INVESTIGAÇÃO DE PROCESSO DE CONVERSÃO AUTOMÁTICA DE TEXTOS ESTRUTURADOS PARA HIPERDOCUMENTOS

ALESSANDRA DORANTE

Dissertação apresentada ao Instituto de Física de São Carlos, da Universidade de São Paulo, para obtenção do título de Mestre em Ciências – "Física Aplicada sub-área Física Computacional"

PROF. DR. ÁLVARO GARCIA NETO
Orientador

São Carlos
1997

Dorante, Alessandra
Investigação de Processo de Conversão Automática de Textos Estruturados para Hiperdocumentos/Alessandra Dorante –
São Carlos, 1997.
142 p.

Dissertação (Mestrado) – Instituto de Física de São Carlos, 1997.

Orientador: Prof. Dr. Álvaro Garcia Neto

1. Hipertexto; 2. Conversão automática. I. Título

UNIVERSIDADE DE SÃO PAULO
Instituto de Física de São Carlos

Av. Dr. Carlos Botelho, 1465
CEP 13560-250 - São Carlos - SP
Brasil

Fone (016) 274-3444
Fax (016) 272-2218

MEMBROS DA COMISSÃO JULGADORA DA DISSERTAÇÃO DE MESTRADO DE **ALESSANDRA DORANTE** APRESENTADA AO INSTITUTO DE FÍSICA DE SÃO CARLOS, DA UNIVERSIDADE DE SÃO PAULO, EM 28 DE NOVEMBRO DE 1997.

COMISSÃO JULGADORA:

Prof. Dr. Álvaro Garcia Neto/IFSC-USP

Prof. Dr. Osvaldo Novais de Oliveira Júnior/IFSC-USP

Profa. Dra. Sandra Maria Aluísio/ICMSC-USP

“Navegar é preciso, viver não é preciso”

ao Álvaro, por acreditar em mim.

AGRADEÇO DE CORAÇÃO

Ao Homero

A Mamy

Ao Grupo de Redes e Hipermídia

Ao Peter

Ao Maikon

A Claudinha

A Wladerez

e a todos meus amigos.

ÍNDICE

LISTA DE FIGURAS v

LISTA DE TABELAS vi

LISTA DE QUADROS viii

LISTA DE ALGORITMOS ix

RESUMO x

ABSTRACT xi

1 INTRODUÇÃO 1

1.1 CONTEXTUALIZAÇÃO 2

1.1.1 O QUE É A CONVERSÃO AUTOMÁTICA 3

1.2 APRESENTAÇÃO DO PROJETO 6

1.3 PROBLEMAS CARACTERÍSTICOS DA ÁREA 8

1.4 SOLUÇÃO PROPOSTA 11

1.4.1 OBJETIVOS 15

1.5 ABORDAGEM EM CAPÍTULOS 16

2 A CONVERSÃO AUTOMÁTICA 18

2.1 PRINCÍPIOS DE HIPERTEXTO 19

2.1.1 CONCEITO DE HIPERTEXTO 20

2.1.2 HIPERTEXTOS "MANUAIS" 21

2.1.3 O ATO DE NAVEGAR 22

2.1.4 A INTERAÇÃO COM A INFORMAÇÃO 23

2.1.5 DIFERENTES PONTOS DE VISTA EM UM HIPERTEXTO: VISÕES 24

2.1.6 DEFINIÇÕES BÁSICAS 24

2.1.7 OS NÓS DE UM HIPERTEXTO 25

2.1.8 OS *LINKS* DE UM HIPERTEXTO 26

2.2 PORQUÊ CONVERTER: PAPEL X HIPERTEXTO 27

2.2.1 DIFICULDADES COM O TEXTO IMPRESSO 27

2.2.2 VANTAGENS DO FORMATO HIPERTEXTO 28

2.3 COMO CONVERTER: CONVERSÃO MANUAL X AUTOMÁTICA 29

2.3.1 CONVERSÃO MANUAL 29

2.3.2 CONVERSÃO AUTOMÁTICA 30

2.4 AS ETAPAS DA CONVERSÃO AUTOMÁTICA 31

2.4.1 SELEÇÃO DE DOCUMENTOS 31

2.4.2 AQUISIÇÃO DE DADOS 32

2.4.2.1 Método 1 – Cópia ou Digitação 33

2.4.2.2 Método 2 – Digitalização 33

2.4.2.3 Método 3 – Texto no Formato Digital 35

2.4.3 PREPARAÇÃO DO TEXTO 37

2.4.3.1 Verificação Ortográfica 37

2.4.3.2 Verificação "Inteligente" 38

2.4.3.3 Verificação Visual 38

2.4.4 APLICAÇÃO DA CONVERSÃO AUTOMÁTICA 38

2.4.5 CODIFICAÇÃO EM LINGUAGEM ESPECÍFICA 40

2.5 TRABALHOS RELACIONADOS 41

3 O PROCESSO DE CONVERSÃO AUTOMÁTICA 44

3.1 TEXTOS ESTRUTURADOS 45

3.2 DESCRIÇÃO GERAL DO PROCESSO 48

3.3 FORMATO DE ENTRADA DO PROCESSO DE CONVERSÃO 49

3.4 SEGMENTAÇÃO 53

3.4.1 PREPARAÇÃO DOS DOCUMENTOS STDNORM 53

3.4.2 ANÁLISE DE PARÁGRAFOS 54

3.5 HIERARQUIZAÇÃO 57

3.6 INFERÊNCIA DE LINKS 59

3.6.1 EXTRAÇÃO DAS CITAÇÕES 60

3.6.2 ENDEREÇAMENTO DAS CITAÇÕES 63

3.6.2.1 Forma Normal para Endereçamento de Segmentos 64

3.6.3 ESTABELECIMENTO DE *LINKS* 66

3.6.3.1 Estrutura de Pares 67

3.6.3.2 Geração do Endereço na Forma Normal 68

3.7 CONSTRUÇÃO DO HIPERTEXTO 70

4 NORMAS: A NATUREZA DA INFORMAÇÃO E SEUS USOS 71

4.1 CONCEITO DE NORMA 72

4.1.1 TIPOS DE NORMAS 72

4.1.2 CARACTERÍSTICAS DE NORMAS 73

4.1.3 PRINCÍPIOS PARA O TRATAMENTO DAS *NEJBS* 75

4.2 ESTRUTURAÇÃO E SEGMENTAÇÃO DE UMA NORMA 76

4.3 DEFINIÇÃO FORMAL DAS *NEJBS* 80

4.3.1 FUNDAMENTAÇÃO TEÓRICA PARA A DEFINIÇÃO FORMAL 80

4.3.2 APLICAÇÃO PARA O CONJUNTO DAS *NEJBS* 84

4.4 REGRAS DE PRODUÇÃO PARA AS *NEJBS* 85

4.4.1 REGRAS DE PRODUÇÃO PARA O PORTUGUÊS DO BRASIL 85

4.4.2 REGRAS DE PRODUÇÃO PARA O CONJUNTO DAS *NEJBS* 87

4.4.3 REGRAS DE VALIDAÇÃO 89

4.5 REGRAS DE PRODUÇÃO PARA AS CITAÇÕES 90

4.6 MODELO DE DADOS PARA *NEJBS* 93

4.6.1 MODELO ENTIDADE RELACIONAMENTO ESTENDIDO 94

4.6.2 ESQUEMA CONCEITUAL DO MODELO DE DADOS PARA *NEJBS* 94

4.6.3 CITAÇÕES NO MODELO DE DADOS PARA *NEJBS* 98

5 A FERRAMENTA WEBIFYLAW 100

5.1 IMPLEMENTAÇÃO 101

5.1.1 LINGUAGEM DE PROGRAMAÇÃO 102

5.1.2 MÓDULOS E BIBLIOTECAS 103

5.2 A GERAÇÃO DE HIPERDOCUMENTOS 104

5.3 USO DA FERRAMENTA *WEBIFYLAW* 108

5.4 RESULTADOS 108

5.4.1 RESULTADOS DA EXTRAÇÃO DAS CITAÇÕES 109

5.4.2 RESULTADOS DA GERAÇÃO DE *LINKS* 111

5.4.3 SITE: PROJETO WEBIFYLAW 116

6 CONCLUSÃO 117

6.1 CONTRIBUIÇÕES DO TRABALHO 118

6.2 SUGESTÕES PARA TRABALHO FUTURO 119

REFERÊNCIAS BIBLIOGRÁFICAS 122

APÊNDICE

LISTA DE FIGURAS

Figura 1 Macro-etapas da Conversão Automática 4

Figura 2 Verificação de Documentos como uma interface 12

Figura 3 Estrutura hierárquica em árvore 45

Figura 4 Exemplo de uma estrutura hierárquica de nível 4 47

Figura 5 Identificação do início de documentos em um documento STDNORM 55

Figura 6 A função da extração das citações 60

Figura 7 Composição de uma *NEJB* 95

Figura 8 Modelo de Dados para *NEJBs* 97

Figura 9 Esquema conceitual para a representação de citações 99

LISTA DE TABELAS

Tabela 1 Padrão STDNORM: formato de entrada de dados no processo de conversão 50

Tabela 2 Algumas categorias presentes nas citações em *NEJBs* 61

Tabela 3 Exemplo de formalização de tipos de citação 62

Tabela 4 Exemplos de endereços na forma normal de endereçamento 65

Tabela 5 Regras para a definição de texto brasileiro, baseado na língua portuguesa 86

Tabela 6 Regras para a definição formal de normas estatutárias jurídicas brasileiras 89

Tabela 7 Regras de Produção para as Citações em *NEJBs* 92

Tabela 8 Construções utilizadas no modelo Entidade Relacionamento Estendido 94

Tabela 9 Exemplos de citações compostas 98

Tabela 10 Opções de configuração 102

Tabela 11 Siglas para os tipos de normas 105

Tabela 12 Resultados da extração das citações por segmento 109

Tabela 13 Casos de extração de citação não coincidentes 109

LISTA DE QUADROS

Quadro 1 Documento de acordo com o padrão STDNORM 52

Quadro 2 Exemplos da estrutura de pares aplicada às *NEJBs* 67

Quadro 3 Exemplo de hiperdocumento gerado automaticamente 106

Quadro 4 Exemplo de um trecho de hiperdocumento com índice interno 107

Quadro 5 Código fonte gerado automaticamente em linguagem HTML 107

Quadro 6 Estatísticas da extração de citações por tipo de citação 110

LISTA DE ALGORITMOS

Algoritmo 1 Preparação de documento padrão STDNORM 54

Algoritmo 2 Etapa de Segmentação no processo de conversão 56

Algoritmo 3 A hierarquização de segmentos no processo de conversão automática 58

Algoritmo 4 Geração de hierarquia 58

Algoritmo 5 Atribuição da ordem do segmento 59

Algoritmo 6 Atribuição do segmento pai 59

Algoritmo 7 Extração das citações na etapa de inferência de *links* 62

Algoritmo 8 Processo de geração de *links* 66

Algoritmo 9 Análise e estabelecimento de *links* 66

RESUMO

Esta dissertação investiga o processo de conversão automática de textos estruturados para hiperdocumentos. Analisa vantagens e desvantagens da utilização de um processo automático. Faz um levantamento detalhado das etapas envolvidas nesta conversão. Como resultado da pesquisa propõe um processo de conversão baseado em definições formais da estrutura dos documentos e das citações. O domínio de aplicação do processo de conversão é o conjunto de *normas estatutárias jurídicas brasileiras*. Outro resultado deste trabalho é a ferramenta *WebifyLaw* que implementa o processo de conversão automática para o conjunto das normas estatutárias jurídicas brasileiras. Os resultados da aplicação da *WebifyLaw* na Constituição Federal, no Código Civil e no Código de Processo Civil e em outras 42 normas são apresentados e discutidos.

ABSTRACT

This work is centered in the research of the automatic conversion of structured texts into hyperdocuments. It presents an analysis concerning the advantages and disadvantages of such automatic process. It also details the steps involved in this conversion. As one of the results it proposes an automatic conversion process, which is based on document structure and citations' formal definitions. The application domain is set as Brazilian statutory norms. Another contribution from this work is a tool called *WebifyLaw* which implements the automatic conversion process for the chosen domain. The tool was applied to the Brazilian Constitution, the Civil Code among other 43 norms. The results obtained in using this application are also presented and discussed.

INVESTIGAÇÃO DE PROCESSO DE CONVERSÃO AUTOMÁTICA DE TEXTOS ESTRUTURADOS PARA HIPERDOCUMENTOS

1 INTRODUÇÃO

Este projeto investiga mecanismos para aproveitamento automatizado de material impresso como fonte de informação para a criação automática de hipertextos. Esta é uma área de bastante relevância devido à grande quantidade de material impresso já existente, uma significativa parte do qual é melhor lido na forma de hipertexto. A ênfase dessa dissertação é no *processo de conversão automática de texto para hipertexto*.

O processo de conversão é de difícil tratamento computacional, e apresenta dois problemas principais: o primeiro é o reconhecimento da formatação e estruturação dos documentos, identificando seus módulos. Outro problema é o tratamento das relações entre documentos, de como utilizar a informação contida nas citações. Estas desempenham

importante papel na construção de relacionamentos entre os textos, que vão resultar em *links* para os hipertextos.

As dificuldades na criação de um processo automatizado se devem, principalmente, à multiplicidade de formatos utilizados na composição de documentos e de citações, exigindo uma flexibilidade na interpretação. Uma parte considerável da informação de certos textos encontra-se em forma não-verbal, incluindo a formatação de títulos e a forma das citações internas.

O domínio de aplicação para teste do processo escolhido foram as *normas estatutárias jurídicas brasileiras – NEJBs*. Esse domínio foi escolhido porque possui várias vantagens: tem sua divulgação, em grande parte, ainda voltada para a mídia impressa e tem surgido grande interesse desde a última década, na colocação desses documentos em meio digital ([27], [28]). Além disso possui uma estrutura e formatação mais formal do que outros textos. Esta iniciativa de conversão foi feita para suprir duas necessidades básicas: alimentar bancos de dados e para serem disponibilizados na forma de hipertexto.

1.1 Contextualização

A tecnologia de hipertexto causou forte impacto no modo de interação e manipulação da informação. A informação mais dinâmica, a divulgação mais ágil, a atualização de material mais eficiente, a facilidade de navegação são algumas entre as muitas vantagens do hipertexto. Os benefícios da utilização de hipertexto já estão comprovados pela popularidade alcançada em diversos domínios de aplicação, como por exemplo *helps* interativos e obras de referência.

Anteriormente divulgada através de CD-ROMs ou mesmo por meio dos discos flexíveis, com o advento da WWW na Internet, a informação também passou a ser divulgada *on-line*. Hoje, a tecnologia Web está cada vez mais presente no cotidiano das pessoas. Há uma

verdadeira explosão no crescimento de *web sites*. Desde aproximadamente janeiro de 1993, o crescimento tem sido exponencial. De acordo com estatísticas, o número de *web sites* dobra a, aproximadamente, cada 6 meses ([30], [18]). Juntamente com o aumento do número de *web sites*, aumentou também a produção de hipertextos.

De acordo com Allan [2], conforme a Web e outros hipertextos crescem e se tornam mais dinâmicos, é cada vez menos provável que a informação seja processada somente com meios manuais.

Apesar de a criação de hipertextos simples não mais exigir uma grande especialização, quando se trata de informação altamente interrelacionada ou em grande volume, essa especialização ainda é necessária. Num processo não-automatizado de criação de hipertextos, os projetos são abordados individualmente, e a cada hiperdocumento a ser produzido todo o processo é repetido [33]. Assim, os benefícios da utilização dessa tecnologia podem ser prejudicados pelo alto custo de criação e manutenção derivados deste modo de produção. Além desses aspectos, existe um grande volume de informação que ainda se encontra armazenada na mídia tradicional: livros e documentos impressos que precisam ser convertidos para hipertexto.

Tal é a área de atuação da conversão automática de texto para hipertexto, cujo objetivo é minimizar as dificuldades inerentes aos fatores mencionados acima. Para otimizar a geração de hipertextos, a conversão automática busca automatizar as tarefas repetitivas e mecânicas, tornando viável a produção de hipertextos de qualidade em um menor intervalo de tempo.

1.1.1 O Que é a Conversão Automática

A conversão automática de texto para hipertexto é um processo de geração de estruturas hipertextuais a partir do texto na sua forma linear. Apesar das evoluções tecnológicas

alcançadas, o processo de conversão não é geralmente um processo automatizado, ainda necessitando da participação humana para confirmar certas inferências geradas automaticamente.

Toda a geração de um hipertexto a partir de um texto já existente pode ser generalizada em cinco macro-etapas [36], ilustradas na **Figura 1**:

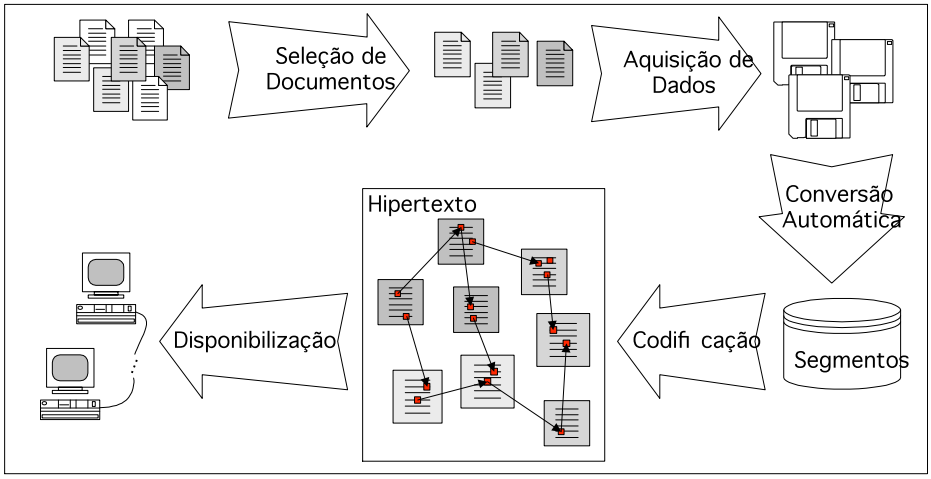


Figura 1 Macro-etapas da Conversão Automática

Seleção de Documentos: Definição do conjunto de documentos que serão convertidos em hiperdocumentos.

