



**UNIVERSIDADE FEDERAL DE SANTA CATARINA  
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA  
E GESTÃO DO CONHECIMENTO**

Samuel Fernandes Ribeiro

**Sistema de Conhecimento para Gestão Documental  
no Setor Judiciário: uma aplicação no Tribunal  
Regional Eleitoral de Santa Catarina**

Dissertação de Mestrado

Florianópolis

2010

Samuel Fernandes Ribeiro

**Sistema de Conhecimento para Gestão Documental  
no Setor Judiciário: uma aplicação no Tribunal  
Regional Eleitoral de Santa Catarina**

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento da Universidade Federal de Santa Catarina, como requisito parcial para obtenção do grau de Mestre em Engenharia e Gestão do Conhecimento.

Orientador: Prof. Dr. Aran Bey Tcholakien Morales

Florianópolis

2010

# **Sistema de Conhecimento para Gestão Documental no Setor Judiciário: uma aplicação no Tribunal Regional Eleitoral de Santa Catarina**

**Samuel Fernandes Ribeiro**

Esta dissertação foi julgada adequada para a obtenção do título de MESTRE em ENGENHARIA E GESTÃO DO CONHECIMENTO e aprovada pelo Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento, em abril de 2010.

## **BANCA EXAMINADORA**

---

Prof. Aran Bey Tcholakian Morales, Dr  
Orientador

---

Prof. Roberto Carlos dos Santos Pacheco, Dr  
EGC/UFSC

---

Prof. José Leomar Todesco, Dr  
EGC/UFSC

---

Prof. Alexandre Leopoldo Gonçalves, Dr  
EGC/UFSC

## DEDICATÓRIA

**À minha esposa,**  
Gizeli, pelo apoio sempre presente e  
compreensão nas minhas ausências.

**Às minhas filhas,**  
Camile e Larissa, vocês dão uma  
razão especial à minha vida.

## **AGRADECIMENTOS**

Agradeço primeiramente ao Deus Criador e Mantenedor de todas as coisas, a verdadeira fonte do conhecimento.

Agradeço aos meus pais, Paulo e Lindomar, por seu exemplo e dedicação que contribuíram diretamente na formação do meu caráter e em minhas conquistas.

Agradeço ao professor Dr. Aran Bey Tcholakien Morales, meu orientador, pela forma eficiente que direcionou o andamento deste trabalho.

Agradeço à direção do Tribunal Regional Eleitoral de Santa Catarina pela oportunidade de desenvolver esta pesquisa não medindo esforços para viabilizar os meios necessários para sua conclusão.

Agradeço aos colegas de trabalho da Coordenadoria de Soluções Corporativas e da Coordenadoria de Gestão da Informação pelo companheirismo do dia a dia e pelo papel fundamental que desempenharam no transcorrer do projeto de desenvolvimento da solução proposta.

## RESUMO

A inclusão tecnológica relacionada à produção e disponibilização de informações eletrônicas vem transformando o cotidiano de instituições públicas e privadas. Sendo que tais documentos constituem uma fração significativa do conhecimento organizacional, percebe-se a relevância quanto ao desenvolvimento e aplicação de tecnologias adequadas para apoiar a gestão deste acervo. Neste sentido, o presente trabalho se propõe a realizar um estudo relacionado à gestão documental no âmbito do Judiciário Eleitoral de Santa Catarina, com foco nas atividades cognitivas deste processo, visando à proposição de um sistema de conhecimento para auxiliar o profissional humano no desempenho de seu papel. A metodologia adotada constituiu-se de pesquisa observacional sobre a situação atual das instituições judiciárias quanto ao tema, em conjunto com a pesquisa bibliográfica e descritiva dos assuntos correlatos. Tendo como base os artefatos identificados no levantamento realizado, é apresentado um novo modelo de negócio, onde está inserido o sistema e-Docs, o qual possibilitou a unificação do repositório de documentos, automação do processo de indexação e classificação temática e a disponibilização de busca semântica de documento. Por fim, são descritos os resultados obtidos pela aplicação do modelo proposto no Tribunal Regional Eleitoral de Santa Catarina e as propostas de pesquisa para trabalhos futuros.

**Palavras-Chave:** Gestão Eletrônica de Documentos. Recuperação da Informação. Tesouro. Pesquisa Semântica de Documentos.

## ABSTRACT

The technology improvement related to production and availability of electronic information is transforming public and private institutions daily. Such documents constitute a significant fraction of organizational knowledge. It's can see the relevance for appropriate technologies development and applications to support this collection management. In this sense, this paper proposes a study related to document management within the Judiciary Electoral of Santa Catarina, with a focus on cognitive tasks of process, aimed at proposing a knowledge system to assist the human professional at performance of your paper. The methodology consisted of observational research on the current situation of the judiciary on the subject, together with descriptive literature about related subjects. Based on the artifacts identified in the survey, is present a new business model, which is inserted the e-Docs system, which led to documents repository unification, indexing and subject classification process automating and the availability of document semantic search. Finally, it's describe the results obtained by applying the model proposed in the Electoral Court of Santa Catarina and research proposals for future work.

**Keywords:** Document Management. Information Retrieval. Thesaurus. Document Semantic Search.

# SUMÁRIO

LISTA DE FIGURAS .....	10
LISTA DE TABELAS .....	11
LISTA DE ABREVIACÕES.....	12
1 INTRODUÇÃO .....	13
1.1 Problema de Pesquisa .....	14
1.2 Perguntas da Pesquisa .....	15
1.3 Objetivos.....	16
1.3.1 Objetivo Geral.....	16
1.3.2 Objetivos Específicos .....	16
1.4 Delimitação da Pesquisa .....	16
1.5 Multidisciplinaridade da Proposta.....	17
1.5.1 Alinhamento com o Programa de Engenharia e Gestão do Conhecimento .....	19
1.6 Estrutura do Trabalho.....	19
2 GESTÃO ELETRÔNICA DE DOCUMENTOS.....	21
2.1 Gestão Eletrônica de Documentos ( <i>Lato Sensu</i> ).....	21
2.1.1 Histórico Evolutivo .....	22
2.1.2 Definições e nomenclaturas .....	25
2.1.3 Ciclo de Vida do Documento.....	28
2.1.4 Principais Tecnologias.....	35
2.2 Gestão Eletrônica de Documentos no Judiciário Brasileiro .....	43
3 RECUPERAÇÃO DA INFORMAÇÃO.....	46
3.1 Definições e Conceitos.....	47
3.2 Representação de Documentos.....	49
3.3 Modelos de Recuperação da Informação .....	52
3.3.1 Modelo Booleano.....	53
3.3.2 Modelo Espaço Vetorial.....	53
3.3.3 Modelo Probabilístico .....	54
3.3.4 Comparação entre os Modelos de Recuperação da Informação .....	55
3.4 Técnicas de Classificação .....	56
3.4.1 Métodos Não Hierárquicos.....	56
3.4.2 Métodos Hierárquicos.....	57
3.5 Mineração de Texto.....	58
3.6 Integração com Base de Conhecimento .....	60
3.6.1 Apresentação da Hierarquia de Conceitos no Documento .....	60
3.6.2 Flexibilidade na Apresentação dos Resultados das Consultas .....	61
3.7 Recuperação Semântica de Documentos.....	61
3.7.1 Indexação .....	63
3.7.2 Tesouros.....	65
3.7.3 Ontologias .....	67

4	METODOLOGIA DO TRABALHO .....	70
4.1	Classificação da Pesquisa .....	70
4.2	Metodologia CommonKads .....	71
4.3	Estudo de Caso .....	73
4.3.1	Modelo de Organização .....	74
4.3.2	Modelo de Tarefa .....	79
4.3.3	Modelo de Agente .....	81
4.3.4	Modelo de Organização, Tarefa e Agente .....	81
4.3.5	Considerações Finais do Estudo de Caso .....	82
5	MODELO DE GESTÃO DOCUMENTAL PROPOSTO.....	83
5.1	Modelo Conceitual.....	83
5.1.1	Repositório Único de Documentos.....	84
5.1.2	Serviço de Indexação .....	85
5.1.3	Serviço de Pesquisa Semântica.....	88
5.1.4	Considerações Finais sobre o Modelo Conceitual .....	89
5.2	Arquitetura Tecnológica .....	90
5.2.1	Interfaces do Sistema e-Docs .....	91
5.3	Resultados Obtidos .....	94
6	CONCLUSÃO .....	96
6.1	Considerações Finais.....	96
6.2	Trabalhos Futuros .....	98
	REFERÊNCIAS.....	101
	ANEXOS.....	106

## LISTA DE FIGURAS

Figura 2.1- Elementos constitutivos do documento.....	26
Figura 2.2 - Níveis de arquivamento. ....	34
Figura 2.3 - Principais tecnologias de GED e o ciclo de vida de documentos.....	35
Figura 3.1 – Dendograma representando dez classes de documentos. ....	57
Figura 3.2 – Níveis de Processamento de Texto.....	59
Figura 3.3 – Ontologia ilustrativa para o Direito Eleitoral. ....	69
Figura 4.1 – Pirâmide metodológica do CommonKads .....	72
Figura 4.2 – Modelos propostos pelo CommonKads.....	73
Figura 4.3 – Processo de Gestão de Documentos Jurisprudenciais do TRESA.....	77
Figura 5.1 – Modelo conceitual do Processo de Gestão Documental Proposto.....	84
Figura 5.2 – Visualização parcial do Tesouro da Justiça Eleitoral.....	85
Figura 5.3 – Visualização parcial do Tesouro da Justiça Eleitoral no formato XML ..	86
Figura 5.4 – Tesouro ilustrativo .....	87
Figura 5.5 – Arquitetura tecnológica.....	91
Figura 5.6 – Interface para cadastro de documentos no repositório .....	92
Figura 5.7 – Interface para consulta ao repositório .....	93
Figura 5.8– Relação dos documentos recuperados a partir de uma consulta.....	93

## **LISTA DE TABELAS**

Tabela 3.1 - Matriz de Incidência Básica.....	49
Tabela 3.2 - Representação Índice Invertido.....	51
Tabela 3.3 – Representação Índice Invertido com pesos de relevância. ....	52
Tabela 4.1 – Modelo Organizacional de Problemas e Oportunidades. ....	74
Tabela 4.2 – Modelo Organizacional dos Aspectos Variantes. ....	75
Tabela 4.3 – Modelo Organizacional da Decomposição dos Processos.....	78
Tabela 4.4 – Modelo Organizacional – Insumos de conhecimento. ....	78
Tabela 4.5 – Modelo Organizacional – Lista para decisão sobre viabilidade. ....	79
Tabela 4.6 – Relação parcial de tarefas desenvolvidas no departamento. ....	79
Tabela 4.7 - Modelo de Tarefa - Análise das Tarefas. ....	80
Tabela 4.8 - Modelo de Tarefa - Itens do conhecimento. ....	80
Tabela 4.9 – Modelo de Agentes – Planilha de Agentes.....	81
Tabela 4.10 – Modelo de Organização, Agente e Tarefa – Planilha de verificação. .	82
Tabela 5.1 – Estrutura de Índice Invertido ilustrativa.....	86
Tabela 5.2 – Estrutura de Índice Invertido sem uso do tesouro .....	87
Tabela 5.3 - Estrutura de Índice Invertido com uso do tesouro .....	88

## **LISTA DE ABREVIações**

**ABNT** – Associação Brasileira de Normas Técnicas

**BPM** – *Business Process Management* (Gerenciamento de Processos de Negócio)

**CNJ** – Conselho Nacional de Justiça

**CONARQ** – Conselho Nacional de Arquivos

**DI** – *Document Imaging* (Imagem do Documento)

**DM** – *Document Management* (Gerenciamento de Documentos)

**ECM** – *Enterprise Content Management* (Gerenciamento do Conteúdo Corporativo)

**EDMS** – *Electronic Document Management Systems*

**ERM** – *Electronic Record Management* (Gerenciamento de Registros Eletrônicos)

**ERepM** – *Electronic Report Management* (Gerenciamento de Relatórios Eletrônicos)

**GED** – Gerenciamento Eletrônica de Documentos

**HSM** – *Hierarchical Storage Management*

**ICR** – *Intelligent Character Recognition* (Reconhecimento Inteligente de Caracteres)

**KDT** – *Knowledge Discovery from Text* (Descoberta de Conhecimento a partir de Textos)

**MoReq-Jus** – Modelo de Requisitos para Sistemas Informatizados de Gestão de Processos e Documentos do Judiciário Brasileiro

**NHR** – *Natural Handwriting Recognition* (Reconhecimento de Escrita Manual)

**OCR** – *Optical Character Recognition* (Reconhecimento Óptico de Caracteres)

**PGD-JE** – Programa de Gestão Documental da Justiça Eleitoral

**PRONAME** – Programa Nacional de Gestão Documental e Memória do Poder Judiciário

**RI** – Recuperação da Informação

**SGED** – Sistemas de Gestão Eletrônica de Documentos

**SRI** – Sistemas de Recuperação da Informação

**TI** – Tecnologia da Informação

**TJE** – Tesouro da Justiça Eleitoral

**TRESC** – Tribunal Regional Eleitoral de Santa Catarina

**TSE** – Tribunal Superior Eleitoral

# 1 INTRODUÇÃO

Os avanços tecnológicos contemporâneos têm proporcionado importantes transformações na sociedade. Assim, como desde os primórdios da civilização, onde existiram fatores que influenciaram o desenvolvimento da sociedade, hoje é possível observar o surgimento de um novo fator que se tornou determinante para o sucesso de pessoas e organizações: o conhecimento.

Inicialmente, quando a economia era essencialmente agropecuária, a terra representava um diferencial competitivo. Quanto maior a extensão de terra com boa qualidade produtiva, maior eram as condições de sobrevivência dos proprietários. Deste fato conclui-se uma das razões dos constantes conflitos e guerras existentes naquela época por posse de terra.

Após a transformação desta sociedade agropecuária em industrial, a terra deixou de ser o fator chave para economia que desde então era influenciada principalmente pelas indústrias. Nesta realidade, as instalações físicas das fábricas, seu maquinário produtivo e linhas de montagem tornaram-se o fator dominante para produção.

Na atualidade, onde a tecnologia tem provido meios cada vez mais efetivos para criação e disseminação da informação, de acordo com O'Hare (2002), o conhecimento tem se tornado fator predominante para produção de riquezas, muito mais do que terras e ferramentas industriais o foram em suas respectivas eras. Neste sentido, as economias mais avançadas tecnologicamente no século XXI são preponderantemente baseadas em conhecimento e cerca de 70% a 80% das economias em crescimento também apresentam um cenário favorável à gestão deste bem intangível.

O conhecimento não é um componente exclusivo da atualidade, pois este sempre foi um fator produtivo e conduziu o desenvolvimento econômico e social em todas as épocas. A sociedade primitiva dependia, além de terras, do conhecimento apropriado para o cultivo da terra. Do mesmo modo, na revolução industrial dependeu-se do conhecimento para operacionalização das fábricas e manufaturas. Entretanto, O'Hare (2002), ressalta que os avanços tecnológicos presentes nas organizações têm transformado fundamentalmente o grau de integração entre o conhecimento e o desempenho efetivo de suas atividades.

Neste cenário, as organizações têm empreendido esforços para valer-se deste patrimônio para melhorias constantes na execução de suas atividades cotidianas, independente de sua natureza. Assim sendo, um princípio básico observado é a identificação dos conhecimentos estratégicos para a organização e onde estes são produzidos, armazenados e, principalmente, utilizados.

Davenport e Prusak (1999) declaram que o conhecimento de uma organização pode ser classificado como tácito, ou seja, aquele que está exclusivamente no domínio das pessoas, como a experiência de vida adquirida pelos anos de execução de uma atividade, por exemplo; e explícito, que é formal, sistemático e facilmente compartilhado e transmitido.

Quanto ao desenvolvimento do conhecimento organizacional, Senge (1990) defende a idéia que cada organização deve propiciar condições para que as pessoas possam expandir seu potencial criativo, ressaltando, em contrapartida, a importância equivalente de também assegurar condições favoráveis para a formalização e compartilhamento deste conhecimento entre as pessoas.

### **1.1 Problema de Pesquisa**

Segundo David (2007), a transformação ocorrida pela inclusão tecnológica, no que tange aos meios digitais e virtuais às organizações, está marcando uma nova era do conhecimento e da cultura humana. As informações produzidas cotidianamente transformam-se de domínio privado para público ou global em questão de segundos. Documentos de voz e eletrônicos têm triplicado os canais de comunicação entre as pessoas.

Em contrapartida, Richardson (2007) expõe que o mundo passa por um momento *sui generis*, onde a organização, agora globalizada, detém tecnologia que disponibiliza continuamente um volume descomunal de informações nas mais diversas mídias e formatos; armazenadas em qualquer parte do mundo e podendo ser acessadas a qualquer momento.

Tais informações constituem, de certa forma, o conhecimento explicitado de uma organização e, segundo Kock (2007), 70% a 80% destas estão disponíveis em meios não estruturados, ou seja, na forma textual como documentos eletrônicos, e-mails e comunicações instantâneas (*chats*).

Diante dos fatos supracitados, percebe-se a proximidade existente entre o sucesso na gestão do conhecimento organizacional e a efetividade de um processo

de gestão documental, sejam estas organizações públicas ou privadas. O desenvolvimento de tecnologias para este fim tem sido de grande utilidade para o adequado aproveitamento do conhecimento incluído nos documentos que permeiam as organizações. Richardson (2007) também apresenta que os documentos têm tanta importância hoje como os equipamentos automatizados o tiveram nas décadas de 1970 e 1980.

Contudo, observa-se ainda um grande descompasso entre os processos de criação e armazenamento de documentos com os processos de recuperação e aproveitamento prático de seus conteúdos. Desta forma, mesmo tendo o conhecimento explicitado na forma de documentos textuais, a exploração deste para agregação de valor estratégico à instituição é limitada em parte pela ausência de uma gestão documental adequada às suas demandas.

No âmbito do Judiciário, onde as atividades desenvolvidas são sintetizadas principalmente pela produção de documentos, percebem-se claramente as dificuldades em gerenciar grandes volumes de documentos. Conforme Brasil (2008), o advento da informatização dos Tribunais e a crescente democratização do acesso à Justiça, fizeram que o aumento da produção documental do Poder Judiciário tenha relação inversamente proporcional à capacidade de gerenciamento de seu acervo.

A ausência de gestão documental ou a falta de efetividade desta no Judiciário causam prejuízos aos jurisdicionados e impactam de forma extremamente negativa no acesso ao conhecimento contido nos documentos jurisdicionais, levando a demoras excessivas na realização de procedimentos que dependem desse acesso, BRASIL (2008).

## **1.2 Perguntas da Pesquisa**

Diante do exposto, tem-se o questionamento que alicerçou os trabalhos da presente pesquisa:

Como aprimorar as atividades intensivas em conhecimento desenvolvidas no processo de gestão de documentos no Judiciário com intuito de facilitar o acesso da sociedade ao conhecimento jurisprudencial?

A pergunta fundamental da pesquisa traz implícitas outras questões mais específicas que semelhantemente precisavam ser elucidadas:

Como é gerenciado atualmente o acervo documental, que constitui a base de conhecimento das instituições jurídicas brasileiras?

Quais são as atividades intensivas em conhecimento envolvidas neste processo e como a engenharia do conhecimento pode contribuir para sua execução?

### **1.3 Objetivos**

As perguntas de pesquisa apresentadas direcionam para o estabelecimento de objetivos com o intuito de apresentar respostas satisfatórias aos questionamentos previamente identificados e principalmente identificar a contribuição do trabalho referente ao tema de estudo proposto.

#### **1.3.1 Objetivo Geral**

Propor um modelo de gestão de documentos para o Judiciário que permita sua devida classificação, indexação, armazenamento e, principalmente, que facilite o acesso da sociedade ao conhecimento jurisprudencial por intermédio de um efetivo mecanismo de recuperação dos documentos contidos em seu acervo eletrônico.

#### **1.3.2 Objetivos Específicos**

Os seguintes objetivos têm como foco contribuir para o cumprimento do objetivo geral do projeto de pesquisa:

- a) Descrever as operações e os métodos utilizados no processo de gestão documental na atualidade nas Instituições Judiciárias, relacionando os problemas encontrados e as necessidades não atendidas pelo modelo utilizado;
- b) Identificar, no processo de gestão documental, as atividades intensivas em conhecimento que possam ser automatizadas;
- c) Identificar técnicas da Engenharia do Conhecimento que possam contribuir para uma efetiva gestão de documentos jurisprudenciais; e
- d) Construir um sistema de conhecimento para validar o modelo de gestão documental proposto.

### **1.4 Delimitação da Pesquisa**

Pela diversidade de pesquisas possíveis sobre os temas de gestão documental e gestão do conhecimento e com o intuito de estabelecer limites no campo de pesquisa, foi tomado como objeto de estudo a gestão de documentos no âmbito do Poder Judiciário, especificamente na Justiça Eleitoral Catarinense.

Com relação ao ciclo de vida do processo de gestão eletrônica de documentos, o modelo proposto abrangerá as fases que detém atividades intensivas em conhecimento, ou seja, as fases de classificação, indexação, armazenamento e recuperação dos documentos.

A validação do modelo será realizada por intermédio de um protótipo que será limitado a documentos jurisprudenciais do tipo Acórdãos e Resoluções do Tribunal Regional Eleitoral de Santa Catarina.

### **1.5 Multidisciplinaridade da Proposta**

Em se tratando de Gestão Eletrônica de Documentos (GED), observa-se a tendência de visualizá-la exclusivamente como uma disciplina da Tecnologia da Informação (TI). Entretanto, a realidade tem apresentado que, longe de ser exclusiva de uma área específica do conhecimento, a GED precisa ser vislumbrada como um tema interdisciplinar.

Uma abordagem multidisciplinar, segundo Fazenda (1993), parte do princípio que nenhuma forma do conhecimento é em si mesma conclusiva, porém busca o diálogo com outras áreas para que a partir desta interação seja possível alcançar uma visão mais próxima da realidade.

Conforme Japiassú (1976), à interdisciplinaridade faz-se mister a intercomunicação entre as disciplinas, implicando em modificações em suas respectivas visões de mundo, por meio de diálogo compreensível e coordenado, uma vez que a simples troca de informações entre organizações disciplinares não constitui um método interdisciplinar.

Em conformidade a isso, Jesus (2002) define interdisciplinaridade como a necessidade de se superar a visão fragmentada do conhecimento, produzindo coerência entre os demais fragmentos que estão postos no acervo de conhecimento da humanidade. Ressalta também que a interdisciplinaridade caracteriza-se pela colaboração entre as disciplinas diversas que são enriquecidas pelas trocas provenientes da reciprocidade.

Desta forma, além de todas as técnicas e ferramentas necessárias para a execução dos processos da GED, faz-se necessária a inclusão de conhecimento de outras áreas para o estabelecimento de uma gestão que contemple as reais necessidades das organizações que, no cenário atual em que estão inseridas, são produtoras em potencial de informações documentais.

Nascimento e Almeida (2006) citam a tríade formada constituída por pessoas, processos e tecnologia, como um fator de sucesso para a implantação de projetos, onde cada uma de suas bases desempenha um papel fundamental. A partir desta abordagem, tem-se que a TI está inserida unicamente na base Tecnologia da pirâmide.

Por sua vez, na base *Processos*, é possível observar disciplinas como Administração, Ciência da Informação e Biblioteconomia trazem contribuições importantes nas definições de processos e métodos para a gestão. Tais conhecimentos, outrora utilizados em realidades distintas e longe de serem obsoletos, permanecem essenciais na realidade tecnológica que vivemos.

Na base *Pessoas*, a Pedagogia e a Neurociência são importantes na definição de critérios e metodologias para que o conhecimento esteja disponível adequadamente para as pessoas que dele necessitem. Ainda nesta base, ressalta-se o uso da Psicologia, pois a instituição de novos métodos de trabalho implica em alterações no cotidiano das pessoas, que muitas vezes apresentam resistência à mudança. Neste sentido, a Psicologia detém o conhecimento de como lidar com tais situações e quais abordagens precisam ser tomadas para que as pessoas, essenciais em qualquer projeto, se tornem seus facilitadores ao invés de seus obstáculos.

Outra disciplina que merece destaque na abordagem proposta é o Direito, pois esta disciplina permeia todas as bases da pirâmide. O conhecimento em Direito é requisito necessário nas definições dos processos de gestão documental. Assim sendo, este conhecimento necessita ser transmitido para as ferramentas tecnológicas que dão suporte a tais processos. E, finalmente, em relação às pessoas, o conhecimento em Direito precisa ser compreendido, estruturado e devidamente formalizado para ser disponível efetivamente à sociedade.

Com base nesta visão interdisciplinar que se pretende desenvolver o trabalho de pesquisa, onde cada disciplina envolvida contribua com suas perspectivas peculiares para compor uma proposta mais abrangente e adequada para a realidade onde está inserida.

### **1.5.1 Alinhamento com o Programa de Engenharia e Gestão do Conhecimento**

A área de concentração escolhida para o projeto é a Engenharia do Conhecimento, pois a proposta busca o estudo de disciplinas deste domínio para aplicação prática destas na gestão do conhecimento no campo jurídico.

Da área de concentração selecionada, optou-se pela linha de pesquisa Engenharia do Conhecimento Aplicada a Governo Eletrônico. Isto pelo fato de que o resultado almejado para o projeto estar diretamente voltada à melhoria dos serviços prestados por uma Instituição Judiciária à Sociedade.

Sendo a Justiça Eleitoral Catarinense também uma fonte geradora de conhecimento jurisprudencial, é evidente a necessidade da aplicação da engenharia do conhecimento na gestão destas informações, fato este que conduziu à decisão pelo tema proposto.

### **1.6 Estrutura do Trabalho**

Com o propósito de facilitar o entendimento dos temas abordados nesta pesquisa, o presente documento foi segmentado em 6 capítulos organizados seguindo uma sequência lógica onde os conceitos fundamentais são apresentados preliminarmente ao tema principal e à proposta objeto deste estudo.

O capítulo 1 apresenta essencialmente a motivação deste trabalho, relacionando o problema inicial, os objetivos almejados, bem como a delimitação da pesquisa e seu alinhamento quanto à interdisciplinaridade e a proposta de pesquisa do Programa de Pós-Graduação de Engenharia e Gestão do Conhecimento.

O embasamento teórico foi fracionado nos capítulos 2 e 3. No capítulo 2, é apresentada uma visão geral sobre Gestão Eletrônica de Documentos seguindo de sua evolução histórica até as principais tecnologias utilizadas na atualidade pelas organizações. Este capítulo é concluído tratando deste mesmo tema, porém, em um âmbito mais restrito, ou seja, no Judiciário Brasileiro, relatando como este tema está inserido em suas instituições.

A Recuperação da Informação é o assunto do capítulo 3. Iniciando com a apresentação das definições e principais conceitos, o texto perpassa as diversas técnicas desenvolvidas nesta área do conhecimento que foram utilizadas na proposição da solução tecnológica da pesquisa.

No capítulo 4 será exposto o método utilizado para o desenvolvimento do trabalho, tratando da classificação da pesquisa, apresentando a metodologia CommonKads e sua aplicação prática em uma Instituição Judiciária.

De posse dos conceitos relacionados nos capítulos anteriores, o capítulo 5 aborda especificamente a solução proposta visando o atendimento dos objetivos iniciais. Neste ponto é apresentado o modelo conceitual, descrevendo cada um dos seus elementos e a relação existente entre eles e, na sequência, a arquitetura tecnológica do sistema desenvolvido, detalhando as tecnologias utilizadas e como estas interagem durante seu funcionamento. O capítulo é encerrado com a exposição dos resultados obtidos após a implantação do sistema na Instituição pesquisada.

Finalizando, no capítulo 6 serão descritas as considerações finais quanto aos temas abordados, relacionando as conclusões obtidas no transcorrer do trabalho, bem como algumas reflexões sobre o estágio atual da gestão eletrônica de documentos nas instituições judiciárias e os caminhos a percorrer para ampliação das pesquisas em trabalhos futuros.

## **2 GESTÃO ELETRÔNICA DE DOCUMENTOS**

A fundamentação teórica é a base sobre a qual todo o trabalho de pesquisa deve ser alicerçado. Assim, buscou-se na literatura científica os principais autores e instituições que atualmente são referências quanto aos temas abordados nesta dissertação.

