadados dos documentos.

### 6.1. TRABALHOS FUTUROS

Como trabalho futuro pretende-se:

- Realizar a integração da Indexação e Busca em um sistema GED;

## Referências Bibliográficas

ALMEIDA, Luis Fernando de; LEMES, Jony Antonio . **Ferramenta para recuperação de informação baseado em arquivos de índices**. 2010. 11 f. Monografia (Especialização) - Curso de Tecnologia da Informação, Ufmg, Belo Horizonte, 2010. Disponível em: <<http://www.lbd.dcc.ufmg.br/colecoes/sbsi/2010/0014.pdf>>. Acesso em: 21 jun. 2014.

ALVES, Rachel Cristina Vesú. **WEB SEMÂNTICA: uma análise focada no uso de metadados**. 2005. 182 f. Monografia (Especialização) - Curso de Ciência da Informação, Unesp, Marília, 2005. Disponível em: <[http://www.marilia.unesp.br/Home/Pos-Graduacao/CienciadaInformacao/Dissertacoes/alves\\_rcv\\_me\\_mar.pdf](http://www.marilia.unesp.br/Home/Pos-Graduacao/CienciadaInformacao/Dissertacoes/alves_rcv_me_mar.pdf)>. Acesso em: 17 jun. 2014.

ARAÚJO JUNIOR, R. H. “**Precisão no processo de busca e recuperação da informação**”. Brasília: Thesaurus, 2007

BALDAM, R., VALLE, R., CAVALCANTI, M. **GED: Gerenciamento Eletrônico de Documentos**. São Paulo: Érica, 2002.

BAEZA-YATES, Ricardo; RIBEIRO-NETO, Berthier. **Modern Information Retrieval**. New York: ACM Press, 1999. 511p. Disponível em: <[ftp://mail.im.tku.edu.tw/seke/slide/baeza-yates/chap10\\_user\\_interfaces\\_and\\_visualization-modern\\_ir.pdf](ftp://mail.im.tku.edu.tw/seke/slide/baeza-yates/chap10_user_interfaces_and_visualization-modern_ir.pdf)> Acesso em: 10 out. 2014

CARDOSO, Olinda Nogueira Paes. **Recuperação de Informação**. INFOCOMP: Journal of Computer Science, Lavras, MG, vol. 2, n. 1, nov. 2000. Disponível em:<<http://professores.dcc.ufla.br/ojs/index.php/INFOCOMP/article/view/46>>. Acesso em: 16 jan. 2015.

CHESTER, Bernard. **Archiving Eletronic Files**. AIIM E - Doc Magazine. 2006. v.20, n.3, p. 63.

DataPark Search Enigne – Key Features (2015). Disponível em:<<http://www.dataparksearch.org/>> Acesso em: 16 jul. 2015

DOS SANTOS, Veronica. **Uma arquitetura suportada por busca semântica para recuperação de fontes de informação em repositórios de metadados**. Master’s thesis, Programa de Pós-Graduação em Informática, Universidade Federal do Estado do Rio de Janeiro, 2011. Disponível em: <[http://www2.uniriotec.br/ppgi/banco-de-dissertacoes-ppgi-unirio/ano-2011/uma-arquitetura-suportada-por-busca-semantica-para-por-busca-semantica-pararecuperacao-de-fontes-de-informacao-em-repositorios-demetadados/at\\_download/file](http://www2.uniriotec.br/ppgi/banco-de-dissertacoes-ppgi-unirio/ano-2011/uma-arquitetura-suportada-por-busca-semantica-para-por-busca-semantica-pararecuperacao-de-fontes-de-informacao-em-repositorios-demetadados/at_download/file)> Acesso em: 23 ago. 2014.

Elastic Search (2015). **Pagina do projeto**. Disponível em: <<https://www.elastic.co/>> Acesso em: 13 jul. 2015.

FANNING, Betsy. **Data,Data,Everywhere Data: Metadata Standards**. AIIM E-Doc Magazine. 2006. v.20, n.3, p. 76.