Aquisição de Dados: Transformação dos documentos selecionados em dados digitais.

Aplicação da Conversão Automática: Segmentação da informação em módulos e criação de hiper-elos ou *links*.

Codificação em Linguagem Específica: Transformação da informação em uma linguagem com suporte a hipertextos.

Disponibilização: Integração a um servidor conectado.

Desde o momento da escolha do texto até o momento no qual o hipertexto está pronto e disponível para interação com o usuário, todo este processo é objeto da conversão automática. As duas primeiras etapas, seleção de documentos e a aquisição de dados,

funcionam como processos externos que fazem o tratamento da informação colocando-a no formato de entrada do processo de conversão.

Assim, a análise dessas etapas preliminares é importante para identificar características do processo como um todo e para que o processo de conversão possa atuar com sucesso sobre a informação desejada.

Depois da aplicação da conversão automática, o resultado é então codificado em uma linguagem específica. Esta linguagem pode ser HTML, SGML, HyperCard ou qualquer outra linguagem que dê suporte para mecanismos de relações entre segmentos de informação em geral.

Após a codificação, a informação deve ser disponibilizada para uso. Atualmente os meios mais comuns para isto são a publicação em CD-ROMs ou a disponibilização *on-line* na WWW da Internet.

O processo de conversão automática engloba as duas penúltimas etapas: aplica a conversão automática e codifica os resultados em uma linguagem específica. A última etapa, a disponibilização, não é automatizada por esse processo de conversão.

Técnicas de diversas áreas, além da área dos conhecimentos específicos em hipertexto, são necessárias para a efetivação de todo o processo: Banco de Dados, Indexação e Recuperação de Informação, entre outras, para que se possa chegar a um resultado satisfatório. A utilização de técnicas mais específicas, de áreas como a Inteligência Artificial e Sistemas Especialistas, é uma prática que pode vir a melhorar o desempenho dos processos de conversão, mas está fora do âmbito deste trabalho. Conhecimentos dessas áreas são utilizados em sistemas para "aprender" a identificar *links* entre documentos a partir de textos marcados por um leitor [6]. Esses sistemas "aprendizes" supõem a intervenção humana na maioria dos casos para confirmar inferências.

1.2 Apresentação do Projeto

Trabalhando com a *conversão automática de texto para hipertexto* e definindo o domínio de aplicação como o conjunto das *normas estatutárias jurídicas brasileiras*, o alvo prático do projeto consiste em estabelecer um *processo automático* e desenvolver uma ferramenta – *WebifyLaw* – através da qual um conjunto de normas possa ser automaticamente convertido para um *hipertexto*. O alvo principal da conversão é gerar informação na forma de documentos HTML para que possa ser facilmente disponibilizada na WWW da Internet.

Apesar do grande número de aplicativos visando a conversão de hipertextos surgido nos últimos anos, estes são normalmente baseados em algum tipo de marcação existente no texto [2]. Esta marcação pode ser basicamente de dois tipos: *descritiva* ou *procedimental* [40].

A *marcação descritiva* faz uso de códigos embutidos no texto, ou etiquetas de marcação, para descrever ou comentar elementos de documentos. Exemplos de linguagem de marcação descritiva são Latex, SGML e HTML. A aplicação Latex2html [13] é um exemplo de interpretação das etiquetas de marcação Latex, e conversão para etiquetas HTML.

A *marcação procedimental*, de acordo com Tittel [40], é "uma forma de descrever conteúdos e *layouts* de documento que destaca o modo de apresentação, em vez de elementos estruturais incorporados nos documentos". Este tipo de marcação é muito comum nos documentos provenientes de processadores de marcas proprietárias, como o Microsoft Word, por exemplo.

Nesses casos de marcação, os aplicativos funcionam como simples conversores de formatos. Interpretam as marcações existentes nos textos e se alguma formatação ou estruturação estiver presente neste formato, esta é convertida para o hipertexto. Caso contrário, o texto é mantido com sua estrutura original linearizada, e fica descaracterizado

como hipertexto. Deixam a desejar, portanto, quando se trata de estabelecer uma segmentação para o documento e de realizar a inferência de *links*, aspectos fundamentais de um hipertexto.

As *normas estatutárias jurídicas* são normalmente apresentadas para consulta como blocos de texto monolíticos condensados em volumes de acordo com a ordem cronológica de sua publicação, como por exemplo, faz o LEX. Este suporte restringe algumas ações, como a busca por palavras-chave, e dificulta outras, no caso de se desejar consultar as citações presentes na norma, por exemplo.

O processo de conversão aqui proposto parte da informação em seu formato original – texto em bloco monolítico, seqüencial, e sem formatação – identifica elementos e atribui seu verdadeiro significado sintático: cabeçalho, seções, subseções, parágrafos, itens e assim por diante, como segmentos componentes de uma norma.

Uma vez identificados de acordo com categorias estabelecidas e armazenados devidamente, os elementos têm suas relações exploradas através da inferência de *links*. Posteriormente são "reagrupados", mas desta vez, em estrutura de hipertexto.

Em face da definição do domínio de aplicação algumas restrições ficam assim estabelecidas em derivação desta, e foram importantes no desenvolvimento da ferramenta *WebifyLaw*:

- A língua para os documentos fonte é o Português do Brasil. Aplicam-se, deste modo, todas as restrições da língua na consideração de caracteres, vocábulos e abreviaturas permitidos. Exceções a esta regra devem ser tratadas como tal.
- Outras normas, que não as estatutárias jurídicas brasileiras não se encaixam neste contexto e assim não são alvo da ferramenta *WebifyLaw* implementada neste trabalho. A conversão automática de informação de outros domínios pode ser feita através da

aplicação do processo de conversão automática proposto no **Capítulo 3**, adaptando cada uma das suas etapas.

- O tipo de informação a ser processada pela ferramenta é a informação textual, que corresponde ao texto das normas acima definidas. Assim, o trabalho proposto se concentra principalmente em técnicas de manipulação destas informações para criar hipertextos.
- Informação não textual é tratada como um *componente externo* ao propósito do trabalho e por conseqüência é também um componente externo ao processo proposto. Esse formato de informação não gera hiper-elos.

1.3 Problemas Característicos da Área

Partindo da informação em seu estado original: seqüencial e sem formatação, para que seja possível chegar ao resultado desejado, os problemas da criação de hipertextos devem ser detalhados. Esta seção descreve os problemas que envolvem o alcance do objetivo final. A seção seguinte apresenta as propostas de solução para cada um deles.

Problema 1: A Aquisição de Dados – Qual deve ser o método para a aquisição de dados e o formato de entrada dos dados?

Da etapa de aquisição de dados resultam os documentos digitalizados, que servirão de entrada para o processo de conversão para os documentos hipertexto. É uma etapa importante pois sem dados de qualidade não é possível criar hipertextos de qualidade. Questões relevantes nessa etapa são: quais os tipos de suporte possíveis para a informação de entrada; quais são os métodos de aquisição de dados; e qual o tempo envolvido em cada um deles. Por exemplo, para um determinado documento pode-se ter uma versão impressa de qualidade e uma versão digital. A princípio, o uso da versão digital pode parecer óbvio. Mas esta decisão depende da disponibilidade de conversores deste formato

digital para o formato aceito pelo processo de conversão. Caso contrário, o método mais adequado para a aquisição de dados passa a ser a digitalização da versão impressa.

O formato de entrada dos dados é outro aspecto importante. O resultado do processo de aquisição de dados está intimamente relacionado com os documentos de entrada. Quanto menor a necessidade de outros processos entre estes, menor o esforço na realização da tarefa. Para tal, deve-se procurar utilizar a saída do processo de aquisição diretamente como entrada do processo de conversão. Isto representa menor esforço, no entanto é um caso raro de ocorrer.

Problema 2: A Segmentação de Dados – Como estabelecer uma segmentação para a informação que atribua o significado sintático de seus elementos, tal como é feito quando executamos a leitura de um documento?

Durante a leitura de um texto, nosso cérebro vai identificando e estruturando suas partes. Pela nossa capacidade de abstração é possível identificar o título de uma seção, o assunto principal de um parágrafo, os itens subordinados ao mesmo, e assim por diante. Como funciona esse mecanismo que ocorre automaticamente sem a ação voluntária? A resposta a esta pergunta é fonte de inspiração para tentar identificar elementos do documento e segmentá-lo corretamente.

A solução para este problema parte da investigação de quais são os elementos significativos do conjunto de documentos em questão, e como tais elementos podem ser distinguidos uns dos outros.

Problema 3: O Armazenamento de Dados – Como fazer o armazenamento dos dados? Qual é a base para a definição de uma base de dados para segmentos de informação interrelacionados?

É um problema relacionado mais diretamente com a tecnologia da informação do que com a tecnologia de hipertexto. No entanto, a solução para esse problema deve buscar adaptar a tecnologia para as características de um hipertexto.

Um dos pontos é que a estrutura de suporte do armazenamento dos dados deve permitir acesso rápido aos itens mais requisitados de uma norma. As soluções devem ser buscadas na área de banco de dados, por já tratar tais problemas. O armazenamento de dados além de representar as construções convencionais como a hierarquia entre elementos de um documento, deve ser projetado para suportar também as outras relações neles presentes. Exemplificando, no caso das *NEJBs*, deve ser oferecido um suporte para estas relações, que seriam as citações.

Problema 4: A Inferência de *Links* – Quais são os tipos de *links* que devem ser estabelecidos no conjunto das informações? Como os *links* são determinados a partir dos segmentos do documento?

A inferência de *links* é um assunto subjetivo. Muitos *links* podem ser criados. No entanto, não é boa prática estabelecer todos os *links* possíveis[1]. E sim, determinar quais são essenciais e representam verdadeiramente uma relação entre dois segmentos de uma norma. Cabe assim, propor limites para que a subjetividade inerente à criação de *links* possa ser revertida em casos objetivos e determinados. Uma vez feito isto, o problema consiste em: como identificar tais relações no texto, uma vez que sua composição parece livre?

[1] Isto vai contra os princípios de composição de um bom hipertexto, gerando um documento confuso.

Problema 5: A Construção do Hipertexto – Como o hipertexto deve ser construído? Quais recursos devem ser acrescentados para a apresentação para o usuário?

Esta é a etapa alvo deste trabalho, construir o hipertexto. Os problemas anteriores trataram de explorar como segmentar adequadamente a informação e descobrir as relações nela presentes. Nesta etapa, a perspectiva se inverte. O objetivo é colocar os segmentos juntos novamente mas, desta vez, no formato de hipertexto. A problemática aqui envolvida é como coordenar todas as informações levantadas na construção de novos documentos.

1.4 Solução Proposta

Uma vez que foram apresentados os problemas, nesta seção as soluções propostas são discutidas.

Solução Proposta para o Problema 1: A Aquisição de Dados

A informação de entrada para o processo de conversão está restrita a praticamente dois suportes. O primeiro é o documento impresso, o mais comum. Outro caso é o documento digital, já armazenado digitalmente, normalmente de acordo com algum formato de processador de texto. Cabe determinar quais são os diversos métodos para a aquisição de acordo com cada suporte e com suas condições particulares e, estabelecer um formato de entrada único para todos os documentos no processo.

Da aquisição, a informação deve ainda, passar por uma verificação numa etapa intermediária para ser validada e aí ser colocada no formato de entrada. Considerando a verificação de documentos como uma interface, conforme ilustra a **Figura 2**, ao determinar um formato de entrada próximo do formato de saída da aquisição de dados, diminui-se o trabalho desta etapa.

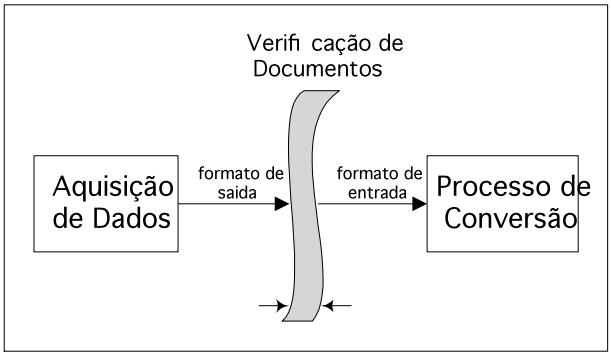


Figura 2 Verificação de Documentos como uma interface

Solução Proposta para o Problema 2: A Segmentação de Dados

A segmentação de uma norma está intimamente relacionada com a identificação dos seus elementos significativos. Os futuros segmentos são os elementos significativos separados em blocos. Assim, para resolver estas questões, a solução proposta é tentar estabelecer uma definição formal da estrutura do documento em questão. Esta definição formal passa a ser responsável por colocar todos os documentos de um conjunto em uma situação uniforme, fazendo com que seja possível aplicar um mesmo método de segmentação para todos eles. Assim, o primeiro passo na solução do problema da segmentação é estabelecer uma única definição da estrutura dos documentos alvo que possa ser seguida durante as fases subseqüentes.

Uma vez que esta estrutura estiver determinada, o passo seguinte é determinar um método para a segmentação, a solução para a segunda parte do problema. A proposta é partir para a realização de análises léxica e sintática baseadas nesta definição formal de normas. A aplicação destas análises vai produzir os segmentos do documento, representando elementos significativos (identificados como, por exemplo: capítulos, seções, e parágrafos). A partir daí, a preocupação passa a ser o armazenamento dos segmentos, o próximo problema.

Solução Proposta para o Problema 3: O Armazenamento de Dados

A solução para o armazenamento de dados é definir um modelo de dados específico para o domínio de aplicação. Assim, o modelo de dados assume a função-chave de uniformizar o armazenamento dos segmentos de normas identificados em etapas precedentes. Outras funções do modelo surgem da consideração das formas básicas de utilização dos segmentos:

- Na recuperação de elementos citados: Considerando que no conteúdo dos documentos, os segmentos são freqüentemente citados – um tipo de relação explícita que tem como suporte as relações estruturais[2]. Na sua forma, a referenciação através de uma citação é variável[3]. Além disto, a citação pode ser feita a documentos externos, ou a segmentos do próprio documento. Tal referenciação, ou o estabelecimento de um *link*, implica na identificação e recuperação do segmento citado.

- E na recuperação de elementos específicos, ou recuperação de conteúdo: Além de citações, documentos podem ser referenciados através dos assuntos nele abordados. Trata-se de um caso de relação implícita[4], a qual também deve ser representada no modelo de dados.

O modelo de dados deve ser projetado para permitir tais utilizações. Em ambos os casos, a recuperação é de responsabilidade do modelo de dados. É possibilitada através das outras etapas: a de segmentação[5] e da inferência de *links*[6].

Na definição do modelo de dados, a fonte de informação principal é a definição formal, parte da solução anterior. Isto porque na definição formal é que estão definidos todos os

[2] A descrição de tipos de relação é abordada no **Capítulo 2,** página 39.
[3] A citação é abordada especificamente na Solução Proposta para o **Problema 4**.
[4] Idem à nota 2.
[5] Solução proposta para o **Problema 2**.
[6] Solução proposta para o **Problema 4**.

elementos significativos dos documentos, com exceção das citações. Do modelo define-se uma base de dados[7] na qual as diversas operações possam ser executadas sobre os segmentos a contento.

Solução Proposta para o Problema 4: A Inferência de *Links*

Uma observação analítica dos documentos permite identificar os três tipos básicos de relações entre elementos: as referências de hierarquia, de citações e de assunto. Cada uma destas exige métodos específicos para sua identificação. As referências de hierarquia são abordadas principalmente durante a construção do hipertexto[8]. As referências relacionadas ao assunto são dependentes da indexação do texto do segmento. As de citações são a preocupação principal nesta etapa.

Uma citação sempre possui um *segmento emissor* ou *hospedeiro* – aquele que contém a citação –, e um *receptor* – aquele que é citado. O uso da citação indica várias situações, as diferentes relações entre os segmentos relacionados[9]: a dependência, a complementação, o esclarecimento, ou até a negação. Para poder estabelecer *links* representando estas citações, é necessária uma análise que as reconheça no texto. A análise é dificultada pela característica de composição livre da citação. A proposta neste caso consiste em examinar uma amostra de documentos, fazendo a análise de *corpus* e extrair exemplos mais comuns de citação. Com base nesta amostragem tenta-se estabelecer uma forma normal e um método para a extração automática de citações. As citações que caírem dentro desta forma normal são reconhecidas pelo processo. Outras formas devem ser tratadas como uma exceção, necessitando da intervenção do usuário.

[7] A base de dados, depois de definida, pode ser implementada utilizando o aplicativo de SGBD que se preferir.
[8] Para a referência mais próxima, veja a Solução Proposta para o **Problema 5**.
[9] A melhor comprovação para esta afirmação ainda é a leitura crítica dos textos de normas.

Solução Proposta para o Problema 5: A Construção do Hipertexto

A construção do hipertexto é uma etapa que permeia todo o desenrolar deste trabalho. É o objetivo final de todas as outras propostas permitir que se chegue a esta etapa com as informações necessárias. Considerando isto, a proposta é construir o hipertexto com base nos segmentos identificados e devidamente armazenados. Através de consultas na base de dados são recuperados os segmentos de uma mesma norma e as informações sobre a hierarquia e outras relações. São criados diversos documentos no formato de saída escolhido, cada um contendo um hipertexto isolado. Para o usuário será apresentada uma espécie de índice no início para representar a estrutura interna do documento e permitir acesso rápido ao ponto desejado, além dos outros *links* presentes no corpo daquele.

A complicação maior está em coordenar todas as informações e produzir um hipertexto representando fielmente a estrutura do documento original. Na solução deste problema, deve ser especificado um método que promova o encaixe perfeito das informações já levantadas: os segmentos e as relações entre eles.

Com relação à manutenção do sistema, a base é a comparação das mudanças a serem feitas com o estado atual do sistema. Na inserção de novos hipertextos ao conjunto devem ser consideradas quais as referências que estes fazem a outros hipertextos já presentes no conjunto. Neste caso, estas referências devem ser estabelecidas.