Nesta fase da pesquisa, o objetivo foi permear os domínios desta área do conhecimento com intuito de adquirir a maturidade necessária para o estabelecimento do pensamento crítico quanto ao problema exposto, bem como desenvolver a argumentação que comprove os objetivos previamente definidos.

Este capítulo está dividido em dois tópicos principais. O primeiro diz respeito à Gestão Eletrônica de Documentos *lato sensu*, onde serão apresentadas: a evolução histórica e a importância deste tema nas organizações; seus conceitos e nomenclaturas; ciclo de vida do documento; e as principais tecnologias utilizadas. O segundo tema, complementar ao primeiro, é a apresentação deste mesmo tema, porém dentro do âmbito do Judiciário Brasileiro, relacionando as iniciativas tomadas nesta área e qual visão de curto e médio prazo para o desenvolvimento de soluções, cujo objetivo seja a disseminação do conhecimento jurisprudencial contido em seu acervo à sociedade.

### **2.1 Gestão Eletrônica de Documentos (*Lato Sensu*)**

Na medida em que a tecnologia proporcionou meios que facilitaram a elaboração eletrônica de documentos, observou-se ao longo dos anos nas organizações, um aumento expressivo na produção de documentos, bem como na capacidade de armazenamento destes em suas infraestruturas tecnológicas.

Em consequência, o aumento na produção documental foi maior do que a capacidade das organizações para gerenciar seus acervos. Desta forma, surgiu a necessidade da criação de metodologias que suprissem esta nova demanda de forma que as instituições continuassem atendendo seus objetivos estratégicos de negócio, (O'HARE, 2002).

Neste contexto, surge a gestão eletrônica de documentos que, segundo Sutton (1996), tem por objetivo compartilhar informações e recursos críticos da organização de forma segura e acessível, disponibilizando mecanismos que facilitem a recuperação e o intercâmbio destes entre os usuários para o efetivo desempenho de suas atividades.

### **2.1.1 Histórico Evolutivo**

Previamente ao estudo que será apresentado sobre Gestão Eletrônica de Documentos, faz-se necessário conhecer sobre a origem e a evolução deste tema na história da humanidade para proporcionar uma visão holística do assunto e melhor entender sua relevância ante o desenvolvimento das organizações contemporâneas.

Desde os primórdios da civilização, o homem buscou formas para armazenar e transmitir o conhecimento. Mesmo as formas mais primitivas de documentação, como as inscrições rupestres e os papiros manuscritos, foram ferramentas importantes para compartilhar com as gerações futuras o conhecimento e crenças de tempos remotos.

De acordo com Lima (2007), antes mesmo da descoberta da escrita havia procedimentos e técnicas desenvolvidos para a transmissão dos conhecimentos classificados como convenientes para a posteridade. Detendo apenas a memória coletiva como meio de armazenamento, utilizava-se de recursos como dramatização e artifícios narrativos onde pessoas mais idosas transmitiam suas experiências para as novas gerações. Ressalta-se também o fato que, tanto o transmissor como o receptor do conhecimento, pertenciam ao mesmo universo de tempo e lugar, além de compartilhar do mesmo domínio semântico de conceitos. Nesta realidade, a transmissão do conhecimento exigia contínuo recomeço. Tal renovação era susceptível a perdas de informação, bem como a alterações em seu significado. O conhecimento dependia fortemente da capacidade de memorização dos membros do grupo social.

Porém, com o advento da escrita, já se observava características mais efetivas para a guarda permanente do conhecimento. A utilização de ideogramas e símbolos equivalentes a fonemas desenhados em papiros e posteriormente em pergaminhos foram os precursores dos documentos e livros na forma que são conhecidos atualmente. Entretanto, a reprodução destas informações demandava um esforço grande, pois era um processo essencialmente manual, além do fato de que apenas um grupo limitado de pessoas - denominado em algumas culturas de escribas - detinha o conhecimento da leitura e escrita e eram encarregados pela cópia e pela guarda dos documentos. Neste tempo, foram desenvolvidas as primeiras técnicas de gestão documental. Distintamente da narrativa oral, a escrita

permitiu que o conhecimento acumulado fosse compartilhados entre pessoas que pertenciam a tempos e lugares distintos.

Outro evento marcante para a história documental foi a invenção da imprensa por Johann Gutenberg em 1436, onde manuscritos da Bíblia foram transformados no primeiro livro impresso em papel da história. Com o advento da imprensa, a utilização do papel difundiu-se no mundo ocidental na medida em que crescia a quantidade de novos livros e cópias. Desta forma, o que no passado era um trabalho manual e custoso tornou-se um processo automatizado e com um custo relativamente baixo. A reprodução documental por meio da mídia impressa estimulou fortemente o registro do conhecimento pelas pessoas, transformando consideravelmente a cultura da época. Outro fato a ser destacado neste período foi que a crescente disponibilização de documentos impressos ampliou o número de pessoas com acesso ao conhecimento agora materializado em papel por intermédio de textos e representações gráficas. Entretanto, os arquivos documentais nesta época sofriam influência da antiguidade sendo dotados de uma visão ainda exclusivista de guarda dos documentos para uso específico da Igreja e do governo, (CORTÊS, 1996).

Calderon (2004) descreve que, em meados do século XIX e a partir do século XX, observa-se uma maior democratização do acesso aos acervos documentais por parte dos cidadãos e pesquisadores de diversas áreas. Em face disto, fez-se necessária o desenvolvimento de profissionais preparados em administrar o crescente acervo de modo a atender a demanda das pessoas que buscavam o conhecimento nele contido.

Em paralelo, o desenvolvimento das indústrias e instituições públicas também fomentou a produção documental fora do universo das bibliotecas e universidades. Com isso, as técnicas arquivísticas utilizadas nestes ambientes foram introduzidas nas organizações que a partir de então eram proprietárias de seu próprio acervo. Destaca-se que até então os documentos tinham por suporte o papel, assim todas as técnicas adotadas visavam a guarda dos documentos em arquivos físicos organizados em estantes, prateleiras, cabines, gavetas e pastas, (SILVA, 2008).

O surgimento dos primeiros computadores em 1945, despontou uma nova era na história da gestão documental. Para Calderon (2004), a revolução provocada pela tecnologia da informação pode ser comparada com a ocorrida por ocasião da invenção da imprensa por Gutenberg.

A princípio, os computadores eram utilizados meramente como “máquinas de escrever eletrônicas”, pois o documento somente existia após sua impressão em papel. Todavia, mesmo esse uso limitado da tecnologia consistia em um grande avanço, pois facilitou muito a elaboração e a reprodução em massa de documentos.

A tecnologia de digitalização de documentos, em meados da década de 1980, possibilitou que documentos, outrora unicamente em suporte físico, pudessem ser convertidos para o suporte eletrônico e, conseqüentemente, acessado a partir dos computadores, a semelhança do que já ocorria com os documentos elaborados internamente pela organização. Os sistemas desenvolvidos para este fim eram denominados *Document Image Processing – DIP* (Processamento de Imagem do Documento). Inicialmente, estes sistemas faziam uma imagem do documento físico, além de possibilitar sua indexação e armazenamento para recuperação posterior. Alguns sistemas mais avançados de DIP também incluíam elementos básicos de *workflow*<sup>1</sup> que permitiam o trâmite destas imagens de documentos entre pessoas e departamentos. Silva (2008) destaca, entretanto, que inicialmente a utilização desta tecnologia era limitada pelo custo elevado dos equipamentos, que eram altamente especializados e exclusivos para digitalização em larga escala. Tal restrição foi eliminada com a evolução tecnológica que permitiu a utilização destes equipamentos acoplados em computadores pessoais, reduzindo assim o custo de implantação desta tecnologia.

A crescente adoção de documentos eletrônicos em instituições públicas e privadas determinou o surgimento de novas demandas como expansão da capacidade de processamento dos computadores, aumento na capacidade de armazenamento e transmissão de informações eletrônicas. Porém, as facilidades advindas com a tecnologia da informação levaram com que as pessoas começassem a produzir muito mais documentos do que eram capazes de gerenciar. Neste contexto, além da administração dos gigantescos arquivos físicos, onde eram armazenados os documentos originais em papel, percebeu-se que de forma análoga existia outro acervo muito maior e mais complexo de se gerenciar, porém invisível aos olhos humanos, os documentos eletrônicos. Estes estavam armazenados nos diversos computadores distribuídos nos pelos setores da organização.

---

<sup>1</sup> Workflow – Conjunto de tecnologias cujo objetivo é gerenciar o fluxo de documentos eletrônicos entre os diversos departamentos da instituição, baseando-se em regras previamente estabelecidas.

Na década de 1990, emergiram os primeiros Sistemas para Gerenciamento Eletrônico de Documentos (SGED) que introduziram novos conceitos para esta área, tais como: integração com ferramentas de automação de escritório, controle de versão e revisão de documentos e permissão de acesso. Neste mesmo período, surgiram as primeiras iniciativas para estabelecimento de padrões para estas tecnologias, como a ISO 15.489 que dispõe sobre gerenciamento de registros eletrônicos.

Atualmente, existem padronizações internacionais consolidadas para gerenciamento eletrônico de documentos e a tendência atual dos fabricantes de softwares é disponibilizar ferramentas que integram em uma única plataforma as diversas demandas relativas à gestão do conteúdo eletrônico corporativa, denominadas *Enterprise Content Management* - ECM (Gerenciamento de Conteúdo da Corporativo).

### **2.1.2 Definições e nomenclaturas**

De acordo com Sutton (1996), um dos primeiros passos para iniciar um processo de gestão documental é determinar o que se pretende gerenciar. Assim, uma pergunta fundamental que necessita de elucidada pela instituição é: O que é um documento em contexto organizacional?

Dada a infinidade de respostas que desta questão podem emergir, esta seção objetiva apresentar um conjunto de definições e nomenclaturas amplamente difundidos nesta área de domínio, estabelecendo assim o contexto conceitual sobre o qual tal pesquisa está alicerçada.

#### **Documento**

O documento é um conceito base para o tema deste capítulo, e a fim de estabelecer uma perspectiva holística, faz-se necessário remeter-se ao passado para estabelecer ligação das características do passado e as utilizadas na atualidade.

Briet (1953) apresenta a definição proposta pela União Francesa de Organismos de Documentação:

“Documento é toda informação registrada em suporte material, passível de ser utilizado para consulta, estudo, prova e pesquisa, pois possui a capacidade de comprovar fatos, fenômenos, formas de vida e pensamentos do homem numa determinada época e lugar.”

Esta definição amplia a abrangência do que geralmente é entendido por documento conforme a perspectiva do observador. Por exemplo, um manuscrito

hebraico para um leigo pode ser apenas um pedaço de papel antigo contendo caracteres ilegíveis, entretanto, para um arqueólogo sem dúvida é um documento, pois dele o pesquisador pode extrair informações preciosas e comprovar fatos. Uma pedra que é lançada ao rio por uma criança, certamente não é um documento, porém esta mesma pedra, sendo analisada por um geólogo torna-se sim um documento, visto que para este profissional, é um objeto de estudo.

Miranda e Simeão (2002) apresentam os elementos constitutivos do documento, conforme figura 2.1, como sendo:

- a) Tipo – Classificação prévia a fim de estabelecer padronização adequada para sua produção, guarda e veiculação;
- b) Conteúdo – Parte substantiva do documento e se constitui das informações propriamente ditas. É também condicionado ao tipo na medida em que está em conformidade com as regras por este preestabelecidas.
- c) Formato – É o modo de concepção e exposição do conteúdo do documento, moldando o conteúdo e tornando compreensível no sentido em que a forma também influencia o significado. Por exemplo, um documento pode ser textual, gráfico, audiovisual, tridimensional, dentre outros formatos possíveis.
- d) Suporte – É a parte manipulável e visível do documento, que devido às características tecnológicas atuais, pode ser físico ou lógico. Um determinado documento pode ser apresentado em diversos suportes, tais como os suportes físicos: as cavernas, o papiro, o couro, o papel, microfilme; e os suportes lógicos: fitas magnéticas, discos magnéticos e discos ópticos.



Figura 2.1- Elementos constitutivos do documento, adaptado de Miranda e Simeão (2002).

Sutton (1996) expõe que a palavra documento origina-se da raiz latina *documentum* – publicação oficial, editais ou diretivas; e derivando tais conceitos em face da realidade tecnológica atual, apresenta as seguintes definições:

“Documentos são registros legalmente sancionados de transações de negócio ou decisões que podem ser observados como uma simples unidade organizacional.”

“Documento eletrônico é uma representação digital de dados fortemente acoplados para uso humano.”

### ***Document Management - DM***

*Document Management* é o termo original para Gerenciamento Eletrônico de Documentos (GED) e, previamente à definição de DM, é importante ressaltar o significado de gerenciamento, principalmente quando aplicado a documentos. Assim, tem-se a definição de Sutton (1996) que apresenta a gestão como um conjunto de processos utilizados em um determinado ambiente com intuito de planejar, acompanhar e controlar pessoas e recursos materiais para alcançar objetivos em comum.

Aplicando a definição supracitada especificamente para documentos eletrônicos pode-se entender por DM como um conjunto de processos de supervisão sobre as transações de negócio da organização, permitindo o devido acompanhamento dos documentos desde sua criação até o devido arquivamento definitivo. Nesta definição o suporte do documento é irrelevante, podendo ser físico (papel) ou eletrônico.

Silva (2008) também conceitua DM como habilidade de controlar a criação e todo o ciclo de vida de todos os documentos eletrônicos manipulados em uma determinada organização.

### ***Enterprise Content Management - ECM***

*Enterprise Content Management* (Gerenciamento do Conteúdo Corporativo) é um conceito mais recente e amplamente difundido em organizações de grande porte. Diz respeito à integração das ferramentas e métodos utilizados nos dois últimos conceitos citados.

Pode-se entender por conteúdo corporativo como a união de todos os documentos elaborados nos mais diversos setores da organização, além de tudo o que é publicado em sua Intranet e Internet.

Boico (2005) ressalta que ECM é um processo para se obter efetivo controle de todo o conteúdo que é produzido, publicado e arquivado dentro de uma instituição, estabelecendo diretrizes de controle produção e acesso, padronização de linguagem e leiaute de apresentação, dentre outros temas relacionados.

Do pressuposto que documentar é uma forma de formalização do conhecimento, um efetivo processo de ECM contribui para que organização tenha controle sobre seus ativos intangíveis, a saber seu conhecimento organizacional.

Adam (2008) amplia a definição de ECM citando como parte deste outros serviços essenciais para a realidade atual das organizações. Em sua definição, além dos serviços de gerenciamento de documentos, conteúdo web e registros eletrônicos, apresenta os serviços de colaboração para troca e compartilhamento de conhecimento entre pessoas e departamentos, gerenciamento e monitoramento dos repositórios de armazenamento e, concluindo, cita também a tecnologia de *workflow* como parte integrante de uma solução ECM.

### ***Electronic Document Management Systems – EDMS***

EDMS refere-se às ferramentas tecnológicas concebidas principalmente para operacionalizar os processos definidos para gerenciamento dos documentos eletrônicos. Estes sistemas disponibilizam mecanismos de controle de acesso e auditoria de operações, controle de versões e revisão de documentos, gerenciamento do armazenamento, da indexação e da recuperação dos documentos.

#### **2.1.3 Ciclo de Vida do Documento**

Para uma implantação de GED, cujo objetivo é melhorar os processos de negócio da instituição, faz-se necessária a identificação das etapas que permeiam o ciclo de vida dos documentos organizacionais. Silva (2008) cita que a habilidade de melhorar os processos de negócio está intimamente relacionada com a integração destes com o ciclo de vida documental.

O ciclo de vida do documento consiste-se de vários estágios que variam de organização para organização. Por exemplo, uma agência de publicidade está mais envolvida na criação e revisão de documentos, do que uma agência de seguros que pela característica do seu negócio manipula muito mais documentos produzidos por terceiros do que produzidos internamente.

Na literatura encontram-se diversas abordagens que definem distintos estágios para o ciclo de vida dos documentos como, por exemplo, os modelos propostos pelo CONARQ (2006) - produção, tramitação, uso, avaliação, arquivamento e destinação – e por Williams, John e Rowland (2009).

Entretanto, apesar da aparente diversidade, tais estágios abrangem os mesmos temas com níveis de especificidades distintos. Assim sendo, foi adotado para esta pesquisa a proposta elaborada por Silva (2008) que identifica apenas quatro macro estágios que envolvem todas as etapas do ciclo de vida dos documentos.

### **Criação**

É o estágio em que o documento passa a existir dentro da organização. O documento pode ser criado utilizando-se de meios analógicos como, por exemplo, a máquina de escrever e a caneta; ou atualmente na maioria dos casos, os documentos já iniciam seu ciclo de vida utilizando o suporte eletrônico, pois são elaborados por meio de um software editor de textos.

Da perspectiva da criação de um documento no contexto organizacional, podem-se observar três formas pelas quais um documento é criado:

- a) Documentos elaborados internamente;
- b) Documentos recebidos de terceiros em suporte eletrônico;
- c) Documentos recebidos de terceiros em suporte físico.

Nas três formas supracitadas, o estágio da criação do documento envolve três etapas básicas: a captura; a classificação; e a indexação.

### Captura

O mecanismo de captura está diretamente relacionado com o suporte utilizado pelos documentos. Por exemplo, documentos em papel recebem tratamento distinto dos documentos eletrônicos, pois estes primeiros necessitam ser “transportados” para o suporte eletrônico antes de sua inclusão no EDMS.

O processo de captura de documentos em suporte físico vai muito além do que passar folhas de papel em um scanner. Dependendo do volume destes documentos, tal processo pode assumir grandes magnitudes na organização que, para garantir bom desempenho dos negócios, são levadas a criar departamentos especializados ou buscar pela terceirização destes serviços.

As etapas da captura de documentos físicos, conforme Silva (2008), são:

- a) Preparação Prévia – realizar os ajustes no documento, preparando-o para a introdução no scanner, tais como a extração de grampos e o alinhamento das folhas;
- b) Digitalização – processo de transformar o conteúdo impresso no papel para o suporte eletrônico;
- c) Controle de Qualidade – verificação realizada após a digitalização para garantir que a imagem gerada está dentro dos padrões predefinidos; e
- d) Preparação Posterior – procedimentos necessários para voltar o documento a sua condição inicial.

Conforme as características da organização e as dimensões do volume de documentos recebidos, a captura pode ser centralizada ou distribuída.

A captura centralizada envolve produção em larga escala utilizando-se de scanners de alto desempenho para tratar grandes volumes de documentos em ambiente especializado. Esta abordagem de captura facilita o controle e a utilização de padrões para controle de imagem e indexação, além de minimizar a utilização dos recursos de conectividade.

Em contrapartida, a captura distribuída constitui-se de pequenas unidades de digitalização com dispositivos de pequeno porte preparados para tratar volumes menores de documentos. Na captura distribuída, o documento não é encaminhado para um local central e especializado, outrossim é digitalizado na unidade onde o documento é recebido. Este tipo de captura exige menos investimento, pois os equipamentos são de pequeno porte e também não existe o custo logístico para encaminhar os documentos para um departamento central de digitalização. Entretanto, neste ambiente tem-se maior dificuldade em se garantir o controle da qualidade e a conformidade com os padrões dos documentos digitalizados.

Os documentos recebidos em meio eletrônico, que não necessitam da digitalização, porém de forma análoga também necessitam passar por um processo simplificado de captura. Na captura de documentos eletrônicos outros requisitos precisam ser atendidos, principalmente aqueles que dizem respeito à padronização tecnológica do acervo eletrônico estabelecido por diretrizes organizacionais, tais como as que normatizam quanto ao formato e o tamanho máximo destes arquivos.

#### Classificação

Nesta fase o documento é tipificado conforme padronização previamente estabelecida. A partir de sua classificação é possível identificar os níveis de acesso

e segurança que serão adotados para o documento. Por exemplo, se o tipo documental escolhido for confidencial, o acesso será limitado a um número restrito de usuários, no entanto, se for um tipo de documento de domínio público, o número de usuários que poderão acessá-lo será muito maior.

Pelo tipo do documento, identifica-se também o mecanismo de armazenamento adotado, ou seja, se aquele documento demanda armazenamento em um dispositivo de rápida recuperação como um disco magnético; ou pode ser um dispositivo de recuperação mais lenta e menos dispendioso como o microfilme. A identificação do dispositivo de armazenamento relaciona-se de igual forma ao tempo que este documento precisará estar disponível no sistema e quais as medidas para seu descarte quando este tempo expirar. Conforme o tempo de permanência definido pelo tipo, devem ser utilizados dispositivos que garantam a permanência segura destes documentos durante o tempo necessário.

### Indexação

A indexação é uma etapa crítica em uma solução de gerenciamento eletrônico de documentos. A qualidade desta etapa está intrinsecamente relacionada com o bom desempenho do sistema como um todo, pois o trabalho desenvolvido nesta etapa será o fundamento utilizado nos processos de busca pelos documentos armazenados. Uma indexação de má qualidade pode impedir ou dificultar o acesso a documentos importantes para o negócio da organização.

Esta etapa consiste em informar ao EDMS atributos que identifiquem o documento que está sendo capturado. Não existem regras quanto à quantidade de atributos por documento, contudo é importante que exista um balanceamento entre o número mínimo necessário para identificação do documento e o custo da alimentação e manutenção destes índices. Por exemplo, a princípio pode-se concluir que quanto maior o número de atributos, mais fácil será a recuperação. Porém, o volume dos documentos a serem capturados determina os custos e o tempo utilizados para alimentar uma quantidade muito grande de índices. Neste contexto é que se faz necessário o balanceamento citado conforme as necessidades no negócio.

Com intuito de otimizar a indexação de documentos, o mercado tem disponibilizado uma série de tecnologias que buscam a automação deste processo durante a própria digitalização, como o reconhecimento de código de barra, a utilização de tinta magnética e leitura de marcas.

## **Revisão**

A fim de garantir a qualidade dos diversos registros dos documentos (tipificação, conteúdo e indexação), estes são constantemente revisados e, por consequência natural deste processo, surgem diversas versões de um mesmo documento. Em um ambiente colaborativo, é imprescindível o gerenciamento do processo de revisão assegurando que um mesmo documento não esteja sendo alterado por duas ou mais pessoas ao mesmo tempo, além de garantir que cada usuário esteja acessando a versão correta do documento.

Neste cenário, foram introduzidos os conceitos de *check in/check out* e controle de versões de documentos. O controle *check in/check out* faz com que um documento somente poderá ser alterado por um único usuário ao mesmo tempo. O segundo conceito diz respeito ao gerenciamento das diversas versões do documento elaboradas durante o processo de revisão.

Outro aspecto importante deste estágio do ciclo de vida é a definição do fluxo de revisão, ou seja, o trâmite que o documento deve percorrer antes de se tornar público. Nesta etapa deve-se seguir uma política que definem as regras e as pessoas responsáveis pela revisão de todos os registros referentes a um determinado tipo documental. Estas políticas facilitam o trabalho do elaborador, que após a devida classificação do documento, deixa os próximos passos a cargo do EDMS onde foi previamente definido o processo de revisão.

Por determinação legal, ou para garantir princípios de segurança eletrônica, certos tipos documentais têm seu processo de revisão concluído após a aplicação da assinatura digital, conceito amplamente difundido na atualidade.

Acevedo (2004) define assinatura digital como tecnologia que aplica modelos matemáticos de criptografia com intuito de garantir os princípios de integridade, autenticidade e não repúdio em documentos eletrônicos. No Brasil (Lei 11.419/2006) e no Chile (Lei 19.799), por exemplo, documentos assinados digitalmente têm presunção de veracidade.

## **Transação**

O estágio de transação inicia-se após a revisão, publicação e disponibilização do documento aos usuários. Os documentos geralmente são criados com um propósito específico. Em uma organização presume-se que a maioria dos

documentos esteja relacionada a um ou mais processos de negócio, pois caso contrário tais documentos não teriam razão para existir neste contexto.

Na realidade tradicional dos documentos em papel é de fácil percepção a necessidade da tramitação física destes em pastas e processos por diversos departamentos até que seus objetivos estejam cumpridos e passem para próximo e último estágio do ciclo de vida, o arquivamento. No paradigma dos documentos eletrônicos, esta necessidade persiste com o mesmo intuito de que o documento cumpra seus desígnios no processo em que está inserido.

A despeito de todos os benefícios advindos desta abordagem como, por exemplo, economia de papel e espaço físico, novos desafios surgem, principalmente quanto à resistência natural à mudança que é inerente ao ser humano. Por séculos os documentos eram visíveis e tramitaram fisicamente, porém no mundo digital, estes mesmos documentos agora tramitam em meios eletrônicos. No entanto, da mesma forma exigem manipulação, demandam atenção e geram esforço de trabalho.

Assim sendo, nesta nova perspectiva faz-se necessária a instituição de regras de tramitação muito bem definidas que estejam em conformidade com fluxo do processo de negócio. Estas regras garantem que o documento permeie corretamente os departamentos e na ordem exigida pelo processo, minimizando assim os problemas de documentos encaminhados para pessoas ou departamentos equivocados.

### **Arquivo**

Após o documento ter cumprido o objetivo para o qual fora criado, passando por todas as etapas previstas no estágio de transação, ainda é imprescindível que este documento continue fazendo parte do acervo documental da organização para consultas futuras ao seu conteúdo, bem como para cumprimento de exigências legais que determinam a guarda temporária ou permanente para tipos específicos de documentos.

Segundo Silva (2008), os arquivos podem ser categorizados pelo tempo necessário para busca e recuperação do documento no repositório corporativo:

- a) *On Line* – Permite a recuperação imediata do documento ou em poucos segundos utilizando para isso discos magnéticos não removíveis.

- b) *Near Line* – Existe um pequeno tempo de espera entre a requisição e a recuperação do documento, contudo o processo todo é automatizado sem a necessidade de intervenção humana ou mecânica.
- c) *Far Line* – Esta categoria exige a intervenção humana ou mecânica para que o documento possa ser recuperado. Como exemplo, tem-se a tecnologia *Jukebox* que utiliza braços mecânicos para manusear as diversas mídias e disponibiliza-las nas leitoras quando solicitado. Devido a esta característica, o tempo de busca pode demorar alguns minutos ou horas dependendo da infraestrutura utilizada.
- d) *Off Line* – Neste caso, o documento não está disponível para o sistema por não ser um documento digitalizado ou pela ausência de indexação adequada para a recuperação.

Existem diversas tecnologias disponíveis para gerenciar este estágio do ciclo de vida dos documentos. As tecnologias diferenciam-se conforme os requisitos relacionados com o tempo de guarda e o tempo de resposta na busca pelo documento.

A figura 2.2 apresenta a relação existente entre os seguintes fatores: capacidade de armazenamento; velocidade de recuperação; custo; e segurança quanto à durabilidade da informação no decorrer do tempo. Deste gráfico conclui-se que quanto maior a velocidade de recuperação, maior é o custo e menor é a capacidade de armazenamento, bem como sua durabilidade. Conseqüentemente, o custo das mídias de armazenamento diminui na mesma proporção que diminui a velocidade de recuperação da informação e aumenta a garantia de durabilidade. Entretanto, cabe é saliente ressaltar que futuramente tal afirmação poderá não ser mais verdadeira devido ao desenvolvimento de novos sistemas de memória em rede.

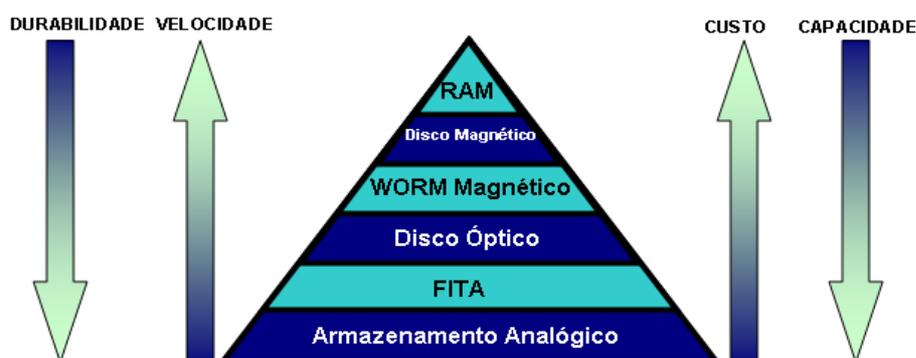


Figura 2.2 - Níveis de arquivamento.