FERNEDA, Edberto. **Análise sobre a contribuição da Ciência da computação para a ciência da informação**. 2003. 147 f. Tese (Doutorado) - Curso de Ciências da Comunicação, Universidade de São Paulo, São Paulo, 2003. Disponível em: <[www.teses.usp.br/teses/disponiveis/27/27143/tde-15032004.../Tese.pdf](http://www.teses.usp.br/teses/disponiveis/27/27143/tde-15032004.../Tese.pdf)>. Acesso em: 20 abr. 2015.

Global Information Locator Service (GILS) - ***Making it easier to find all the information***. Disponível em: <<http://www.gils.net/>>. Acesso em: 09 abr. 2015.

GOSPODNETIC, Otis; HATCHER, Erik; MCCANDLESS, Michael. **Lucene in Action, 2.ed**, Greenwich: Manning Publications, 2009

HtDig - **Features and System requirements (2015)**. Disponível em: <<http://www.htdig.org/>> Acesso em: 15 jul. 2015.

LAGOZE, C. **The Warwick Framework – A container Architecture for Aggregating Stes os Metadata**. 1996.

LANCASTER, F.W. **Indexação e resumos**. Brasília: Briquet de Lemos/Livros, 1993.

LEMES, Jony Antonio; ALMEIDA, Luis Fernando de; FILHO, Eurico Arruda, **APLICAÇÃO DE ARQUIVOS DE ÍNDICES PARA RECUPERAÇÃO DE INFORMAÇÃO NO CADERNO DO JUDICIÁRIO DO TRIBUNAL REGIONAL DO TRABALHO**. Cruzeiro: Revista Ciências Exatas – Universidade de Taubaté (unitau) – Brasil, v. 17, n. 2, 2011. Mensal. Disponível em: <<http://periodicos.unitau.br/ojs-2.2/index.php/exatas/article/viewFile/1335/913>>. Acesso em: 12 jul. 2015.

LOPES, Ilza Leite. **Estratégia de busca na recuperação da informação: revisão da literatura**. 2002. 10 f. Tese (Doutorado) - Curso de Informática, Unb, Brasília, 2002. Disponível em: <<http://www.scielo.br/pdf/ci/v31n2/12909.pdf>>. Acesso em: 19 jun. 2014.

LUCCA, Giana; CHARÃO, Andrea Schwertner; STEIN, Benhur de Oliveira. **METADADOS PARA UM SISTEMA DE GESTÃO ELETRÔNICA DE DOCUMENTOS ARQUIVÍSTICOS**. 2009. 15 f. Artigo- Curso de Computação, UFSM, Santa Maria, 2009. Disponível em: <[http://www.brapci.inf.br/\\_repositorio/2009/11/pdf\\_38361a2090\\_0006734.pdf](http://www.brapci.inf.br/_repositorio/2009/11/pdf_38361a2090_0006734.pdf)>. Acesso em: 17 jun. 2014.

Lucene Project (2014). **Página do projeto Lucene**. Disponível em:<<http://lucene.apache.org>> Acesso em: 20 set. 2014.

M

MARCONDES, Carlos Henrique; SAYÃO, Luis Fernando . **Integração e interoperabilidade: a proposta da Biblioteca Digital Brasileira**- Curso: Ciência da Informação , Brasília, v. 30, n. 3, p. 24-33, 2001.

MARQUES, Eduardo Zanoni. **Aplicação da Busca por Informação via Texto em um Sistema de Recuperação de Imagens por Conteúdo**. 2006. 69 f. TCC (Graduação) - Curso de Ciência da Computação, Uel, Londrina, 2006.

Mno Go Search (2015). **Página da ferramenta**. Disponível em: <<http://www.mnogosearch.org/>> Acesso em: 14 jul. 2015.

MODESTO, Fernando. **Metadados**. 2. ed. São Paulo: Usp, 2005. 34 p. Disponível em: <<http://www.eca.usp.br/prof/fmodesto/textos/livrometadados.pdf>>. Acesso em: 13 jun. 2014.

NISO. **Understanding metadata**. Disponível em: <<http://www.niso.org/publications/press/UnderstandingMetadata.pdf>> Acesso em: 21 jul. 2015, 2015.