1.4.1 Objetivos

Os objetivos principais deste trabalho se concentram em:

1. Especificar <u>métodos para a aquisição de dados</u> para cada suporte de acordo com as condições particulares.

2. Apresentar diretivas para a realização da etapa de verificação dos dados, antes que estes atinjam o formato de entrada.

3. Especificar um formato para entrada de dados, e apresentar diretivas de como alcançá-lo partindo da aquisição dos dados.

4. Estabelecer uma definição formal para a estruturação de *normas estatutárias jurídicas brasileiras*.

5. Definir e implementar um método para a segmentação de normas com base na especificação formal.

6. Especificar um modelo de dados, utilizando o modelo entidade-relacionamento estendido como representação formal.

7. Definir uma forma normal para o reconhecimento e a representação de citações .

8. Definir formalmente um método para a extração automática de citações com base na forma normal para citações e especificá-lo em termos de um algoritmo que possibilite sua implementação.

9. Especificar e implementar um método para a construção de documentos hipertexto a partir do modelo de dados, que apresente documentos em HTML como formato de saída. Índices por categoria e por ordem cronológica também devem ser apresentados como resultado da aplicação do método.

1.5 Abordagem em Capítulos

Esta seção descreve a estrutura desta dissertação. O **Capítulo 2** investiga o tema "Conversão Automática". Discute outros sistemas desenvolvidos que foram utilizados como embasamento teórico para o desenvolvimento do projeto proposto. Também explora as vantagens e desvantagens da utilização de um processo para a conversão automática, e

apresenta diretivas de como executar uma conversão de texto para hipertexto generalizada.

O *processo de conversão* é discutido no **Capítulo 3**. Esse capítulo apresenta uma definição detalhada do processo de modo generalizado, sem aludir a um domínio específico de aplicação.

As *normas estatutárias jurídicas brasileiras* são revistas no **Capítulo 4**, definindo a natureza da informação, como é criada, estruturada, e seus usos atuais. A partir daí, forma-se a base para levantamento de questões pertinentes à estruturação de normas em geral. Como solução são apresentadas duas formalizações: uma definição formal da estruturação e um modelo de dados para o armazenamento de *NEJBs*.

O **Capítulo 5** aborda o desenvolvimento, a implementação e o uso da ferramenta *WebifyLaw* que emprega os conceitos discutidos nos **Capítulos 3** e **4**, passando por todos os passos da conversão automática.

Finalmente, o **Capítulo 6** apresenta as conclusões, as contribuições do trabalho e as sugestões para trabalhos futuros.

2 A CONVERSÃO AUTOMÁTICA

O grande desenvolvimento alcançado nos anos 80 pelos processadores de texto mostrou que o texto na sua forma eletrônica pode ser manipulado e visualizado de uma maneira muito mais flexível do que o texto na sua forma impressa [25]. Os processadores de texto conquistaram espaço revolucionando o processo de editoração da informação através da exibição dos resultados diretamente na tela, simulando o uso do papel no computador.

O armazenamento e consumo de textos em forma eletrônica permite remeter rápida e eficientemente o leitor de uma parte para outra do documento. Esta capacidade de navegação, que liberta o leitor das restrições impostas pelo formato físico da encadernação, é uma das bases do mecanismo de hipertexto.

Seguindo a analogia, com o hipertexto, pode-se explorar uma simulação do ato físico de seguir referências de uma citação para outra. Domínios específicos caracterizados fortemente como obras de referência e de grande volume de informação só têm a se beneficiar desta tecnologia. Mesmo assim, ainda hoje, grande parte do material de referência ainda é consultado na forma impressa, talvez porque as dificuldades de disponibilizar este material no formato digital sejam grandes.

Esse capítulo introduz a tecnologia básica em hipertexto, que é o resultado final e alvo do processo de conversão. Também faz uma análise dos processos de conversão automática, explorando as razões que justificam o interesse nessa conversão e investigando as dificuldades inerentes ao processo.

2.1 Princípios de Hipertexto

Antes mesmo de o termo hipertexto ser utilizado pela primeira vez[10], os princípios relacionados a ele já tinham sido definidos por Vannevar Bush em 1945, no seu artigo "*As we may think*" [8]. Bush defendia a utilização de meios de manipular a informação inspirados na mente humana, nos processos mentais. O meio atual de apresentação ou manipulação da informação ainda nos remete à linearidade imposta pela cultura impressa [26]. A mente humana por sua vez opera por associações, relacionando as diversas informações que acumulamos durante o período de vida. As informações que nossa mente manipula provém da interação com o mundo experimentado por nós: interações com (grupos de) indivíduos, interações com artefatos; consistindo-se de fontes naturalmente multimídia, ricas em imagens, sons, figuras e outros elementos de linguagem. Nós estamos continuamente transformando, gerando esse universo, produzido pela comunicação, na qual somos atores; não podemos escolher ficar fora, esse universo nos invade. Ao ouvir

[10] O termo *hipertexto* foi instituído por Ted Nelson em 1965.

uma palavra, é ativada em nossa mente uma rede de infinitas conexões com outras palavras e também com sons, odores, imagens remotas e imaginárias, sensações e lembranças .

Longe de tentar ser uma duplicação da mente humana, o hipertexto tem na mente humana a sua fonte de inspiração, materialização de parte da idealização de Bush.

2.1.1 Conceito de Hipertexto

Hiper. [do gr. Hypér.] pref. = 'posição superior'; 'além'; 'excesso'.
Texto. [do lat. Textu, 'tecido'.] s.m. 1. Conjunto de palavras, de frases escritas[11].

O termo hipertexto surgiu, em contraste com o texto tradicional, como um tipo de texto que excede, supera ou vai além do texto tradicional. Um texto tradicional é seqüencial, uma característica que define o modo de construção e manipulação do texto. Há uma única seqüência definindo a ordem de leitura: uma página após a outra.

A imposição de uma ordem também está presente no processo de escrita. Mesmo raramente utilizando um raciocínio linear, o autor é obrigado a linearizar o texto para a sua publicação.

O hipertexto é essencialmente não seqüencial, não existe seqüência única de escrita ou leitura [31]. Analogias a "hipertextos manuais" serão feitas adiante, mas o hipertexto, tal como o concebemos hoje, depende do computador para sua existência. O hipertexto só pôde ser concretizado quando no cenário da informática conseguiu-se colocar em um mesmo ambiente, de modo integrado, diversas características:

- evolução do hardware com aumento da capacidade de processamento e de utilização de memória, diminuição do custo de equipamentos;
- representação da informação por interfaces gráficas utilizando ícones;

- abstração dos comandos em ações executadas com o uso de menus e ícones, eliminando a necessidade de conhecimento de comandos;
- utilização do mouse como dispositivo indicador, permitindo agir de forma 'mais' intuitiva;
- evolução no campo das comunicações, permitindo a larga utilização das redes;
- evolução da estruturação da informação permitindo bancos de dados distribuídos e acesso concorrente da informação.

Encontrar uma única definição para a concepção do que seria um hipertexto é uma tarefa difícil. Conforme esta tecnologia é utilizada em maior escala, suas possibilidades são expandidas, e assim o hipertexto é redefinido. Também não encontraremos sistemas que incorporem todas as características que serão abordadas aqui. As características são incorporadas ou não aos sistemas atendendo a propósitos específicos.

2.1.2 Hipertextos "Manuais"

De algum modo todos nós já tivemos contato com o princípio que rege o hipertexto. Simplificando o conceito de hipertexto, o que pode ser chamado de "hipertexto manual" nos rodeia de alguma forma: cartões e livros de referência como manuais, dicionários e enciclopédias [10]. Através deste tipo de manipulação da informação já podemos identificar algumas características:

- a informação está estruturada em blocos separados;
- existe sempre uma ordem de apresentação adotada, normalmente alfabética;
- a consulta é feita em um pequeno número de blocos dentro do todo; os blocos possuem referências a outros blocos relacionados;

[11] Definições dadas pelo Dicionário Aurélio [15].

- o leitor procura por outros itens de acordo com seu interesse ou necessidade.

Dentro de um texto seqüencial, também existem outras estruturas, como índices, sumários, notas de rodapé, e referências bibliográficas que possibilitam acesso à informação de modo não-seqüencial.

Quando utilizamos um livro, esquecemos que este tipo de "dispositivo lógico" nem sempre existiu. Mencionando "algumas interfaces da escrita", Lévy [26] nota que estamos "tão habituados com esta interface que nem notamos mais que existe". Mas quando do seu surgimento na substituição do manuscrito, pessoas provavelmente também passaram por uma fase da adaptação. Nós nos encontramos hoje em uma fase de transição semelhante. A interface mais comum fornecida pelo computador ainda se restringe à tela, o que não permite a interação do modo que estamos habituados a realizar com livros ou jornais. Meios de manipulação específicos devem ser desenvolvidos para esta interface [45]. E outras interfaces irão surgir, e mudar nosso modo de interagir, ver e viver o mundo.

2.1.3 O Ato de Navegar

Em um hipertexto, o usuário é livre para caminhar pela informação disponível, através dos *links* estabelecidos, ou estabelecendo novos. Este caminho é percorrido conforme o curso ou interesse determinado pelo "comandante" que é o leitor do hipertexto.

A navegação possível em um hipertexto abre novas possibilidades de exploração da informação. O que antes, em um livro também era possível através do ato de simplesmente folhear páginas ou lendo trechos de parágrafos, no hipertexto assume um novo caráter. A velocidade em que a informação é trazida diante da tela é que aproxima a comparação com a navegação. Técnicas de busca se encarregam de selecionar informação de interesse, quando a navegação é direcionada. Já o ato de navegar curiosamente pelo hipertexto, como o folhear de um jornal é expressado pela ação de *browsing*.

Devido à vastidão de informação disponível em alguns hipertextos, os diversos caminhos possíveis, e o fato de somente uma seção deste caminho ser exibida por vez, a chance de se tornar "perdido no hiperespaço" aumenta consideravelmente. Na navegação no hipertexto, mecanismos para a orientação são fundamentais. Permitem ao leitor maior 'segurança' na navegação. Os mecanismos de orientação no hipertexto como a gravação do caminho percorrido – *backtracking* –, para consulta posterior, e visualização da área sendo visitada por mapas, facilitam a localização e situam o leitor.

2.1.4 A Interação com a Informação

Imaginando um hipertexto, de acordo com o idealizado por Vannevar Bush e Ted Nelson, como algo passível de representação física, teríamos um espaço comparável a uma megabiblioteca [26]. Um espaço onde fosse possível armazenar uma gama de informações abrangendo todo o universo de conhecimento. Interagindo neste espaço, pessoas "navegam" fazendo consultas, anotações, inserindo comentários em textos, imagens, filmes. Esta é uma representação imaginária da interação em um 'hipertexto físico' na sua totalidade.

Com a informação impressa, realizamos outras formas de interação. Os sublinhados e destaques, entre outros, freqüentemente feitos nas bordas e entrelinhas de um livro e as anotações paralelas auxiliam no entendimento, traduzem a forma de raciocínio e até explicitam parte dos processos cognitivos durante a leitura. Estes registros utilizados muitas vezes para 'reavivar a memória' representam modos de interagir com a informação. No hipertexto, tais registros também são possíveis, e se estendem quando possibilitam o estabelecimento de *links* pessoais, relacionando novas informações, acrescentando dados; criando e recriando o espaço hipertextual.

2.1.5 Diferentes Pontos de Vista em um Hipertexto: Visões

As visões são estabelecidas em um hipertexto para que a mesma informação possa ser consultada de maneiras diferentes por grupos diferentes de leitores.

Assim, considerando um hipertexto educacional, a informação é consultada por educadores e por alunos. Utilizando-se do conceito de visão, o educador tem permissão, por exemplo, para editar o conteúdo da informação, visualizar as anotações de todos os alunos e incluir correções [19]. Os alunos possuem limitações com relação à edição da informação existente, mas podem ser autorizados a estabelecer *links* próprios, inserir novas informações, comentários e até talvez examinar anotações e soluções de outros (após a data da entrega, por exemplo).

Para o atendimento de grupos com diferentes níveis de conhecimento, pode-se estabelecer visões determinando níveis como introdutório, normal ou avançado. As tarefas propostas, ou a complexidade do assunto para cada grupo são determinadas de acordo com a visão.

Um único assunto em um hipertexto pode assumir diferentes funções com o uso das visões. No caso de um manual, podem ser estabelecidas visões para: treinamento completo, manutenção de rotina, solução de problemas urgentes – *troubleshooting*, ou até guia do proprietário [36].

2.1.6 Definições Básicas

Um hipertexto é estruturado pela divisão da informação em pedaços interligados entre si. Os "pedaços" de informação são objetos básicos de um sistema hipertexto. Não possuem tamanho pré-determinado; a intenção é que estes armazenem idéias concisas. A nomenclatura para estes pedaços de informação difere de acordo com a metáfora adotada pelo sistema de autoria em questão: nós, molduras, cartões, componentes, segmentos. Nós

e segmentos são utilizados de modo generalizado, representando cartões ou molduras de acordo com o hipertexto. Componente é utilizado, em especial pelo Modelo de Dexter [20], mas se referindo a qualquer componente do hipertexto, sendo um nó ou *link*.

Um nó pode possuir relações com vários nós. Os *links* representam uma ligação entre nós possuindo algum relacionamento lógico entre si: associação, explicação, especificação, especialização, classificação, generalização, exemplificação, numeração; não existindo um número fixo de *links* em um nó.

Um *link* possui uma locação de ativação definida e uma locação destino. O ponto de ativação é a âncora do *link*. A âncora consiste de uma região do nó, podendo ser: uma palavra destacada, um botão da interface, ou qualquer ícone. O destino de um *link* é normalmente um outro nó.

O conjunto de nós e *links* determinam a rede de nós do hipertexto. A estrutura da rede de nós é a base do hipertexto, estabelece a semântica e determina as possibilidades de navegação do hipertexto.

2.1.7 Os Nós de um Hipertexto

A metáfora adotada para a definição do nó influencia no modo de manipulação — armazenamento e acesso — do hipertexto. Para que esta influência seja entendida é necessário o conceito de granularidade de nós. A granularidade de um nó corresponde ao seu "tamanho". Mas já que não podemos falar de informação grande, o tamanho do nó é tratado em termos do refinamento. Assim, um hipertexto com granularidade fina possui nós pequenos e por isso em maior volume. Do mesmo modo, a granularidade grossa implica em um pequeno número de grandes nós.

Os cartões normalmente armazenam pequena quantidade de informação textual e gráfica, correspondendo à granularidade fina. A informação é dividida em nós menores, aumentando a quantidade de nós do hipertexto.

Considerando a metáfora de moldura, a mesma informação pode ser dividida em blocos maiores, pois não há limitação sobre o que apresentar em uma tela; a granularidade grossa é adotada. No entanto esta divisão da informação deve considerar a capacidade de absorção da informação pelo leitor. Muita informação exibida de uma só vez pode causar o que é chamado de *overhead cognitivo* [4].

Por outro lado a granularidade muito fina com grande quantidade de pequenos nós pode gerar problemas na manutenção da rede de nós. O ideal é adotar uma opção balanceada.

2.1.8 Os *Links* de um Hipertexto

Uma das grandes potencialidades do hipertexto é permitir o estabelecimento de relações entre documentos diferentes e situados remotamente em máquinas diferentes. Os *links* implementados em um hipertexto se diferem basicamente pelo local para onde "apontam".

Links intra-documentos são *links* possuindo sua locação destino localmente, no mesmo documento/hipertexto. Representam referências a partes correlatas do documento.

Links entre documentos apontam para uma locação destino que não necessariamente está presente localmente, podendo estar em algum ponto de uma rede de comunicação entre outros computadores. Exigem uma estrutura de suporte para este tipo de ligação. *Links* entre documentos são muito utilizados na WWW. Esse tipo de *link* implica em uma compatibilidade no formato de arquivos ou um suporte a diversos tipos. O sistema de responsável pelo *browsing* do hipertexto deve ser capaz de entender os formatos dos documentos armazenados remotamente para poder apresentá-los ao leitor.

2.2 Porquê Converter: Papel x Hipertexto

Há várias razões que incentivam a conversão de documentos do formato impresso para o hipertexto. Essa seção analisa as dificuldades de utilização do texto impresso, analisando em seguida as vantagens do hipertexto.

2.2.1 Dificuldades com o Texto Impresso

O texto impresso tem algumas vantagens sobre o texto eletrônico. Determinadas categorias de texto ainda são melhor lidas quando impressas, como por exemplo os romances. Outro ponto é que as limitações eletrônicas dos CRTs ou monitores fazem com que a leitura extensiva em uma tela de computador ainda seja um pouco desconfortável, mesmo porque o ambiente em torno de um computador normalmente não está preparado para "convidar" o usuário a permanecer grande tempo. Há grandes avanços tecnológicos nessa área: a miniaturização dos atuais *notebooks*, a evolução da tecnologia das telas de cristal líquido com matriz ativa e o aumento da capacidade de resistência dos mecanismos de armazenamento através de DVDs. Esses avanços permitem prever uma maior facilidade, no futuro, para a leitura ergonômica diretamente no computador. A tecnologia atual, entretanto, ainda não permite isso na maioria dos casos.

As dificuldades com o texto impresso foram irrelevantes durante muito tempo pela falta de alternativas, mas as limitações do texto impresso que devem ser consideradas são [11]:

1. Há dificuldade em realizar procura entre grandes volumes de material impresso. Após seguir um certo número de referências cruzadas, se o leitor não estiver anotando tudo se confunde facilmente;

2. A tabela de conteúdo e o índice são ferramentas que facilitam a procura de tópicos, mas o número de referências cruzadas é mínimo;

3. O índice impresso é limitado pelo tamanho e pelo critério de seleção dos autores e nem sempre é direcionado para a informação relevante para o leitor. Dependendo do domínio de aplicação, coletâneas são largamente utilizadas. Mas mesmo assim, editoras esperam um certo intervalo para a publicação de novas edições;

4. A razão mais tocante, a informação não pode ser atualizada periodicamente. Em determinados âmbitos de aplicação este é um fator crítico, senão essencial. Alguns destes casos são: legislação, manuais de procedimentos de operação/instrução; estes exigem que a informação sendo consultada esteja sempre atualizada.