Concluindo, ressalta-se que a abordagem quanto ao arquivamento está diretamente relacionada com a classificação realizada no momento da captura. Conforme a realidade da instituição, pode-se adotar mecanismos distintos para o arquivamento. Assim, o estágio de arquivo pode ser complexo e oneroso para documentos diretamente vinculados ao negócio da instituição, tendo, em contrapartida, um processo mais simplificado e de baixo custo para outros documentos.

#### 2.1.4 Principais Tecnologias

No transcorrer da história da Gestão Eletrônica de Documentos, diversas foram as tecnologias utilizadas como ferramenta de apoio e controle em um ou mais estágios do ciclo de vida dos documentos, porém de forma isolada e com baixa integração entre os estágios.

Contudo, com a expansão do uso de documentos eletrônico nas organizações e com a evolução tecnológica, atualmente já é possível gerenciar de forma centralizada e integrada todas as etapas relativas à manipulação de grandes volumes de documentos eletrônicos a partir de uma única arquitetura. A figura 2.3 apresenta as principais tecnologias disponíveis na atualidade e onde tais tecnologias estão inseridas no ciclo de vida documental.

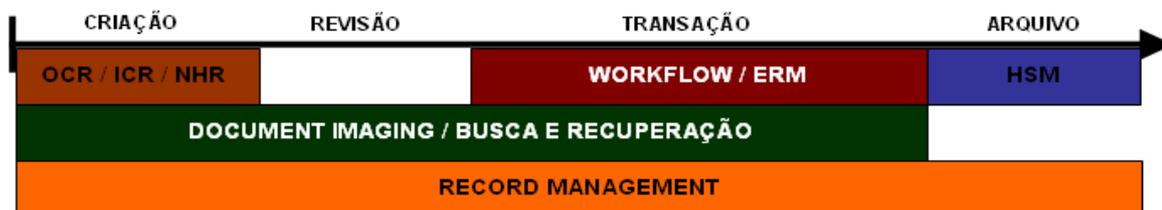


Figura 2.3 - Principais tecnologias de GED e sua inserção no ciclo de vida de documentos.

#### Reconhecimento Automático de Caracteres

O reconhecimento automático de caracteres foi desenvolvido para possibilitar a manipulação eletrônica do conteúdo existente nos documentos em suporte físico após sua digitalização. Com a aplicação dessas tecnologias, é possível realizar as mesmas operações realizadas em documentos já elaborados em meio eletrônico, tais como: pesquisar, alterar, copiar, excluir, dentre outras.

Silva (2008) destaca algumas especializações das tecnologias de reconhecimento de acordo com a natureza do documento físico digitalizado:

##### Optical Character Recognition – OCR

É tecnologia que possibilita aos computadores efetuarem a leitura do conteúdo de documentos impressos em papel por meio de mecanismos não manuais de impressão como, por exemplo, máquinas de escrever, impressoras e edições gráficas. Este tipo de reconhecimento busca identificar padrões de impressão e convertê-los para a apropriada codificação eletrônica correspondente.

#### Intelligent Character Recognition – ICR

Intelligent Character Recognition utiliza-se do mesmo princípio que o OCR porém para documentos escritos manualmente com letras de forma, ou seja, escrita não cursiva. Mesmo tendo como premissa a utilização de um padrão único de escrita, existe uma infinidade de pequenas diferenças que impossibilitam em muitos casos o reconhecimento automático.

Para minimizar a ocorrência de erros, os softwares que implementam estas tecnologias utilizam-se de dicionários de palavras, além de dispor interfaces gráficas para intervenção humana quando da impossibilidade de reconhecer automaticamente determinada parte do texto.

#### Natural Handwriting Recognition – NHR

O reconhecimento de documentos escritos de forma cursiva é um processo ainda mais complexo que o ICR, devido ao fato que um mesmo autor pode possuir diversos padrões de escrita distintos, além de que nenhuma pessoa possui a mesma forma de escrita utilizada por outra.

Devido a estes desafios, o NHR utiliza-se de diferentes abordagens para executar o reconhecimento dependendo da natureza do documento e da língua utilizada. Um texto escrito em uma língua ocidental como o português, inglês e espanhol não pode ser reconhecido com os mesmos mecanismos de um texto em língua oriental que possui características de escrita muito peculiares.

Assim, tendo identificado as propriedades do conteúdo, este é decomposto em conjuntos de arquétipos que podem ser convertidos para a linguagem de máquina e devidamente reconhecidos. Neste processamento também se utiliza dicionários de palavras e a intervenção humana.

#### ***Document Imaging***

*Document Imaging* é um conjunto de tecnologias que tem por objetivo fazer a digitalização documentos que se originaram em suporte físico, com intuito de disponibilizar a imagem deste documento para os usuários de um sistema GED.

Este é um importante estágio do ciclo de vida documental, não só pela economia do espaço físico, antes destinado ao tradicional arquivo-morto, mas principalmente pela velocidade e facilidade em disponibilizar um determinado documento a diversas pessoas simultaneamente. (ADAM, 2008).

Este conjunto de tecnologias abrange softwares e hardwares especializados para o tratamento de imagens, essencialmente por que a qualidade do produto deste processo que é a imagem digital do documento físico está diretamente relacionada com a qualidade do sistema como um todo. Devido a isso, há uma busca infindável pela indústria de TI por novas tecnologias que garantam bom desempenho ante às crescentes necessidades corporativas, sem perder a qualidade mínima aceitável das imagens produzidas.

Quanto à qualidade das imagens produzidas, Silva (2008) apresenta alguns fatores devem ser levados em consideração, tais como:

- a) Resolução – quantidade de pontos por polegadas utilizados para desenhar a imagem seja no monitor ou quando esta é impressa em papel;
- b) Cor – o número de cores de um documento determina o espaço de armazenamento reservado para esta definição. Quanto à cor um documento pode ser: *bitonal* (preto e branco); *escala de cinza* (apenas utilizando variações da cor cinza); *colorido*; *halftone* (semelhante ao bitonal, porém com a aparência de um documento em escala de cinza).
- c) Compressão – refere-se ao algoritmo de compressão utilizado para o armazenamento da imagem. Este é um fator importante no que diz respeito à economia de espaço lógico, entretanto deve-se haver um balanceamento entre o nível de compressão e a qualidade mínima desejado, visto que estes dois elementos são inversamente proporcionais.

Quanto à definição do hardware utilizado para digitalização, ressalta-se a importância das seguintes variáveis no momento de sua especificação:

- a) Papel – identificar as características físicas do papel em que estão impressos os documentos para digitalização é um fator essencial, porém, muitas vezes não levado em consideração. Deve-se atentar para a dimensão máxima e mínima do papel utilizado nos documentos, sua gramatura, acabamento e cor. Todos estes aspectos são subsídios para identificar o hardware necessário para atender as demandas de digitalização;

- b) Velocidade – a velocidade de digitalização necessária é determinada pela razão entre quantidade de documentos para digitalização e o tempo máximo exigido para a conclusão do processo. Atualmente existem muitas opções quanto à velocidade, tanto para digitalizações de pequeno porte onde o valor de 10 ppm (páginas por minuto) é suficiente, bem como para ambientes de produção em larga escala os quais necessitam de equipamentos que produzam mais do que 100 ppm;
- c) Digitalização Duplex – é a capacidade do scanner de digitalizar os dois lados do documento em uma única digitalização, aumentando o desempenho do processo. Esta funcionalidade faz-se necessária quando existe uma quantidade considerável de documentos com informações nos dois lados do papel.

Silva (2008) ressalta que dentro do escopo de *Document Imaging* também estão os serviços de indexação, armazenamento e recuperação dos documentos captados por intermédio da digitalização.

### ***Electronic Record Management - ERM***

Adam (2008) define *registro* como toda evidência ou parte de informação utilizada para descrever um fato ocorrido no passado. Em se tratando especificamente de registros eletrônicos, apresenta registro como a infraestrutura tecnológica utilizada para evidenciar um fato ou parte de um fato ocorrido no passado. Neste sentido, um registro eletrônico provê as informações factuais de um incidente ocorrido, conseqüentemente não podem sofrer modificações.

Um registro eletrônico pode ser constituído de diversos tipos de mídias como, por exemplo, de registros em papel capturados para um Sistema de Gerenciamento de Registros Eletrônicos, fax, e-mails e documentos eletrônicos; desde que o agrupamento destes diferentes documentos sirva para comprovar fatos do passado.

Assim sendo, *Electronic Record Management* (Gerenciamento de Registros Eletrônicos) é o conjunto de técnicas e ferramentas utilizadas manipular os diversos registros eletrônicos existentes em uma organização. Um efetivo processo de ERM, conforme Adam (2008), deve prover controle para:

- a) Repositório Central de Registros - onde as informações são arquivadas;

- b) Estrutura de Pastas – servem para categorizar o arquivo dos registros eletrônicos hierarquicamente com vistas à organização destes conforme a estrutura de negócio da organização;
- c) Classificação, Indexação e Meta Dados – mecanismos adicionais de organização utilizados pelos mecanismos de busca para recuperar as informações armazenadas no repositório;
- d) Retenção e Descarte de Registros – gerencia o tempo de guarda e descarte dos registros com base na tabela de temporalidade<sup>2</sup> previamente estabelecida;
- e) Segurança – estabelece políticas de acesso aos registros do repositório;
- f) Gerenciamento de Registros Físicos – controle dos registros que, por exigência legal ou importância histórica, exigem o tradicional arquivamento físico.

### ***Business Process Management – BPM***

*Business Process Management* (Gerenciamento de Processos de Negócio), também conhecido como tecnologia de *Workflow*, é utilizado para o gerenciamento do fluxo das informações que permeiam uma organização conforme as regras e os requisitos de seus processos de negócio. (ADAM, 2008).

Cruz (2004) define BPM como o conjunto de ferramentas e mecanismos cuja finalidade é automatizar processos de negócios, racionalizando-os e consequentemente aumentando sua produtividade por meio de dois componentes básicos: organização e tecnologia. Krammes (2008) ressalta que os principais benefícios para as organizações seriam: a otimização dos processos críticos de negócio, maior qualidade nos serviços prestados, melhoria da comunicação entre as pessoas e maior confiabilidade.

Para Silva (2008), BPM, além de automatizar os processos de negócio, habilita os usuários a controlar a lógica destes processos. A capacidade de se controlar diversos processos e o controle de seu conteúdo e integridade pelos SGED permite que as aplicações essenciais para o negócio centradas em documentos operem em um ambiente fácil de implementar e gerenciar.

---

<sup>2</sup> Tabela de Temporalidade – Define o tempo de guarda e as ações de descarte de documentos. (CONARC, 2006).

Um sistema de BPM basicamente é formado por três elementos primários também conhecidos como 3Rs: roles, routes e rules (papéis, rotas e regras). Sendo que os papéis são as atribuições e competências de determinado usuário no sistema. As rotas são as definições do fluxo de tramitação das informações ou etapas que necessitam ser cumpridas para andamento do processo de negócio. As regras são padronizações que normatizam as atividades que compõe o processo de negócio em conformidade com as diretrizes organizacionais. (CRUZ, 2004).

A abrangência da tecnologia de BPM não se restringe unicamente a tramitação de documentos eletrônicos. Entretanto, com a popularização do uso deste tipo de informação pelas pessoas, tal tecnologia tem se apresentado como uma importante ferramenta para transação de grandes volumes de documentos em consonância com as necessidades de negócio e diretrizes estabelecidas pela organização.

### ***Enterprise Report Management – ERepM***

Conforme Silva (2008), o Gerenciamento Corporativo de Relatórios (ERepM), anteriormente conhecido como *Computer Output Laser Disk (COLD)*, é uma solução integrada de software e hardware que armazena e indexa a produção formatada de computador (relatórios) em disco ótico, disco magnético ou fita magnética como uma alternativa à impressão de grandes volumes de papel ou em microfimes.

Estes relatórios consistem basicamente de listagens de transações, extratos e faturas, que são utilizados pelos usuários no desempenho de suas atividades. A estrutura e o formato dessa produção são conhecidos e as informações são focadas em um determinado período de tempo. Estes relatórios formam imensos volumes de papel, demandam tempo excessivo para impressão, e são de difícil manipulação, o que fazia com as pessoas gastassem muito tempo procurando pelas informações necessárias.

Com o surgimento das tecnologias de ERM, estes mesmos relatórios são gerados no suporte eletrônico e detêm todas as facilidades de manipulação e acesso rápido à informação características de um documento eletrônico.

Em essência, um processo ERM envolve dois procedimentos: gravação, onde é realizada a indexação do relatório e seu armazenamento no sistema GED; e recuperação, quando o relatório é disponibilizado aos usuários, (SILVA, 2008). O procedimento de gravação dos relatórios pode constituir-se de diversas atividades

complexas, dependendo da origem das informações e do destino, ou seja, onde o relatório eletrônico será armazenado. Quanto à origem das informações, a complexidade pode advir quando os sistemas legados da organização não dispõem de interfaces para integração com outros sistemas; ou mesmo quando estes sistemas são executados em plataformas tecnológicas distintas. Quanto ao destino da gravação, a mídia adotada é um fator determinante para o processo. Em mídias on-line o procedimento é mais simplificado, pois a mídia sempre está disponível, entretanto as mídias *near-line* ou *far-line* exigem um controle maior quanto ao conteúdo nelas armazenados e também quanto à sua localização física.

### ***Hierarchical Storage Management – HSM***

Os níveis de armazenamento expostos na figura 2.2 apresentam a diversidade de mecanismos disponíveis para o arquivamento de informações. Cada um deles possui especificidades que demandam um trabalho criterioso pelos profissionais de ECM para decidir uma política de arquivo que atenda as necessidades organizacionais dentro de sua disponibilidade orçamentária.

Em síntese, os fatores levados em consideração nesta análise são: quantidade e tempo máximo de armazenamento; projeção de crescimento do acervo; e o tempo máximo aceitável para busca e recuperação. Estas informações mínimas subsidiam a tomada de decisão quanto aos mecanismos de arquivo adotados.

Dá-se ênfase ao fato que a quantidade de acesso a um documento específico diminui com o decorrer do tempo, ou seja, nos três primeiros estágios de seu ciclo de vida, os documentos são muito mais acessados do que no estágio de arquivo.

Assim, pode-se concluir que para os documentos em trâmite na organização, faz-se necessário a utilização de mídias que tenham um tempo de resposta imediata às solicitações de busca, pois em caso contrário haverá perda de produtividade. Este mesmo princípio pode ser levado em consideração quanto aos documentos recém arquivados.

Quanto aos documentos de arquivo temporário, que constitui a maior parte do acervo, pode-se pensar na utilização de uma mídia mais lenta e de um custo mais baixo.

Como citado previamente, determinados documentos devem permanecer no acervo permanentemente. Neste caso, mesmo dispondo de diversas mídias, para

armazenamentos de longo prazo a mídia mais segura é o microfilme, que é uma tecnologia analógica de baixo custo, porém o tempo de acesso é lento se comparado com as mídias magnéticas.

Neste cenário, onde diversas variáveis condicionam a política de armazenamento, a tecnologia denominada *Hierarchical Storage Management* (HSM) é uma ferramenta para automatizar o estágio de arquivo documental. O HSM automatiza o processo de transferência de documentos entre as mídias de alto custo e com alta performance e as mídias de baixo custo com baixa performance. Tal transferência é realizada nos dois sentidos: o primeiro e mais comum que é quando um documento atinge o tempo predefinido para acesso imediato (*on-line*) e é movido para uma mídia de acesso lento (*near-line* ou *far-line*); o segundo, quando um documento que já tramitou e foi movido para uma mídia lenta, necessita voltar para o estágio de transação e voltar a ser acesso frequentemente.

A utilização de HSM otimiza a utilização dos recursos de armazenamento, permitindo que a organização gerencie grande volume de documentos conforme demanda seu negócio a um custo dentro de suas possibilidades orçamentárias.

### **Mecanismos de Pesquisa**

Conforme explanação anterior, no estágio de captura de documentos é realizada a classificação e indexação destes com intuito de facilitar o acesso a seu conteúdo. Assim sendo, o repositório de documentos necessita ser sistematicamente organizado de acordo estes processos.

Conforme Adam (2008), uma característica essencial para um sistema GED é disponibilizar diversos mecanismos para localizar documentos tais como navegação pela estrutura de pastas, pesquisas básicas e pesquisas avançadas.

O mecanismo de navegação pela estrutura pastas do repositório é um mecanismo importante de pesquisa por adotar a mesma abordagem já utilizada pelos sistemas operacionais e de amplo conhecimento por parte dos usuários. Entretanto este tipo de pesquisa exige do usuário maior conhecimento com relação ao documento a ser localizado e, principalmente, de como está organizado logicamente o repositório de documentos.

A pesquisa básica permite ao usuário a recuperação de documentos a partir de algumas palavras-chave. Assim, o sistema percorre os índices do repositório

procurando por documentos que contenham estas palavras em seus metadados ou em seu conteúdo.

No que tange a pesquisa avançada, além do que é disponibilizado na pesquisa básica, existe a possibilidade de localizar documentos por diferentes tipos de atributos, tais como: um campo específico de seus metadados, período de datas, autor, departamento etc. Na pesquisa avançada também é possível utilizar-se de combinações de campos de busca, bem como de operadores lógicos. Um exemplo de pesquisa avançada pode ser quando um usuário busca por uma sentença proferida por determinado magistrado, contendo determinada frase em seu conteúdo e que ainda esteja no estágio de transação.

Ainda no domínio da pesquisa avançada, existem mecanismos que realizam a identificação do contexto da busca, cujo objetivo é reduzir a quantidade de documentos resultantes da busca com base em uma perspectiva semântica preestabelecida.

## **2.2 Gestão Eletrônica de Documentos no Judiciário Brasileiro**

De acordo com Faustino (2009), a década de 70 foi de fundamental importância no que tange à gestão documental e arquivística no Brasil. Neste período, foram criadas as graduações em nível superior de Arquivologia e instituído o Sistema Nacional de Arquivos. Tais iniciativas foram um marco introdutório para a gestão documental no país.

Em 1991, a Lei n. 8.159 de 08.01.1991, que dispõe sobre a política nacional de arquivos públicos e privados, regulamentou a gestão documental em âmbito nacional e criou o Conselho Nacional de Arquivos (CONARQ). Esta Instituição é um órgão colegiado vinculado ao Arquivo Nacional da Casa Civil da Presidência da República, que tem por finalidade definir a política nacional de arquivos públicos e privados, como órgão central de um Sistema Nacional de Arquivos, bem como exercer orientação normativa visando à gestão documental e à proteção especial aos documentos de arquivo, BRASIL (2009).

Dentre as diversas ações, Faustino (2009) destaca a Resolução n. 26 de 06.05.2008, que estabelece diretrizes básicas de gestão de documentos a serem adotadas nos arquivos do Poder Judiciário. Esta norma regulamenta o uso dos conceitos fundamentais como o Plano de Classificação, Tabelas de Temporalidade e Descarte de Documentos. Estes fundamentos ainda que estejam no âmbito na

Biblioteconomia e Ciência da Informação são requisitos essenciais e obrigatórios para o desenvolvimento de um efetivo SGED.

Ainda em 2008 foi firmado um acordo de cooperação técnica entre o CONARQ e o Conselho Nacional de Justiça (CNJ) para instituir o Programa Nacional de Gestão Documental e Memória do Poder Judiciário – PRONAME. Este programa tem como principal objetivo implantar uma política nacional de gestão documental e de preservação da memória do Poder Judiciário. As ações do PRONAME são voltadas à integração dos Tribunais, à padronização e utilização das melhores práticas de gestão documental, visando à acessibilidade e à preservação das informações contidas nos autos judiciais a fim de melhor suportar a prestação dos serviços jurisdicionais e a utilização dos acervos judiciais na construção da História, BRASIL (2008).

Como resultado deste acordo de cooperação técnica foi editado o Modelo de Requisitos para Sistemas Informatizados de Gestão de Processos e Documentos do Judiciário brasileiro (MoReq-Jus). Este modelo, conforme Brasil (2009), tem por objetivo fornecer especificações técnicas e funcionais, para orientar a aquisição, o detalhamento e o desenvolvimento de sistemas de gestão de processos e documentos no âmbito do Judiciário brasileiro. Além disso, também apresenta critérios para certificação do grau de aderência ao modelo.

No âmbito da Justiça Eleitoral, em 2007, foi lançado o Programa de Gestão Documental da Justiça Eleitoral (PGD-JE) que já está em conformidade com as normas estabelecidas pelo CONARQ e CNJ.

Apesar das iniciativas citadas, o Poder Judiciário ainda não detém um SGED que atenda aos requisitos definidos nessas normatizações. Porém, o esforço que a princípio estava sendo empreendido para o domínio das definições de regras e modelos de referências, está sendo encaminhado para a aquisição e o desenvolvimento de sistemas que suportem as demandas identificadas nestas instituições judiciárias. Assim sendo, espera-se que a utilização de sistemas informatizados que sigam os padrões de segurança, interoperabilidade, preservação temporal, dentre os outros, a sociedade possa ter acesso ágil e seguro ao acervo público de documentos judiciais.

Em se tratando de GED tanto no aspecto *lato sensu*, como em perspectivas específicas como a supracitada do Poder Judiciário, o desenvolvimento de ferramentas tecnológicas nesta área vem agregando um importante papel nas

organizações pelo crescimento exponencial das informações não estruturadas que são produzidas cotidianamente. Neste contexto, a Engenharia do Conhecimento dispõe de uma área de pesquisa, cujo enfoque é exclusivamente a recuperação destas informações. Assim, tendo em vista o objetivo deste trabalho em disseminar o conhecimento jurisdicional contido nos acervos do Judiciário, o próximo capítulo versará sobre *Information Retrieval* (Recuperação da Informação).

### 3 RECUPERAÇÃO DA INFORMAÇÃO

Em sua gênese, a atividade de recuperação estava restrita a poucas pessoas como: bibliotecários, pesquisadores, e outros similares. Porém, as mudanças advindas com a socialização da TI facilitaram o acesso de outras pessoas, e atualmente é possível observar um número crescente de pessoas envolvidas nestas atividades, seja pesquisando documentos, e-mails ou mensagens instantâneas. Nesta senda, a Recuperação da Informação (RI) tem se tornado um dos principais mecanismos para acesso à informação. (Manning, Raghavan e Schütze, 2008).

As definições apresentadas abaixo estabelecem em suma que RI é uma área da ciência cujo foco de estudo é desenvolver técnicas para que uma pessoa, com uma determinada necessidade de informação, possa recuperá-la por intermédio de uma interface simplificada de forma ágil.

De acordo com Calderon (2004), as primeiras técnicas de RI vieram muito antes do surgimento dos computadores em meados do século XIX onde se iniciou a formação de profissionais específicos para administração do crescente acervo até então em papel.

Com relação à moderna RI, ou seja, a RI computadorizada, Cleverdon (1991) apresenta que estes estudos iniciaram apenas ao final da década de 1940. Um artigo publicado por Bush (1945) serviu de inspiração para o novo campo de pesquisa, neste tem-se:

“Considere no futuro um dispositivo para uso individual, o qual seja capaz de armazenar livros e bibliotecas eletrônicas... Sendo um dispositivo automatizado, poderá disponibilizar consultas a livros, registros e comunicações com excelente velocidade e flexibilidade.”

A crescente produção de literatura científica em conjunto com a disponibilidade de computadores rebuscou o interesse em RI, porém com um novo foco: desenvolvimento de técnicas de automação. Naqueles dias, a RI era limitada a pesquisas pelo título, autor e poucas palavras-chave previamente cadastradas nos catálogos eletrônicos. Pesquisas ao inteiro teor dos documentos somente foram disponíveis anos mais tarde.

Em complemento à GED, esta seção tem por objetivo apresentar os estudos e as técnicas atualmente adotadas para recuperação da informação armazenada em acervos de documentos eletrônicos, servindo como subsídio para a aplicação proposta neste trabalho de pesquisa.

### 3.1 Definições e Conceitos

Manning, Raghavan e Schütze (2008) definem RI como:

“A busca por material (geralmente documentos) de natureza não estruturada (geralmente textos livres) que satisfaça uma necessidade de informação a partir de um grande acervo de documentos (geralmente armazenadas em microcomputadores).”

Os autores também expõem que o termo *informação não estruturada* diz respeito a toda informação sem estrutura predefinida que possa ser compreendida por um computador. Um exemplo clássico de *informação estruturada* são as tabelas de um banco de dados relacional, onde as colunas são seus atributos e as linhas são os dados propriamente ditos. Um documento textual não segue uma estrutura predefinida por isso é classificado como informação não estruturada. A RI pode trabalhar ainda com um terceiro conceito que se coloca entre os dois apresentados anteriormente, denominado *informação semi-estruturada*. Neste tipo de informação existe alguma estrutura predefinida como, por exemplo, título, autor, resumo e palavras-chave que podem ser recuperadas por um Sistema de Recuperação da Informação (SRI).

A definição acima apresenta outro conceito essencial em RI, *necessidade de informação*, ou *informação necessária*, que é um tópico ou assunto sobre o qual uma pessoa está buscando obter maior conhecimento. Esta é uma questão relevante em RI, pois saber o que se está procurando, apesar de ser um requisito básico para uma pesquisa, nem sempre é algo claro na mente dos usuários dos mecanismos de pesquisa automatizada. Além disso, a conversão da necessidade de informação em uma expressão de pesquisa é uma tarefa subjetiva e depende muito da expertise do usuário em selecionar os termos relevantes para encontrar os documentos que versem sobre o tema por ele desejado.

A definição elaborada por Greengrass (2000) introduz o conceito de *query*, que justamente é uma abstração da necessidade da informação na tentativa de reduzi-la, ou melhor, converte-la a uma sentença que possa ser compreensível a um SRI:

“Recuperação da Informação é a disciplina que estuda a recuperação de dados não estruturados, especialmente documentos textuais, em resposta a uma expressão de consulta (*query*) que, por sua vez, também pode ser não estruturada como um texto livre ou estruturada como uma expressão booleana.”

Em relação a SRI, Manning, Raghavan e Schütze (2008) os categorizam em três escalas. A primeira delas é a mais abrangente e atualmente a mais difundida na atualidade que é a *pesquisa de informação web*, onde os sistemas provêm

pesquisa em bilhões de documentos armazenados em milhões de computadores espalhados ao redor do mundo. Estes requisitos demandam o desenvolvimento de técnicas para viabilizar pesquisas rápidas em um acervo heterogêneo e descentralizado de documentos. A segunda escala em abrangência, *pesquisa de informação corporativa*, onde o universo de busca se restringe à rede privativa de uma determinada organização. Nesta escala observa-se uma redução considerável na abrangência pois os objetos de pesquisa são constituídos de documentos internos da organização armazenados, via de regra em um sistema de arquivos centralizado e homogêneo. A terceira escala, a *pesquisa de informação pessoal*, está presente em todos os computadores, disponível nos sistemas operacionais, gerenciadores de arquivos e clientes de e-mails. Estes são sistemas cuja abrangência se resume a um único computador ou no máximo a alguns computadores em uma pequena rede privativa.

Baeza-Yates e Ribeiro-Neto (1999) apresentam um conceito mais abrangente de RI quando a definem como a área de pesquisa onde os estudos abrangem os processos de representação, armazenamento, organização e acesso à informação. Onde os três primeiros processos (representação, armazenamento e organização) visam ao estabelecimento de uma infraestrutura tecnológica que possibilite aos usuários fácil acesso às informações que tem interesse. Percebe-se aqui uma diferença significativa entre *recuperação de dados*, baseado exclusivamente na lógica relacional que recupera todos os dados que conferem com os parâmetros de busca, e *recuperação da informação*, cujo foco é recuperar prioritariamente os documentos cujos conteúdos mais se aproximem da necessidade de informação do usuário.

Neste sentido, o conceito de *efetividade* em SRI não está relacionado com a quantidade de documentos que satisfaçam a *query*, mas sim com a proximidade do conteúdo dos documentos recuperados com o tema sobre o qual o usuário tem interesse. Para dimensionar o grau de efetividade de um SRI, Manning, Raghavan e Schütze (2008), relacionam dois indicadores básicos: *precisão* e *cobertura*. A precisão indica qual a fração dos documentos recuperados na pesquisa são relevantes para a necessidade de informação do usuário, quanto o indicador *cobertura* diz respeito à fração dos documentos relevantes do acervo que foram recuperados pelo SRI.

Com intuito de garantir valores satisfatórios para os indicadores de precisão e retorno sem prejudicar a performance das pesquisas, faz-se necessária a representação dos documentos eletrônicos em estruturas auxiliares, possibilitando assim a otimização e flexibilização dos processos de pesquisa automatizada em grandes acervos documentais. Estas estruturas auxiliares são chamadas de *índices* e na próxima seção serão abordadas algumas técnicas de representação de documentos utilizadas na RI.