NOGUEIRA, Adriana Hypólito; ROSETTO, Marcia. **APLICAÇÃO DE ELEMENTOS METADADOS DUBLIN CORE PARA DESCRIÇÃO DE DADOS BIBLIOGRÁFICOS ONLINE DA BIBLIOTECA DIGITAL DE TESES DA USP**. 2009. 13 f. Monografia (Especialização) - Curso de Sistemas Integrado de Bibliotecas, Departamento de Técnico, Usp, São Paulo, 2009. Disponível em: <[http://www.liber.ufpe.br/tg/modules/tg/docs/aplicacao de metadados.pdf](http://www.liber.ufpe.br/tg/modules/tg/docs/aplicacao%20de%20metadados.pdf)>. Acesso em: 17 jun. 2014.

OLIVEIRA, Luiz Henrique Gonçalves de. **Extração de Metadados utilizando uma ontologia de domínio**. 2009. 67 f. Dissertação (Mestrado) - Curso de Ciência da Computação, Ufrs, Porto Alegre, 2009. Disponível em: <<http://www.lume.ufrgs.br/bitstream/handle/10183/22814/000740674.pdf?sequence=1>>. Acesso em: 15 jun. 2014.

OLIVEIRA, Ricardo. **RECONHECIMENTO AUTOMÁTICO DE BLOCOS PARA AUXILIAR A INDEXAÇÃO EM SOLUÇÕES GED**. 2014. 45 f. TCC (Graduação) - Curso de Sistemas de Informação, Centro de Ciências Tecnológicas, Universidade Estadual do Norte do Paraná, Bandeirantes, 2014.

PEREIRA, A. M. ; RIBEIRO JÚNIOR, D. I. ; NEVES, G. L. C. **Metadados para a descrição de recursos da Internet: as novas tecnologias desenvolvidas para o padrão Dublin Core e sua utilização**. Revista ACB: Biblioteconomia em Santa Catarina. v. 10, n. 1, p. 241-249, jan./dez., 2005.

REIS, Marco. **Como indexar os arquivos do seu computador com Lucene: As funcionalidades de um buscador: indexação e busca**. 2013. Disponível em: <<http://imasters.com.br/front-end/como-indexar-os-arquivos-do-seu-computador-com-lucene/>>. Acesso em: 20 set. 2014.

ROBREDO, Jaime. **Indexação e Recuperação da Informação na Era das Publicações Virtuais**. 1999. 15 f. Monografia (Especialização) - Curso de Ciência da Computação, Unb, Brasília, 1999.

ROCHA, Igor Pessoa. **Indexação e Referências Textuais: Um Estudo de Caso com Implementação de Ferramenta para o Programa Nacional de Atividades Espaciais**. 2014. 63 f. TCC (Graduação) - Curso de Computação, Instituto de Ciências Exatas Departamento de Ciência da Computação, Universidade de Brasília, Brasília, 2014. Disponível em: <[http://bdm.unb.br/bitstream/10483/7731/1/2013\\_IgorPessoaRocha.pdf](http://bdm.unb.br/bitstream/10483/7731/1/2013_IgorPessoaRocha.pdf)>. Acesso em: 1 ago. 2015.

SANTOS, V. B. **Gestão de documentos eletrônicos: uma visão arquivística**. Brasília: ABARQ, 2002.

SMILEY, D ; PUGH, E. **Solr 1.4 Enterprise Search Server**. Birmingham, Mumbai: Packt Publishing, 2009. 336 p.

SOUZA, Renato Rocha. **Sistemas de recuperação de informações e mecanismos de busca na web: panorama atual e tendências**. 2006. 8 f. Dissertação (Mestrado) - Curso de Engenheiro de Sistemas, UFMG, Belo Horizonte, 2006. Disponível em: <[http://www.scielo.br/scielo.php?script=sci\\_arttext&pid=S1413-99362006000200002](http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1413-99362006000200002)>. Acesso em: 20 jun. 2014.

SOUZA, Terezinha Batista de; CATARINO, Maria Elisabete; SANTOS, Paulo César dos. **Metadados: Catalogando Dados na Internet**. 1997. 13 f. Dissertação (Mestrado) - Curso de Biblioteconomia, Puccamp, Campinas, 1997.