2.2.2 Vantagens do Formato Hipertexto

Apesar de não ser mais necessário comprová-las, é válido mencionar as vantagens do formato hipertexto ([32],[11]):

- O formato hipertexto suporta boas características de navegação: dinamismo na seqüência das referências cruzadas e procura por palavras-chave. Mesmo na ausência de páginas de índices, existem vários mecanismos de busca disponíveis na WWW, além do mecanismo básico oferecido pelos *browsers*. Alguns exemplos são: Yahoo[12], Altavista[13], Cadê?[14] entre muitos outros.
- O suporte digital do hipertexto fornece mecanismos para uma maior diversidade da mídia utilizada na divulgação da informação. Além de texto, fotografia e imagens gráficas, o meio eletrônico pode fornecer áudio e vídeo.
- A maioria dos usuários procura nos dicionários, enciclopédias e manuais de treinamento informação para completar uma tarefa maior e é de grande utilidade poder gravar o resultado de uma pesquisa para uso futuro em atividades similares [3].

[12] Endereço: http://www.yahoo.com.
[13] Endereço: http://www.altavista.digital.com.
[14] Endereço: http://www.cade.com.br.

2.3 Como Converter: Conversão Manual x Automática

2.3.1 Conversão Manual

A primeira solução na conversão de texto para hipertexto seria fazer a conversão manualmente utilizando um dos vários *softwares* de autoria de hipertexto já disponíveis[15]. Alguns aspectos são intrínsecos a este tipo de conversão e são examinados a seguir [36]:

Interpretação individual: decisão sobre o que deve constituir um *link*, onde um *link* deve apontar e o que constitui um nó do hipertexto, estão sujeitas à interpretação e política individual do editor. Isso pode ser desejável ou aceitável em alguns casos específicos, mas não se aplica no geral. Informações importantes sobre outros aspectos podem ficar sem relações estabelecidas por passarem despercebidos pelo editor, dificultando o acesso do leitor e diminuindo a qualidade do hipertexto.

Processo Repetitivo: o processo de conversão é uma tarefa repetitiva consistindo de três passos:

1. determinação da informação que vai constituir um nó;
2. verificação de relações entre o nó em questão e outros nós já existentes
3. estabelecimento de *links*, alterando a rede de nós.

Considerando um grande volume de informação, a conversão se torna mais sujeita a erros humanos, principalmente com relação à manipulação da rede de nós.

Aspecto Temporal: o processo de conversão pode se tornar um processo de longa duração especialmente na conversão de textos extensos, inviabilizando a disponibilização da informação rapidamente.

[15] Alguns exemplos são o FrontPage® da *Microsoft*, o HTMLEd® da *W3Internet* e o HotDog® da *Sausage Software*.

Atualização da Informação: na inserção de novos dados toda a rede de nós deve ser reconsiderada. Considerando que a informação pode alcançar grande volume e uma rede de nós altamente interconectada, aumenta a dificuldade de atualização da informação contida no hipertexto de modo coerente e a manter a integridade da rede.

Esses aspectos são menos críticos quando o volume de informação é pequeno. No entanto, no caso de um grande volume de informação que é atualizada periodicamente e com grande número de relações, a conversão manual se torna praticamente inviável. Na prática, normalmente a maior parte das obras de consulta e referência pertence à segunda categoria, de grande volume de informação.

2.3.2 Conversão Automática

A conversão automática apresenta vantagens sobre a conversão manual na medida em que procurar minimizar o efeito dos aspectos mencionados na seção anterior:

- Executando as tarefas repetitivas automaticamente, principalmente a segmentação da informação;
- Estabelecendo os *links* de acordo com critérios bem estabelecidos;
- Fornecendo meios para a atualização da informação;
- Executando essas tarefas no menor intervalo de tempo possível.

Considerando os problemas apresentados com a conversão manual, as vantagens da conversão automática e a crescente demanda de documentos que necessitam conversão, vem a pergunta: Por que é que não existem muitos sistemas de conversão de hipertexto? A primeira razão: eles são difíceis de implementar, e outra: são limitados no seu escopo de aplicação [13]. Este trabalho vem tentar reverter estas afirmações, propondo um processo de conversão voltado para a geração de documentos para a WWW.

2.4 As Etapas da Conversão Automática

As etapas envolvidas na conversão automática já foram mencionadas no **Capítulo 1** e aqui passam por uma breve análise antes de o processo de conversão e as etapas subseqüentes serem abordados detalhadamente no **Capítulo 3**.

2.4.1 Seleção de Documentos

Na tentativa de estabelecer diretivas para a seleção de textos que melhor se adeqüem ao processo de conversão automática, algumas características do texto devem ser observadas. Nem todos os tipos de textos se prestam igualmente à conversão automática [16]. Muitos textos são escritos e concebidos para terem uma conotação linear mesmo na leitura, no sentido de que o próprio texto tem seu significado alterado se for transformado em hipertexto.

Os melhores candidatos para a conversão automática são documentos que possuem informação com grande volume de referências cruzadas, com uma estrutura complexa, hierárquica e bem definida. No entanto, mesmo que não em tão grande volume como uma enciclopédia, a maioria dos documentos que são consultados "aos poucos" ou "em passos" se aplicam à conversão automática. De acordo com algumas regras chamadas de "*Golden Rules of Hypertext*" [38], um texto é adequado à conversão automática se:

- O texto possui grande volume de informação organizada em fragmentos[16];
- Os fragmentos se relacionam entre si[17];
- O leitor a qualquer momento necessita somente de parte pequena da informação.

[16] Estes *fragmentos* são partes do texto que possuem certa coesão semântica, exploram um determinado assunto e podem ser consultados de modo independente do resto. Por ser uma definição muito abstrata, está intimamente ligada ao contexto no qual a informação é apresentada. Assim, um fragmento pode tanto ser, numa enciclopédia uma definição de um item e, num livro um capítulo.

[17] Esta relação entre *fragmentos* pode ser tanto explícita – quando ocorre a citação literal de outro fragmento, quanto implícita caracterizada pela afinidade de assunto entre fragmentos.

Durante o processo de seleção de texto estas características podem ser mais bem explicitadas através da tentativa de responder às questões propostas por Riner [36]:

O uso do texto tende a um hipertexto?

Existe uma estrutura implícita ou explícita no documento?

O texto tem uma tabela de conteúdo ou esta pode ser extraída?

O texto possui um índice, glossário, etc.?

O texto possui muitas referências cruzadas interna ou externamente?

Existem unidades definíveis que possam ser relacionadas umas com as outras?

É difícil afirmar que textos mais bem adaptados a estas condições resultam em melhores hipertextos, principalmente porque não há um meio de se julgar se um hipertexto está correto ou não. No entanto, pode-se afirmar que textos com essas características irão melhor se adequar a um processo de conversão automática. Um texto caracteristicamente monolítico e seqüencial dificilmente vai desfrutar das vantagens de tal processo.

Concluindo, a seleção de textos é uma etapa que depende totalmente da tomada de decisão humana arbitrária. Caso esteja interessado em empregar mecanismos de conversão automática, o editor deve privilegiar textos cujas características se prestam à conversão.

2.4.2 Aquisição de Dados

Apesar de não ser objetivo deste trabalho explorar em detalhes como se dá o processo de aquisição de dados, é uma preocupação estabelecer padrões para a qualidade dos dados e para o formato de entrada destes no processo proposto. Assim, algumas diretivas para guiar esta etapa são sugeridas. Nesse escopo, os dados a serem adquiridos estão restritos a texto. Assim *aquisição de dados* e *aquisição de texto* são utilizados como sinônimos.

A aquisição de texto pode ser feita de acordo com diversos métodos, variando conforme as condições físicas dos documentos, a disponibilidade dos mesmos e o acesso a dispositivos adequados.

O enfoque desta etapa está em três métodos distintos de aquisição de texto, cada um deles baseado em um tipo de suporte: manuscrito, impresso ou já no formato digital.

2.4.2.1 Método 1 – Cópia ou Digitação

Método que exige a dedicação integral de pessoas durante o processo. Trata-se do método mais demorado e mais trabalhoso. Tem a vantagem de eliminar alguns dos problemas particulares dos métodos de digitalização descritos na seção seguinte. As desvantagens deste método são basicamente duas:

1. Não elimina a necessidade de correção ortográfica para garantir a qualidade do documento de entrada. Erros de digitação podem ocorrer e devem ser corrigidos.
2. Ainda há necessidade de um processo de aquisição dos componentes considerados "externos" ao texto, normalmente presentes nos anexos: figuras, tabelas, gráficos e outros similares, que não podem ser digitados.

Este método deve ser adotado em casos que, devido à presença de algum fator crítico, a utilização dos outros dois métodos descritos se tornar inadequada. Qualidade pobre do texto impresso, texto impresso em meio não adequado ao dispositivo de *scanning*, ou a ausência deste dispositivo, são alguns desses fatores.

2.4.2.2 Método 2 – Digitalização

Método que explora a criação de uma imagem digital do texto para depois ser convertida em texto passível de ser lido e editado em um processador de textos, em uma etapa

subseqüente. O texto impresso passa por diversas etapas antes de atingir o padrão para a entrada [33],[36]:

1. *Scanning*: Apesar de ter evoluído consideravelmente, o processo de digitalização de páginas impressas de texto ainda é um processo que demanda tempo e dedicação. No entanto, pode ser uma opção mais razoável do que a digitação do texto impresso.

   A evolução alcançada nos *scanners* permite que a presença humana seja dispensada quando o dispositivo possui o recurso de alimentação automática, por exemplo. Mesmo sem a alimentação automática, o avanço tecnológico dos dispositivos resultou em um considerável aumento de velocidade na varredura e no armazenamento das páginas. Esta ainda é a melhor opção quando a fonte de informação é o documento impresso.

2. OCR – *Reconhecimento Óptico de Caracteres*: Influencia muito na qualidade do texto que será utilizado nas etapas subseqüentes. Quanto melhor (mais acurado) for o processo de OCR, menor a quantidade de erros nas etapas futuras e, por conseqüência, menor o esforço em corrigi-los.

   Dificuldades ainda são encontradas pelo OCR em reconhecer texto quando os caracteres são acentuados, estão muito próximos uns aos outros ou ainda na presença de caracteres especiais. Para tentar solucionar alguns dos erros do OCR existem abordagens diferentes para cada tipo de problema.

   Para a correção de erros no reconhecimento de palavras, a verificação mais comum é através do uso de corretores ortográficos, realizada na etapa subseqüente à preparação do texto. Outras correções podem ser feitas através de diversas verificações. Em alguns casos pode-se utilizar de um recurso semi-automático, abordado na **Seção 2.4.3.2**. Outros casos requerem a leitura comparativa do texto, uma tarefa exaustiva, abordada na **Seção 2.4.3.3**, página 38.

Uma situação crítica ocorre, às vezes, quando o reconhecimento gera palavras que apesar de sintaticamente corretas, são diferentes das originalmente utilizadas no texto. Este problema específico se não for detectado pode gerar confusões, devido à possibilidade de mudança da semântica do texto. Outras características inerentes ao texto impresso não reconhecidas pelo OCR devem ter tratamento especial e são tratadas no passo seguinte.

3. Preparação do texto: deve ser responsável por eliminar a inclusão de caracteres de formatação pelos *softwares* OCR e a utilização da hifenização.

   O primeiro caso caracteriza-se principalmente pela inclusão de marcas de parágrafo "¶" ao final de linhas [36] sem corresponder ao final do parágrafo propriamente dito. O segundo caso, a hifenização, corresponde à separação de sílabas comumente utilizada nos textos impressos.

   A correção nestes casos pode ser executada manualmente utilizando um processador de texto em um processo repetitivo de procura e troca de caracteres. Um outro meio é através da aplicação de filtros, consistindo de processos automáticos para tal tratamento.

2.4.2.3 Método 3 – Texto no Formato Digital

Este método de aquisição evita a necessidade de aquisição da informação. Entretanto, este só é possível de ser executado quando o texto já está disponível digitalmente, devido a uma aquisição anterior.

A possibilidade de a informação ser disponibilizada já no formato digital e de acordo com o formato de entrada do processo de conversão seria a situação ideal para a aquisição de dados. Elimina a necessidade da realização dos métodos anteriormente mencionados, reduzindo assim, a duração desta etapa ao longo do tempo. Em contrapartida, exige do

profissional ou da equipe trabalhando com a informação, a pesquisa da existência ou não destes documentos no formato digital. E também da disponibilidade destes, principalmente devido aos direitos autorais e de reprodução do material.

Apesar da vantagem temporal em relação aos outros métodos de aquisição de texto, o texto digitalizado pode também apresentar algumas dificuldades, dependendo do formato no qual o texto está armazenado [33]. Exige muitas vezes, um filtro conversor para o formato de entrada do processo de conversão.

O processo de conversão ideal deveria eliminar a necessidade da aplicação de tais filtros aceitando diversos formatos de arquivo como entrada. No entanto, para isto seria necessária a incorporação de diversos desses conversores de formatos de representação de texto. Fato semelhante ocorre atualmente com diversos pacotes de *software*, em especial com os processadores de texto. Com o surgimento de uma multiplicidade de formatos de representação, os processadores de texto passaram a incorporar tais conversores. A inconveniência desta abordagem é que há uma multiplicidade de formatos para serem incorporados que exigem trabalho extra e acabam por "engordar" o *software*.

Em virtude disso, no processo de conversão, tais conversores serão mantidos externos. Assim, no caso do texto eletrônico estar armazenado em um formato não reconhecido pelo processo de conversão, é necessária a aplicação de um destes conversores.

Concluindo, para que o processo de conversão possa ser executado com sucesso é necessário que a fase de aquisição de dados, principalmente nas etapas que envolvem a produção de arquivos obtenha resultados de alta qualidade. Dificuldades nesta etapa são inevitáveis, independentemente do método selecionado. O processo de conversão será beneficiado com melhoria das tecnologias envolvidas nesta etapa.

2.4.3 Preparação do Texto

A preparação do texto é um dos passos abordados na **Seção 2.4.2.2**. Aqui este tópico é analisado mais profundamente devido à sua grande influência na qualidade dos dados de entrada.

A preparação do texto é responsável por transformar o texto de entrada no *texto validado*, formato adequado para a entrada nas etapas de segmentação e subseqüentes. Consiste de várias verificações, e exige a participação de uma pessoa na coordenação das etapas e na tomada de decisão, já que nem todas as etapas precisam ser necessariamente executadas. A decisão, neste caso, fica na dependência da qualidade do texto de entrada.

2.4.3.1 Verificação Ortográfica

Executada com o auxílio de um processador de textos, o maior requisito é a presença de verificadores de ortografia para a língua portuguesa do Brasil. Entre os que atendem a tal quesito, estão o Microsoft Word e o Redator da Itautec/Philco [42]. O papel da verificação ortográfica é corrigir erros provenientes da aquisição de texto. É de importância especial quando a aquisição é feita por digitação do texto. Neste caso, a verificação ortográfica pode atuar em paralelo à digitação do texto, marcando palavras incorretas e, em certos casos, corrigindo-as de modo automático.

Vale mencionar que a presença de um verificador gramatical também pode auxiliar a manter a qualidade do texto. Tal característica vem sendo incorporada às últimas versões dos processadores de texto. Um exemplo é o trabalho pioneiro de desenvolvimento de um corretor gramatical para a língua portuguesa que está presente no Redator da Itautec/Philco que emprega tecnologia desenvolvida no Departamento de Física e Informática (DFI) e no Instituto de Ciências Matemáticas de São Carlos (ICMSC) da USP [42].

2.4.3.2 Verificação "Inteligente"

Alguns erros comuns da aquisição de texto, normalmente decorrentes da digitalização, nem sempre são detectados através da verificação ortográfica por não se tratarem de erros de grafia de palavras. São erros ocorridos devido à incapacidade de reconhecer certos símbolos. No entanto, por serem erros recorrentes podem ser corrigidos através da observação e correção automática. Esta correção automática pode ser feita utilizando o recurso de procura e troca de qualquer editor de textos. Em alguns exemplos observados durante a aquisição de normas, a partir da aplicação do método de digitalização, ocorre a troca de seqüências de caracteres como[18]:

> A presença de "**9 20**" ou "**9 2O**" em de "**§ 2º**"; de "**Art. lo**" em de "**Art. 1º**"; de "**Art. 1 1**" no lugar de "**Art. 11**"; e de "**~**" no lugar de "**§**".

2.4.3.3 Verificação Visual

A verificação visual é o último recurso para a verificação do texto. Em se tratando de um grande conjunto de textos, e de volume crescente, nem sempre esta é possível. A vantagem é que o texto não precisa ser lido com atenção. Às vezes, através de uma observação crítica da diagramação do texto é possível identificar falhas que passaram despercebidas nas verificações anteriores. Em casos extremos, a releitura do texto, acompanhada da edição, deve ser feita para que o texto atinja a qualidade desejada.

2.4.4 Aplicação da Conversão Automática

A aplicação da conversão automática é a etapa na qual o processo de conversão entra finalmente em ação. Antes da realização desta etapa o texto ainda se mantém como foi concebido, em seu formato seqüencial, como um bloco monolítico. Ao se aplicar a

[18] Os exemplos são somente ilustrativos, e não têm nenhuma relação com a freqüência de erros no texto.

conversão automática, o resultado será o documento segmentado, com as ligações estabelecidas, já como um hipertexto.