### 3.2 Representação de Documentos

O processamento linear no conteúdo dos documentos para selecionar aqueles que atendem a uma determinada *query* demanda um elevado custo computacional que, devido à dimensão dos acervos eletrônicos disponíveis, torna inviável esse tipo de abordagem.

Em face disso, foram desenvolvidas formas alternativas para representar os documentos de maneira que fosse viável o processamento de pesquisas automatizadas em grandes acervos com um nível de desempenho aceitável.

Manning, Raghavan e Schütze (2008) citam a *matriz de incidência* como uma forma básica de representação de documentos. Esta matriz é composta pela relação de ocorrência de uma lista de palavras - geralmente chamadas termos de indexação ou apenas termos - em uma lista de documentos de um determinado acervo, também chamado de coleção. Esta representação é considerada básica, pois adota apenas valores binários para indicar a ocorrência (1) ou ausência (0) de um termo em um documento.

TERMOS	DOC. 1	DOC. 2	DOC. 3	DOC. 4	DOC. 5	DOC. N
gestão	0	0	1	1	0	1
documental	1	1	0	0	1	0
abordagem	0	0	1	0	1	0
ênfase	0	0	1	1	0	0
disseminação	1	0	0	1	1	0
conhecimento	1	1	0	1	0	1
judiciário	0	1	0	1	1	0

Tabela 3.1 - Matriz de Incidência Básica.

Ao conjunto de termos que compõe uma tabela de incidência é dado o nome de dicionário que quanto maior, maior será o custo computacional necessário para seu processamento. Assim sendo, com intuito de reduzir o número de termos do dicionário da matriz de incidência, usualmente são realizadas diversas operações

sobre o dicionário de termos, tais como: exclusão de *tags*, exclusão de *stopwords*, *stemming* e uso de tesouros. Estas operações serão detalhadas na seção 3.5 que tratará sobre mineração de texto.

A tabela 3.1 apresenta uma matriz de incidência fictícia que representa a ocorrência de onze termos relacionados com seis documentos. Segundo Baeza-Yates e Ribeiro-Neto (1999) esta é uma forma clássica de representação de documentos, os quais podem ser descritos por um conjunto relevante de termos cujo objetivo é ajudar na identificação semântica dos assuntos abordados nos documentos.

Retomando o fator tamanho dos acervos eletrônicos, conclui-se que este tipo de representação tem melhor desempenho que a leitura linear dos documentos. Porém, a matriz de incidência ainda não resolve o problema do elevado custo computacional demandado no processamento de pesquisas textuais, visto que a matriz pode assumir valores gigantescos tanto na dimensão dos termos como na dimensão dos documentos. Tal característica é comum porque esta representação armazena informações de ocorrência ou ausência para todas as relações [termo, documento] existentes na matriz, resultando sempre no produto cartesiano das dimensões.

Uma alternativa para reduzir o tamanho desta matriz é apresentada por Frakes e Baeza-Yates (1992), onde propõem uma representação que armazene somente as informações relativas à ocorrência dos termos nos documentos, ignorando as informações relativas à ausência. Esta representação é denominada *Índice Invertido*.

Na representação Índice Invertido, à semelhança do que ocorre na matriz de incidência, existe a dimensão dos termos, ou seja, o dicionário. Entretanto, para cada termo do dicionário existe uma lista ordenada contendo unicamente os documentos que contém este termo em seu conteúdo, a qual é denominada de *postings*. A tabela 3.2 demonstra uma representação de índice invertido contendo à direita o dicionário da tabela 1 após exclusão dos termos *stopwords* e à esquerda os *postings* correspondentes a cada termo do dicionário.

<b>conhecimento</b>	Doc. 1	Doc. 3	Doc. 5	Doc. 8	Doc. 9
<b>disseminação</b>	Doc. 2	Doc. 4			
<b>documental</b>	Doc. 1	Doc. 2	Doc. 5	Doc. 8	
<b>Gestão</b>	Doc. 2	Doc. 5	Doc. 9		
<b>Judiciário</b>	Doc. 4				

Tabela 3.2 - Representação Índice Invertido, composta pelo dicionário e os postings.

Observa-se que o dicionário é ordenado alfabeticamente, enquanto que os *postings* são ordenados pelo código identificador de cada documento. Este tipo de ordenação faz-se necessária para garantir o desempenho desejado no processamento das pesquisas textuais nesta representação.

Desta forma, para realizar a pesquisa na coleção que retorne os documentos contendo os termos *gestão* e *conhecimento* na estrutura acima, basta selecionar os *postings* respectivos dos termos informados na *query*, e após fazer o cruzamento destas listas obtêm-se a lista de documentos desejada: *Doc. 5 e Doc. 9*.

As duas representações acima formam a base para o modelo de recuperação booleana, que será detalhado em seção subsequente. Porém, Frakes e Baeza-Yates (1992) ressaltam que este tipo de abordagem não permite qualquer tipo de classificação do resultado da pesquisa quanto ao grau de relevância dos documentos em relação à *query* em questão. Assim sendo, dependendo da *query* elaborada o sistema pode recuperar uma coleção imensa de documentos que atendem aos requisitos da pesquisa, porém, sem relação direta com a real necessidade de informação. Este é um dos principais problemas encontrados nos mecanismos de busca na *web* que geralmente encontram respostas às pesquisas, entretanto, tais respostas nem sempre contemplam os anseios do pesquisador.

Segundo Manning, Raghavan e Schütze (2008), este problema ocorre por que a matriz de incidência e a estrutura de índice invertido atribuem um mesmo valor de relevância para todos os termos do documento, o que impede qualquer tipo de ordenação dos documentos recuperados quanto ao seu grau de aderência à *query*. Neste tipo de abordagem existem apenas dois estados: o documento atende ou não atende à pesquisa; não existe a opção *atende parcialmente*. Todavia, os autores destacam a importância de um SRI apresentar os resultados ordenados pelo grau de proximidade do documento à necessidade de informação inicial. Nesta abordagem, os documentos que aparecem no topo desta lista ordenada são os considerados

mais relevantes, permitindo assim ao usuário acesso direto à informação desejada sem a necessidade de navegação por uma lista imensa de documentos sem qualquer critério de ordenação.

Uma forma de viabilizar a classificação pela relevância e apresentada por diversos autores seria representar os documentos como vetores de termos ( $t_1, t_2, t_3, t_4, \dots, t_n$ ) e para cada vetor é atribuído um valor de relevância deste em relação ao conteúdo do documento, chamado de *peso*. Baeza-Yates e Ribeiro-Neto (1999) citam que o valor do peso de cada termo no documento é calculado a partir de cálculos estatísticos, partindo-se do pressuposto onde um termo que ocorre diversas vezes em um número restrito de documentos tem maior relevância para estes do que outro termo que ocorre na maioria dos documentos de uma coleção.

<b>conhecimento</b>	Doc. 1   0,25	Doc. 3   0,70	Doc. 5   0,50	Doc. 8   0,10	Doc. 9   0,40
<b>disseminação</b>	Doc. 2   0,65	Doc. 4   0,2			
<b>documental</b>	Doc. 1   0,80	Doc. 2   0,40	Doc. 5   0,15	Doc.8   0,25	
<b>gestão</b>	Doc. 2   0,10	Doc. 5   0,85	Doc. 9   0,30		
<b>judiciário</b>	Doc. 4   0,60				

Tabela 3.3 - Representação ilustrativa de Índice Invertido com os respectivos pesos de relevância.

Voltando a pesquisa de documentos pelos termos *gestão* e *conhecimento*, com base nos pesos dos atribuídos apresentados na representação da tabela 3.3, conclui-se que *Doc. 5* tem mais relevância quanto à *query* do que *Doc. 9*.

A representação vetorial é o fundamento sob o qual foi construído o modelo de RI chamado Vetor Espacial, explanado na sequência.

### 3.3 Modelos de Recuperação da Informação

Para Baeza-Yates e Ribeiro-Neto (1999), as diferentes abordagens e premissas para recuperação e identificação de documentos relevantes em uma coleção levaram ao desenvolvimento de diferentes modelos de RI. Cada modelo estabelece sua forma de determinar o que é e o que não é relevante em uma consulta, implementando seu próprio conceito de relevância nos respectivos SRIs. O propósito desta seção é fazer uma abordagem teórica dos três modelos clássicos em RI: Booleano, Vetorial e Probabilístico.

### 3.3.1 Modelo Booleano

O modelo booleano é um modelo simples de RI que está baseado na teoria dos conjuntos e álgebra booleana, provendo uma estrutura lógica comum à maioria dos usuários de SRI. Neste modelo as *queries* são especificadas como expressões booleanas utilizando-se dos operadores lógicos AND, OR e NOT, (BORDOGNA, 1993).

Neste modelo os documentos são representados utilizando a estrutura de índices invertidos (tabela 3.2). Assim sendo, como nesta representação não existe distinção de relevância entre os termos, ou seja, a seleção dos documentos é realizada baseando-se sempre em critérios binários, onde o documento é considerado relevante ou não relevante, sem qualquer tratamento de relevância parcial. Conclui-se então que este modelo é muito mais um modelo de recuperação de dados do que recuperação de informação propriamente dita.

Sua simplicidade de implementação e claro formalismo fizeram com que este modelo se tornasse muito popular e fortemente utilizado nos SRIs comerciais. Em contrapartida, Baeza-Yates e Ribeiro-Neto (1999), destacam que este modelo apresenta algumas desvantagens em relação aos demais. Primeiro, nem sempre é fácil traduzir a necessidade informação para uma expressão booleana, este fator restringe o escopo das pesquisas disponíveis nos sistemas. Segundo, por se tratar de um modelo binário, não existe tratamento de parcialidade de relevância, assim, um documento muito próximo do que o usuário deseja pode não ser recuperado. Por fim, este modelo não possibilita qualquer tipo de ordenação quanto à relevância dos resultados recuperados, desta forma, os documentos mais importantes podem estar no final da lista, prejudicando assim o acesso do pesquisador.

### 3.3.2 Modelo Espaço Vetorial

De acordo com Baeza-Yates e Ribeiro-Neto (1999), as limitações existentes no modelo binário, onde não é possível a comparação parcial dos termos de pesquisa para recuperação de documentos, fomentaram as pesquisas para o desenvolvimento de um modelo de RI que representasse o documento conforme a relevância dos termos que o compõe para atender as necessidades não contempladas no modelo binário. Assim, Gerard Salton propôs o modelo Espaço Vetorial para ser utilizado em um sistema chamada SMART.

O modelo Espaço Vetorial representa uma coleção de documentos como vetores de termos com seus respectivos pesos de relevância (Tabela 3.3) em um mesmo ambiente vetorial com objetivo de viabilizar operações como o cálculo de relevância, a classificação e o agrupamento de documentos semelhantes, (MANNING, RAGHAVAN e SCHÜTZE, 2008) .

Neste modelo, o cálculo de similaridade entre dois documentos é realizado por meio de cálculos vetoriais como, por exemplo, a *similaridade por co-seno*, que calcula o co-seno do ângulo entre os seus respectivos vetores determinando a distância existente entre estes; onde, quanto menor for a distância, maior é o seu grau de semelhança. De forma análoga, em pesquisas documentais, a *query* de pesquisa também é convertida em um vetor de termos que, quando colocado no mesmo espaço vetorial, possibilita o cálculo de similaridade entre o vetor pesquisa e vetores dos documentos. Com isso, os documentos cujos vetores forem mais próximos do vetor pesquisa são classificados como os mais relevantes para a pesquisa em questão.

Ao contrário do que ocorre no modelo booleano, onde não existe parcialidade na recuperação por se tratar de um modelo binário, no modelo espaço vetorial é ampliado o escopo de busca pelo fato de que este modelo, ao aplicar os cálculos de similaridade, pode recuperar também documentos que atendem parcialmente aos termos de pesquisa informados.

### **3.3.3 Modelo Probabilístico**

O modelo Probabilístico, proposto inicialmente em 1976 por Roberston e Sparck Jones, é assim chamado por ser baseado nos conceitos oriundos da probabilidade e estatística. Assim, ao contrário do que ocorre no modelo Espaço Vetorial e análogo ao modelo Booleano, o modelo Probabilístico não armazena previamente informações relativas à relevância dos termos em seus respectivos documentos. A relevância dos documentos recuperados em relação aos termos da *query* é calculada dinamicamente no momento da pesquisa, (WANG e NG, 2003).

Conforme Baeza-Yates e Ribeiro-Neto (1999), o modelo busca resolver o problema de RI sob a perspectiva do princípio probabilístico, onde dada uma determinada *query*  $q$  e um documento  $d$ , o sistema estimará a probabilidade do documento  $d$  ser relevante para a necessidade da informação expressa na *query*. O

modelo assume que a probabilidade de relevância depende única e exclusivamente da *query* e do documento.

A proposta do modelo segue o seguinte raciocínio: a partir de uma *query* é possível recuperar um conjunto de todos os documentos relevantes para os termos de pesquisa informados. Este conjunto resultado é chamado de *conjunto resposta ideal*. Entretanto, faz-se necessário identificar, para cada *query*, as características dos documentos que deverão compor este conjunto ideal de resultado. Tem-se aqui outro problema, pois a princípio não é possível conhecer tais características. Com isso, partindo do princípio que existem determinados termos capazes de representar semanticamente o conteúdo dos documentos, pode-se afirmar que as propriedades do conjunto resposta ideal de uma *query* também pode ser composto por um conjunto de termos. Desta forma, o modelo propõe a geração de um conjunto de termos que tenham maior probabilidade de representar o conjunto resposta ideal. Com este primeiro conjunto ideal, é recuperado um conjunto inicial de documentos, iniciando assim uma operação interativa com o usuário que seleciona os documentos relevantes, descartando os sem relevância. Com base as respostas do usuário, o sistema ajusta as características do conjunto ideal e refaz a pesquisa. Pela repetição destes passos, espera-se a descrição do conjunto ideal fique muito próxima da realidade, recuperando assim os documentos mais relevantes.

Pelas características deste modelo, o processo e a estrutura de indexação são mais simplificados, pois não é necessário o cálculo prévio de relevância dos termos como ocorre no modelo espaço vetorial. Entretanto, como a proposta parte de um processo interativo com o usuário para melhorar o cálculo probabilístico em identificar o conjunto ideal, o modelo exige um tempo de aprendizagem para alcançar seu desempenho ótimo.

### **3.3.4 Comparação entre os Modelos de Recuperação da Informação**

Dos três modelos apresentados, o modelo booleano é considerado o mais limitado dentre os modelos clássicos. A maior desvantagem apontada seria o fator não permitir a comparação parcial entre os termos da *query* e os documentos o que leva uma busca com baixo desempenho, (MANNING, RAGHAVAN e SCHÜTZE, 2008).

Existe certa controvérsia em relação ao desempenho dos modelos Espaço Vetorial e o Probabilístico. Alguns autores citam experimentos onde o modelo

probabilístico tem melhor desempenho. Em contrapartida, outros autores também apresentam outros cenários em que o modelo Espaço Vetorial se apresentou mais adequado.

Porém, como o modelo probabilístico necessita de um tempo de “aprendizagem” para identificar corretamente o conjunto resposta ideal para as *queries*, percebe-se uma predominância no uso do modelo Espaço Vetorial, mesmo este demandando um custo maior de manutenção e armazenamento da estrutura de indexação.

### **3.4 Técnicas de Classificação**

A classificação de objetos em classes consiste de técnicas de análise capazes de agrupar itens em categorias conforme o grau de similaridade existente entre eles, (FRAKES e BAEZA-YATES, 1992). Em RI, estas técnicas são utilizadas na criação de grupos de documentos com o objetivo de melhorar a efetividade das pesquisas ou mesmo para realizar o agrupamento das publicações de uma determinada área do conhecimento de forma automatizada.

Segundo Ribeiro (2009), os métodos de classificação são categorizados em dois grandes grupos: *hierárquicos* e *não hierárquicos* ou *flat*. Os métodos hierárquicos agrupam são capazes de agrupar documentos em categorias relacionadas entre si, estabelecendo relações de hierarquia. Os métodos não hierárquicos de forma análoga agrupam uma coleção de documentos em categorias pela similaridade, porém sem estabelecer relações entre elas.

A classificação dos documentos em classes é realizada por meio de medidas comparativas de documentos, como a medida de distância de vetores, medidas de similaridade ou dissimilaridade, tais como: distância euclidiana, coeficiente de Dice e o coeficiente do co-seno, (FRAKES e BAEZA-YATES, 1992).

#### **3.4.1 Métodos Não Hierárquicos**

Os métodos não hierárquicos exigem um menor custo computacional para execução dos agrupamentos. São métodos basicamente heurísticos, onde são necessárias definições prévias relativas ao número de agrupamentos e seus respectivos tamanhos, critério de pertinência ou não pertinência de cada grupo e sua forma de representação.

Ribeiro (2009) cita alguns dos algoritmos utilizados para o agrupamento não hierárquicos. Alguns se baseiam na teoria de grafos que, a partir de uma matriz de

similaridade, agrega ou divide documentos. Outros são embasados na teoria da informação, densidade de pontos, ou aqueles que se utilizando do conceito de centróide, como o K-means.

Conforme o tamanho da coleção, a quantidade de combinações de documentos nos grupos torna inviável a busca por uma solução ótima. Assim sendo, os métodos não hierárquicos buscam encontrar uma solução próxima do ideal, utilizando usualmente o particionamento do conjunto de documentos para iniciar um processo de realocação até que os critérios de proximidade do ideal sejam satisfatórios.

Não obstante seu baixo custo computacional dos métodos não hierárquicos, os métodos hierárquicos são os mais utilizados atualmente pela estrutura relacional dos dados que estes métodos permitem obter.

### 3.4.2 Métodos Hierárquicos

Os métodos hierárquicos são capazes de produzir um alinhamento da coleção de documentos de forma a identificar relações entre documentos e entre categorias, permitindo uma visualização da estrutura hierárquica de uma coleção de documentos como, por exemplo, o dendograma apresentado na figura 3.1.

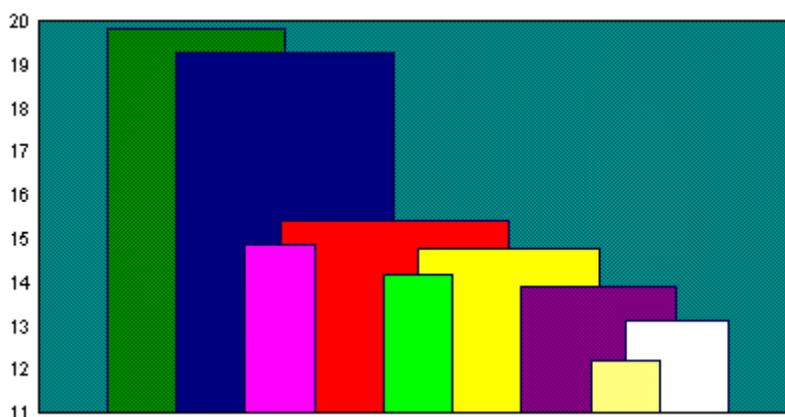


Figura 3.1 – Dendograma representando o relacionamento de dez classes de documentos.

Os principais algoritmos para agrupamentos hierárquicos, conforme Frakes e Baeza-Yates, (1992) são: *single link*, *complete link*, *group average link* e *Ward's method*. Apesar destes algoritmos partirem do mesmo princípio, ou seja, identificar relações entre documentos e classes, o que diferencia um algoritmo do outro é o cálculo da distância entre os objetos, ou seja, quão semelhantes ou quão diferentes eles são. Por exemplo, no algoritmo *single link* o cálculo é feito pela menor distância;

em contrapartida, no *complete link*, é utilizada a maior distância para identificar as relações entre as classes.

No início do processo de agrupamento, cada documento é considerado um grupo. No próximo passo, os dois documentos mais próximos, ou similares, se unem formando um novo grupo e uma nova iteração se inicia. Durante a execução deste método, percebe-se que em cada iteração o número de grupos é reduzido até finalmente todos os documentos façam parte de um grande grupo conforme apresentado previamente na figura 3.1.

### **3.5 Mineração de Texto**

Mineração de Texto, também conhecido por *Knowledge Discovery from Text* (KDT), diz respeito ao processo de extração de padrões de uma base de dados textual para possibilitar a descoberta de conhecimento. Com este intuito, a mineração de texto utiliza as mesmas funções de análise utilizadas em mineração de dados, bem como no processamento de linguagem natural e recuperação de informação, (MAIMON e ROKACH, 2005).

Feldman e Sanger (2006) definem Mineração de Texto “como um processo intensivo de conhecimento onde o usuário interage com uma coleção de documentos a todo tempo, através da utilização de um conjunto de ferramentas de análises”. Procura extrair informações úteis de fontes de dados através da identificação e exploração de padrões.

Para Silva (2005), a tecnologia de Mineração de Texto possibilita a identificação dos conceitos presentes nos textos. Tais conceitos representam entidades do mundo real (pessoas, organizações, eventos, sentimentos) e permitem entender os temas ou assuntos presentes direta ou indiretamente nos textos.

Concluída a extração de conceitos dos textos são aplicadas técnicas estatísticas sobre estes. Este processo analisa a frequência que estes conceitos aparecem na base textual analisada, além de verificar a relação existente entre eles, descobrindo assim possíveis associações ou dependências.

Para analisar e descobrir ocorrências de certos eventos dentro de uma determinada coleção de documentos são utilizadas três tipos comuns de funções: distribuições, frequências e associações. Feldman e Sanger (2006) explicam que essas funções oferecem a capacidade de descobrir mais de um tipo de padrões, e a mineração de textos possibilita aos usuários a habilidade de permutar as exibições

de diferentes tipos de padrões para um determinado conceito ou conjunto de conceitos.

Mladenic (2005) apresenta seis níveis para o processamento de textos, conforme apresentado na figura 3.2:

- a) Nível de palavras;
- b) Nível de sentenças;
- c) Nível de documento;
- d) Nível de coleção de documentos;
- e) Nível de coleção de documentos relacionados; e
- f) Nível de aplicação.

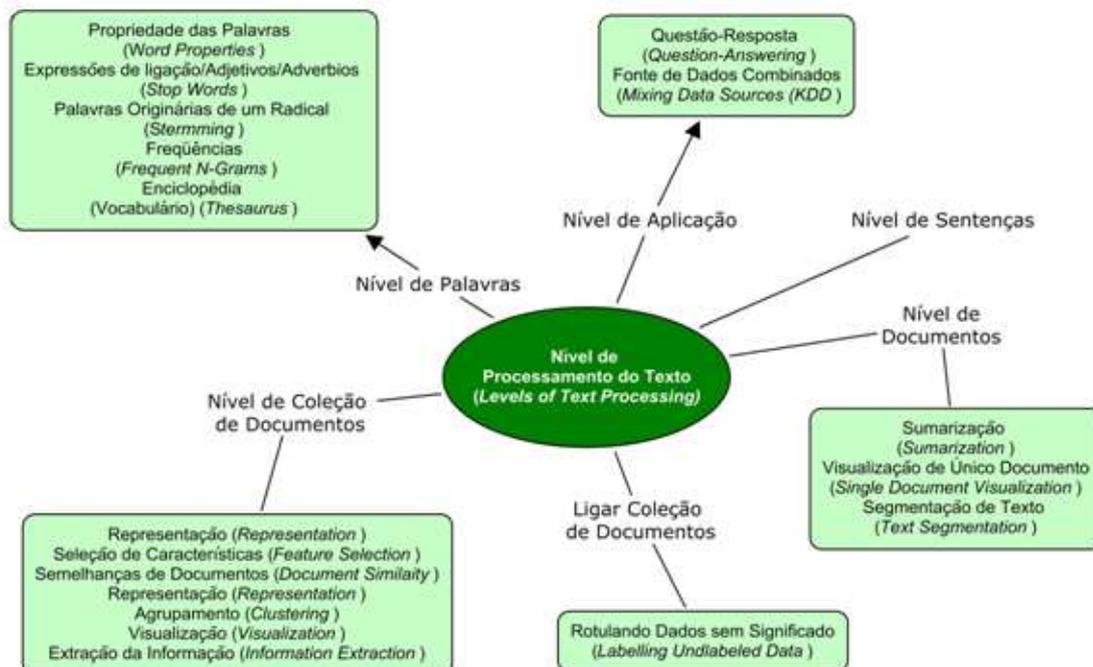


Figura 3.2 – Níveis de Processamento de Texto, adaptado de Mladenic (2005).

Por ser a base do processo, o tratamento no nível de palavras tem fundamental importância para os demais níveis. Berry (2004) relaciona algumas técnicas utilizadas neste nível de processamento. Inicialmente, é realizada a “limpeza dos dados”, onde são eliminadas informações adicionais que não fazem parte do conteúdo do documento, como: rótulos de linguagens como *html* e *xml* e caracteres especiais. Ainda nesta fase inicial, também são eliminadas as chamadas *stopwords*, que são palavras que não agregam valor semântico ao conteúdo dos documentos, tais como artigos, preposições, advérbios, dentre outros. Após esta fase, são aplicados algoritmos de *stemming*, cujo objetivo é identificar a raiz da palavra, descartando seus prefixos e sufixos. Assim, os termos agravado e

agravante são transformados para a mesma raiz agravo. Outra técnica apresentada no nível de palavras, é a identificação dos termos compostos, ou seja, termos que sempre aparecem seguidos um do outro como, por exemplo, agravo regimental. Concluindo esta fase, ocorre a normalização de termos, onde se utiliza de tesouros para fazer a substituição dos termos buscando a padronização da linguagem de indexação.

Os demais níveis de processamento utilizam-se da relação de termos remanescente para realização de análise estatística, associações e agrupamentos de documentos, identificação de relações de co-ocorrência de termos, visualização gráfica dos termos do documento, sumarização, dentre outras aplicações, (MLADENIC, 2005).

O processo de análise de documentos para extração de assuntos e termos para indexação está diretamente relacionado com praticamente todos os níveis de processamento supracitados.

### **3.6 Integração com Base de Conhecimento**

Os SRIs podem trazer uma quantidade imensa de informações como resultado de pesquisa aos termos solicitados. Entretanto, tais informações geralmente são apresentadas fora do contexto da pesquisa realizada. A integração de bases de conhecimento a estes sistemas pode ser aplicada com o intuito de melhorar os serviços de busca de forma a retornar somente conteúdos de uma determinada área.

A integração de uma base de conhecimento no processo de RI tem por objetivo restringir o espaço da busca documental para uma determinada área, sem levar em consideração documentos não relacionados com os conceitos formalizados nesta base.

Conforme Feldman e Sanger (2006), os conceitos contidos em documentos podem pertencer não somente ao seu contexto especificamente, porém ao domínio de uma determinada área do conhecimento. Assim, podem ser levantadas informações de bases de conhecimento com intuito de potencializar a execução das buscas para recuperação da informação necessária.

#### **3.6.1 Apresentação da Hierarquia de Conceitos no Documento**

A análise de documentos baseada em bases de conhecimento possibilita a visualização dos conceitos identificados como uma “árvore hierárquica”, conforme as

regras previamente estabelecidas. Este tipo de visualização é muito eficaz, visto que é apresentado um resumo visual do conteúdo dos documentos. Desta forma, é possível estabelecer relações entre os conceitos locais do documento e os conceitos, atributos e regras constantes na base de conhecimento, ampliando assim o conteúdo descrito no documento.

### **3.6.2 Flexibilidade na Apresentação dos Resultados das Consultas**

Uma abordagem possível a partir a integração descrita nesta seção é a apresentação dos resultados em forma de árvore hierárquica de conceitos, ou seja, ao invés de relacionar os documentos que contenham as palavras-chaves da pesquisa, apresentar a hierarquia de conceitos relacionados a tais palavras. Nesta mudança de paradigma o acesso aos documentos é realizado a partir dos conceitos de uma base de conhecimento (busca semântica) e não simplesmente através de palavras-chaves (busca sintática).

De posse das relações existentes entre os conceitos, é possível também identificar as relações existentes entre documentos. Desta forma, o usuário ao ler um determinado documento pode solicitar ao sistema a recuperação de outros documentos relacionados com o mesmo assunto do documento inicial, permitindo assim a navegação pelos documentos do acervo conforme se percorre os conceitos formalizados na base de conhecimento.