SPRAGUE JR., Ralph H.; D'OLIVEIRA E SILVA, Flávio Luiz. **Gerenciamento Eletrônico de Documentos (GED): Natureza, Princípio e Aplicações**. Cuiabá, 2001. Disponível em: <[https://www.passeidireto.com/arquivo/2154473/ged\\_natureza\\_principios\\_aplicacao](https://www.passeidireto.com/arquivo/2154473/ged_natureza_principios_aplicacao)> Acesso em 15 jun 2015.

TSearch2 (2015) - **Full text extension for PostgreSQL**. Disponível em: <<http://www.sai.msu.su/~megeera/postgres/gist/tsearch/V2/>> Acesso em: 13 jul. 2015.

TEIXEIRA, Marco Alexandre Figueira. **Indexação de Multimédia com base no SOLR**. 2010. 145 f. Dissertação (Mestrado) - Curso de Engenharia Informática, Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa, Lisboa, 2010.

THOMAZ, Kátia P; SANTOS, Vilma Moreira. **Metadados para o gerenciamento eletrônico de documentos de caráter arquivístico – GED/A: estudo comparativo de modelos e formulação de uma proposta preliminar**. Revista DataGramaZero. v. 4, n. 4, ago/2003. Disponível em: <[http://www.dgz.br/ago03/Ind\\_art.htm](http://www.dgz.br/ago03/Ind_art.htm)> Acesso em: 16 jun. 2014.

WEBB, Collin. **The role of preservation and the library of the future**. National Library of Australia, 2000. Disponível em: < [www.nla.gov.au/nla/staffpaper/cwebb9.html](http://www.nla.gov.au/nla/staffpaper/cwebb9.html) > Acesso em: 10 mai. 2014.

WEIBEL, S. **The Dublin core: a simple content description model for electronic resources**. Bulletin of the American Society for Information Science, p.9-11, Oct./Nov. 1997.

Xapian - Xapian Features (2015). Disponível em:< <http://xapian.org/features>> Acesso em: 14 jul. 2015.

## APÊNDICE A – Imagens de diferentes tipos de Documentos da UENP

The image shows a document from the Universidade Estadual do Norte do Paraná (UENP) with several fields highlighted in red boxes to illustrate its metadata structure. At the top center is the UENP logo, which includes the text 'UNIVERSIDADE ESTADUAL DO NORTE DO PARANÁ' and 'ORDETE LUCEM TUA'. Below the logo, the following fields are defined:

- Título:** Ordem de Serviço 006/2014
- Súmula:** Determina a instalação de software livre gratuito na UENP.
- Introdução:** A reitora da Universidade Estadual do Norte do Paraná, no uso de suas atribuições regimentais, com fundamento no art. 57, inciso V, do Regimento da Reitoria, RESOLVE:
- Corpo do Texto:**
  - Art. 1º Determinar a instalação de software livre em todos os computadores dos setores administrativos da UENP.
  - Parágrafo único. Caso haja necessidade de contratação de licença de software proprietário, os Diretores de Campi e Diretores de Órgãos da Reitoria deverão encaminhar justificativa pormenorizada ao Núcleo de Tecnologia da Informação para providências.
  - Art. 2º Fixa o prazo de 20 (vinte) dias úteis para o cumprimento desta ordem de serviço.
- Local:** Jacarezinho, 31 de outubro de 2014.
- Autor:** Fátima Aparecida da Cruz Padoan, Reitora.

At the bottom of the page, there is a footer with contact information: 'Criada pela Lei Estadual 15.300/2006 - Autorizado pelo Decreto Estadual nº 3909/2008 - CNPJ 08.885.100/0001-54 Av. Getúlio Vargas, 850 - CEP 86.400-000 - Jacarezinho/PR - fone/fax 43 3525 3589 - www.uenp.edu.br'

Figura. Estrutura do Documento do tipo Ordem de Serviço e seus respectivos metadados definidos – Fonte: Site UENP





**PORTARIA Nº 206/2015**

**Título**

**Súmula**

**SÚMULA:** Altera o regime de trabalho do Professor Robinson Osipe, de Tempo Integral e Dedicção Exclusiva – TIDE – para T-40.