Para que o texto possa ser convertido automaticamente em hipertexto é necessário que o processo de conversão identifique elementos básicos e explore as relações entre estes. Algumas dessas relações podem ser identificadas tendo como base a ocorrência de palavras chaves, a estruturação implícita do documento e a formatação de parágrafos. Relações que devem ser exploradas são:

- *Relações Estruturais*: Uma vez identificadas, permitem verificar a estruturação do documento, sua hierarquia implícita. Por exemplo, em um documento com uma estrutura hierárquica, o texto pode estar dividido em partes, as partes em seções, e as seções em sub-seções. As relações estruturais mantêm as características desta estrutura também no hipertexto, fornecendo hiper-elos entre os segmentos ou nós. Estes casos são caracterizados pelas relações entre a tabela de conteúdo e as seções, e entre estas e seus subcomponentes como parágrafos e itens.

- *Relações Explícitas*: Podem ser identificadas como relações inter-textuais, ou relações internas, caracterizadas por relações entre o texto e as referências bibliográficas e citações feitas a outras partes do texto, como por exemplo, a frase "veja a seção X";

- *Relações Implícitas*: Normalmente caracterizadas por relações determinadas de acordo com o assunto do qual o texto trata. Por se tratar de uma área que deve explorar o significado das expressões envolvidas, é das três, o tipo de relação mais complexo de ser explorado.

Esta etapa consiste basicamente de duas etapas que fazem: a segmentação do documento e outra que realiza a associação entre segmentos ou a inferência de *links*[19].

A segmentação do documento é feita através da identificação de elementos significativos do documento. Tipicamente os textos possuem vários elementos como parágrafos, seções, subseções, capítulos, índices, listas e tabelas. Desses elementos, os estruturais são aqueles que formam a estrutura básica do texto e definem relações de pertinência entre os outros elementos. O reconhecimento destes elementos é de fundamental importância para a realização da conversão automática. Sua realização está fortemente vinculada à utilização das análises léxica e sintática.

Já a etapa de inferência de *links* exige técnicas específicas, voltadas para cada domínio de aplicação, analisando a forma como é estruturada a citação.

2.4.5 Codificação em Linguagem Específica

Esta etapa utiliza os resultados da aplicação da conversão automática para gerar uma versão do hipertexto que possa ser utilizada pelos usuários finais. Para o processo de conversão esta etapa realiza tarefa análoga aos exportadores para diferentes formatos freqüentemente presentes em pacotes de software. Resumidamente, a tarefa da etapa de codificação é interpretar a informação no formato interno ao processo de conversão e gerar uma saída codificada de acordo com as regras da linguagem escolhida.

[19] As associações serão referidas ao longo deste texto como *links*, utilizando assim, a palavra original em inglês.

2.5 Trabalhos Relacionados

Três trabalhos na área de conversão automática em especial possuem estreita relação com o processo aqui proposto.

O sistema *Justus* [44] é um sistema de recuperação de informação de hipertexto[20] que tem enfoque para o uso de documentos legais em uma base de dados para hipertexto. Complementa um sistema de hipertexto usado largamente na Inglaterra, chamado *Lexis*, em vários aspectos: fornece uma interface gráfica de fácil utilização pelos usuários alvos, que são na sua maioria advogados e implementa diversas novas relações entre os dados não exploradas no sistema *Lexis* permitindo acesso mais fácil aos dados. Foi implementado utilizando o sistema de hipertexto *Guide* tendo como base os hiperdocumentos do sistema *Lexis*.

O sistema *Lexis* armazena o texto completo de documentos considerados primários: casos e estatutos. A recuperação da informação neste sistema é através do uso de pesquisa booleana por termos-chave em campos específicos, sendo portanto, limitada.

Wilson [44] sugere que relações explícitas normalmente possuem algum tipo de "marcador semântico", palavras-chave que ajudam na identificação destas e confirma a necessidade de um processo de estabelecimento automático destas relações. Para tal, trabalha com o reconhecimento de padrões. Quando possuem o mesmo significado um padrão é mapeado para o nó adequado. Trata-se de um mecanismo disponível no sistema Guide chamado de *aliasing*.

Um aspecto importante do sistema Justus é que as relações descobertas são marcadas como "*links* potencialmente ativos". Este conceito de "potencialmente ativo" foi criado porque não se pode garantir que o documento ou nó referido pelo link está presente na

[20] Tradução do original: *hypertext information retrieval system*.

base de dados de hiperdocumentos. Através deste conceito é possível definir que ações o sistema deve tomar no caso de *links potencialmente ativos* não resultarem em links efetivos, evitando a frustração do usuário ao tentar segui-lo.

Bernstein propõe um sistema "aprendiz" que descobre *links* entre hipertextos [6] com enfoque aos *links* de conteúdo – *links* que exploram as relações implícitas. O trabalho é voltado para textos que serão intencionalmente lidos e não consultados. Enfoca hipertextos que abordam um único assunto, incluindo monografias e livros texto. O "aprendiz" procura extrair *links* que não são descobertos via métodos tradicionais de análise léxica, mas não trata os *links* derivados de relações estruturais. O aprendiz é um sistema definido por Bernstein como "um assistente computacional trabalhando sobre a supervisão imediata de um especialista em sistemas hipertexto, ao invés de um agente autônomo trabalhando independentemente".

James Allan [2] trabalha com processos automáticos para tentar determinar o tipo de relações entre dois documentos ou trechos de um documento. Este trabalho lida com coleções arbitrárias de documentos e automaticamente detecta *links* sem a intervenção do usuário. Para estabelecer os *links* é feita uma análise da similaridade entre documentos ou trechos de documentos. Esta análise é baseada em um modelo chamado *vector space model* [37]. Consultas e documentos são representados em termos de vetores, que são pares de conceitos e pesos não-negativos. A conversão para estes vetores permite a comparação e a determinação de um nível de similaridade entre os documentos comparados. Existe um nível mínimo para que os documentos sejam considerados relacionados e, então, o *link* pode ser automaticamente estabelecido entre as partes. Utiliza da área de análise de discurso para determinar que tipos de *links* são úteis em um hipertexto. Este trabalho é particularmente importante para determinar as chamadas relações implícitas entre coleções de documentos. Estabelece relações entre documentos

baseadas em suas afinidades por assunto. Trata-se de um trabalho que vem a complementar a atuação deste projeto.

3 O PROCESSO DE CONVERSÃO AUTOMÁTICA

Este capítulo explora o *processo de conversão*, que abrange as etapas de *aplicação da conversão automática* e *codificação em linguagem específica*, conforme o especificado no **Capítulo 1**.

Uma descrição geral do processo é apresentada na **Seção 3.2**. As fases de aquisição de texto e de preparação do texto foram abordadas no **Capítulo 2** e são consideradas as preliminares do processo de conversão. O texto validado resultante destas fases deve ser colocado no formato de entrada do processo de conversão para que possa ser processado. Esse formato é discutido na **Seção 3.3**.

O processo de conversão, contribuição desse trabalho, foi desenvolvido para ser aplicado em textos com estruturação hierárquica. Estes estão definidos na **Seção 3.1** a seguir.

3.1 Textos Estruturados

O processo de conversão proposto se aplica a textos estruturados. A estrutura de textos adotada e interpretada neste processo de conversão é a *estrutura hierárquica em árvore*, conforme ilustra a **Figura 3**. Esta estrutura foi escolhida por ser muito comum em documentos técnicos, normas e manuais de instrução em geral. Assim, o documento para ser convertido por este processo de conversão, deve possuir estrutura interna hierárquica em árvore.

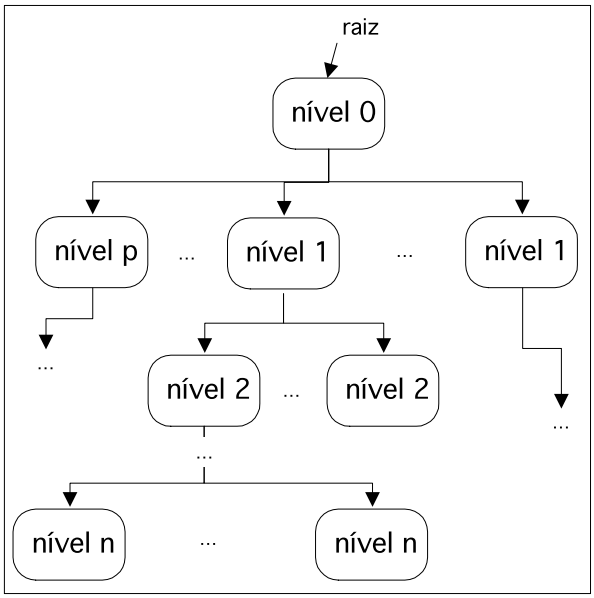


Figura 3 Estrutura hierárquica em árvore

A estrutura hierárquica em árvore considera que todo documento possui um *cabeçalho* ou *título* que marca seu início. O *cabeçalho* é o segmento inicial, também chamado de *raiz* do documento, é o *nível 0*. É a partir dele que segmentos de outros níveis são estruturados.

Outras restrições estruturais aplicáveis a todos os segmentos do documento são:

- um segmento possui <u>somente um</u> segmento pai;
- um segmento de nível *i* só pode conter segmentos de nível hierárquico maior que *i*.

De acordo com o conjunto de documentos outras restrições podem ser aplicadas.

O nível de hierarquia máximo *n* é determinado de acordo com cada tipo de documento e de acordo com o nível de granularidade de segmentação que se deseja atingir. Por exemplo, em um conjunto de documentos no qual a granularidade atinge o nível de blocos de texto, o nível máximo pode ser estabelecido como *n=4*, onde:

Nível 0 → Título ou Cabeçalho

Nível 1 → Capítulo

Nível 2 → Seção

Nível 3 → Sub-Seção

Nível 4 → Blocos de Texto

A estrutura hierárquica para este exemplo é ilustrada na **Figura 4**.

A escolha do nível de granularidade máximo tem implicações diretas na composição dos níveis antecedentes. Continuando com o exemplo acima, e definindo a granularidade ao nível de palavras, os níveis superiores também seriam segmentados em palavras e definidos como uma composição destas.

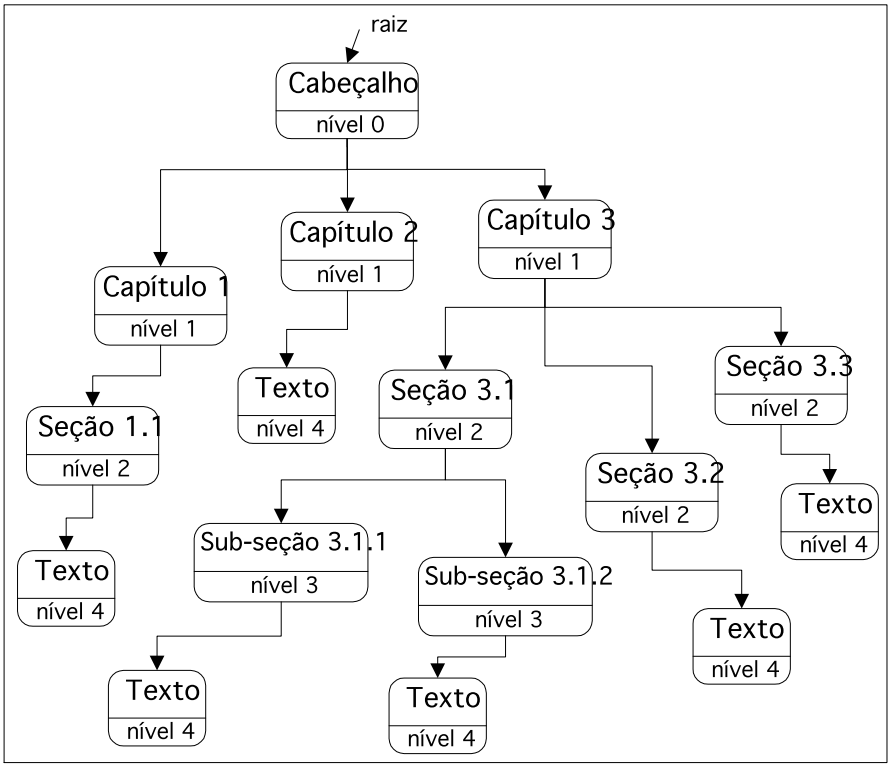


Figura 4 Exemplo de uma estrutura hierárquica de nível 4

Devido a essas implicações é adequado procurar um nível de granularidade que evite a segmentação excessiva do documento e que não gere segmentos de granularidade muito grande. Segmentos muito pequenos correm o risco de ser insignificantes do ponto de vista sintático e semântico, e segmentos muito grandes limitam o acesso a estruturas menores internas. A granularidade adotada na segmentação também influencia no reconhecimento e estabelecimento das relações explícitas do documento.

Considerando essas afirmações, o processo de conversão proposto adota a granularidade no nível de parágrafos de texto. Isto porque parágrafos de texto normalmente contêm sentenças sintaticamente corretas dentro da língua portuguesa e correspondem ao *segmento mínimo citável* na maioria dos documentos estruturados. Ou seja, as citações entre documentos normalmente citam segmentos do nível de parágrafos ou superior. Dificilmente se encontram citações mais específicas que cheguem a se referir, por exemplo, um conjunto de termos ou palavras isoladas do texto.

3.2 Descrição Geral do Processo

Partindo do documento no formato de entrada, o processo de conversão consiste em realizar uma etapa para a *segmentação* do documento, seguida de uma etapa para a *hierarquização* dos segmentos.

Na segmentação o documento deixa de ser um bloco de texto monolítico e passa a ser um conjunto de segmentos. Essa etapa foi elaborada para se basear em uma definição formal da informação. A partir desta definição formal são construídas regras específicas de análise do texto que aplicadas à informação identificam os diversos segmentos.

Na hierarquização são exploradas as relações estruturais entre os segmentos, recompondo a estrutura na forma de relacionamentos entre os dados. Nessa etapa, um modelo de dados para a informação deve ser utilizado para determinar a estrutura das tabelas e fornecer armazenamento adequado dos dados.

Após estas duas etapas, inicia-se a *inferência de links*, que explora as relações explícitas ou as *citações* presentes nos segmentos. A inferência de *links* consiste de três sub-etapas: a *extração das citações*, a *identificação* ou *endereçamento das citações* e o *estabelecimento dos links*, etapas a serem exploradas em detalhe nas **Seções 3.6.1**, **3.6.2** e **3.6.3**. Para realizar estas tarefas, a inferência de *links* utiliza uma definição formal das citações para o tipo de informação sendo convertida. A definição formal para as citações é utilizada na criação das regras para extração e estabelecimento de *links*.

A etapa final do processo de conversão, a *construção do hipertexto*, é realizada fazendo consultas aos dados gerados pelas etapas anteriores. Nesta etapa são explorados os relacionamentos estruturais identificados durante a segmentação e hierarquização e armazenado no banco de dados.

3.3 Formato de Entrada do Processo de Conversão

Para que o processo de conversão possa processar os documentos e convertê-los de um modo confiável é necessário que todos os documentos estejam num mesmo formato de entrada. No processo de conversão proposto neste trabalho o formato de entrada é chamado de padrão STDNORM.

O padrão STDNORM foi definido para incluir informações complementares aos dados a serem submetidos ao processo de conversão. Esses dados complementares são importantes para realizar estatísticas e para fornecer informações para um banco de dados. Informam ao processo de conversão por exemplo, a língua, o tipo e a organização responsável.

Estas e outras informações são fornecidas através de uma série de etiquetas que têm os seus valores preenchidos de acordo com dados do documento sendo submetido para a conversão.

Algumas regras foram estabelecidas para o padrão STDNORM, e estão especificadas abaixo:

- **Suporte**: o texto está digitalizado, e o conjunto de caracteres segue o especificado pela ISO 8859-1 [23], também conhecida como *ISO-Latin*.
- **Cabeçalho**: O texto deve ser submetido para processamento antecedido por um cabeçalho padrão. As etiquetas do cabeçalho estão especificadas na **Tabela 1**. Dentre as etiquetas apresentadas a seguir, as exigidas em todo documento submetido para o processo de conversão estão marcadas com um × na última coluna. Outras etiquetas são opcionais.

Etiqueta	Definição	Escopo de Valores	
BEGIN:STDNORM	Marca o início do arquivo de dados.	–	×
NAME:	Nome do documento atribuído por quem o submete.	Qualquer seqüência de caracteres.	
VER:	Versão dos documentos incluídos no formato X.X	De 1.0 a 9.9	
VER;LOG:	Observações sobre a versão do documento.	Qualquer seqüência de caracteres.	
VER;DATE:	Data da geração da versão do documento.	Qualquer data válida	
LANG:	Língua padrão do documento com etiquetas definidas pela RFC nº 1766 [34]. Para o processo de conversão este valor é fixo para o português do Brasil.	pt-BR	×
INCLUDE;DATE:	Data da entrada do documento para processamento, gerada automaticamente pelo processo.	Qualquer data válida.	
INCLUDE;RESP; VCARD/INLINE:	Pessoa responsável pela submissão do documento. Esta etiqueta possui um modificador que especifica o tipo do valor: INCLUDE;RESP;VCARD ou INCLUDE;RESP;INLINE.	VCARD: cartão de visitas eletrônico especificado pelo padrão versão 2.1 [41] INLINE: nome do responsável por extenso	
INCLUDE;SOURCE:	Endereço na internet do documento fonte, caso este esteja disponível para *download*.	Qualquer endereço *http* ou *ftp*.	
TYPE:	Tipo do texto do documento. Somente esses 2 formatos são aceitos.	TXT: texto sem quebras de linha, formato *Text-only*. ETEXT: texto no padrão Gutenberg, "*wrapped*" com uma marca de parágrafo ao final de linhas e 2 marcas no final dos parágrafos.	×
ORG;FED;EST;MUN;INST:	Organização responsável pela custódia dos documentos. A etiqueta deve ser composta por, no mínimo o modificador FED. Outros modificadores são acrescentados conforme o caso. Para uma instituição não-governamental estadual a etiqueta é: ORG;FED;EST;INST.	FED: sigla no formato INT, para o Brasil, sempre: BR; EST: sigla de um dos estados brasileiros; MUN: nome do município INST[-nomeExt]: sigla da instituição, com opcional para extenso.	×
STDNORM: 1.0	Versão do formato de entrada padrão utilizado no documento, por enquanto, este valor é sempre 1.0.	1.0	
START:	Marca o início do texto dos documentos.	–	×
<texto do documento>		–	×
END:STDNORM	Marca fim do texto dos documentos e fim do cabeçalho.	–	×

Tabela 1 Padrão STDNORM: formato de entrada de dados no processo de conversão

- **Múltiplos Textos**: Vários textos de diversos documentos podem ser submetidos tendo um único cabeçalho. A única restrição conceitual é que todos os documentos submetidos devem "seguir" os dados contidos nas etiquetas do cabeçalho padrão. Por exemplo, esta restrição implica que ao submeter documentos onde a etiqueta ORG;FED;INST tem valor BR;IBAMA, todos os textos incluídos entre as etiquetas START e END devem ser de custódia do IBAMA.