### **3.7 Recuperação Semântica de Documentos**

Em contrapartida à crescente produção documental observada atualmente nas organizações, percebe-se que as estratégias tradicionais de recuperação de documentos têm apresentado limitações inerentes aos problemas da linguagem natural tais como a ambigüidade, sinonímia, polissemia. Neste cenário, Lee, Hendler e Lassila (2001) propuseram o conceito de Web Semântica, cujo propósito é atribuir conteúdo semântico e formal aos documentos publicados na Internet, permitindo que agentes de softwares tenham a capacidade de interpretar o conteúdo destes documentos e realizar as operações para as quais foram designados.

Segundo Izza, Vincent e Burlat (2006), a semântica pode ser definida como um ramo do conhecimento, cujo objeto de estudo é o significado produzido pelas diversas formas de uma língua, abrangendo tanto o significado dos elementos que constituem as palavras (prefixo, radical e sufixo) como as próprias palavras como um todo e as expressões (frases) onde estas são inseridas. Destaca também alguns

conceitos da semântica que são de grande relevância para a construção de mecanismos de recuperação semântica de documentos:

- a) Sinônimos: formas lingüísticas que apresentam o mesmo significado, por exemplo: cansado e fatigado;
- b) Polissemia: propriedade em que um conceito pode ser expresso por diversas palavras, por exemplo: embargo (conceito jurídico) também pode ser entendido como impedimento;
- c) Ambigüidade: propriedade de uma palavra poder expressar conceitos diferentes, por exemplo: manga (fruta) e manga (parte de uma camisa).

Dos conceitos de semântica supracitados percebe-se a complexidade em conceber um mecanismo de busca capaz de atender a todas as minúcias semânticas existentes em uma determinada língua. Entretanto, Quillian (1968) cita o conceito de memória semântica que é a base para a construção do modelo computacional da memória representativa humana com intuito de utilizá-lo para obtenção do significado a partir de um determinado contexto.

Dentre os modelos propostos pela memória semântica, Arbib (2002) cita o modelo da rede semântica como um grafo direcionado, sendo que os conceitos representados pelos vértices e as arestas deste grafo representam as relações semântica existente entre os conceitos. Este modelo tem sido um importante fundamento teórico para o desenvolvimento de soluções tecnológicas com intuito de resolver os problemas da semântica na recuperação da informação.

Assim sendo, pode-se concluir que para transformar um mecanismo tradicional de busca que analisa o unicamente conteúdo sintático dos documentos para um mecanismo de busca semântica, faz-se necessário o acoplamento de mais um elemento ao sistema, um componente onde estejam representados os conceitos de uma determinada área do conhecimento, bem como suas inter-relações. Com isso, em um processo de recuperação, tal instrumento será capaz de localizar os documentos por sua essência temática, descartando os demais. Nesta proposta, a pessoa pode até receber um volume menor de informação como resultado de sua consulta, porém com maior qualidade, no sentido que estas informações estão relacionadas com a necessidade informacional do usuário consulente.

Nas próximas seções serão apresentados alguns instrumentos propostos com fim de formalizar domínios semânticos visando à recuperação semântica de documentos.

### **3.7.1 Indexação**

A indexação de documentos é uma prática comum e há tempos muito utilizada no cotidiano profissional dos que lidam com volumes expressivos de documentos. Esta consiste em uma atividade intensiva em conhecimento, cujo objetivo é identificar o conteúdo temático do documento por intermédio da seleção de palavras-chaves, ou termos de indexação, visando à rápida recuperação da informação nele contida pelo usuário.

#### **Indexação Clássica**

A NBR 12.676 (1992), da Associação Brasileira de Normas Técnicas (ABNT), conceitua a indexação como “o ato de identificar e descrever o conteúdo de um documento com termos representativos dos seus assuntos e que constituem uma linguagem de indexação”. Assim, a temática de um documento é definida pelo conjunto de conceitos presentes em seu conteúdo e a relação existente entre eles. Desta forma, a qualidade de uma indexação pode ser observada quando esta possibilita ao usuário identificar um número restrito de documentos que tenham total aderência ao assunto de interesse.

Brasil (2008) ressalta que a indexação não tem por objetivo apresentar um resumo do documento ou mesmo dispensar sua leitura, entretanto, fornece subsídios eficazes para que o usuário possa identificar se os assuntos tratados no documento é o que está se buscando, para então proceder com o exame do inteiro teor.

Dentre as ferramentas disponíveis ao profissional indexador de documentos, Cavalcanti (1978) apresenta o tesauro como “uma lista estruturada de termos associados, empregada por analistas de informação e indexadores para descrever um documento com a desejada especificidade, em nível de entrada, e para permitir aos pesquisadores a recuperação da informação que procuram”. Desta definição, pode-se concluir que um tesauro busca a padronização dos termos de indexação de uma determinada área de domínio. Um tesauro também tem a função de definir prioridade ao uso de termos mais comuns e conhecidos da área de domínio - chamados termos descritores - em relação a outros que, mesmo tendo semântica semelhante, não sejam de uso comum ou conhecido – termos não-descritores, Brasil (2008).

#### **Indexação Jurisprudencial**

Um documento jurisprudencial é o dispositivo onde são registradas as decisões pronunciadas pelos órgãos judiciários no efetivo desenvolvimento de sua função jurisdicional. (GUIMARÃES, 1994). Este tipo de documento serve como meio de prova de um determinado ato judicial ou mesmo como fonte para pesquisas futuras. Assim sendo, ao ser analisar um documento jurisprudencial, é importante identificar alguns tópicos que são peculiares, tais como o fato ocorrido, o Direito discutido, o posicionamento judicial e seu fundamento. (BRASIL, 2008).

Guimarães (1994) propõe uma metodologia para indexação de jurisprudência que se constitui em duas etapas básicas: *análise temática ou conceitual e tradução para linguagem controlada*.

#### Análise Temática ou Conceitual

Esta é a principal e mais complexa atividade para a indexação. Exige do indexador conhecimento do assunto a ser indexado, compreensão do texto, concisão e principalmente a capacidade de colocar-se no lugar do usuário que, no futuro, estará buscando por estes temas. (CAVALCANTI, 1978).

A atividade da análise temática inicia com a leitura minuciosa do documento, onde se busca identificar no texto cada um dos tópicos supracitados, peculiares a um documento jurisprudencial. Após a identificação dos tópicos no texto, é realizada a extração dos conceitos contidos em cada tópico, selecionando destes, apenas os que sejam relevantes para consultas futuras.

#### Tradução para Linguagem Controlada

Nesta etapa busca-se a representação dos assuntos identificados anteriormente em uma linguagem controlada e padronizada com intuito de facilitar a recuperação da informação. Esta linguagem também é chamada de artificial, no sentido de que é utilizado um vocabulário de termos – tesouro - de sintaxe própria com vistas à máxima uniformização dos termos de uma área de domínio. (BRASIL, 2008).

Nesta fase da indexação é feita a substituição dos termos não-descritores, menos comuns, desconhecidos ou obsoletos, pelos termos descritores, prezando sempre em manter preservado seu valor semântico.

#### **Métodos e Técnicas**

No processo tradicional de indexação, Brasil (2008) apresenta uma metodologia para execução desta atividade agrupamento o trabalho em três etapas.

Na primeira etapa sugere-se a leitura atenta e detalhada do documento, buscando identificar os conceitos presentes nos fatos narrados pelo autor. Após a conclusão desta etapa, o indexador precisa ter uma percepção concreta dos fatos ocorridos, do Direito em questão, das alegações das partes interessadas, o entendimento dos magistrados e os fundamentos da decisão que constituem basicamente as categorias de análise propostas nesta metodologia.

- a) Fato – Situação concreta que deu origem à questão em julgamento;
- b) Instituto Jurídico – É o tipo de Direito em discussão no âmbito dos fatos ocorridos;
- c) Entendimento – É a ligação que os magistrados estabelecem entre os fatos ocorridos e o instituto jurídico; e
- d) Fundamentação – É o conjunto de fundamentos sob os quais foram tomadas as devidas decisões relativas aos fatos descritos.

Assim sendo, a segunda etapa da metodologia de indexação apresentada é a identificação exaustiva dos conceitos existentes em cada uma das categorias supracitadas. O objetivo desta etapa é levantar informações que poderão contribuir fortemente para que o documento seja facilmente localizado em pesquisas que tratem especificamente de seu assunto. Isso geralmente exige uma atenção intensa do indexador para que nenhum conceito relevante seja ignorado.

Finalmente, de posse da relação de conceitos presentes no documento, a etapa conclusiva consiste em se selecionar desta lista, os conceitos mais relevantes para sua recuperação. Esta é uma atividade intensiva em conhecimento, visto que a decisão pela relevância de um termo em relação a outro dependerá da experiência do indexador. Este, com o passar do tempo na execução desta tarefa, irá conhecer em detalhes o acervo documental indexado, podendo assim ter subsídios mais concretos para as decisões necessárias neste processo.

### **3.7.2 Tesouros**

Conforme Eckert, Pfeffer e Stuckenschmidt (2008), é comum o uso de indexação baseada em tesouros para melhorar o resultado da recuperação de documentos. Ressaltam também que a atual produção documental tem tornado inviável a tradicional indexação manual, tornando os métodos estatísticos de indexação automática uma opção econômica e atrativa.

Entretanto, qualquer processo de indexação automática sem a utilização de uma base semântica tem uma abrangência essencialmente sintática, característica presente na grande maioria dos mecanismos de busca disponíveis atualmente, ou seja, analisa as palavras isoladamente não levando em consideração seu significado no contexto do documento. Neste contexto, Souza et al. (2008) apresentam iniciativas científicas que se utilizam de tesouros para implementação de estratégias semânticas de recuperação da informação.

O termo tesouro, é uma tradução do termo grego *thesaurus* cujo significado é “estoque de tesouros”. Inicialmente era utilizado como sinônimo de dicionário, ou seja, meramente uma relação de palavras e seus significados. Tal forma de utilização acabou entrando em desuso após a clássica obra de Peter Roget no século XIX, que estabeleceu o uso corrente do termo utilizado para fins de indexação documental, onde continha uma relação estruturada de conceitos e seus respectivos termos descritores. (MURAKAMI, 2005). Nesta abordagem, existe uma relação hierárquica entre os termos e seus descritores, bem como seus significados em uma determinada área de domínio. Souza et al (2008) acrescentam que os tesouros constituem uma relação de idéias ou conceitos afins, sem contudo deter o caráter estabelecer definições, característica fundamental dos dicionários. Courteau (2010), por sua vez, sumariza tesouro como um livro onde estão agrupadas palavras com significados similares ou semanticamente relacionados.

No final da década de 1950, surgiu o primeiro tesouro documental que era utilizado como apoio à atividade de indexação para extensão de seus vocabulários e padronização da linguagem utilizada para extração da essência temática dos documentos, tendo como objetivo facilitar sua recuperação em um processo de busca, até então realizado manualmente.

Quanto à construção de tesouros, Schneider (2005) apresenta três perspectivas:

- a) **Manual** – realizada por especialistas de uma determinada área do conhecimento, sendo estes responsáveis na criação dos termos deste domínio;
- b) **Automática** - realizada por softwares que com base em uma coleção de um domínio específico e utilizando-se de técnicas estatísticas e ocorrência e co-ocorrência dos termos nestes documentos, estabelecem o tesouro;

- c) **Semi-Automática** – realizada por especialistas utilizando-se de ferramentas de software para auxiliar na tarefa de identificação dos termos.

No âmbito da ciência da computação, Souza et al (2008), Eckert, Pfeffer e Stuckenschmidt (2008) destacam que o tesauro representa uma base de dados contendo tópicos semanticamente ortogonais e que pode ser utilizado em mecanismos de busca automatizados. As tecnologias contemporâneas têm ampliado as opções de uso e permitido a integração deste importante instrumento, inicialmente específico da área de biblioteconomia, com ferramentas para apoiar o exercício profissional de diversas áreas. Arano (2005) destaca os benefícios advindos com a integração do tesauro no meio eletrônico:

- a) O hipertexto facilitou ao usuário a navegação entre os conceitos definidos no tesauro, permitindo um rápido acesso a todos os elementos presentes no tesauro (termos relacionados, descrições, notas, referências);
- b) Redução no custo de atualização e manutenção do conteúdo do tesauro, por meio de ferramentas específicas para este fim, tornando obsoleto o uso do tesauro em meio físico (papel);
- c) Maior interação do usuário no processo de criação, gerenciamento e otimização do tesauro por meio de testes de usabilidade e técnicas de modelagem conceitual;
- d) Possibilidade de integração com outras ferramentas permitindo o reuso e a interoperabilidade do conhecimento explicitado como apoio para execução de atividades diversas.

### **3.7.3 Ontologias**

A palavra ontologia origina-se do grego e é formada pela junção das palavras *ontos* (ser) com *logos* (palavra). Este termo, introduzido por Aristóteles, fora utilizado inicialmente como uma área da filosofia que estuda a natureza e a organização do ser e suas propriedades. Da filosofia tem-se a definição dada por Corazzon (2000), onde “Ontologia é a teoria dos objetos e seus vínculos. Provê critérios para distinguir os vários tipos de objetos (concretos e abstratos; existentes e não existentes; real e ideal; independente e dependente) e seus vínculos (relações, dependências e predicados)”.

Atualmente, o termo ontologia também está inserido na área da Inteligência Artificial como meio de representação formal do conhecimento, com o propósito de facilitar o compartilhamento deste entre pessoas e sistemas. (FENSEL, 2002).

Para Neches (1991), “Uma ontologia define os termos básicos e relações compreendendo o vocabulário de uma área específica, bem como as regras para combinação entre termos e relações para definir extensões do vocabulário”.

Gruber (1993) complementa, afirmando que “Uma ontologia é uma especificação formal e explícita para um conceito compartilhado”.

Da definição de Gruber, Fensel (2002) ressalta os termos formal, explícita e compartilhada. Formal, pelo fato de que uma ontologia deve ser compreendida e processada por sistemas. Explícita significa que os conceitos utilizados, bem como as restrições sobre seu uso, são explicitamente definidos. Compartilhado no sentido que uma ontologia reflete o conhecimento consensual sobre um determinado assunto por uma comunidade.

A ontologia pode ser compreendida de forma análoga ao conceito de mapa conceitual, oriundo das ciências cognitivas, que representa graficamente um conjunto de termos de uma determinada área e as relações existentes entre eles. Esta forma de explicitar o conhecimento permite a criação de uma taxonomia dos termos, estabelecendo relações entre conceitos gerais e abstratos até conceitos concretos e específicos.

A figura 3.3 apresenta uma definição ilustrativa de uma ontologia do Direito Eleitoral Brasileiro. Além de apresentar os conceitos gerais e específicos, é possível visualizar os diferentes tipos de relacionamentos entre eles. Por exemplo, o conceito *Governador* tem um relacionamento com o conceito *Cargo Público* do tipo “é um”, assim pode-se concluir: “Governador é um cargo público”; já, entre os conceitos *Coligação* e *Partido Político*, existe um relacionamento do tipo *é composto por*, desta relação tem-se: “uma coligação é composta por partidos políticos”.

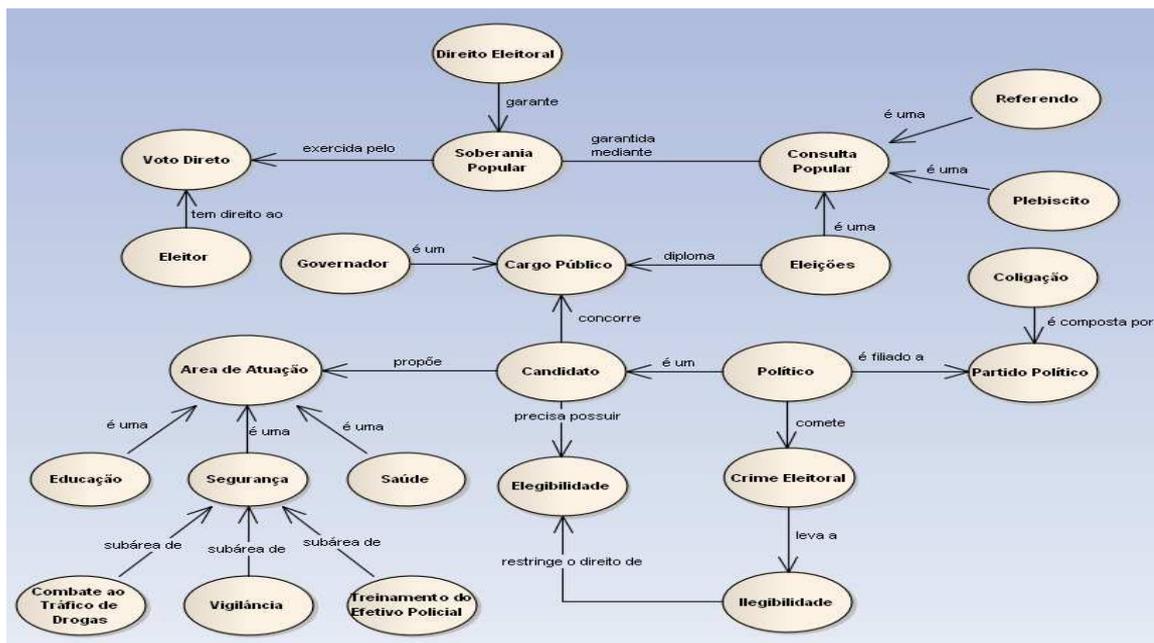


Figura 3.3 – Ontologia ilustrativa para o Direito Eleitoral.

Além da representação dos conceitos e suas relações, uma ontologia também é constituída de axiomas, ou seja, regras que possibilitam inferências sobre os conceitos. Por meio dos axiomas é possível elaborar sentenças verdadeiras entre conceitos que não possuam uma ligação direta. Um exemplo de inferência baseada em axioma pode ser observado na ontologia ilustrativa da figura 3.3. Das relações existentes conclui-se: que um candidato precisa possuir elegibilidade para concorrer a um cargo público; e que na situação em que um Político é condenado por um crime eleitoral, este se torna inelegível. Desta forma, mesmo não existindo uma relação direta entre os conceitos *Candidato* e *Crime Eleitoral* é possível para um agente computacional, por inferência, concluir que um crime eleitoral impede a candidatura de um político.

## **4 METODOLOGIA DO TRABALHO**

Este capítulo versará sobre os procedimentos metodológicos realizados para o desenvolvimento da presente pesquisa, com intuito de estabelecer um alinhamento entre o problema exposto e as atividades realizadas para a apresentação da solução proposta.

Inicia-se então com um sucinto levantamento bibliográfico relativo à metodologia científica visando delimitar a classificação deste trabalho dentro das diversas perspectivas existentes no campo científico de pesquisa.

Em seguida, será abordado sobre a metodologia CommonKADS, concebida para o desenvolvimento de soluções intensivas em conhecimento, apresentando como estudo de caso inicial da pesquisa a aplicação desta metodologia no Tribunal Regional Eleitoral de Santa Catarina.

### **4.1 Classificação da Pesquisa**

O conhecimento científico é caracterizado por ser obtido por meio da aplicação de um método. Este, definido como o conjunto de etapas e processos executados durante a investigação dos fatos ou na procura da verdade, é a linha de raciocínio adotada no processo de pesquisa e pode ser dedutivo, indutivo, hipotético-dedutivo e dialético. (RUIZ, 1991).

O método utilizado neste trabalho se classifica como hipotético-dedutivo, pois a partir de conhecimento prévio sobre o assunto e dadas as dificuldades existentes no campo da gestão documental no âmbito do Judiciário foi constituído o problema de pesquisa. Este foi o fundamento para a formulação das conjecturas e hipóteses elaboradas para comprovar fatos e verdades, complementando a lacuna de conhecimento identificada.

A investigação do problema foi efetuada por intermédio de um estudo de caso, aplicando-se a metodologia CommonKads, onde foi realizada uma análise observacional dos processos de GED junto ao Tribunal Regional Eleitoral de Santa Catarina (TRESC). Assim sendo, a natureza do estudo é classificada como pesquisa aplicada, pois visa à geração de conhecimento para aplicação prática dirigidos à solução de problemas específicos. (GIL, 1991).

Do ponto de vista da abordagem do problema, caracteriza-se como pesquisa qualitativa no sentido que, com base nas observações realizadas, buscou-se

identificar as relações entre pessoas, ferramentas e processos para validação das hipóteses identificadas.

Além dos métodos e técnicas mencionados, foi realizada uma pesquisa bibliográfica para levantamento das publicações técnico-científicas relacionadas aos temas abordados neste trabalho. Para tal atividade, buscou-se selecionar os principais autores de cada área do conhecimento com o intuito de disponibilizar uma visão ampla e atualizada dos temas, acrescida do senso crítico para interconectar os temas para elaboração do modelo de gestão documental proposto.

Com base na pesquisa bibliográfica e visando à validação do modelo foi desenvolvido um protótipo de uma ferramenta tecnológica utilizando-se de técnicas de GED e da Engenharia do Conhecimento. Tal protótipo foi aplicado junto ao TRESK, e tratou especificamente dos documentos jurisprudenciais desta organização.

## **4.2 Metodologia CommonKads**

A metodologia CommonKads foi proposta inicialmente 1983 por Schreiber et al. (2002) objetivando atender as demandas das indústrias por sistemas baseados em conhecimento. A proposta dos autores era estabelecer um conjunto de procedimentos e técnicas específicas de engenharia do conhecimento que possibilitassem o desenvolvimento destes sistemas em larga escala de uma maneira estruturada, controlada e repetível; aos moldes que do já estava ocorrendo no campo da engenharia de software. Assim sendo, parte-se do pressuposto que o conhecimento pode ser modelado, ou seja, a partir da experiência de um especialista é possível formalizar seu conhecimento e utilizá-lo em sistemas de apoio à decisão.

Nesta perspectiva, para o CommonKads o conhecimento é considerado como um componente de software que, assim sendo, poderá ser compartilhado, distribuído e reaproveitado por diferentes aplicações até que se torne obsoleto.

Freitas (2003) apresenta a metodologia CommonKads como tendo características de outras metodologias baseadas em modelos que abrange diversos aspectos de um projeto de sistema de conhecimento: análise organizacional, representação e modelagem do conhecimento, comunicação, integração e implementação. Neste sentido, Schreiber et al. (2002) acrescenta que a metodologia estrutura-se em cinco pilares que constituem a pirâmide metodológica que

perpassam desde a visão de mundo, sobre a qual estão fundamentados seus conceitos, até o uso onde são efetivamente aplicados os sistemas desenvolvidos.

Na figura 4.1 estão relacionados os cinco pilares do CommonKads sendo que esta pirâmide tem por fundamento a visão da engenharia do conhecimento baseada em modelos e na reutilização de padrões de conhecimento. Subindo os pilares, tem-se uma estrutura teórica que disponibiliza uma série de notações gráficas e textuais como planilhas e documentos pré-estruturados, seguidos por métodos que definem o modelo ciclo de vida do conhecimento, modelagem do processo de desenvolvimento com diretrizes e técnicas de elucidação. No topo desta estrutura estão as ferramentas de desenvolvimento e os ambientes de implementação, para então chegarmos ao objetivo principal: a aplicação prática do produto de todo esse arsenal metodológico, ou seja, um sistema intensivo em conhecimento.

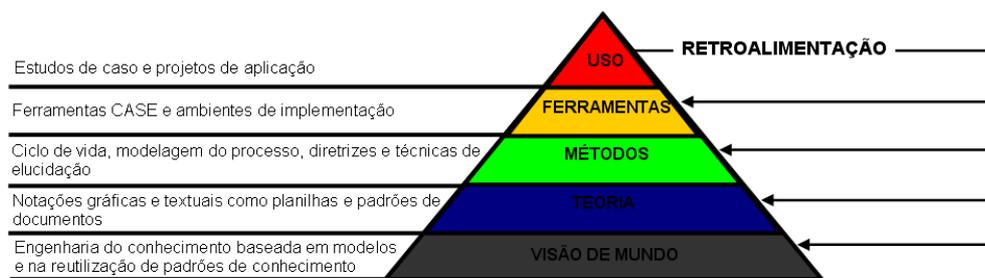


Figura 4.1 – Pirâmide metodológica do CommonKads, adaptado de Schreiber et al., (2002).

Sendo uma metodologia baseada em modelos, o CommonKADS faz um agrupamento dos modelos em três categorias, na medida em que estes modelos contribuam para o esclarecimento de três importantes questionamentos que devem ser elucidados durante o ciclo de desenvolvimento dos sistemas de conhecimento:

- a) Por quê?
- b) O que?
- c) Como?

A figura 4.2 ilustra como estão agrupados os modelos do CommonKads. O primeiro grupo, contextual, constituído dos modelos de organização, tarefa e agente, tem por objetivo responder ao primeiro questionamento, “Por que?”. O grupo conceitual – modelo do conhecimento e de comunicação buscam a resposta para o “O que?” e o grupo artefato responde ao “Como?”.

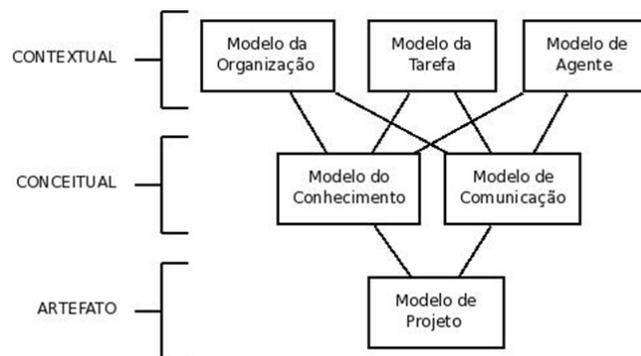


Figura 4.2 – Modelos propostos pelo CommonKads, Schreiber et al., (2002).

Na fase Contextual da metodologia é analisado o contexto organizacional e o ambiente onde, em princípio, existe a necessidade de um sistema de conhecimento. A questão principal é identificar o problema a ser resolvido, buscando entender por que se faz necessária a implantação de uma solução desta natureza e quais os impactos na organização e nas pessoas envolvidas para, a partir de então, decidir continuidade ou não do projeto de desenvolvimento.

Tendo sido identificado o contexto e optado pela continuidade do desenvolvimento, na fase Conceitual o foco é compreender a natureza do conhecimento e das comunicações envolvidas no âmbito organizacional. Assim, busca-se uma descrição conceitual do conhecimento utilizado pelos agentes na execução de suas tarefas e como se dá a comunicação entre estes agentes.

Finalmente, na fase de Artefato são levados em consideração os aspectos técnicos necessários para o projeto e a implementação do sistema em questão, ou seja, como formalizar o conhecimento identificado de maneira que este sirva de insumo para que uma ferramenta tecnológica seja capaz de resolver ou auxiliar na resolução dos problemas identificados nas etapas anteriores.

### 4.3 Estudo de Caso

Como parte do processo metodológico adotado nesta pesquisa, foi utilizada a metodologia CommonKads para elucidação do ambiente organizacional e cultural do TRESA, onde foram identificadas as atividades intensivas em conhecimento desempenhadas na gestão dos documentos jurisprudenciais, buscando identificar os problemas e as oportunidades de melhoria nestes processos e decidir se o desenvolvimento de uma solução baseada em conhecimento pode contribuir efetivamente para melhoria nos serviços prestados.

Neste sentido, serão aplicados os três modelos contextuais do CommonKads, utilizando-se dos documentos e planilhas propostos por Schreiber et al. (2002).

### 4.3.1 Modelo de Organização

O modelo de organização apóia a análise das principais características da organização, a fim de identificar os problemas e oportunidades e decidir pela viabilidade ou não do desenvolvimento de um sistema baseado em conhecimento. São objetos deste estudo: a estrutura e cultura organizacional; os processos de negócios; as atividades desempenhadas; as pessoas; e os recursos envolvidos.

Como resultado deste modelo, espera-se pela viabilidade do projeto de desenvolvimento, pelo mapeamento dos conhecimentos envolvidos e as atividades onde estes são aplicados, bem como pela identificação das pessoas que detêm tais conhecimentos e deles se utilizam.