**Introdução**

O Vice-Reitor da Universidade Estadual do Norte do Paraná – UENP, Fabiano Gonçalves Costa, nomeado pelo decreto nº 12.191, de 17 de setembro de 2014, do Governo do Estado do Paraná, no uso de suas atribuições legais e regimentais, considerando protocolo nº 12001-476/2015

**RESOLVE**

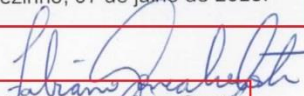
**Art. 1º** Alterar, a partir de 01 de julho de 2015, o regime de trabalho do Professor Robinson Osipe, RG nº 1.518.794-8-SSP/PR, de Tempo Integral e Dedicção Exclusiva – TIDE – para T-40.

**Art. 2º** Esta Portaria entra em vigor na data de sua publicação, revogadas as disposições em contrário.

**PUBLIQUE-SE.**

Gabinete do Vice-Reitor da UENP em,  
Jacarezinho, 07 de julho de 2015.

**Local**

  
**Fabiano Gonçalves Costa**  
Vice-Reitor

**Autor**

**Figura. Estrutura do Documento do tipo Portaria e seus respectivos metadados definidos – Fonte: Site UENP**

## APÊNDICE B – Descrição dos Casos de uso

### Classificar os blocos de texto

**Atores:** Administrador.

**Descrição:** Este caso de uso tem como objetivo classificar os metadados dos blocos de texto conforme seu gênero.

**Pré-condição:** O administrador deve ter efetuado o *login* para executar este caso de uso.

#### **Fluxo Principal:**

1. Por meio de um programa externo os blocos de texto de um arquivo são extraídos;
2. O administrador classifica estes blocos de texto; (O bloco de título, que contém o Título do documento é classificado como: Título)
3. O administrador reunirá todos os blocos que fazem parte do mesmo documento em uma pasta.

#### **Fluxo Alternativo:**

Não existe.

### Extrair os metadados

**Atores:** Administrador.

**Descrição:** Este caso de uso tem como objetivo extrair os dados que são realmente importantes, contidos nos blocos de texto.

**Pré-condição:** O administrador deve ter efetuado o *login* para executar este caso de uso.

#### **Fluxo Principal:**

1. O administrador extrairá quais blocos de texto e quais dados contidos nos blocos de texto são relevantes para uma pesquisa;

#### **Fluxo Alternativo:**

Não existe.

### Indexar dados

**Atores:** Administrador, *SOLR*.

**Descrição:** Este caso de uso tem como objetivo o recebimento de dados contidos nos blocos de texto e permitir o fornecimento dos mesmos posteriormente.

**Pré-condição:** O administrador deve ter efetuado o *login* e o *SOLR* deve estar inicializado para executar este caso de uso.

**Fluxo Principal:**

1. Acessa a página de administração do *SOLR*;
2. Acessa a aba *Documents*;
3. Escolhe o tipo do documento que será indexado;
4. Procura o documento que será indexado em seu computador;
5. O documento é indexado.

**Fluxo Alternativo:**

5.a. O tipo do documento não é compatível com os documentos que na Biblioteca *SOLR*

5.a.1. O sistema informa ao administrador que não é possível indexar tal documento

5.a.2. O sistema retorna ao passo 2 do caso de uso Indexar dados.

**Solicitar uma pesquisa**

**Atores:** Usuário.

**Descrição:** Este caso de uso tem como objetivo fornecer dados ao usuário através de uma pesquisa.

**Pré-condição:** O usuário deve ter efetuado o *login* para executar este caso de uso.

**Fluxo Principal:**

1. Acessa a página inicial do sistema;
2. Escolher o tipo de pesquisa que deseja fazer.
3. Digitar as palavras para serem pesquisadas.

**Fluxo Alternativo:**

3.a. As palavras pesquisadas não são encontradas

3.a.1. O sistema informa ao usuário que as palavras buscadas não existem em nenhum documento

3.a.2. O sistema retorna ao passo 1 do caso de uso Solicitar Pesquisa