- **Grafia das etiquetas**: As etiquetas são sempre escritas em letras maiúsculas. A terminação das etiquetas é feita com o caracter de dois pontos ":". Este serve como separador entre etiqueta/valor. As exceções desta regra são as etiquetas de início e fim, respectivamente BEGIN:STDNORM e END:STDNORM, inspiradas nas etiquetas do padrão vCard [41]. O caracter ponto e vírgula ";" é um separador entre etiqueta e seus modificadores.
- **Ordem das etiquetas**: O padrão STDNORM é iniciado pela etiqueta inicial BEGIN:STDNORM e termina com a etiqueta final END:STDNORM. A etiqueta START que marca o início do texto do documento sendo submetido é sempre a penúltima etiqueta do padrão, sendo seguida somente pelo texto e pela etiqueta final. As demais etiquetas do padrão não possuem restrições de ordenação, podem aparecer em ordem aleatória. As etiquetas apresentadas na **Tabela 1** acima foram descritas na ordem que seguem estas restrições e pode ser utilizada como sugestão de ordenação.

A etiqueta INCLUDE;RESP; com o modificador VCARD (INCLUDE;RESP;VCARD) da especificação do padrão STDNORM possibilita que um vCard seja incluído quando referenciando dados pessoais. Esta característica foi incluída visando adaptação às tendências em troca de informação.

O **Quadro 1** a seguir mostra um exemplo de um documento para entrada no processo de conversão seguindo o padrão STDNORM.

```
BEGIN:STDNORM
NAME: Decretos Federais
LANG: pt-BR
TYPE: ETEXT
VER: 1.0
VER;DATE: 11/08/1997
INCLUDE;RESP;INLINE: Alessandra
INCLUDE;DATE: 12/08/1997
ORG;FED: BR
STDNORM: 1.0
OBS: Texto não passou por correção ortográfica.
START:
DECRETO  N° 66, de 18  DE  MARÇO  DE  1991

Promulga a Convenção Para a Conservação das Focas Antárticas, concluída em Londres, a
1° de Junho de 1972.

O Presidente da República, usando da atribuição que lhe confere o Artigo 84, inciso 4°, da
Constituição e

Considerando que a Convenção Para a Conservação das Focas Antárticas foi adotada em
Londres, a 1° de Junho de 1972, sob a égide dos princípios estabelecidos  no Tratado
Sobre a Antártica, concluído em Washington, a 1° de dezembro de 1959;

Considerando que o Congresso Nacional aprovou a Convenção, por meio do Decreto
Legislativo 37, de 26 de Outubro de 1990;

Considerando que a Carta de Adesão à Convenção ora promulgada, foi depositada em 11 de
fevereiro de 1991;

Considerando que a Convenção Para Conservação das Focas Antárticas entrará em vigor,
para o Brasil, em 13 de Março de 1991, na forma de seu Artigo 13, Inciso 2;

DECRETA:

Artigo 1° - A  Convenção  para  a  Conservação  das Forças Antárticas, apensa por cópia ao
presente Decreto, será executada e cumprida tão inteiramente como nela se contém.

Artigo 2° - Este Decreto entra em vigor na data de sua publicação.

Brasília, 18 de Março de 1991;  170° da Independência e 103° da República.

FERNANDO COLLOR

Francisco Rezek.

END:STDNORM
```

Quadro 1 Documento de acordo com o padrão STDNORM

O padrão STDNORM é uma contribuição deste trabalho que visa facilitar a troca de informações quando da submissão de documentos para o processo de conversão. Pretende também possibilitar contribuições remotas para uma base de documentos quando o processo for implementado *on-line* e com acesso remoto, via a WWW da Internet, por exemplo.

3.4 Segmentação

A segmentação de um documento tem estreita ligação com a estruturação hierárquica vinculada ao tipo de documento. É baseada em uma definição formal da estrutura do documento para definir as regras de reconhecimento dos diversos tipos de segmentos. Nessa etapa estas regras são aplicadas diretamente na fase de Análise de Parágrafos, descrita na **Seção 3.4.2**.

3.4.1 Preparação dos Documentos STDNORM

A fase de preparação dos documentos de entrada para o processo de conversão faz a leitura dos dados a serem processados, a partir de um arquivo com dados no formato STDNORM. Esta preparação consiste da aplicação de filtros de acordo com o tipo do texto determinado pela etiqueta TYPE.

No caso do tipo do documento ser TXT é aplicado um filtro específico para este tipo de dado, o filtroTXT. Este filtro faz a leitura seqüencial de parágrafos do texto, aplicando o seguinte tratamento ao texto:

- Elimina linhas em branco
- Elimina espaços e marcas de tabulação do início e do fim dos parágrafos.

O segundo tipo de documento aceito pelo padrão STDNORM é chamado aqui de texto no formato ETEXT, como uma referência ao Projeto Gutemberg[21]. Neste formato, as linhas do texto estão separadas pela presença de uma quebra de linha, e os parágrafos por duas quebras de linha. Os documentos nesse formato passam por um filtro específico, chamado filtroETEXT. Este filtro atua de modo similar ao filtroTXT e adicionalmente junta linhas separadas por quebras em um único parágrafo. Na sintaxe da linguagem C ou Perl, elimina

[21] Duas referências sobre o Projeto Gutemberg são: http://promo.net/pg/history.html e http://www.etext.org/about.html.

a ocorrência do caracter de fim de linha ("\n") do final das linhas, e mantém somente um caracter de fim de linha, no final do parágrafo.

O algoritmo para essa fase da conversão automática é apresentado a seguir.

```
sub preparacaoDoc ($arqEntrada)
{
  @docEntrada = Leitura_padrao_STDNORM ($arqEntrada);

  for $documento (@docEntrada) {
    $tipo_do_documento = $documento{TYPE};
    if ($tipo_do_documento eq "ETEXT") {
      $documento{TEXTO} = &filtroETEXT($documento{TEXTO});
    }
    elsif ($tipo_do_documento eq "TXT") {       # aplica o filtro TXT: só separa os parágrafos
      $documento{TEXTO} = &filtroTXT($documento{TEXTO});
    }
    else {  # este formato ainda não está tratado pelo sistema, ignorar
      print "Documento foi ignorado: formato não tratado pelo sistema";
    }
  }
  return @docEntrada;
}
```

Algoritmo 1 Preparação de documento padrão STDNORM[22]

Como resultado da preparação dos documentos está a informação separada em parágrafos pronta para uma análise identificando o tipo de cada um deles.

No processo de conversão proposto o fato de o segmento ser a unidade textual parágrafo facilita nesta primeira segmentação. A partir da aplicação dos filtros nesta preparação já se tem a informação segmentada, apesar de ainda não tipificada.

3.4.2 Análise de Parágrafos

É nesta fase que a descrição formal da estrutura do documento passa a ser essencial. Esta é utilizada na identificação do tipo do documento sendo convertido e na análise do segmento.

A definição formal para a estruturação do documento deve conter pelo menos uma regra inicial para identificar o cabeçalho que marca o início do documento dentro do conjunto sendo analisado. Por exemplo, submetendo-se um documento no padrão STDNORM,

contendo os textos de três manuais de operação, esta regra de início do documento faz a identificação do segmento que marca o início de cada documento, ou a *raiz* de cada um deles, conforme ilustra a **Figura 5**.

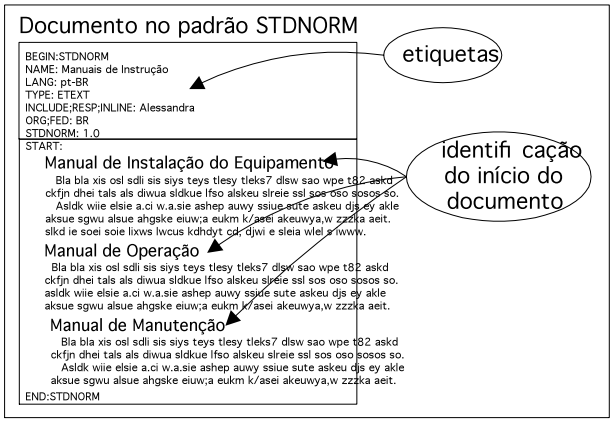


Figura 5 Identificação do início de documentos em um documento STDNORM

O passo seguinte é a identificação do tipo do documento. A tipificação de documentos é inerente ao conjunto ao qual o documento pertence. Por exemplo, dentro do conjunto de normas estatutárias jurídicas os documentos podem ser tipificados em categorias como Lei, Decreto e Portaria. A identificação do tipo do documento é feita quando é detectado o início de um novo documento.

Assim a etapa da segmentação consiste dos seguintes passos:

1. Fazer a preparação do documento no padrão STDNORM
2. Definir o reconhecimento do início de documentos dentro de um bloco de texto, a partir da regra inicial da definição formal da estrutura do documento.
3. Determinar o tipo de cada documento identificado no passo anterior, se o conjunto de documentos tiver tipificação.
4. Analisar os segmentos componentes do documento.

O **Algoritmo 2** a seguir apresenta a segmentação de modo estruturado.

[22] Os algoritmos apresentados neste capítulo estão escritos em pseudocódigo com a sintaxe utilizada na linguagem Perl [43].

```
sub segmentacao ($arqSTDNORMS, $arqLog)
{
  @docsSTDNORM = &preparacaoDoc($arqSTDNORMS);

  foreach $documento (@docsSTDNORM) {
    next if ($documento{TYPE} ne "TXT" or $documento{TYPE} ne "ETEXT"); # ignora outros tipos de documentos
    # análise de parágrafos
    foreach $paragrafo ($documento)  {
      if ( &InicioDeDocumento($paragrafo) ){      # o parágrafo inicia um novo documento
        @dadosNovoDocumento = &identificaTipoDocumento ($paragrafo);
        &storeDados (@dadosNovoDocumento, @dadosSegmentados);
      }
      else {                                      # o parágrafo é um segmento dentro do documento atual
        @dadosSegmento = &analisaSegmento ($paragrafo);
        &storeDados (@dadosSegmentos, @dadosSegmentados);
      }
    }
  }
  return @dadosSegmentados;
}
```

Algoritmo 2 Etapa de Segmentação no processo de conversão

A análise de segmentos é uma sub-etapa da segmentação particularmente importante para documentos que possuem diversos tipos de segmentos. Baseada na definição formal da estrutura, esta etapa é o cerne da segmentação. Utiliza as regras da definição formal para construir instruções de identificação de cada um dos tipos de segmentos.

Por exemplo, para o conjunto de normas estatutárias jurídicas brasileiras alguns tipos de segmentos são *livro*, *capítulo*, *artigo*, *inciso* e *alínea*. Meios de identificar estes segmentos são determinados pelas regras da definição formal e são baseados na presença de etiquetas ou de palavras que são consideradas como chave quando utilizadas em contexto adequado. Exemplificando, novamente para o conjunto de normas estatutárias jurídicas, o segmento *artigo* é identificado pela presença da palavra "Artigo" seguida de um numeral arábico ou ordinal. Esta seqüência de vocábulos é a <u>etiqueta</u> para o segmento artigo. Quando esta etiqueta é reconhecida no início de um parágrafo o segmento é reconhecido como tipo *artigo*.

Alguns tipos de segmentos podem estar vinculados a uma análise mais elaborada do contexto no qual estão inseridos. Esta análise é então retomada na hierarquização para eliminar ambigüidades. No conjunto de normas estatutárias jurídicas, casos particulares são

os segmentos do tipo *apelido*, *disposição inicial* e *escopo normativo*. Estes segmentos não possuem uma etiqueta padrão e devem ser identificados através da análise de conteúdo e do contexto. A análise de contexto é feita pela aplicação da regra que determina, por exemplo, que um apelido é sempre antecedido pelo cabeçalho de uma norma. Nesta etapa esses tipos de segmentos são marcados como *parágrafos genéricos*.

O resultado da etapa de segmentação são os segmentos com os tipos identificados. Estes dados devem ser armazenados de acordo com um modelo de dados para o conjunto de documentos sobre o qual a conversão está sendo aplicada.

Não é propósito deste processo de conversão apresentar restrições com relação à ferramenta utilizada na implementação. No entanto, sugere-se que seja utilizada uma linguagem de programação que dê suporte ao uso de expressões regulares. Estas são particularmente úteis nas análises executadas durante o processo. Entre as linguagens que suportam esta característica estão *Perl* [43] e *awk* [9].

3.5 Hierarquização

Os segmentos gerados na etapa de segmentação possuem seus tipos identificados, mas estão todos num mesmo nível. A etapa de hierarquização atribui níveis para cada tipo de segmento refletindo a estrutura hierárquica do documento. Depois desta atribuição de níveis é possível identificar, para cada segmento, seu segmento pai, e a sua ordem dentro de um conjunto de segmentos do mesmo tipo.

Para a realização desta etapa é necessário que todos os segmentos tenham seus tipos bem definidos. Para isto, segmentos que não foram identificados na etapa anterior passam por uma segunda análise, chamada de *verificação de tipo*, antes da etapa de *atribuição de nível* e da *geração da hierarquia*. O **Algoritmo 3** apresenta a hierarquização e suas etapas.

```
sub hierarquizacao (@dadosSegmentados)
{

  @dadosSegmentados = &verificaTipo (@dadosSegmentados);
  @dadosSegmentados = &atribuiNivel (@dadosSegmentados);
  @dadosHierarquizados = &geraHierarquia (@dadosSegmentados);

  return @dadosHierarquizados;
}
```

Algoritmo 3 A hierarquização de segmentos no processo de conversão automática

A etapa de atribuição de nível utiliza uma escala de níveis para os tipos de segmentos. Esta escala pode ser abstraída a partir da definição formal da estrutura do documento. No exemplo exibido na **Figura 4**, a escala de níveis está bem definida de acordo com os tipos de segmentos.

Após a atribuição de nível, a geração de hierarquia é feita incluindo junto com os dados de cada segmento duas informações novas: a ordem do segmento dentro do conjunto de segmentos de mesmo nível e o pai do segmento. Estas informações, quando armazenadas no banco de dados, funcionam como um mapeamento das relações entre os segmentos.

```
sub geraHierarquia (@dadosSegmentados)
{    #        A Hierarquia entre os segmentos é determinada através de 2 dados:
     #   a ordem dos segmentos dentro de um grupo de segmentos de um mesmo nível
     #   e o pai do segmento
     foreach $segmento (@dadosSegmentados) {
          $segmento{ORDEM} = &achaOrdem ($segmento);
     }
      for $segmento (@dadosSegmentados) {
          $segmento{PAI} = &achaPai($segmento);
     }
}
```

Algoritmo 4 Geração de hierarquia

Ao determinar a ordem do segmento dentro do contexto onde ele aparece é possível saber, por exemplo, qual o *n-ésimo* capítulo de um documento. Ou, por exemplo, qual é o 3º artigo de uma determinada norma.

```
sub achaOrdem ($segmento)
{   #    Acha a ordem do segmento dentro de um grupo de segmentos do mesmo tipo.
    #    Esta ordem deve ser determina a partir da comparação do nível do segmento atual
    # com o nível dos  segmentos que o antecedem.
    if ($segmento{NIVEL} == 0) {  # é início de um novo documento
        return 0;                 # a ordem é reiniciada
    }

    # enquanto o nível do segmento atual for maior que o nível do segmento anterior
    # continua procurando segmentos anteriores
    while ($segmento{NIVEL} > $segmentoAnterior{NIVEL} )  {
        $segmentoAnterior--;
    }
    if ($segmento{NIVEL} < $segmentoAnterior{NIVEL}) {
        return 1;                     # segmento atual é o primeiro da lista
    }
    if ($segmento{NIVEL} == $segmentoAnterior{NIVEL}) {
        # segmento atual e o segmento anterior são do mesmo nível, continua a ordem anterior
        return $segmentoAnterior{ORDEM}+1;                    # acrescenta 1 na ordem
    }
}
```

Algoritmo 5 Atribuição da ordem do segmento

É através da identificação do pai do segmento que a hierarquização é finalmente realizada.

```
sub achaPai ($segmento)
{   #    Acha o segmento pai do segmento atual.
    #    O pai do segmento é o primeiro segmento que o antecede
    #   e que possui nível hierárquico maior que o nível do segmento

    if ( $segmento{NIVEL}==0 )    {    # É início de um documento, ele é o seu próprio pai
        return $segmento;
    }
    while ($segmento{NIVEL} > $segmentoAnterior{NIVEL})  { # O nível do segmento é maior que o do anterior
        $segmentoAnterior--;                                 # procura nos segmentos anteriores
    }
    if ($segmento{NIVEL} < $segmentoAnterior{NIVEL}) {  # Se o segmento anterior é de nível maior,
        return $segmentoAnterior;                       # ele é o pai do segmento.
    }
    if ($segmento{NIVEL} == $segmentoAnterior{NIVEL})  # Se os segmentos são de mesmo nível,
        return $segmentoAnterior{PAI};                  # eles tem o mesmo pai
    }
}
```

Algoritmo 6 Atribuição do segmento pai

3.6 Inferência de Links

Diferentemente das etapas de segmentação e hierarquização, a *inferência de links* não trabalha com os *segmentos*, mas com o conteúdo destes. Esta etapa é a mais importante para alcançar o objetivo final que é a *geração do hipertexto*. Sem esta etapa, os documentos estariam segmentados, mas as relações não-estruturais, ficariam sem exploração.