Modelo de Organização	Problemas e Oportunidades OM – 1
Problemas e Oportunidades	<p>Não existe uma base de dados única para os documentos jurisprudenciais.</p> <p>Os documentos, após aprovados, são impressos para assinatura dos magistrados e digitalizados para disponibilização do documento com a assinatura manuscrita.</p> <p>A classificação dos documentos é realizada por agente humano e exige a leitura dos documentos antes de sua publicação.</p> <p>O formato de publicação dos documentos impossibilita a pesquisa por seu conteúdo.</p>
Contexto Organizacional	<p><b>Negócio:</b> Processo Eleitoral para Soberania Popular.</p> <p><b>Missão:</b> Assegurar a legitimidade do processo eleitoral, visando ao exercício da soberania popular.</p> <p><b>Visão:</b> Ser reconhecida pela sociedade como uma instituição inovadora, confiável, independente e efetiva na consolidação da soberania popular.</p> <p><b>Valores:</b></p> <ul style="list-style-type: none"> <li>● <b>Acessibilidade:</b> viabilizar de forma pro ativa o acesso à Justiça Eleitoral de Santa Catarina, para oferecer à sociedade a certeza de um meio seguro para resolver as questões a ela submetidas.</li> <li>● <b>Cidadania:</b> respeitar o conjunto de liberdades e obrigações relativo aos direitos individuais, políticos, sociais e econômicos.</li> <li>● <b>Conscientização política:</b> estimular o pensamento crítico para que os indivíduos participem ativamente das decisões políticas da comunidade.</li> <li>● <b>Desenvolvimento Humano:</b> incentivar o aperfeiçoamento profissional e pessoal, por meio da capacitação e da formação de um ambiente que permita ao servidor uma vida criativa e produtiva.</li> <li>● <b>Ética:</b> conduzir as ações da Justiça Eleitoral de Santa Catarina na busca do interesse público com responsabilidade social, respeitando os princípios que norteiam a administração pública e os valores morais da sociedade.</li> <li>● <b>Eficiência:</b> realizar as ações com emprego criterioso e otimizado de recursos, atingindo resultados com excelência.</li> <li>● <b>Excelência:</b> atuar de maneira planejada e inovadora na busca da satisfação do cidadão e da sociedade na condição de usuários e destinatários dos serviços prestados pela Justiça Eleitoral de Santa Catarina.</li> <li>● <b>Imparcialidade:</b> agir de forma isenta, sem distinções ilegítimas.</li> <li>● <b>Integração:</b> propiciar um ambiente de informação, diálogo, cooperação e interação entre pessoas, unidades e áreas da Justiça Eleitoral de Santa Catarina, visando à constante troca de conhecimentos e à valorização da participação de cada indivíduo como parte de um único corpo funcional.</li> <li>● <b>Transparência:</b> dar visibilidade às ações da Justiça Eleitoral de Santa Catarina, permitindo à sociedade verificar se a instituição está atuando em consonância com os princípios e valores da administração pública.</li> </ul> <p><b>Fatores Externos:</b> Instituição reconhecida pela sociedade pela inovação, excelência e transparência na condução do processo eleitoral.</p>
Soluções	<p>Criar um repositório único para os documentos jurisprudenciais.</p> <p>Implantar a tecnologia de assinatura digital no gabinete dos magistrados.</p> <p>Implementar uma solução tecnológica que efetue a atividade de classificação, indexação e publicação dos documentos.</p>

Tabela 4.1 – Modelo Organizacional de Problemas e Oportunidades.

Na tabela 4.1 estão representados os problemas e oportunidades, além da missão da Instituição e seus valores estratégicos. As informações deste contexto em geral são invariantes, e decisões posteriores devem ser julgadas sob esta perspectiva.

Os aspectos variantes da organização estão relacionados na tabela 4.2, onde são identificadas as pessoas e os processos que podem ser afetados ao término da implantação da metodologia. Desta feita a estrutura orgânica da Instituição está contemplada no anexo I e a figura 4.3 apresenta o processo chave deste estudo de caso que delimita o escopo de aplicação da pesquisa.

Além dos fatores supracitados, a planilha de aspectos variantes introduz a identificação dos conhecimentos envolvidos na execução do processo de negócio objeto da aplicação e apresenta algumas questões relativas à cultura organizacional, onde tais conhecimentos são utilizados.

A cultura e os recursos disponíveis são fatores de grande influência no dia a dia de uma Instituição. Estes podem contribuir ou ser um grande obstáculo para o desenvolvimento de uma nova solução. Assim sendo, é imprescindível levar estas questões em consideração desde o início do trabalho, para delinear as decisões que serão tomadas durante o ciclo de vida do desenvolvimento da solução.

Modelo de Organização	Planilha de Aspectos Variantes OM – 2
Estrutura	Anexo 1
Processo	Figura 1
Pessoas	<p>O quadro de pessoal da instituição é formado por algumas categorias distintas:</p> <p><b>Juízes Eleitorais:</b> Compõe o quadro de outros tribunais e exercem a magistratura eleitoral por períodos temporários de 2 a 4 anos.</p> <p><b>Servidores:</b> Pessoas que ingressaram na justiça eleitoral por meio de concurso público.</p> <p><b>Requisitados:</b> Pessoas que estão exercendo atividades no TRESA que são cedidos por outras instituições públicas.</p> <p><b>Terceirizados:</b> Pessoas que prestam serviços aos TRESA, mas que são contratadas por outra empresa prestadora de serviços.</p> <p>A maioria das pessoas tem formação ou ampla experiência na área jurídica. Entretanto, existem pessoas de diversas áreas atuando em conjunto.</p>
Recursos	<p>Possui um amplo parque tecnológico (computadores e softwares) atualizado com as tendências do mercado.</p> <p>A evolução da conectividade dos sistemas tem demandado a ampliação da infraestrutura de rede com os cartórios eleitorais do interior. Esta solução está em processo de contratação.</p>
Conhecimento	<p>O conhecimento fim do TRESA é o Direito Eleitoral, entretanto, existem outras áreas chamadas de áreas meio, que se utilizam de conhecimentos em Engenharia Civil, Medicina, Odontologia, Tecnologia da Informação e Comunicações.</p>
Cultura e Poder	<p>Os servidores têm um plano de carreiras e, após avaliação em estágio probatório, adquirem estabilidade no emprego, fator traz segurança aos servidores refletindo no desempenho de suas atividades.</p>

Tabela 4.2 – Modelo Organizacional dos Aspectos Variantes.

A figura 4.3 representa graficamente o processo de gestão dos documentos jurisprudenciais. Como citado, além de delimitar as fronteiras de atuação, este processo apresenta uma perspectiva inicial das atividades desenvolvidas e facilita a

identificação dos problemas e oportunidades de melhoria. O primeiro aspecto está no processo de captura dos documentos; pelo fato que estes são produzidos por diversas unidades funcionais, não existe um repositório único de armazenamento, ou seja, cada documento é armazenado na unidade que o originou. Devido à necessidade legal, os documentos devem ser assinados pelos magistrados, assim sendo, após a materialização e devidas assinaturas, o documento é encaminhado para uma unidade intermediária procede com a digitalização e publicação na Internet. A centralização e a guarda permanente dos documentos são realizadas fisicamente em uma estrutura devidamente preparada para este fim, a Seção de Arquivo.

Com objetivo de facilitar a recuperação dos documentos, é realizada uma seleção prévia quanto à importância dos documentos e somente os selecionados passam pela indexação manual, onde um serventuário, após a leitura, identifica os termos essenciais do texto e atualiza uma base chamada de *Índice Temático*. Neste ponto é possível vislumbrar uma oportunidade de melhoria, pois em decorrência do volume de documentos produzidos, a automação desta atividade minimizará o tempo entre o recebimento do documento e sua indexação.

Outro aspecto a ser ressaltado é que, apesar de existir uma ferramenta de busca de documentos por conteúdo, esta não abrange todas as categorias de documentos produzidos. Isso ocorre, pois o documento disponibilizado na Internet é uma cópia digitalizada de um documento físico, o que dificulta a implementação da pesquisa, pois sendo uma digitalização, o conteúdo é tratado como uma imagem e não mais como texto, impedindo a indexação automatizada. A ausência desse mecanismo limita consideravelmente o acesso ao conhecimento contido neste acervo.

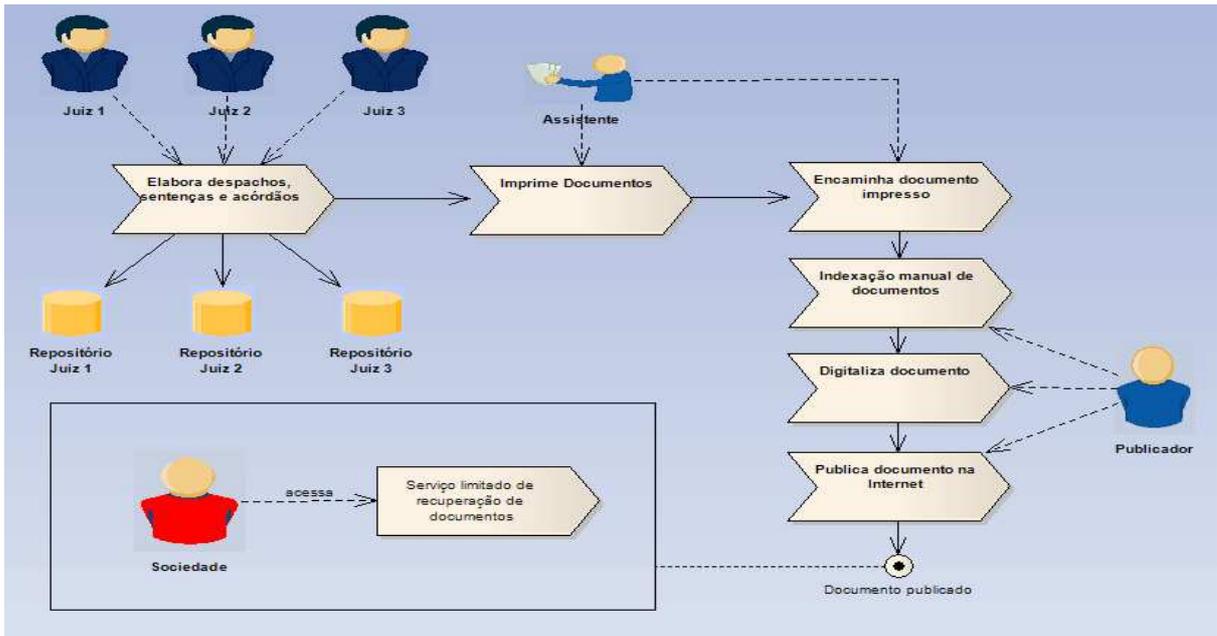


Figura 4.3 – Processo de Gestão de Documentos Jurisprudenciais do TRESA.

Com base no processo definido, a tabela 4.3 faz um detalhamento das etapas que constituem a gestão documental em questão, identificando as pessoas envolvidas, onde as atividades são executadas, quais são os insumos de conhecimento e qual a relevância de cada atividade para o processo como um todo. Uma função importante fase etapa é realizar a distinção entre as atividades intensivas em conhecimento das demais, visto que o foco de abrangência desta metodologia limita-se a tarefas desta natureza.

Em se tratando especificamente das tarefas intensivas em conhecimento, a tabela 4.4 complementa as informações anteriores apresentando, para cada atividade previamente classificada: quais os tipos de conhecimento necessários para sua execução; quem detém estes conhecimentos; além de informar o resultado da análise que verifica se estas atividades estão sendo realizadas da forma, momento e lugar corretos, por pessoas que possuam competência para tal; e, por fim, se a qualidade do produto resultado destas atividades está dentro de um padrão aceitável.

Com base nestas análises é possível identificar claramente as atividades que poderiam sofrer intervenção por meio de automação advinda com um sistema de conhecimento. Conseqüentemente, também é possível vislumbrar os conhecimentos que deverão ser modelados e formalizados como componentes deste futuro sistema. Conforme o resultado deste estudo apresentado na tabela 4.4, é possível concluir que a atividade desempenhada pelos juizes e promotores para prolatar as

sentenças, acórdãos e despachos, além de ser de difícil automação, estão sendo executadas conforme os padrões estabelecidos pela metodologia. Entretanto, a indexação e classificação de documentos para recuperação posterior, pode ser objeto de um sistema de conhecimento visando à melhoria da qualidade do serviço prestado.

Modelo da Organização		Decomposição do Processo – OM-3			
Nome da Tarefa	Executada por	Onde	Insumos de Conhecimento	Intensiva em conhecimento	Relevância
Elaboração de despachos, sentenças e acórdãos	Juiz e Promotor Eleitoral	Gabinete dos Juizes	Direito Eleitoral - decisão judicial	Sim	5
Impressão de documentos	Técnico Auxiliar	Gabinete dos Juizes		Não	1
Indexação e Classificação de documentos	Classificador de jurisprudência	Seção de Jurisprudência	Direito Eleitoral - classificação de documentos	Sim	4
Digitalização de documentos	Digitalizador	Seção de Gestão de Documentos		Não	2
Publicação de documentos na Internet	Gerente de Conteúdo Web	Seção de Gestão de Documentos		Não	2

Tabela 4.3 – Modelo Organizacional da Decomposição dos Processos.

Modelo da Organização			Insumos de Conhecimento – OM-4			
Conhecimento	Possuído por	Usado em	Forma correta?	Lugar correto?	Momento correto?	Qualidade correta?
Direito Eleitoral para decisão judicial	Juiz e Promotor Eleitoral	Elaboração de despachos, sentenças e acórdãos	Sim	Sim	Sim	Sim
Direito Eleitoral para classificação de documentos	Classificador de jurisprudência	Indexação e Classificação de documentos	Não: Manual → Automática	Sim	Não	Sim

Tabela 4.4 – Modelo Organizacional – Insumos de conhecimento.

Neste ponto de aplicação da metodologia, é possível apresentar um resumo das principais implicações para o desenvolvimento de sistema baseado em conhecimento nesta organização. Na tabela 4.5 é apresentada a viabilidade do projeto de várias perspectivas, além de propor ações para continuação do desenvolvimento do projeto.

Sob o ponto de vista do negócio da Instituição, percebe-se que a melhoria do serviço de disponibilização dos documentos jurisprudenciais para a Sociedade está diretamente relacionada com o contexto organizacional.

Dentro da perspectiva técnica, baseando-se em pesquisas na área de RI e nos recursos de conhecimento que a Instituição detém e os utiliza de forma manual, conclui-se da viabilidade técnica para a criação de solução tecnológica. Seu objetivo seria auxiliar a execução da atividade de indexação e classificação automática de documentos utilizando-se do conhecimento que está atualmente no domínio do profissional classificador de jurisprudência.

Modelo da Organização	Checklist para decisão sobre viabilidade – OM-5
Viabilidade do Negócio	A implantação de um Sistema de Gestão do Conhecimento como ferramenta de automação da tarefa de classificação e indexação dos documentos jurisprudenciais é de fundamental relevância para diminuir o tempo de publicação, bem como melhorar a qualidade da indexação com o intuito de melhorar os serviços de buscas ao conhecimento contido nestes documentos.

Viabilidade Técnica	Para automação da classificação e indexação dos documentos fazem-se necessários a implementação ou aquisição de uma ferramenta de mineração de texto ( <i>text mining</i> ) e a definição de um tesouro para a Justiça Eleitoral, tecnologias viáveis e acessíveis para a instituição que possui profissionais técnicos capacidades nas áreas de Tecnologias de Informação e Direito Eleitoral.
Viabilidade do Projeto	Projeto de acordo com o planejamento estratégico da Instituição. A instituição deter o conhecimento necessário para automatizar a tarefa de classificação e indexação. Existe a disposição orçamentária para o projeto.
Ações Propostas	Definir a equipe do projeto. Elaborar o planejamento para execução do projeto. Decidir pelas ferramentas de mineração de texto e edição de ontologias.

Tabela 4.5 – Modelo Organizacional – Lista para decisão sobre viabilidade.

### 4.3.2 Modelo de Tarefa

Seguindo a metodologia de Schreiber et al (2002), passa-se a focar o modelo de tarefa. As tarefas são partes relevantes de um processo de negócio. O modelo de tarefa analisa o leiaute da tarefa global, suas entradas, saídas, pré-condições e critérios de execução, bem como os recursos e as competências necessárias. Especificamente, é importante destacar nas tarefas a serem realizadas a indicação desta ser intensiva em conhecimento, o grau de prioridade e a competência associada.

Na tabela 4.6 são apresentados os detalhamentos da análise das tarefas identificadas previamente como intensivas em conhecimento, retomando algumas informações já identificadas nas etapas anteriores, acrescentando porém novas minúcias como a forma de manuseio do conhecimento, agentes envolvidos, recursos alocados, e qualidade de execução requerida.

Tarefa	Subtarefa	Meta	Intensiva em conhecimento	Prioridade	Competência
Elaboração de Sentenças, Acórdãos e Despachos	Analisar o processo	Obter subsídios para fundamentar uma decisão.	Sim	Alta	Juiz e Promotor Eleitoral
	Elaborar texto preliminar da decisão	Com base nas conclusões sobre o processo, escolher um modelo de documento pré-determinado para a decisão tomada.	Sim	Baixa	Assessor
	Aprovar e assinar documento decisório	Finalizar o processo de elaboração de sentença.	Sim	Média	Juiz e Promotor Eleitoral
	Encaminhar documento para publicação	Publicar decisão proferida.	Não	Baixa	Assessor
Classificar e Indexar Documentos	Tomar ciência do conteúdo do documento.	Identificar o assunto que se trata o documento.	Sim	Alta	Classificador de Documentos Jurisprudenciais
	Classificar o documento	Com base no conteúdo documento, definir em qual das categorias pré-existentes este tem maior aderência.	Sim	Alta	Idem
	Atualizar índice temático	Após classificação, atualizar a tabela com a indexação dos termos relevantes do documento.	Sim	Alta	Idem
	Verificar surgimento de novos termos	Acrescentar novos termos na base de índice temático.	Sim	Média	Idem
	Publicar documento	Disponibilizar documento para consulta na Internet.	Não	Baixa	Idem

Tabela 4.6 – Relação parcial de tarefas desenvolvidas no departamento.

Nas tabelas 4.7 e 4.8, o enfoque passa unicamente para as atividades identificadas como objeto de automação, no caso desta aplicação, a atividade *Classificar e Indexar Documentos*. Nestes quadros é analisado o detalhamento da tarefa relacionada com a natureza, formato e disponibilidade do conhecimento, relatando também um indicativo quanto aos gargalos e oportunidades de melhorias.

Modelo de Tarefa	Análise de Tarefas – TM-1
Tarefa	Classificar e Indexar Documentos
Organização	Coordenadoria de Gestão da Informação
Objetivo ou Valor Agregado	<b>Objetivo:</b> Automatizar o processo de classificação e indexação de documentos para disponibilizar um serviço satisfatório de busca ao conhecimento jurisprudencial. <b>Valor Agregado:</b> Diminui significativamente o tempo entre a conclusão do documento e sua disponibilização para consulta. Permite também que esta atividade seja realizada utilizando critérios semelhantes para todos os documentos, evitando problemas de interpretação pessoal.
Dependências e Fluxo	<b>Entrada:</b> Documento da decisão judicial finalizado e assinado pelo magistrado. <b>Saída:</b> Documento devidamente categorizado nos grupos pré-existentes e índices de busca atualizados.
Objetos manuseados	<ul style="list-style-type: none"> <li>• Documentos jurisprudenciais eletrônicos</li> <li>• Tesouro da Justiça Eleitoral</li> </ul>
Tempo e controles	Após ser disponibilizado um documento finalizado de uma determinada decisão jurídica, este é recebido pelo classificador que inicia o processo de classificação e indexação automaticamente. Para validação, de tempos em tempos são selecionados alguns documentos classificação para verificar se a atividade foi realizada adequadamente.
Agentes	Classificador de Documentos Jurisprudenciais
Conhecimento e competência	<ul style="list-style-type: none"> <li>• Critérios para Classificação de Documentos</li> <li>• Direito Eleitoral</li> </ul>
Recursos	<ul style="list-style-type: none"> <li>• Servidor da Base de Conhecimento</li> <li>• Sistema para Classificação de Documentos Jurisprudenciais</li> </ul>
Qualidade e Performance	O fator principal para qualidade da atividade está mais relacionado com a correta classificação do que com o tempo em que esta é realizada. Assim, a avaliação da qualidade será realizada periodicamente verificando o resultado da classificação e a indexação de documentos aleatórios e já processada.

Tabela 4.7 - Modelo de Tarefa - Análise das Tarefas.

Modelo de Tarefa	Itens do Conhecimento – TM-2	
Nome	Critérios para classificação de documentos jurisprudenciais	
Pertence a	Classificador de Documentos Jurisprudenciais	
Usado em	Classificar e Indexar Documentos	
Domínio	Direito Eleitoral	
Natureza do Conhecimento		Gargalos/Melhorias
Formal, Rigoroso	X	
Empírico, quantitativo		
Heurístico, Regras	X	
Altamente especializada	X	
Baseado em experiência	X	X
Baseado em atividades		
Incompleto		
Incerto, pode estar incorreto.	X	X
Mudando rapidamente		
Difícil de verificar		
Tácito, difícil de transmitir.		
Formato do Conhecimento		Gargalos/Para ser melhorado
Mente	X	X
Papel		
Formato Eletrônico		
Habilidade da ação	X	
Outros		
Disponibilidade do Conhecimento		Gargalos/Para ser melhorado
Limitações de Tempo	X	X
Limitações de espaço		
Limitações de acesso		
Limitações de qualidade	X	X
Limitações de forma	X	X

Tabela 4.8 - Modelo de Tarefa - Itens do conhecimento.

Tendo como fundamento as informações levantadas até o presente momento, já é possível visualizar com propriedade e clareza, os pontos específicos dentro do escopo de abrangência da atividade de classificação e indexação de documentos que necessitam ser melhorados. Tais oportunidades de melhoria estão diretamente relacionadas com a natureza do conhecimento que, até então, é tácito e, assim sendo, de difícil automação e compartilhamento. Este fato dificulta significativamente a utilização de ferramental tecnológico para otimizar o tempo, bem como a qualidade de sua execução.

#### 4.3.3 Modelo de Agente

O propósito do modelo do agente, conforme a metodologia, é entender melhor as atribuições e competências que os atores tem ao realizar as tarefas. Na verdade o modelo do agente é apresentado na tabela 4.9 e consiste de uma reorganização de informações oriundas das etapas anteriores, porém apresentadas sob um prisma diferente. Esta apresentação pode auxiliar a avaliar impacto de decisões aos agentes.

Modelo de Agente	Planilha de Agentes – AM-1
Nome	Classificador de documentos jurisprudenciais
Organização	Coordenadoria de Gestão da Informação
Envolvido em	Classificar e indexar documentos jurisprudenciais
Comunicação com	Gabinetes dos Juízes e Promotores Eleitorais
Conhecimento	Conhecimento jurídico para distinguir e classificar o documento com base em seu conteúdo e identificar os termos relevantes para compor o índice temático.
Outras competências	Facilidade em identificar novos padrões; Conhecimento no <i>software</i> para gestão de conteúdo <i>web</i> .
Responsabilidades e restrições	Classificar e indexar os documentos e proceder com sua publicação na Internet.

Tabela 4.9 – Modelo de Agentes – Planilha de Agentes.

#### 4.3.4 Modelo de Organização, Tarefa e Agente

Neste modelo último modelo elaborado no estudo de caso é realizada uma verificação, tabela 4.10, em que se apresentam ações propostas para a busca de melhorias. Essas ações podem não estar diretamente relacionadas com o do sistema de conhecimento proposto, porém, são muito importantes para assegurar o compromisso e o apoio por parte dos atores de relevância da organização. As principais questões de tomada de decisão são:

- a) Mudanças organizacionais recomendadas;
- b) Medidas a serem implementadas com relação a tarefas específicas e trabalhadores envolvidos; em particular, que melhoramentos são possíveis no uso e disponibilidade do conhecimento; e
- c) Apoio às mudanças das pessoas envolvidas.

Modelo de Organização, Tarefa e Agentes	Planilha para Verificação de Impacto e Melhoramentos – OTA-1
Impactos e mudanças na organização	A implementação de um sistema baseado em conhecimento como apoio à gestão do conhecimento jurisprudencial possibilitará a <u>disponibilização de um serviço mais ágil e eficaz</u> de socialização do conhecimento, tanto para uso interno na organização, mas principalmente para a sociedade que é o “cliente” primordial das instituições governamentais. Para viabilizar esta inovação, são necessárias mudanças principalmente nos processos de elaboração e disponibilização do conteúdo jurídico. Alterando os processos, altera-se, conseqüentemente, os recursos e as pessoas. Quanto aos recursos, serão disponibilizados novos elementos como, por exemplo, a base de conhecimento e, em contrapartida, as pessoas poderão ser alocadas para outras atividades que demandem maior esforço cognitivo, cuja automação atualmente não seja viável.
Impactos e mudanças específicos a tarefas / agentes	A <u>tarefa</u> de disponibilização do conteúdo jurídico será fortemente afetada pela proposta de automação da classificação e indexação dos documentos e composição do serviço de índice temático. Atualmente esta atividade é totalmente dependente do elemento humano, demandando tempo e susceptível a avaliações subjetivas.  O <u>agente</u> que atualmente se envolve diretamente na atividade de classificação e indexação, após a automação, terá um papel de supervisão no processo, com intuito de garantir que o serviço automatizado está sendo realizado com a qualidade desejada.
Atitudes e Compromissos	Para viabilização deste sistema de conhecimento, faz-se necessário o compromisso da administração da instituição no apoio e garantia da aplicação das mudanças propostas nos processos de disponibilização do conhecimento. As pessoas, como sempre, têm papel fundamental no sucesso do projeto, pois são elas as detentoras do conhecimento que se busca formalizar.
Ações Propostas	Definição de um gerente para o projeto; definição da equipe do projeto; Elaboração do planejamento das atividades para implementação do projeto; Análise da viabilidade técnica e jurídica da proposta

Tabela 4.10 – Modelo de Organização, Agente e Tarefa – Planilha de verificação.

#### 4.3.5 Considerações Finais do Estudo de Caso

Concluindo esta etapa da pesquisa, pode-se observar que o uso da metodologia CommonKADS se mostrou uma forma sistemática de organizar as informações da Instituição, tendo em vista explicitar os processos e sugerir mudanças considerando necessariamente o impacto destas no contexto organizacional. O uso da metodologia permitiu a visão holística da organização, apresentando um panorama geral de seus processos, pessoas, recursos e, principalmente, os pontos onde a engenharia do conhecimento pode ser aplicada para contribuir no desempenho de suas atividades.

Do ponto de vista do conhecimento, a metodologia possibilitou identificar a tarefa intensiva do conhecimento que será automatizada, e levantar as estruturas de inferência relativas ao conhecimento.

O uso desta metodologia em organizações que sofrem o problema de acúmulo de informações não estruturadas, mostrou-se de grande valia, identificando pontos de gargalo de processo e norteador ações para melhoria do desempenho tendo em mente o negócio estratégico de instituições públicas, levantado desde o início do processo, servir bem a sociedade.

## 5 MODELO DE GESTÃO DOCUMENTAL PROPOSTO

Baseando-se na fundamentação teórica compilada nos capítulos prévios, em conjunto com os resultados obtidos pela aplicação dos modelos contextuais da metodologia CommonKads no TRESA, será exposto um modelo de gestão documental cujo intuito é responder os questionamentos iniciais da pesquisa e preencher as lacunas de conhecimento relativos ao tema.

Inicialmente, será detalhado o modelo conceitual da proposta que contempla as mesmas atividades identificadas no processo de gestão documental até então utilizado, porém, com o acoplamento de um novo elemento ao processo, o Sistema e-Docs. Na seqüência, será descrita a arquitetura tecnológica sob a qual o sistema foi concebido e, por fim, serão relacionados os resultados obtidos com a aplicação deste modelo.

### 5.1 Modelo Conceitual

O modelo conceitual tem por objetivo apresentar os principais elementos envolvidos no processo de gestão documental (figura 4.3) trabalhando conjuntamente com o Sistema e-Docs. O foco da interação destes elementos é fazer com que as atividades intensivas em conhecimento, desempenhadas preliminarmente por profissionais humanos, possam ser substituídas por outras, executadas de forma automatizada pelo sistema, sem que haja prejuízo na qualidade nos serviços prestados.

A partir dos resultados obtidos com a aplicação dos modelos contextuais da metodologia CommonKads foi possível relacionar os elementos objetos de estudo para proposição de melhorias no processo atual, apresentadas na figura 5.1. O processo original constituía-se de seis macro atividades, abrangendo tarefas meramente operacionais como impressão e digitalização de documentos, além de outras, mais complexas e intensivas em conhecimento, como a elaboração de documentos contendo decisões judiciais, por exemplo.

Quanto às fases do ciclo de vida documental adotado, apresentado por Silva (2008), o processo abrange basicamente a *captura* e o *arquivo*. As fases de *revisão* e *transição* foram retiradas do escopo do modelo, pois estas fases ocorrem nos gabinetes dos magistrados e tais controles permanecerão sob sua responsabilidade já que estão relacionados com a atividade *Elaboração de Despachos, Sentenças e Acórdãos* e esta, por definição de projeto, não será abordada.

Das duas atividades intensivas em conhecimento identificadas, foi observado que o conhecimento utilizado na classificação prévia à publicação dos documentos poderia ser automatizado visando ao aperfeiçoamento do processo, conforme as tabelas 4.7 e 4.8.

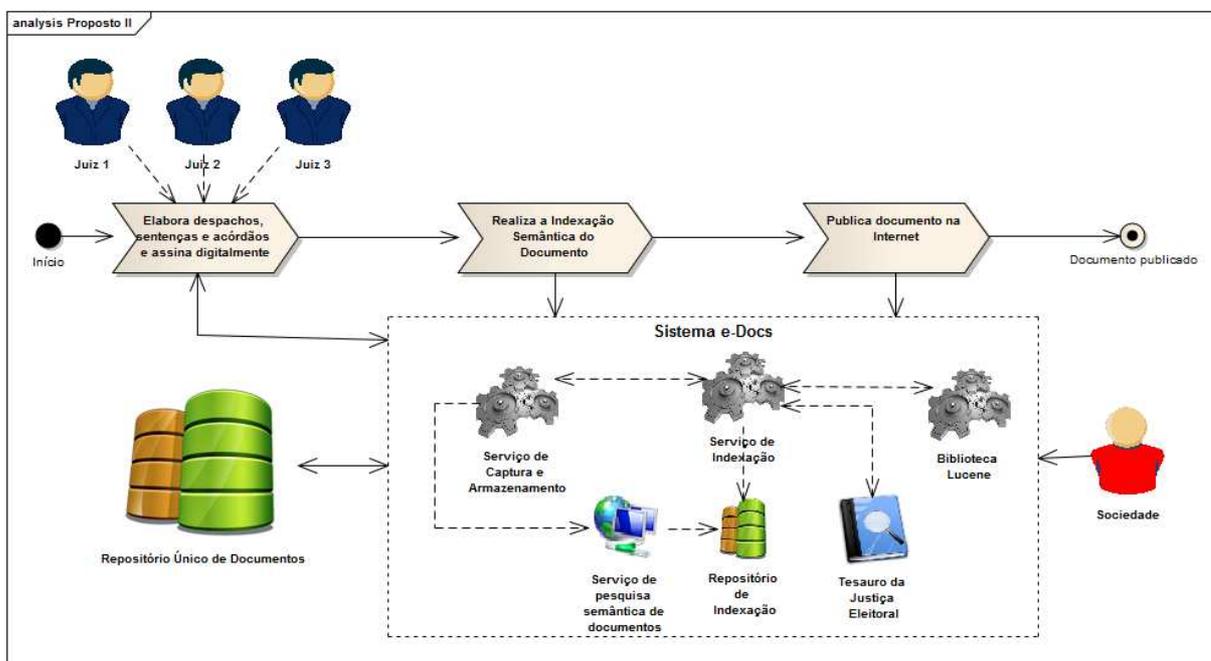


Figura 5.1 – Modelo conceitual do Processo de Gestão Documental Proposto.

### 5.1.1 Repositório Único de Documentos

Independente da natureza das atividades envolvidas, um fator que merece destaque diz respeito aos diversos repositórios utilizados para armazenar documentos. Esta abordagem inviabilizava a implantação de qualquer serviço de pesquisa informacional, visto que cada unidade elaboradora de documentos detinha seus próprios mecanismos de armazenamento e recuperação. Assim sendo, a criação do *Repositório Único de Documentos* é uma medida fundamental e necessária para a implantação de um sistema efetivo de recuperação do conhecimento jurídico contidos nos documentos do acervo institucional. A unificação do repositório de documentos, além dos benefícios da efetividade na busca, objetiva:

- a) a padronização da fase de captura, visto que os usuários detêm a mesma interface para realizar o cadastro de seus documentos no acervo;
- b) a racionalização dos recursos de infraestrutura tecnológica ao reduzir a redundância de documentos entre diversos repositórios; e

- c) a redução da complexidade do processo de cópias de segurança (*backup*) dos documentos, pois este estará limitado ao repositório unificado, sem a necessidade de gerenciar diversos repositórios setoriais.

### 5.1.2 Serviço de Indexação

Contrapondo o conceito da indexação clássica, o qual a descreve como uma representação do conteúdo do documento por meio de termos previamente selecionados, a indexação proposta no *Serviço de Indexação* envolve todos os termos constantes nos documentos, que são inseridos em estruturas de dados análogas às apresentadas no capítulo 3 que tratou sobre Recuperação da Informação.

Tendo em vista um desempenho adequado do *Serviço de Pesquisa*, antes do processo de indexação propriamente dito, são aplicados processamentos em nível de palavras no conteúdo dos documentos, tais como: eliminação de *stopwords* e rótulos (*tags*) específicos do formato; *stemming*. Na sequência, é realizada a normalização dos termos com a introdução do Tesouro da Justiça Eleitoral.

O Tesouro da Justiça Eleitoral (TJE) foi elaborado pelo Tribunal Superior Eleitoral e destina-se à padronização da linguagem durante a indexação dos documentos, sendo peça fundamental para viabilizar a implementação da busca semântica a documentos. As figuras 5.2 e 5.3 trazem duas visualizações do TJE: a primeira ilustra o documento do tesouro tal como é utilizado no processo de indexação tradicional; e a segunda apresenta o mesmo conteúdo no formato *XML*, utilizado no serviço de indexação automática proposto.

<p><b>Embargos</b> TR</p> <ul style="list-style-type: none"> <li>• Embargado</li> <li>• Impedimento</li> </ul> <p>Embargos declaratórios USE</p> <ul style="list-style-type: none"> <li>• <b>Embargos de declaração</b></li> </ul> <p><b>Embargos de declaração</b> UP</p> <ul style="list-style-type: none"> <li>• Embargos declaratórios</li> </ul> <p>TR</p> <ul style="list-style-type: none"> <li>• Efeito modificativo</li> </ul> <p><b>Embargos de divergência</b> TR</p> <ul style="list-style-type: none"> <li>• Interpretação superada</li> </ul> <p>Embargos de nulidade USE</p> <ul style="list-style-type: none"> <li>• <b>Embargos infringentes</b></li> </ul>	<p><b>Emendatio libelli</b> NOTA</p> <ul style="list-style-type: none"> <li>• Hipótese do art. 383 do Código de Processo Penal, que faculta ao juiz a desclassificação do fato criminoso, narrado na queixa ou denúncia, ainda quando dela resulte pena mais grave para o réu</li> </ul> <p>TR</p> <ul style="list-style-type: none"> <li>• Mutatio libelli</li> </ul> <p><b>Ementa</b> TR</p> <ul style="list-style-type: none"> <li>• Resumo</li> </ul> <p><b>Emergência</b> <b>(EMFA)</b> UP</p> <ul style="list-style-type: none"> <li>• Estado Maior das Forças Armadas</li> </ul> <p><b>Emissão</b> TR</p> <ul style="list-style-type: none"> <li>• Expedição</li> </ul>
--	--

Figura 5.2 – Visualização parcial do Tesouro da Justiça Eleitoral

```

- <Termo Nome="Embargos">
  - <termos-relacionados>
    <termo-relacionado>Embargado</termo-relacionado>
    <termo-relacionado>Impedimento</termo-relacionado>
  </termos-relacionados>
</Termo>
- <Termo Nome="Embargos declaratórios">
  - <termos-relacionados>
    <termo-relacionado>Embargos de declaração</termo-relacionado>
  </termos-relacionados>
</Termo>
- <Termo Nome="Embargos de declaração">
  - <termos-relacionados>
    <termo-relacionado>Efeito modificativo</termo-relacionado>
  </termos-relacionados>
</Termo>
- <Termo Nome="Embargos de divergência">
  - <termos-relacionados>
    <termo-relacionado>Interpretação superada</termo-relacionado>
  </termos-relacionados>
</Termo>

```

Figura 5.3 – Visualização parcial do Tesouro da Justiça Eleitoral no formato XML

Assim sendo, após a conclusão do processamento prévio de mineração de texto, o processo de indexação segue para a etapa de construção da estrutura de índice invertido (tabela 3.2). Neste estágio é realizado o batimento entre os termos do documento e os termos do acervo, no intuito de atualizar as informações da estrutura de indexação com base no documento que está sendo indexado. O TJE é então introduzido com vistas à preparação da infraestrutura necessária para a busca semântica.

Com objetivo de apresentar uma visão geral das etapas supracitadas, será descrito um exemplo meramente ilustrativo deste processo.

Partindo-se da estrutura de índice invertido inicial da tabela 5.1 e do tesouro ilustrativo da figura 5.4, busca-se a indexação de um determinado documento (Doc. 10) com o seguinte conteúdo: “A *gestão documental é fundamental para a disseminação do conhecimento no Judiciário.*”.

<b>conhecimento</b>	Doc. 1	Doc. 3	Doc. 4	Doc. 5	Doc. 9
<b>disseminação</b>	Doc. 2	Doc. 4			
<b>documental</b>	Doc. 1	Doc. 2	Doc. 5	Doc.8	
<b>gestão</b>	Doc. 2	Doc. 5	Doc. 9		
<b>judiciário</b>	Doc. 4	Doc. 5			
<b>propagação</b>	Doc. 5				

Tabela 5.1 – Estrutura de índice invertido ilustrativa

```

<?xml version="1.0" encoding="windows-1252" ?>
- <Thesaurus>
- <Termos>
- <Termo Nome="Disseminação">
- <termos-relacionados>
<termo-relacionado>Propagação</termo-relacionado>
<termo-relacionado>Difusão</termo-relacionado>
</termos-relacionados>
</Termo>
</Termos>
</Thesaurus>

```

Figura 5.4 – Tesauro ilustrativo

Inicialmente são excluídas do texto as chamadas *stopwords*, ou seja, termos sem relevância para recuperação futura, que resulta no seguinte conteúdo:

*“gestão documental fundamental disseminação conhecimento judiciário”.*

Para simplificar o exemplo, não será realizada a redução ao radical comum (*stemming*). Assim sendo, a próxima etapa é a atualização da estrutura de indexação. Observa-se, com base na tabela 5.2, que a estrutura agora inclui os dados relativos ao Doc. 10, ou seja, para cada termo de seu conteúdo, foi incluída uma referência ao documento.

<b>conhecimento</b>	Doc. 1	Doc. 3	Doc. 4	Doc. 5	Doc. 9	Doc. 10
<b>disseminação</b>	Doc. 2	Doc. 4	Doc. 10			
<b>documental</b>	Doc. 1	Doc. 2	Doc. 5	Doc.8	Doc. 10	
<b>gestão</b>	Doc. 2	Doc. 5	Doc. 9	Doc. 10		
<b>judiciário</b>	Doc. 4	Doc. 5	Doc. 10			
<b>propagação</b>	Doc. 5					

Tabela 5.2 – Estrutura de índice invertido após indexação do Doc. 10, sem uso do tesauro

Neste ponto do processo, é possível vislumbrar uma estrutura de indexação já apropriada para recuperação de documentos, porém, sem nenhum tratamento semântico quanto ao conteúdo, o que não desoneraria a atividade de indexação clássica, que é o objeto de estudo para melhoria do processo de gestão documental.

Desta forma, fundamentado no referencial teórico pesquisado quanto à busca semântica e tendo em vista o objetivo pelo qual foi criado o TJE, introduzindo o tesauro ilustrativo da figura 5.4 ao exemplo, tem-se agora uma estrutura de indexação que amplia o espaço de busca com base no conhecimento formalizado nesta ferramenta, conforme tabela 5.3.

<b>conhecimento</b>	Doc. 1	Doc. 3	Doc. 4	Doc. 5	Doc. 9	Doc. 10
<b>disseminação</b>	Doc. 2	Doc. 4	Doc. 10			
<b>documental</b>	Doc. 1	Doc. 2	Doc. 5	Doc.8	Doc. 10	
<b>gestão</b>	Doc. 2	Doc. 5	Doc. 9	Doc. 10		
<b>judiciário</b>	Doc. 4	Doc. 5	Doc. 10			
<b>propagação</b>	Doc. 5	<b>Doc. 10</b>				

Tabela 5.3 - Estrutura de índice invertido após indexação do Doc. 10, com uso do tesauro

Observa-se na tabela 5.3, que existe uma nova referência do Doc. 10 junto ao termo *propagação*, mesmo não fazendo parte do seu conteúdo original. Porém, visto que o tesauro estabelece semântica entre termos, pode-se afirmar que existe sim o conceito de propagação no texto e, assim sendo, faz-se necessária a ligação entre o termo e o documento em questão.

Em suma, o processo de indexação consiste em manter esta estrutura de dados utilizada essencialmente pelo *Serviço de Pesquisa Semântica de Documentos*.

### 5.1.3 Serviço de Pesquisa Semântica

Conforme Adam (2008), um fator de sucesso na implantação de um SGED é a disponibilização de um mecanismo de recuperação de documentos de qualidade. Partindo desta premissa, buscou-se estabelecer requisitos bem definidos visando proporcionar ao usuário uma interface simplificada, porém, com grande efetividade na recuperação de documentos, principalmente relacionada a seu conteúdo semântico.

Assim sendo, com base da estrutura de indexação descrita anteriormente, o *Serviço de Pesquisa Semântica* constitui-se de um mecanismo capaz de realizar a leitura nos índices do acervo, fazendo o devido batimento com os parâmetros informados pelos usuários, recuperando assim os documentos que atenderem aos requisitos da consulta.

Ressalta-se, entretanto, que a pesquisa semântica proposta constitui-se de uma expansão vetorial dos termos de pesquisa, por meio do conhecimento formalizado no tesauro. Não será abordado neste projeto, o tratamento quanto aos problemas de ambigüidade e inferências no conteúdo do tesauro em uso. Tais temas são sugeridos como trabalhos futuros.

Durante o processo de recuperação, são utilizados mecanismos análogos aos utilizados na indexação como, por exemplo, *stemming* e a eliminação das *stopwords*, para aproximar o termo da consulta à linguagem de indexação utilizada. Concluídas estas operações básicas, a *query* informada é convertida em um vetor de termos para então ser processada a pesquisa, conforme os modelos de recuperação da informação explanados anteriormente.

Voltando ao exemplo da seção 5.1.2, uma consulta ao acervo ilustrativo pelos termos: “*disseminação do conhecimento judiciário*” retornaria os seguintes documentos: Doc. 4, Doc. 5 e Doc. 10; visto que, de acordo com a indexação prévia, são estes os documentos que contém os conceitos expressos do termo da pesquisa.

Na abordagem apresentada, com intuito de melhorar o desempenho das consultas, o tesauro é utilizado exclusivamente na fase de indexação, visto que neste processo as referências aos conceitos do documento são propagadas para todos os termos semanticamente relacionados no tesauro. Assim sendo, é desnecessária a consulta ao tesauro no processamento da consulta.

Por fim, ressalta-se que a qualidade da busca semântica está diretamente realizada com a qualidade do conhecimento formalizado no tesauro utilizado. Se a definição dos conceitos estiver incompleta nesta ferramenta, conseqüentemente a busca também será limitada.

#### **5.1.4 Considerações Finais sobre o Modelo Conceitual**

Contrapondo o modelo inicial da figura 4.3 com o modelo proposto na figura 5.1, percebe-se uma mudança conceitual no papel do Classificador de Documentos Jurisprudenciais e na atividade intensiva em conhecimento que executava.

No modelo anterior, existia a necessidade da indexação manual para viabilizar a consulta com base nos termos de indexação selecionados. Entretanto, visto que a automação do processo possibilita a indexação completa do conteúdo dos documentos, garantindo a padronização da linguagem de indexação com a integração com o tesauro, esta atividade manual pode ser substituída sem redução na qualidade do serviço prestado.

Assim sendo, a atividade intensiva em conhecimento desempenhada por um profissional, atualmente é executada de forma automatizada, utilizando para isso o conhecimento previamente formalizado no tesauro.

Nesta perspectiva, o papel do profissional indexador tem novo enfoque, pois está mais voltado à formalização do conhecimento de indexação de documentos jurisprudenciais junto ao tesouro, visando à melhoria contínua da qualidade do conteúdo modelado neste instrumento.

## **5.2 Arquitetura Tecnológica**

A arquitetura tecnológica do Sistema e-Docs está fundamentada sob a plataforma Java J2EE que disponibiliza diversas tecnologias e estabelece uma série de padrões para o desenvolvimento de softwares corporativos em multicamadas. A plataforma busca simplificar a implementação de sistemas por meio da utilização de componentes modulares de softwares, provendo um conjunto extenso de serviços que determinam grande parte do comportamento básico das aplicações sem a necessidade de complexa programação. (SUN, 2010).

A figura 5.5 apresenta a arquitetura tecnológica, relacionando as cinco camadas que trabalhando em conjunto formam o SGED proposto nesta pesquisa:

- a) e-Docs Web – Constitui-se de componentes cujo fim é disponibilizar as interfaces do sistema que serão utilizadas pelos usuários.
- b) Utilitários – É composta por um conjunto de customizações realizadas a partir de determinadas tecnologias J2EE visando melhorias na padronização, segurança e produtividade quanto ao desenvolvimento interno de sistemas desta natureza.
- c) Componentes Externos – São pacotes de softwares desenvolvidos por terceiros para objetivos peculiares. Nesta camada, destacam-se os elementos que realizam a indexação automática propriamente dita e o tratamento de formatos específicos de documentos.
- d) e-Docs Core – Nesta camada estão implementadas as regras de negócio que determinam como o sistema desempenhará a atividade de gerenciamento do acervo documental. Dispõe de todos os serviços propostos pelo sistema, servindo como ponte entre as camadas de interface do usuário e persistência de informações.
- e) Persistência – É uma camada estática e essencialmente representa o repositório de documentos e informações utilizados pelo sistema. Os documentos são armazenados em um servidor de arquivos, já o catálogo

das informações relativas aos documentos, ou seja, os metadados do acervo documental, é armazenado no servidor de banco de dados.

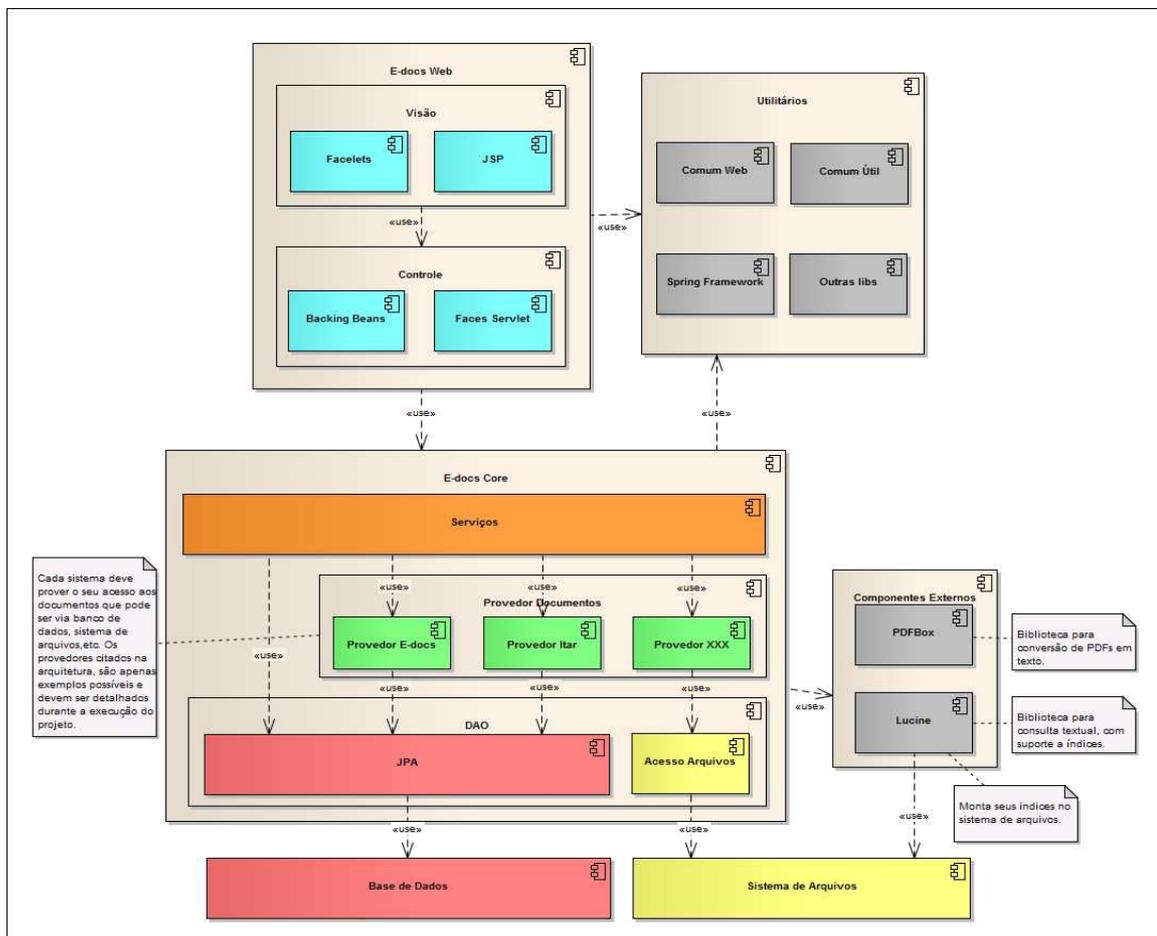


Figura 5.5 – Arquitetura tecnológica

As principais tecnologias utilizadas na arquitetura tecnológica do sistema e-Docs estão apresentadas com maior detalhamento no Anexo II.

### 5.2.1 Interfaces do Sistema e-Docs

As interfaces do sistema são compostas basicamente por três módulos:

- a) Captura;
- b) Consulta; e
- c) Administração.

O captura dos documentos é realizada por meio da interface apresentada pela figura 5.6. A partir deste cadastro, o usuário informa os metadados e envia o respectivo arquivo eletrônico. Todos os dados preenchidos neste módulo servem de subsídios para a recuperação posterior do documento no acervo.

Mesmo tendo um repositório unificado, o sistema estabelece áreas lógicas de armazenamento conforme a área de atuação dos usuários. Assim sendo, foram

definidos perfis de usuário para cada área e para cada perfil, uma relação de usuários responsáveis pelo gerenciamento de seus documentos. Esta abordagem garante que somente as pessoas devidamente autorizadas possam inserir, alterar e excluir documentos de suas respectivas áreas.

The screenshot shows the 'Cadastro de Documentos' (Document Registration) interface. At the top, it identifies the user as 'Samuel Fernandes Ribeiro' and the system as 'e-Docs Pesquisa Textual em Documentos'. The form includes the following fields:

- \* Tipo de Documento: Dropdown menu with 'Portaria DG' selected.
- \* Arquivo: Text input field containing 'C:\ci-sti.pdf' and a 'Browse...' button.
- \* Ano: Text input field containing '2010'.
- \* Número: Text input field containing '15'.
- Indexador: Large text area for entering the indexer's name.
- Observações: Large text area for entering additional notes.

A 'Cadastrar' button is located at the bottom center of the form. The footer of the page reads 'Pesquisa Textual em Documentos - 1.1.1'.

Figura 5.6 – Interface para cadastro de documentos no repositório

A figura 5.7 trata da interface de consulta aos documentos. Este módulo apresenta duas opções: a consulta básica, contendo apenas um campo genérico para informação dos termos da consulta (*query*); e a consulta avançada, que dispõe deste mesmo recurso, agregando, porém, outras opções de consulta por meio de metadados específicos como: ano, número, tipo etc.

Os resultados das consultas são apresentados conforme a figura 5.8. A partir desta interface o usuário é capaz de navegar entre os documentos que atenderam aos requisitos definidos pelos termos da consulta. Para facilitar a navegação, o sistema apresenta um resumo do documento, desta forma, não é necessário abrir o documento para ter conhecimento de seu conteúdo.


**TRIBUNAL REGIONAL ELEITORAL SANTA CATARINA**
**e-Docs**  
Pesquisa Textual em Documentos





---

**Consulta avançada**

Palavras a pesquisar:  Consulta básica

Pesquisar por palavras similares

**Pesquisar**

Ano: 
 Tipo de documento:  Todos

Número Inicial:  Número Final:

Data Inicial:  Data Final:

Observações:

Indexação Manual:

- Ata de Sessão Plenária
- Manual de Prática Cartorária
- Ofício Circular
- Portaria DG
- Portaria Vice-Presidência
- Comunicação Interna
- Mensagem Circular
- Ordem de Serviço
- Portaria Presidência
- Provisimento

Pesquisa Textual em Documentos - 1.1.1

Figura 5.7 – Interface para consulta ao repositório


**TRIBUNAL REGIONAL ELEITORAL SANTA CATARINA**
**e-Docs**  
Pesquisa Textual em Documentos

Usuário: Samuel Fernandes Ribeiro
 




Sessão expira em 10min40s

---

**Consulta básica**

Palavras a pesquisar:  Consulta avançada

Pesquisar por palavras similares

**Pesquisar**

Alguns caracteres estranhos podem ser exibidos na coluna Resumo. Isto ocorre porque o processo de digitalização de documentos não é capaz de identificar todas as palavras de um documento impresso. No entanto, o documento em formato PDF será apresentado corretamente.

Tipo de Documento	Ano	Número	Resumo	Número de ocorrências	Documento	Indexado Manualmente
Portaria Presidência	2010	34	PORTARIA PN. 34/2010 O Presidente do Tribunal Regional Eleitoral de Santa Catarina, no uso da atribuição que lhe é conferida pelo inciso XIII do art. 20 da Resolução TRESO n. 7.357, de 17.12.2003, <b>Prorrogar</b> , a pedido e sem ônus para este Tribunal, ao servidor EMERSON CARGNIN, ocupante	1		Sim
Portaria DG	1994	231	DA PORTARIA DG N2 229, DE 14/09/94, TENDO EM VISTA O <b>ADIAMENTO</b> DA VIAGEM DE RETORNO* DE BRASÍLIA-DF	1		Não
Portaria DG	1999	247	(uma) diária, a título de complementação, em função de <b>adiamento</b> do retorno deste servidor para o dia 20	1		Não
Portaria DG	1995	366	do requerimento protocolado sob ne 000076, de 09/01/96, onde a titular requer <b>adiamento</b> de/s	1		Não
Portaria DG	2001	203	, por motivo de <b>adiamento</b> de férias da titular. Dê-se ciência e cumpra-se. Tribunal Regional	1		Não
Portaria Presidência	2004	185	ao <b>adiamento</b> ou antecipação dos períodos de férias agendados para os meses de junho e julho do presente	1		Não
Portaria DG	1987	096	da Secretaria deste Tribunal, que requereram <b>adiamento</b> das suas fe_rias referentes ao corrente	1		Não
Mensagem Circular	2008	57	incompatíveis com as informações do cadastro de eleitores. Exemplo: "nome do eleitor <b>difere</b> do cadastro	1		Não
Portaria Presidência	2008	0395	OAtiãuaicd *H*eaio***leããa/de Gfaw/a fyt/avwna PORTARIA P.N. 395/2008 Dispõe sobre a <b>prorrogação</b> ... a <b>prorrogação</b> da licença à gestante na esfera da Administração Pública; - considerando o disposto... SGP n. 615/2008, RESOLVE: Art. 15 Esta Portaria dispõe sobre a <b>prorrogação</b> da licença à gestante	9		Não
Portaria Presidência	2010	16	Valores <b>Diferidos</b> - Valores <b>Diferidos</b> Depósitos Exigíveis a Longo Prazo <b>INSUFICIÊNCIA</b> .	2		Sim

1
2
3
4
5
6
7
8
9
10

Figura 5.8– Relação dos documentos recuperados a partir de uma consulta

O módulo de administração tem por objetivo servir de apoio ao gerenciamento técnico do repositório. Nesta interface é permitido ao administrador atualizar a

estrutura de indexação do sistema a qualquer momento. A indexação é um processo automático que é executado em cada novo cadastro de documento. Porém, para garantir sua integridade e a disponibilidade do sistema, este processo também é executado periodicamente por meio de um serviço pré-agendado no servidor de aplicações.