Para que os *links* possam ser estabelecidos automaticamente algumas etapas devem ser seguidas. O primeiro passo é fazer a *extração da citação*, isolando-a do texto do segmento. Em seguida é necessário analisar a citação extraída e identificar o segmento sendo citado. Uma vez identificado o segmento, é necessário verificar se o documento ao qual este segmento pertence foi convertido e se o segmento está presente na base de dados. Devido a esta necessidade de consultar dados de segmentos pertencentes a vários documentos, a etapa de inferência de *links* é executada somente após a segmentação de todos os documentos desejados, e consulta os dados diretamente da base de dados.

3.6.1 Extração das Citações

Esta etapa é responsável por identificar as relações explícitas entre segmentos de um mesmo documento ou de documentos diferentes. Conforme descrito no **Capítulo 1**, as relações explícitas são caracterizadas por referenciarem o elemento citado diretamente.

A função da etapa de extração de citações é delimitar a citação dentro do texto do segmento, para que esta possa ser analisada nas fases subseqüentes. Para exemplificar, algumas citações presentes na Constituição Federal estão marcadas na **Figura 6** a seguir, ilustrando a função da fase de extração das citações.

Exemplos de citação presentes na Constituição

artigo 5º, inciso I:
"homens e mulheres são iguais em direitos e obrigações, nos termos desta Constituição";

artigo 15, inciso IV:
"recusa de cumprir obrigação a todos imposta ou prestação alternativa, nos termos do art. 5º, VIII";

artigo 15, inciso V:
"improbidade administrativa, nos termos do art. 37, § 4º".

artigo 241:
"Aos delegados de polícia de carreira aplica-se o princípio do art. 39, § 1º, correspondente às carreiras disciplinadas no art. 135 desta Constituição".

artigo 223, §1º:
"O Congresso Nacional apreciará o ato no prazo do art. 64, §§ 2º e 4º, a contar do recebimento da mensagem".

Figura 6 A função da extração das citações

A extração das citações é baseada em um conjunto de tipos de citação que deve ser determinado para cada conjunto de documentos sobre o qual se deseja aplicar a conversão automática. Este conjunto de tipos de citação é determinado através da análise das citações presentes nos documentos do conjunto e da definição de categorias para as palavras significativas presentes nas citações.

As palavras significativas possuem estreita ligação com o conjunto de documentos sendo analisado. Devido à diversidade na grafia de palavras e ao uso de abreviações, categorias de palavras são utilizadas para o reconhecimento das citações. Uma vez determinadas as categorias, estas são expressas em *tokens*. Exemplificando, para as normas estatutárias jurídicas brasileiras, algumas categorias e seus *tokens* são apresentadas na **Tabela 2**[23].

CATEGORIA	GRAFIAS	SINGULAR	PLURAL	TOKEN
CAPÍTULO	cap.; capítulo	⊗		capS
	caps.; capítulos		⊗	capP
ARTIGO	art.; artigo	⊗		artS
	arts.; artigos		⊗	artP
PARÁGRAFO	Parágrafo; par.; §	⊗		parS
	Parágrafos; pars.; §§		⊗	parP
INCISO	Inciso; item; número; nº	⊗		numS
	Incisos; itens; nºs, números		⊗	numP

Tabela 2 Algumas categorias presentes nas citações em *NEJBs*

Os *tokens* são determinados de acordo com as categorias e substituem as ocorrências de palavras de acordo com a grafia estipulada. Para o processo de conversão proposto são gerados diferentes *tokens* para as categorias quando aparecem no singular e plural.

Ao utilizar o mecanismo de *tokens*, consegue-se uma abstração na identificação dos tipos das citações. Assim é possível escrever um tipo de citação utilizando uma abstração, sem a preocupação com uma grafia particular da palavra. Por exemplo, alguns tipos de citações para as *NEJBs*, com uma formalização simples usando *tokens*:

Formalização	**Tipo de citação abrangida**

[23] Os *tokens* utilizados nesta tabela são somente ilustrativos.

artS numero, *parS numero*	artigo 5º, § 1º
numS/parP numero deste *artS*	inciso II deste artigo § 2º deste artigo

Tabela 3 Exemplo de formalização de tipos de citação

Na prática, os tipos de segmentos devem ser definidos formalmente por um conjunto de regras detalhadas. No processo de conversão proposto, as regras para os tipos de citação são definidas utilizando recursos da área de gramáticas para linguagens formais [21] que permitem generalizações. Esta formalização é abordada em detalhes no **Capítulo 4**, que apresenta a definição formal dos tipos de citação para o conjunto das *NEJBs*.

Definindo o conjunto de regras formais para os tipos de citação, a extração de citações é feita utilizando diversas técnicas da área de compiladores. Em um primeiro passo é feito um *parse* do texto do segmento trocando as ocorrências de palavras significativas por *tokens*. Em seguida, é feita uma análise léxica da seqüência presente no texto (gerada no passo anterior) verificando se esta corresponde a alguma das regras formais. Verificada a correspondência, significa que o texto possui uma citação de acordo com as regras estipuladas. A citação reconhecida é então marcada para que possa ser analisada na próxima fase da inferência de *links*. As marcas utilizadas no processo de conversão seguem a estrutura de etiquetas HTML e são: <cit> para o início da citação, e </cit> marcando o fim. A fase seguinte vai procurar por estas marcações quando for analisar as citações. O **Algoritmo 7** a seguir apresenta a fase de extração das citações de modo estruturado.

```
sub extraiCitacao ( $segmento )
{
    @TiposCitacao = &InicializaTiposDeCitacao;
    foreach $tipoCitacao (@TiposCitacao) {                    # para cada tipo de citação existente
        $segmento{TOKENS} = &parseTokens($segmento);          # faz a substituição por tokens
                                                              # se encontrou, marca a citação com <cit>..</cit>
        $segmento{CIT} = &analiseLexica($segmento{TOKENS}, $tipoCitacao);
    }
    return $segmento{CIT};                # retorna o segmento com a citação marcada/extraída
}
```

Algoritmo 7 Extração das citações na etapa de inferência de *links*

Como resultado desta fase de extração das citações, o texto dos segmentos é marcado quando possui citações conforme definidas pelas regras de tipos de citação.

3.6.2 Endereçamento das Citações

Do ponto de vista semântico, a citação é um "endereço", e assim sendo designa onde encontrar determinado objeto. Não se trata de um endereço físico absoluto, como o uso de termos como "linha" ou "página" designando posição física dentro da obra original. A citação é um endereço com dois significados:

- Como endereço relativo que faz referência a posições dadas pela estruturação do texto, tais como "item", "parágrafo", etc.
- Como um "endereço classificatório", estabelecendo onde o item se localiza entre os demais.

Nesse trabalho propomos fazer uma tradução do "endereço" presente na citação de modo informal para uma forma normal de endereçamento. Nesta tradução o conceito de nível de segmentos é muito importante, porque uma citação apesar de referenciar segmentos de vários níveis, é estruturada semanticamente para indicar o segmento ao qual se refere em última instância, o segmento de menor nível na hierarquia. Por exemplo, a citação "artigo 20, parágrafo único, da Lei n° 4.771/65" extraída do Decreto nº 1.282, se refere em última instância ao parágrafo único, e menciona informações de outros segmentos para possibilitar a localização do mesmo.

Assim, uma citação pode se referir a mais de um segmento e neste caso é considerada uma *citação composta*. O tratamento para as citações compostas é dividi-la em várias *citações unívocas*. Aqui, quando se menciona citação, é uma referência direta à *citação unívoca*, que cita sempre um único segmento.

Idealmente, qualquer citação feita a um segmento de nível *k* pode ser traduzida, com equivalência do ponto de vista semântico, para um endereço na notação proposta na seção seguinte.

3.6.2.1 Forma Normal para Endereçamento de Segmentos

A forma normal aqui proposta procura refletir a hierarquia de um segmento dentro de um texto estruturado. Considerando que o nível *K* é o nível do segmento, o seu endereço na forma normal é definido pelo conjunto de tuplas:

$$\{C_N , ID_N; C_{N-1}, ID_{N-1}; \ldots; C_{K+1}, ID_{K+1}; C_K, ID_K\}$$

Onde:

- C_K representa o tipo ou a categoria do segmento de nível *K* ;
- ID_K é o identificador do segmento de nível *K*;
- C_{K+1} representa o tipo do segmento que hospeda C_K ;
- ID_{K+1} é o identificador do segmento de nível *K+1*;
- e assim sucessivamente

O limite superior é C_N e designa o universo do conjunto em questão. Por exemplo, "universo das normas brasileiras", ou algo menos abrangente, por exemplo, "normas do CONAMA" ou ainda considerando um documento, N seria *nível 0* na estrutura hierárquica.

Aplicando a forma normal para as citações, uma citação inicia em C_K, pois o seu significado remete a um módulo de granularidade *k*. Se fosse exigido, por exemplo, que a forma normal iniciasse sempre por um valor fixo, diga-se C_1, a referência seria sempre feita a detalhes, ao módulo mínimo (módulos de granularidade *K*=1), sem a possibilidade de se referir a um segmento de maior nível na hierarquia (módulos de granularidade *K*>1), o que não representa fielmente a semântica da citação.

A escolha dos valores para os limites inferior e superior (*K* e *N*) é que determina a hierarquia implícita na citação. Entretanto, esta estrutura hierárquica adotada (as correspondências para os valores de C_K) e a escolha do limite superior da estrutura hierárquica (o valor para C_N) são decisões que devem ser tomadas *a priori*. A escolha deve ser direcionada para refletir a hierarquia existente no tipo de documento em questão e assim, garantir a preservação da semântica da citação. Por exemplo, com relação ao conjunto de *NEJBs*, poderia adotar-se: *K*=4 para alíneas, *K*=3 para itens, *K*=2 para parágrafos, *K*=1 para artigos, e assim por diante.

As regras para a forma normal de endereçamento – valores para C_N e C_K – devem ser selecionadas de modo a melhor cumprir as convenções de redação ou com a estruturação do conjunto de documentos em questão.

Na aplicação direta da forma normal de endereçamento a **Tabela 4** mostra alguns exemplos de endereçamento de segmentos e de tradução de citações para as *NEJBs*.

Hospedeiro	Forma Normal	Citação[24]	Forma Normal
Escopo Normativo do Decreto nº 3.079	{D,3079;EN}	"O PRESIDENTE DA REPÚBLICA, usando da atribuição que lhe confere o **Artigo 74, Letra "a" da Constituição**,"	{CF;T,4;C,1;S,9;A,74;AL,1}
Inciso VI, artigo 1º do Decreto nº 3.079	{D,3079;A,1; IR,6}	"Certidão dos documentos referidos na **Letra "b", I.**"	{D,3079;A,1;IR,1;AL,2}
Artigo 3º, §2º, do Decreto nº 1.752	{D,1752;C,3; A,3;P,2}	"Os especialistas referidos no **inciso I** serão indicados pelo …"	{D,1752;C,3;A,3;IR,1}

Tabela 4 Exemplos de endereços na forma normal de endereçamento

A forma normal de endereçamento é utilizada para gerar o endereço de cada segmento reconhecido, independentemente de este ser citado ou não. Este endereço é útil também na base de dados, funcionando como um campo de identificação-chave único para cada segmento.

[24] A citação está em negrito. As vizinhanças foram mantidas para exemplificação.

3.6.3 Estabelecimento de *Links*

A etapa de *estabelecimento de links* vai utilizar os resultados da etapa de extração das citações juntamente com o conceito de endereçamento de segmentos para identificar os segmentos citados e estabelecer o *link*. Esta etapa pode ser estruturada conforme mostra o **Algoritmo 8** a seguir.

```
sub geraLinks
{
    foreach $segmento ( @dados ) {     # para cada segmento
        foreach $citacao ( &extraiCitacao($segmento) ) {
            $segmento{CIT} = &analisaLink($segmento, $citacao);
            $segmento{TEXTO} = &substituiTokens($segmento);          # substitui os tokens pelo texto original
        }
    }
}
```

Algoritmo 8 Processo de geração de *links*

A fase mais importante nesta etapa é a *análise de links*. Partindo das marcações feitas pela extração de *links*, a fase de análise de *links* vai decompor a citação em uma <u>estrutura de pares</u>, abordada na **Seção 3.6.3.1**. A partir desta estrutura, esta etapa gera o endereço da citação na forma normal, abordado na **Seção 3.6.3.2**. Em seu último passo, pesquisa pelo segmento na base de dados e estabelece o *link*. O **Algoritmo 9** apresenta a análise de *links* estruturada.

```
sub analisaLink ( $segmento, $citacao )
{
    # Analisa a citacao, e se encontrou o segmento citado estabelece o link.
    # Caso contrário, estabelece link com uma mensagem.
    @pares = &compoePar($citacao);
    @enderecoCitacao = &FormaNormalEndereco(@pares);      # gera endereco na Forma Normal
    &comparaFormaNormal (@enderecoSegmento, @enderecoCitacao);

    if ( &FindNorma(@enderecoCitacao) ) {                 # se o segmento citado está no banco de dados
        print "Estabelece LINK com o segmento citado";    # estabelece o link
        return &estabeleceLink($segmento, $citacao);
    }
    else {
        print "Estabelece link com mensagem";
        $mensagem = "O segmento citado não está no Banco de Dados";
        return &estabeleceLink($segmento, $mensagem);
    }
}
```

Algoritmo 9 Análise e estabelecimento de *links*

Caso o segmento procurado não esteja presente na base de dados, o link é estabelecido com uma mensagem padrão.

3.6.3.1 Estrutura de Pares

A estrutura de pares é uma estrutura intermediária entre a citação e a forma normal de endereçamento. É composta por pares formados pelo tipo do segmento e seu identificador quando presente. Cada nível de segmento mencionado na citação possui seu par.

A estrutura de pares suporta casos nos quais o identificador está ausente ou é determinado pela presença de algum modificador. No caso das *NEJBs*, os modificadores mais comuns são: *deste(a), neste(a), este(a), anterior, único*. Modificadores como estes serão analisados na tradução para o endereço na forma normal e convertidos para identificadores. Alguns exemplos para a estrutura de pares estão no quadro abaixo.

Hospedeiro: {D,1282;EN,2}
1.1 Citação: "Artigo 84, Inciso IV, da Constituição"
Pares:
[nivel 0] TIPO: Constituição ID: ausente
[nivel 8] TIPO: Artigo ID: 84
[nivel 9] TIPO: Inciso ID: IV

1.2 Citação: "Lei n° 4.771, de 15 setembro de 1965"
Par:
[nivel 0]: TIPO: Lei ID: 4771
Hospedeiro: {L,3017;PT,2;L,4;T,3;C,8;A,1704}
2.1 Citação: "art. 1.679"
Pares:
[nivel 8] TIPO: Artigo ID: 1679
Hospedeiro: {D,1752;C,3;A,3;P,5}
3.1 Citação: "inciso IV deste artigo"
Pares:
[nivel 8] TIPO: Artigo ID: modificado-deste
[nivel 9] TIPO: Inciso ID: IV
Hospedeiro: {D,1282;C,2;A,8;P,2}
4.1 Citação: "parágrafo anterior"
Pares:
[nivel 9] TIPO: Parágrafo ID: modificado-anterior

Quadro 2 Exemplos da estrutura de pares aplicada às *NEJBs*

A estrutura em pares respeita o nível hierárquico de cada parte da citação e ordena-os de modo a identificar facilmente o menor nível na hierarquia (que corresponde ao elemento citado). Esta estruturação é o primeiro passo na interpretação da composição livre das citações para um formato estruturado.

3.6.3.2 Geração do Endereço na Forma Normal

Conforme mostram os exemplos anteriores, a estrutura de pares incorpora somente os segmentos explicitamente presentes na citação. A tradução da estrutura de pares para o endereço na forma normal trata de três casos em especial, abordados a seguir.

Caso 1 – A citação refere-se a um documento externo: Quando ocorre esta situação sempre existe o *nível 0* na estrutura de pares, e os dados deste nível são diferentes do *nível 0* do hospedeiro. Neste caso, a estrutura de pares nem sempre reflete toda a hierarquia de níveis do documento citado.

Por exemplo, a estrutura de pares para citação 1.1 no **Quadro 2** acima contém somente três níveis, correspondendo aos níveis mencionados. Colocando esta estrutura diretamente na forma normal, tem-se: {CF;A,84;IR,4}. No entanto, o endereço na forma normal para o segmento mencionado é {CF;T,4;C,2;S,2;A,84;IR,4}.

Para a tradução para a forma normal é necessário procurar o segmento na base de dados, a partir dos dados disponíveis. No exemplo acima, seria feita uma pesquisa por um segmento do tipo "A" de artigo e com identificação igual a 84. Ao encontrar este segmento os dados faltantes na forma normal são completados gerando o endereço correto.

Caso 2 – A citação faz referência a um segmento do mesmo documento: Nestes casos, normalmente a referência ao *nível 0* está ausente ou, estando presente, refere-se ao mesmo documento do hospedeiro. Isto ocorre porque a citação utiliza de seu contexto abreviando o número de níveis citados. Esta noção de contexto é particularmente importante.

O contexto da citação é derivado do endereço do segmento no qual a citação se encontra. Assim, o *nível 0* da citação é derivado diretamente do segmento que a hospeda. A citação 2.1 mostra esta situação. O endereço da citação traduzido diretamente para a forma

normal gera: {A,1674}. O contexto derivado do hospedeiro complementa o endereço gerando**:** {L,3017; A,1674}. O endereço ainda precisa ser completado com os outros níveis e procede-se do mesmo modo que o passo anterior. O artigo 1674 é procurado entre os segmentos da Lei nº 3.017 e o endereço completo passa a ser: {L,3017;PT,2;L,4;T,3;C,7;A,1679}.

Caso 3 – A citação possui modificadores: Este caso é uma variação do anterior. O contexto da citação também é referenciado, mas aparece modificado pela presença dos modificadores mencionados. Neste caso, a tradução para a forma normal exige um cálculo adicional sobre os dados do contexto para determinar os identificadores da citação. As citações 3.1 e 4.1 mostram esta situação.