### **5.3 Resultados Obtidos**

Com a implantação da solução proposta no TRESA, foi possível perceber uma série de melhorias à Instituição quanto ao modelo anterior. Destas, destacam-se: a unificação do repositório; a redução no tempo do processo de indexação; e a ampliação da abrangência dos documentos indexados.

A unificação do repositório foi um fator determinante para a efetividade da solução proposta. Atualmente, todos os documentos normativos estão disponíveis a partir do sistema. Assim sendo, os usuários detêm um mecanismo unificado para consulta que disponibiliza funcionalidades importantes para facilitar a recuperação documental. A unificação também contribuiu na gestão da infraestrutura tecnológica, facilitando os processos de cópia de segurança e controle de acesso e integridade.

A automação da indexação integrada com o TJE, além de reduzir o tempo deste processo, possibilitou uma nova perspectiva quanto ao papel do profissional que detém o conhecimento para indexação de documentos jurisprudenciais. Inicialmente suas atividades envolviam a indexação clássica, onde era realizada a leitura dos documentos, selecionando também nesta fase as palavras que melhor expressavam seu conteúdo semântico. A lista de palavras selecionada era então comparada com o TJE para garantir a padronização da linguagem de indexação, conforme o objetivo inicial do uso desta ferramenta. Finalizando o processo, a lista com as palavras traduzidas eram inseridas em um sistema de busca que se limitava aos termos de indexação cadastrados.

Sendo uma atividade que demanda um tempo considerável do profissional e a quantidade de documentos elaborados é muito maior que a capacidade de indexação manual, era realizada uma seleção prévia de documentos e somente estes eram indexados. Com a implantação do Sistema e-Docs, o processo de indexação agora automatizado abrange todo conteúdo dos documentos. Além disso, cada palavra é comparada com o tesauro na busca por outros termos que expressem o mesmo conceito com objetivo de ampliar o espaço de busca.

A partir de então, com o advento da automação, todos os documentos passam a ser indexados eliminando a necessidade da seleção prévia. Além disso, percebe-se um incremento considerável quanto à abrangência dos termos de indexação que atualmente consistem das palavras do documento e todos os termos relacionados pelo tesouro.

Após dois meses de uso do sistema, implantado apenas em algumas unidades da Instituição, tem-se um repositório com 24.000 documentos devidamente indexados para possibilitar a busca semântica proposta. Neste mesmo período, constatou-se a execução de aproximadamente 2.000 consultas ao repositório, realizadas por 8 profissionais lotados nos setores nos quais o sistema está implantado.

Apesar da ausência de estatísticas quando à efetividade do modelo anterior, os números obtidos quando ao uso do sistema e-Docs, mesmo em uma implantação parcial, vem demonstrando que a solução está se tornando uma importante ferramenta para recuperação do conhecimento na Instituição. Isso pode ser observado tanto no modelo clássico de busca sintática, mas principalmente pela inclusão da funcionalidade semântica que permite a expansão vetorial na consulta, obtida pela integração com o conhecimento formalizado no TJE.

Por fim, destaca-se o fato ocorrido quando da necessidade em se resgatar um assunto tratado na década de 1940. O modelo anterior exigiria que fosse realizada uma busca manual documento por documento, o que inviabilizaria a busca e o conhecimento seria perdido. Porém, após com a inclusão destes documentos no Sistema e-Docs, a recuperação do conteúdo necessário se deu em minutos.

## **6 CONCLUSÃO**

Este capítulo tem por objetivo fazer o fechamento da presente dissertação, relacionando as conclusões obtidas no decorrer do desenvolvimento dos trabalhos, bem como apresentar propostas de temas correlatos não abordados neste escopo, com intuito de fomentar pesquisas futuras nesta área do conhecimento.

### **6.1 Considerações Finais**

Contraopondo o problema inicial e os questionamentos de pesquisa ante aos objetivos propostos que, em suma, orientam à melhoria da gestão do acervo eletrônico de documentos para facilitar a recuperação do conhecimento organizacional, durante o transcorrer dos trabalhos, buscou-se discorrer sobre os diversos elementos que, de uma forma ou de outra, influenciam este processo.

Para tanto, o uso da metodologia CommonKads apresentou importante contribuição, pois sua aplicação facilitou a identificação clara dos elementos supracitados, bem como suas respectivas relevâncias e os papéis que desempenhavam. A metodologia proporcionou uma perspectiva sistêmica tanto do processo em si, destacando suas atividades e, principalmente, as pessoas envolvidas. Com base neste estudo, foi possível desenvolver a proposta dentro de expectativas concretas previamente analisadas.

Os aspectos tecnológicos que, de certa maneira, estão presentes nas causas do problema proposto, certamente desempenharam papel fundamental na proposição da solução. Diz-se “influenciaram na causa do problema”, pois a disseminação do uso da tecnologia da informação contribuiu diretamente para crescimento da produção documental sem que as organizações estivessem preparadas para gerenciá-los. Em contrapartida, o uso apropriado da tecnologia é um fator determinante para o sucesso do gerenciamento efetivo de acervos que crescem constantemente em um contexto socioeconômico onde a velocidade da busca pelo conhecimento tem grande influência no cotidiano das instituições. No cenário atual, pode-se afirmar da inviabilidade de qualquer solução que não envolva este aspecto, ressalta-se, entretanto, que a tecnologia em si não é capaz de preencher as lacunas existentes no que tange à GED. Nesta senda, faz-se necessária uma abordagem sistêmica de outros aspectos correlatos.

Na abordagem quanto às pessoas, o olhar sob o prisma dos envolvidos no processo é outro ponto merecedor de destaque. Pouco contribuiria a implantação de

um arsenal tecnológico adequado, sem pessoas motivadas e comprometidas com o papel que exercem na Instituição no cumprimento de sua missão. Assim, o acompanhamento próximo por parte das pessoas detentoras do conhecimento de negócio foi essencial na construção da proposta. Os aspectos humanos muitas vezes são complexos de tratamento, visto que cada pessoa possui uma carga cultural, emocional, física que lhe são peculiares. Sendo assim, precisa-se de meios distintos de tratamento para tornar as pessoas aliadas à proposta, pois em sentido oposto, podem impor séria resistência que prejudicariam em muito a implantação.

Sendo vislumbrados os elementos envolvidos, outro aspecto a ser tratado é quanto à integração harmônica destes para o processo alcançasse os objetivos propostos. De posse dos requisitos funcionais necessários para atendimento das expectativas do negócio, aliando-os com a visão das pessoas envolvidas, o resultado obtido foi a disponibilização de um mecanismo que atualmente responde a contento às demandas do negócio, além de ser reconhecido positivamente pelas pessoas que o utilizam. Isso somente foi possível pelo uso adequado da tecnologia, onde se buscou por métodos e técnicas tendo como prioridade o suprimento das necessidades institucionais, dando ênfase semelhante à usabilidade das interfaces disponíveis aos usuários para simplificação das tarefas operacionais (cadastro, classificação, indexação e recuperação de documento), com intuito de dispor de mais de tempo no desempenho de atividades puramente intelectuais.

Quanto aos processos de gestão documental em uso nas instituições judiciárias na atualidade, ainda é perceptível a ausência de processos bem definidos, bem como de ferramentas para gestão documental. Nestas entidades, grande parte das atividades correlatas a esse tema é realizada de forma *ad hoc*, ou seja, sem padronização ou normatização prévia. Esta realidade acarreta em prejuízos significativos, principalmente no que tange à recuperação do conhecimento institucional. Averiguou-se, entretanto, boas iniciativas do governo brasileiro por parte do CONARQ e do CNJ cujo foco está alinhado com os objetivos deste trabalho no sentido de disponibilizar mecanismos apropriados para gerenciamento do acervo eletrônico. Tais medidas em curto e médio prazo tendem a melhorar expressivamente os serviços prestados à Sociedade relativos a documentos em meio eletrônico.

Durante a aplicação da metodologia CommonKads, identificou-se duas atividades intensivas em conhecimento: Elaboração de Sentenças, Acórdãos e

Despachos e Classificar e Indexar Documentos. Destas atividades, em conformidade com os recursos disponíveis, foi selecionada a segunda como passível de ser automatizada. Assim sendo, foram pesquisadas diversas técnicas da Engenharia do Conhecimento com intuito de conceber uma ferramenta de apoio à sua execução e, como concretização desta pesquisa, foi formalizado um novo modelo de gestão documental conceitual para a Instituição e desenvolvido o Sistema e-Docs como ferramenta tecnológica.

Assim sendo, conclui-se que os objetivos desta pesquisa foram alcançados, pois o modelo implantado, além de envolver os tipos documentais inicialmente propostos, permitiu a abrangência total destes documentos, algo inviável no modelo anterior, sem que para isso houvesse aumento de pessoal ou carga de trabalho. Acrescenta-se também que o Sistema e-Docs tem se tornado a ferramenta institucional para recuperação de documentos e gradativamente está sendo ampliado seu escopo para atendimento de demanda análoga em outros setores.

## **6.2 Trabalhos Futuros**

No transcorrer do trabalho foram identificados outros tópicos também relacionados com o presente tema, porém, em consonância com o escopo preliminarmente definido para a pesquisa, tais assuntos não foram abordados nesta dissertação. Entretanto, por sua relevância, tais assuntos serão apresentados como sugestões para o desenvolvimento de pesquisas futuras.

### Conversão de Tesauro em Ontologia

A utilização do Tesauro da Justiça Eleitoral viabilizou a disponibilização de um mecanismo de pesquisa semântica, visto que, conforme exposto nos capítulos anteriores, o tesauro não se resume a um dicionário de sinônimos, mas sim tem seu enfoque em relacionar termos com proximidade semântica.

Entretanto, o tesauro em uso permite unicamente um tratamento generalizado com respeito às diversas relações semânticas existente entre os termos, assim sendo, não existe diferenciação alguma quanto ao tipo de ligação que une os conceitos formalizados neste instrumento.

Desta forma, um projeto de pesquisa que estivesse centralizado na conversão deste tesauro em uma ontologia para a Justiça Eleitoral, seria de grande relevância para esta área de atuação. Nesta abordagem seria possível ampliar a abrangência das funcionalidades do mecanismo de busca ao permitir a pesquisa documental pelo

tipo da relação semântica entre os conceitos formalizados na ontologia. Além disso, a disponibilização de uma ontologia do Direito Eleitoral abriria frente para a construção de novos sistemas baseados em conhecimento neste domínio.

#### Pesquisa Documental por Semelhança de Conteúdo

Outra linha de pesquisa proposta seria o aprofundamento do tema relativo à classificação de documentos baseando-se nos modelos de recuperação da informação - Espaço Vetorial e Probabilístico - tendo em vista a disponibilização de outras formas de busca documental a partir de um determinado documento jurisprudencial.

Sendo que as matérias julgadas pelos magistrados geralmente envolvem assunto já tratados preliminarmente, tais documentos podem ser utilizados para fundamentação e defesa de uma idéia relacionada com uma situação atual. Desta forma, a disposição de um ferramental que permita a navegação por documentos com conteúdos semanticamente relacionados, consistiria em outra importante ferramenta para disseminação do conhecimento judiciário.

#### Filtragem Colaborativa para o Judiciário

Como forma de aprimoramento para o modelo proposto neste documento, propõe-se também o estudo relacionado a mecanismos de filtragem colaborativa para personificação dos resultados das pesquisas nos documentos judiciários. Este tema permite a proposição de outras ferramentas efetivas direcionadas para a disseminação do conhecimento, por meio de informações pessoais armazenadas durante o processo de pesquisa.

Esta tecnologia permite definir perfis de usuários pelo conjunto de informações de pesquisa armazenadas. Assim sendo, seria possível ao sistema de gestão documental sugerir outros documentos de interesse do usuário com base nas características de seu perfil.

#### Tratamento de Ambigüidade Semântica

Dentre os problemas da pesquisa semântica, não foi abordado no escopo deste projeto, a ambigüidade – característica de uma palavra poder expressar conceitos diferentes. Entretanto, o desenvolvimento de uma solução para este problema seria mais uma forma de apurar a pesquisa semântica proposta neste trabalho.

Esta linha de pesquisa está relacionada com a primeira proposta para trabalhos futuros, visto que, uma ontologia devidamente formalizada pode ser uma ferramenta essencial para o tratamento da ambigüidade.

## REFERÊNCIAS

- ACEVEDO, FERNANDO J. FERNANDEZ. **El documento electrónico en el derecho civil chileno: Análisis de la Ley 19.799**. Ius et Praxis, Talca, v. 10, n. 2, 2004.
- ADAM, AZAD. **Implementing Electronic Document and Record Management Systems**. Boca Raton: Auerbach Publications, 2008.
- ARANO, SILVIA. **Thesauruses and Ontologies**. Hipertext.net, num. 3, 2005.
- ARBIB, MICHAEL A. **Semantic networks**. Em: The Handbook of Brain Theory and Neural Networks (2nd ed.), Cambridge, MA: MIT Press, 2002.
- ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. **Métodos para Análise de Documentos - Determinação de seus Assuntos e seleção de termos de indexação**. NBR 12676. Rio de Janeiro, 1992. p. 4.
- BAEZA-YATES, RICARDO E RIBEIRO-NETO, BERTHIER. **Modern Information Retrieval**. New York: ACM Press, 1999.
- BERRY, M. W. **Survey of Text Mining Clustering, Classification, and Retrieval Scanned by Velocity**. New York: Springer-Verlag, 2004.
- BISWAS, RAHUL E ORT, ED. **The Java Persistence API - A Simpler Programming Model for Entity Persistence**. Disponível em: <http://java.sun.com/developer/technicalArticles/J2EE/jpa>. Acesso: 0/03/2010.
- BOICO, BOB. **Content Management Bible, 2nd Edition**. Indianapolis: Wiley Publishing, 2005.
- BORDOGNA, GLORIA. **A fuzzy linguistic approach generalizing Boolean Information Retrieval: A model and its evaluation**. Journal of the American Society for Information Science Vol: 44 Issue: 2 Pages: 70-82, mar. 1993.
- BRAGA, GILDA MARIA. **Informação, ciência da informação: breves reflexões em três tempos**. Ciência da Informação, v. 24, n. 1, 1995.
- BRASIL. **Tribunal Superior Eleitoral. Thesaurus**. – 6. ed. rev. ampl. – Brasília: TSE/SDI, 2006. p. 260.
- BRASIL. Conselho Nacional de Justiça. **Programa Nacional de Gestão Documental e Memória do Poder Judiciário**. Plano de Trabalho. Brasília, 2008.
- BRASIL. Conselho Nacional de Justiça. **Modelo de Requisitos para Sistemas Informatizados de Gestão de Processos e Documentos do Judiciário brasileiro**. Brasília, 2009.
- BRASIL. Conselho Nacional de Arquivos. **CONARQ**. 2009. Disponível em: <http://www.conarq.arquivonacional.gov.br/cgi/cgilua.exe/sys/start.htm>. Acesso: 10/12/2009.
- BRIET, SUZANNE. **Qu'est-ce que la documentation?** Paris: Presses Universitaires de France, 1953.

- BUSH, VANNEVAR. **As we may think**. The Atlantic Monthly, Washington, jul. 1945.
- CALDERON, WILMARA RODRIGUES; et al. **O processo de gestão documental e da informação arquivística no ambiente universitário**. Ciência da Informação, Brasília, v. 33, n. 3, p.97-104, set./dez. 2004.
- CAVALCANTI, CORDÉLIA R. **Indexação & Tesouro: Metodologia & Técnicas**. Ed. Preliminar. Brasília: Associação de Bibliotecários do Distrito Federal, abr. 1978.
- CLEVERDON, CYRILW. **The significance of the Cranfield tests on index languages**. Em Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 3-12, 1991.
- CONARQ. **Modelo de Requisitos para Sistemas Informatizados de Gestão Arquivística de Documentos**. Brasília: E-Arq Brasil, 2006.
- CORAZZON, RAUL. **Descriptive and formal ontology**. 2000. Disponível em: <http://www.formalontology.it>. Acesso: 02/05/2008.
- CÔRTEZ, MARIA REGINA PERSECHINI ARMOND. **Arquivo público e informação: acesso à informação nos arquivos públicos estaduais do Brasil**. Belo Horizonte: Escola de Biblioteconomia da UFMG, 1996.
- COURTEAU, SARAH L. **A Word by Any Other Name**. Wilson Quarterly, Vol. 34 Issue 2, p105-106, 2p, 2010.
- CRUZ, TADEU. **Wokflow: A Tecnologia que Revolucionou Processos**. Rio de Janeiro: Epapers Serviços Editoriais Ltda., 2004.
- O'HARE, DANIEL. **The Emerging Knowledge Society**. Information Society Commission of Ireland, dez. 2002.
- DAVENPORT, T. H e PRUSAK, L. **Conhecimento Empresarial**. São Paulo: Publifolha, 1999.
- DAVID, EDUARDO. **Do papel ao bit**. Document Management, São Paulo, n.1, p.3, ago. 2007.
- ECKERT, KAI; PFEFFER, MAGNUS e STUCKENSCHMIDT, HEINER. **Assessing Thesaurus-Based Annotations for Semantic Search Applications**. International Journal of Metadata, Semantics and Ontologies, vol. 3, issue 1, jan. 2008.
- FAUSTINO, MARIA NATÁLIA CORREA. **Relatório do programa de gestão documental da Justiça Eleitoral**. Florianópolis: Tribunal Regional Eleitoral de Santa Catarina, 2009.
- FAZENDA, IVANI. **Interação e Interdisciplinaridade no Ensino Brasileiro: Afetividade ou Ideologia**. São Paulo: Loyola, 1993.
- FELDMAN, R.; SANGER, J. **The Text Mining: advanced approaches in analyzing unstructured data**. Cambridge University Press, 2006.
- FENZEL, D.; DAVIS, J.; HARMELEN, F. **Towards the Semantic Web: Ontology-driven Knowledge Management**. Hardcover, 2002.

FRAKES, WILIAM B. E BAEZA-YATES, RICARDO. **Information Retrieval Data Structures & Algorithms**. New York: ACM Press, 1992.

FREITAS JÚNIOR, OLIVAL DE GUSMÃO. **Um Modelo de Sistema de Gestão do Conhecimento para Grupos de Pesquisa e Desenvolvimento**. Tese de Doutorado em Engenharia de Produção - Programa de Pós-Graduação em Engenharia de Produção, Florianópolis: UFSC, 2003.

GIL, ANTÔNIO C. **Métodos e técnicas em pesquisa social**. São Paulo: Atlas, 1999.

GOSPODNETIC, OTIS E HATCHER, ERIK. **Lucene in Action**. Manning Publications, 2004.

GUIMARÃES, JOSÉ AUGUSTO CHAVES. **Análise documentária em jurisprudência: subsídios para uma metodologia de indexação de acórdãos trabalhistas brasileiros**. Tese (Doutorado em Ciências da Comunicação - área Biblioteconomia). São Paulo: USP, 1994.

GRUBER, THOMAS R. **A Translation Approach to Portable Ontology Specifications**. Knowledge Acquisition, vol. 5, Issue 2, Pages: 199 – 220, 1993.

IZZA, SAÏD; VINCENT, LUCIEN e BURLAT, PATRICK. **A Framework for Semantic Enterprise Integration**. Interoperability of Enterprise Software and Applications, Springer London: Londres, 2006.

KOCK, WALTER W. **Os 20 anos da gestão eletrônica de documentos no Brasil**. Document Management, São Paulo, n.1, p.3, ago. 2007.

KRAMMES, ALEXANDRE GOLIN. **Aplicação de Workflow em Processos Judiciais Eletrônicos**. Dissertação (Mestrado em Engenharia e Gestão do Conhecimento) - Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento. Florianópolis: UFSC, 2008. p. 124.

JAPIASSU, Hilton. **Interdisciplinaridade e patologia do saber**. Rio de Janeiro: Imago, 1976. p. 220.

JESUS, LEANDRO MARTINS DE. **Interdisciplinaridade**. Recanto das Letras: São Paulo, fev. 2002.

JOHNSON, ROD. **Introduction to the Spring Framework**. The ServerSide.com, mai. 2005. Disponível em: <http://www.theserverside.com/news/1364527/Introduction-to-the-Spring-Framework>. Acesso: 01/12/2009.

LEE, TIM BERNERS; HENDLER, JAMES E LASSILA, ORA. **The Semantic Web**. Scientific American, mai. 2001.

LIMA, GERCINA ÂNGELA BORÉM. **A transmissão do conhecimento através do tempo: da tradição oral ao hipertexto**. Revista Interamericana de Bibliotecologia, Jul.- Dic. 2007, vol. 30, no. 2, p. 275-285.

MAIMON, ODED; ROKACH, LIOR. **Data Mining and Knowledge Discovery Handbook**. Springer Science, 2005.

MANNING, CHRISTOPHER D.; RAGHAVAN, PRABHAKAR E SCHÜTZE, HINRICH. **An Introduction to Information Retrieval**. Cambridge: Cambridge University Press, 2008.

McGARRY, K. **O contexto dinâmico da informação: uma análise introdutória**. Trad. Helena Vilar de Lemos. Brasília, DF: Briquet de Lemos/Livros, p. 72, 1999.

MIRANDA, ANTONIO; SIMEÃO, ELMIRA. **A conceituação de massa documental e o ciclo de interação entre tecnologia e o registro do conhecimento**. Datagramazero, v.3 n.4 ago. 2002.

MLADENIC, DUNJA. **Text Mining in Action**. Em: Proceedings of the 29th Annual Conference of the Gesellschaft für Klassifikation e.V. University of Magdeburg, Mar. 2005.

MURAKAMI, TIAGO RODRIGO MARÇAL. **Tesauros e a World Wide Web**. São Paulo: USP, 2005.

NASCIMENTO, JULIANO DINIZ DO e ALMEIDA, ADIEL TEXEIRA DE. **Método de Gestão de Conhecimento para a Gestão de Projetos de Fábrica de Software**. Em: XIII SIMPEP: Bauru, 2006.

NECHES, ROBERT; et al. **Enabling Technology for Knowledge Sharing**. AI Magazine, Vol. 12, Issue 3, p. 16-36, 1991.

OLIVEIRA, MARTA K. DE. **Escolarização e organização do pensamento**. Revista Brasileira de Educação, n. 3. São Paulo: Anped, 1996.

QUILLIAN, ROSS. **Semantic Memory**. Em: Semantic Information Processing. Cambridge, Mass.:M.I.T..Minsky, Marvin (ed). Ed. The MIT, 1968, p. 227-270.

RIBEIRO, MARCELO NUNES. **Seleção Local de Características em Agrupamento Hierárquico de Documentos**. Dissertação (Mestrado em Ciência da Computação) – Centro de Informática. Pernambuco: UFP, 2009. p. 66.

RICHARDSON, PETER. **A Crise Documental**. Document Management, São Paulo, n.1, p.3, ago. 2007.

RUIZ, JOÃO A. **Metodologia científica**. 3ª ed. São Paulo: Atlas, 1991.

SCHREIBER, GUUS et al. **Knowledge Engineering and Management: the CommonKADS Methodology**. Cambridge: MIT Press, 2002.

SENGER, PETER. **A quinta disciplina: arte, teoria e prática das organizações da aprendizagem**. São Paulo: Best Seller, 1990.

SILVA, J. U. **Text Mining com uma aplicação dos registros de ocorrências policiais na região da grande**. Dissertação de Mestrado em Ciência da Computação, Universidade Federal de Santa Catarina, Florianópolis, 2005.

SOUZA, ALEKSANDRO BARBOZA DE; et al. **Recuperação Semântica de Objetos de Aprendizagem: Uma Abordagem Baseada em Tesauros de Propósito Genérico**. In: XIX Simpósio Brasileiro de Informática na Educação: Fortaleza, 2008.

SUN DEVELOPER NETWORK. **Java 2 Platform, Enterprise Edition (J2EE) Overview**. 2010. Disponível em: <http://java.sun.com/j2ee/overview.html>. Acesso: 03/03/2010.

SUTTON, MICHEL J. D. **Document Management for the Enterprise: Principles, techniques, and applications**. New York: Wiley Computer Publishing, 1996.

SILVA, ANTONIO PAULO DE ANDRADE E. **Como montar projetos de GED Gerenciamento Eletrônico de Documentos – Linhas mestras para análise, seleção e implantação**. São Paulo: CENADEM, 2008.

TAFNER, ELISABETH P. & SILVA, RENATA. **Metodologia Científica**. Itajaí: ASSEVIM, 2007.

W3C, WORLD WIDE WEB CONSORTIUM. **Leading the Web to Its Full Potential**. 2001. Disponível em: <http://www.w3c.org>. Acesso: 02/05/2008.

WANG, QUAN e NG, YIU-KAI. **An Ontology-Based Binary-Categorization Approach for Recognizing Multiple-Record Web Documents Using a Probabilistic Retrieval Model**. Information Retrieval Vol. 6, Numbers 3-4, Sep, 2003.

WILLIAMS, PETER; JOHN, JEREMY LEIGHTON e ROWLAND, IAN. **The personal curation of digital objects: A lifecycle approach**. Aslib Proceedings, Vol. 61 Issue 4, p340-363, 2009.



## **Anexo II – Tecnologias Utilizadas**

Neste anexo serão descritas de forma sucinta as principais tecnologias presentes na composição das camadas do sistema e-Docs com objetivo de dar amparo técnico necessário para melhor entendimento das funcionalidades disponíveis e quanto ao funcionamento do sistema internamente.

### ***Java Server Faces***

*Java Server Faces* (JSF) tem por objetivo prover meios que viabilize o mapeamento entre os componentes da interface com o usuário e seus respectivos objetos de controle em uma aplicação Java.

Nesta perspectiva, a tecnologia permite o controle do estado dos componentes de interface, o tratamento de eventos ocorridos na camada cliente, a validação das informações providas pelos usuários, facilita a navegação entre as páginas da aplicação e a implementação de mecanismos de acessibilidade e internacionalização.

Uma característica relevante do JSF é a utilização de padrões de projeto consolidados que, dentre outros benefícios, facilitam o desenvolvimento de aplicações com interoperabilidade e que independam do dispositivo que está sendo utilizado pelo usuário. (SUN, 2010).

### ***Spring Framework***

*Spring* é uma arquitetura voltada para facilitar o uso das tecnologias propostas pela plataforma J2EE e promover as boas práticas de programação, principalmente no que diz respeito à camada de negócio de uma aplicação Java. (JOHNSON, 2005).

Assim sendo, a utilização dos padrões preconizados nesta arquitetura reduz o custo de implementação, pois o *Spring* possibilita o baixo acoplamento entre a configuração e a especificação de dependências e a lógica de programação em si. Também provê integração com arquiteturas de persistência de dados, tais como o JPA, facilitando o fluxo de informações entre as camadas de persistência, negócio e cliente.

### ***Biblioteca Lucene***

Segundo Gospodnetic e Hatcher (2004), *Lucene* é uma biblioteca de Recuperação de Informação de alto desempenho projetada para ser agregada a

sistemas de indexação e pesquisas textuais em acervos de documentos eletrônicos. É uma ferramenta de código aberto, implementada em Java e atualmente faz parte da família de soluções *Apache Jakarta*.

A *Lucene* possibilita a indexação e posterior recuperação de qualquer dado do qual possa ser extraído texto, não importando o formato. Isso significa que textos contidos em documentos *PDF*, *Microsoft Word*, *XML* ou *HTML* podem ser pesquisados após um processo de indexação.

Além de garantir um efetivo tempo de resposta, mesmo em consultas a grandes acervos, a *Lucene* permite a ordenação dos documentos recuperados pela relevância destes em relação aos termos de pesquisa, conforme o modelo *Espaço Vetorial* para representação de documentos.

A *Lucene* também disponibiliza mecanismos avançados de RI como busca fonética, redução de termos ao radical comum (*stemming*) e busca semântica por meio de dicionário de termos, funcionalidade que permitirá a integração do tesouro da Justiça Eleitoral com a pesquisa textual proposta neste sistema.

### ***Java Persistence API***

A biblioteca *Java Persistence API* provê um mecanismo para realizar o mapeamento entre as informações armazenadas em banco de dados relacionais para os objetos de negócio em uma aplicação desenvolvida sob a plataforma J2EE. (BISWAS e ORT, 2006).

A biblioteca facilita o processo de persistência, ou seja, gravação das informações manipuladas por sistemas em um servidor de banco de dados, independente do fabricante. Assim sendo, propõe uma abstração quanto a esta importante atividade, garantindo os princípios de segurança da informação e desempenho, necessários para o desenvolvimento de um sistema computacional com qualidade.