Na citação 3.1 o contexto é {D,1752;C,3;A,3}, o modificador "deste" no nível de artigo faz com que todo o endereço até o nível de artigo seja herdado pela citação e, assim, o endereço na forma normal para a citação passa a ser: {D,1752;C,3;A,3;IR,6}.

Na citação 4.1, o contexto é {D,1282;C,2;A,8;P,2}, o modificador "anterior" no nível 9 força uma comparação com o mesmo nível do contexto {P,2}, resultando em {P,1} para o anterior. A partir da aplicação da intersecção o endereço da citação passa a ser: {D,1282;C,2;A,8;P,1}.

O endereço da citação na forma normal é utilizado como mecanismo de busca do segmento na base de dados. Uma vez encontrado, o link é estabelecido e as informações sobre o segmento referenciado são armazenadas na base de dados de acordo com o modelo para o tipo do documento.

3.7 Construção do Hipertexto

Para a construção do hipertexto são realizadas repetidas consultas nos dados para recuperar segmentos e seus segmentos filhos. Uma vez recuperados, os dados são codificados em uma linguagem específica. A restrição é que a linguagem escolhida tenha suporte para o estabelecimento de *links* entre os segmentos.

A escolha da linguagem para a codificação dos dados também está vinculada ao meio de divulgação que se deseja utilizar. Para a divulgação de dados na WWW da Internet é adequado que os dados sejam codificados em HTML. Já para a distribuição em CD-ROMs outras alternativas podem ser consideradas. A base de dados resultante do processo de conversão pode ser utilizada para fornecer diversos tipos de consultas: por assunto, por ordem cronológica ou por tipo de documento.

4 NORMAS: A NATUREZA DA INFORMAÇÃO E SEUS USOS

O conceito de norma abrange um amplo conjunto de idéias. Este capítulo investiga as *normas estatutárias jurídicas brasileiras*, que são o domínio de aplicação do trabalho, em suas diversas características: a natureza intrínseca a sua composição, seus usos atuais e tendências para estabelecer definições precisas sobre as quais atuar. Assim, este capítulo está dedicado a:

- Estabelecer uma conceituação sobre este domínio de aplicação.
- Investigar a estruturação e segmentação de uma norma.
- Apresentar uma definição formal de normas e um modelo de dados para normas, como resultado das investigações.

Essa investigação é necessária porque o conjunto que compõe as normas estatutárias jurídicas brasileiras é o domínio de aplicação da ferramenta *WebifyLaw* desenvolvida nesse trabalho.

4.1 Conceito de Norma

> Norma é "aquilo que se estabelece como base ou medida para a realização ou a avaliação de alguma coisa; princípio, preceito, regra, lei; Modelo, padrão"[25].

4.1.1 Tipos de Normas

Normas estatutárias jurídicas são um conjunto resultante da interseção de outros dois conjuntos: o de *normas estatutárias* e o de *normas jurídicas*. *Normas jurídicas*, no âmbito deste trabalho, são resultantes de decisões do poder Legislativo e são normas cuja *agência formuladora*[26] é o Estado, ou uma subdivisão deste. Além disto, uma norma jurídica estabelece, implícita ou explicitamente um *contrato*[27] entre membros da comunidade.

As normas que não estabelecem um contrato são consideradas *normas não-jurídicas*. Encaixam-se nesta definição as normas formuladas por comissões não-governamentais, de organizações públicas ou privadas, como por exemplo as *normas técnicas* de organizações como a ABNT no Brasil e a ISO no meio internacional.

---

[25] Dicionário Aurélio [15].

[26] A *agência formuladora* é uma organização, ou parte de uma, responsável pelo registro dos interesses consensuais de uma dada comunidade, grupo, organização ou consórcio de organizações que adota a norma. É também responsável por manter a custódia dos documentos originais da norma.

[27] Um *contrato* é a legitimação de um acordo entre partes. Numa norma jurídica, o *contrato* é conhecido como contrato social, e corresponde às regras que devem ser cumpridas para a integração na comunidade. Quando tal comunidade engloba todos os membros do Estado, esta comunidade corresponde à sociedade. Deste modo, a constituição de uma nação estabelece as regras – direitos e deveres – dos membros da comunidade, formada por todos os cidadãos.

Normas estatutárias são documentos previstos e determinados por um estatuto. Sua função é regulamentar o funcionamento da comunidade regida por tal estatuto. Algumas propriedades gerais, que as distinguem das normas não-estatutárias são:

1. O documento é gerado com a finalidade principal de registrar uma norma.
2. Possui uma única agência formuladora que a formula e atualiza.
3. São documentos livres de acréscimos não-relativos à norma. O documento na íntegra consiste no texto da norma. Mesmo quando possui apêndices, estes também são partes constituintes da norma.

Por fim, as normas estatutárias jurídicas – *NEJs*, pela aplicação da interseção, são documentos pertencentes a um estatuto e, com valor de tal, formulados através de agências estatais que definem regras para os membros do Estado ou subdivisão deste, de acordo com a agência formuladora. Assim, uma norma formulada por um município tem como agência formuladora os órgãos Legislativo ou Executivo do município. E por princípio, se aplica aos membros de tal comunidade, que correspondem aos cidadãos do mesmo município.

4.1.2 Características de Normas

Ao procurar estabelecer características das *NEJBs* foi feita uma pesquisa por meta-normas[28] que definissem a composição e estruturação de normas deste conjunto. No entanto, foi verificado que essas meta-normas nunca foram formalizadas. Buscava-se uma norma semelhante à NBR 6822 – "Preparo e Apresentação de Normas Brasileiras" – norma da ABNT que define a composição de normas técnicas. Tal norma, mesmo apresentando certas definições e restrições particulares às normas técnicas, apresenta um arcabouço aplicável a normas gerais. Entretanto, na falta dessa formalização, as características particulares de

textos de normas foram estabelecidas a partir da consulta e observação dos textos de normas e da pesquisa em literatura especializada.

Sobre a redação de normas, duas colocações extraídas da NBR 6822 são de grande importância:

> "A redação da norma tem estilo próprio, lingüisticamente correto, sem preocupações literárias e tanto quanto possível uniforme. A qualidade essencial é a clareza do texto, para evitar interpretações diferentes."
>
> [NBR 6822, *Princípios Gerais*, seção 5.1.1]
>
> "As prescrições de uma norma devem ser redigidas de maneira impessoal e sem apreciações subjetivas, …"
>
> [NBR 6822, *Prescrições*, seção 5.2.1]

Destas afirmações podemos extrair características básicas para as normas estatutárias jurídicas:

- Corretude Lingüística: O texto de uma norma passa por verificações lingüísticas e revisões antes de ser aprovado e publicado.
- Uniformidade: Garante que todas as normas seguem as mesmas diretivas e apresentam-se uniformes com relação a forma e estilo. Apesar da ausência de meta-normas, a uniformidade é garantida pela forte influência que a tradição representa na redação de normas jurídicas. Esta tradição faz com que ao escrever um texto de normas procure-se seguir a mesma estruturação hierárquica utilizada nos textos mais antigos.
- Clareza: Visando atingir esta característica as frases do texto de uma norma são preferencialmente construídas em ordem direta e são curtas "para facilitar o perfeito entendimento". São empregadas palavras de uso corrente, sentido preciso e não-ambíguo geralmente pertencente a um conjunto de vocábulos restrito. Essa característica garante que cada palavra possui somente um único significado dentro

[28] Também conhecidas neste âmbito como normas de formatação.

de uma norma. Quando esta regra não é cumprida, há referência explícita a este fato no próprio texto da norma[29].

- Colocações Diretas: Esta característica reforça a clareza do texto, e facilita o tratamento automatizado por processamento da linguagem natural – PLN.

4.1.3 Princípios para o Tratamento das *NEJBs*

As características mencionadas na seção anterior permitem identificar três princípios que indicam que o tratamento automatizado de normas é possível. Esses princípios são discutidos a seguir.

O conteúdo de uma *norma* é suficiente para o entendimento e interpretação da mesma.

Outras normas que podem estar citadas durante o texto de uma norma, assim são com a intenção de fornecer informações extras, cuja ausência não compromete o entendimento da mesma. De acordo com a NBR 6822, o mesmo pode ser aplicado às notas:

> "As notas de rodapé são de caráter informativo, sendo utilizadas para prestar esclarecimentos, justificativas, deduções, demonstrações, explicações, chamar atenção para alguma particularidade, fazer referência a uma prescrição de outra seção ou de outra norma, com tal conteúdo que a sua eliminação não torne a norma incompleta."
>
> [NBR 6822, *Notas*, seção 4.3.5.1].

Normas são redigidas em uma linguagem que difere da linguagem natural por ter um nível de ambigüidade menor e controlado.

Uma das justificativas para isto é que a linguagem utilizada na redação de normas tem um estilo próprio, em oposição ao estilo livre ou a ausência de estilo que rege a linguagem natural. O tratamento automatizado de normas jurídicas é, portanto, um problema de complexidade computacional e lingüística muito menor que o tratamento de linguagem natural.

[29] Por exemplo, no Código das Águas, o termo 'corrente' é utilizado em referência a correntes de água e este é o significado único, embora o vocábulo em si possa assumir outros significados em diferentes contextos.

Uma *norma jurídica* está inserida em um contexto de normas maior com relações particulares entre si.

Este contexto maior é referido na teoria do Direito como *ordenamento jurídico* [Bobbio]. Uma *norma jurídica* apesar de ter conteúdo suficiente para o seu entendimento, não existe isoladamente. A análise das relações entre normas é importante para um entendimento maior da totalidade de sua ação dentro do domínio do *Direito*. De acordo com Bobbio [7]:

> "a palavra 'direito', entre seus vários sentidos, tem também o de 'ordenamento jurídico', por exemplo nas expressões 'Direito romano', 'Direito canônico'"
>
> [7], p.19

Outras características de normas são provenientes da sua estruturação e serão abordadas na próxima seção.

4.2 Estruturação e Segmentação de uma Norma

A principal justificativa para o interesse no tratamento automatizado de norma, entretanto, não são as características apresentadas na **Seção 4.1.3**. O atrativo principal decorre da constatação de que hipertextos são uma forma natural e eficiente para a representação e uso de normas. Essa constatação advém do fato que normas são textos altamente estruturados. A exploração dessa estrutura permite a construção automatizada de hipertexto.

Há quatro elementos básicos na estruturação de normas: sub-divisão do texto, estrutura hierárquica, correlação entre as estruturas semânticas e explícita, e segmentação.

O texto de uma norma pode ser sub-dividido e apresenta uma segmentação interna inerente à sua composição.

Com relação a sua estruturação geral, os textos de normas "são em geral modulares, explicitamente segmentados" . De acordo com a NBR 6822, tal propriedade é intrínseca a qualquer norma, e independentemente do seu tipo apresentam a estruturação geral a seguir:

> "Independentemente do seu tipo, uma norma compreende, na sua apresentação mais completa, de:
>
> a) identificação;
> b) elementos preliminares;
> c) texto subdividido em capítulos;
> d) elementos complementares."
>
> [NBR 6822, *Estrutura das Normas*, seção 4]

Esta citação vem reforçar outro aspecto na estruturação de normas que é a segmentação interna, particularmente através do item c. Embora a experimentação mostre que diversas normas empregam divisão em outras variantes além de simplesmente capítulos[30], tal constatação não torna seu propósito falso, mas em contrapartida, vem reforçá-lo. Por exemplo, no conjunto de normas estatutárias jurídicas é comum a presença de seções como *parte*, *livro*, *título* e *capítulo*.

A estrutura típica de uma *norma estatutária jurídica* é a estrutura hierárquica.

A subdivisão do texto mencionada acima não define uma estrutura específica, que poderia ter a forma de um grafo sem restrições ou assumir uma estrutura mais formal, como a estrutura na forma de árvore. A observação de textos de normas permite notar que os mesmos são normalmente segmentados de modo a manter uma *estrutura interna hierárquica*, conforme define a NBR 6822:

> "O texto contém as prescrições da norma e apresenta-se subdivido em seções e eventualmente, alíneas e subalíneas, incluindo também, figuras, tabelas, notas e anexos."
>
> [NBR 6822, *Estrutura do Texto*, seção 4.3]

Na composição de normas, para tentar atingir as características de significado claro e não-ambígüo, cada elemento explora um assunto único. Quando há necessidade de mencionar outro assunto, outro elemento é criado com esta função. De acordo com as relações entre os assuntos, estes elementos passam a ou pertencer a uma mesma classe, ou um passa a estar subordinado ao outro. Por exemplo, o artigo 1º do Código das Águas, Decreto nº 24.4643, diz:

[30] No conjunto de *NEJBs* , notou-se também a presença de seções nomeadas como "livro", "titulo" e "seção", além do tradicional "capítulo".

"As águas públicas podem ser de uso comum ou domiciliares."

Esta afirmação necessita da especificação do que são consideradas águas de uso comum. Tal informação trata de um assunto derivado deste, e que por ter importância relativa, passa a ser apresentada no próximo artigo, que vai especificar todos os tipos de águas comuns.

Em outros casos, a informação complementar pode ser convertida em parágrafos para aquele artigo ou mesmo alíneas. Analogamente, esta divisão ocorre também para as seções e suas sub-seções quando o assunto é mais abrangente.

Por exemplo, o código civil na Parte Especial, Livro I, o Título I é dedicado ao casamento e assim intitulado: "Do Casamento". Com relação a este assunto os vários tópicos estão divididos em capítulos, nomeando-os: "Das Formalidades Preliminares", "Dos Impedimentos", "Da Oposição dos Impedimentos", "Da Celebração do Casamento", "Das Provas do Casamento", "Do Casamento Nulo e Anulável", e "Disposições Penais". Em outros casos, a divisão continua a existir nos capítulos em seções e sub-seções.

Através deste exemplo, fica claro que a estrutura implícita no texto de uma norma é uma estrutura que partindo de uma generalidade ramifica-se em vários tópicos sobre o assunto principal, na forma de uma estrutura hierárquica.

Normas possuem uma relação direta entre a sua estrutura semântica e sua estrutura explícita. Tal relação é proporcional à clareza semântica e à não-ambigüidade de seus elementos.

Adotando-se a postura de que uma norma possui uma estrutura hierárquica, e que cada elemento trata somente de um único assunto, pode-se considerar o texto de uma norma como composto por elementos hierarquizados com significados semânticos distintos. Através de um mapeamento, é possível identificar uma estrutura que representa a segmentação semântica do texto. Indo além, este mapeamento corresponde à segmentação explícita do texto, a suas divisões entre seções, artigos, parágrafos, itens e

alíneas. Através do exemplo anterior, verifica-se que a segmentação por assunto corresponde à segmentação hierárquica.

O texto de uma *norma estatutária jurídica* pode ser organizado em *segmentos*.

A *segmentação do texto* de uma *norma* é um caso particular da *modularização*. Na *segmentação*, é aplicado um único critério de modularização sobre todos os módulos obtendo elementos com um mesmo nível de granularidade, chamados de *segmentos*. A *segmentação* proposta considera a *frase* como o elemento mínimo de uma norma, definindo-o como um *segmento de uma norma*. Esta definição o determina também, como sendo o *elemento mínimo citável*[31]. Tendo como base a estrutura explícita da norma é possível separá-la em diversos *segmentos*, com a certeza de que estes possuem significado semântico e coesão interna.

Uma vez que uma norma possui uma estruturação explícita (essencial para a geração do hipertexto) e que sua segmentação pode realmente ser realizada do ponto de vista semântico (sem danos para seu significado), para realizar tal segmentação este projeto vai se apoiar em uma definição formal desta estrutura. A partir dessa definição, elementos de uma norma podem ser comparados com os elementos da definição e identificados como segmentos. Com a definição de segmentos, que é uma proposta original desse trabalho, é possível realizar um tratamento formal das *NEJBs*. A **Seção 4.3** a seguir detalha esse tratamento.

[31] O elemento mínimo citável é uma definição importante para a identificação das citações presentes nos segmentos de normas, pois corresponde ao menor elemento que pode ser referenciado por uma citação. Essa definição impede, por exemplo, que uma linha ou palavra sejam referenciadas em uma citação.

4.3 Definição Formal das *NEJBs*

Para suprir a ausência de meta-normas para a composição e verificação do conjunto de *NEJBs* foi estabelecida esta definição formal através da análise de diversos documentos pertencentes a este conjunto.

A definição formal proposta utiliza dois recursos para especificar a estruturação explícita e a sintática do conjunto de *NEJBs*:

- Uma especificação de regras de produção para o conjunto de *NEJBs*, apresentada na **Seção 4.4**.
- Um conjunto de regras de validação especificado através de descrições informais em linguagem natural explicitando algumas restrições envolvendo a semântica dos elementos, apresentadas na **Seção 4.4.3**.

4.3.1 Fundamentação Teórica para a Definição Formal

Gramáticas, especificações finitas para linguagens, resultaram do estudo para estabelecer meios de fornecer descrições estruturais das sentenças. Este estudo foi iniciado por lingüístas e inicialmente voltado para as linguagens naturais [21]. O conceito de gramática é baseado nas seguintes definições:

- *Símbolos não-terminais*: Um símbolo que representa uma construção gramatical que pode ser expressa através de regras em termos de construções menores [12]. Originalmente representados entre os símbolos “<” e “>”, aqui estes foram abolidos para facilitar a leitura.
- *Símbolos terminais*: Palavras que pertencem à sentença final. Na definição aqui adotada estão sempre representados entre aspas duplas “”, tendo seu significado

literal. O texto é indiferente para letras maiúsculas e minúsculas[32], a menos que seja explicitamente mencionado o contrário.

- *Produções* ou *Regras de produção*: São relacionamentos existentes entre *variáveis* e símbolos *terminais*, indicando meios de derivar os diversos componentes.
- *Símbolo inicial*: É o símbolo que dá início às regras de produção. Na definição formal para *NEJBs* este símbolo é **Norma**, especificada na regra 1, na **Seção 4.4**.

Uma gramática é formalmente definida como uma tupla G (V_N, V_T, P, S) onde:

- V_N representa o conjunto dos *símbolos não-terminais* ou *variáveis;*
- V_T representa o conjunto de *símbolos terminais;*
- P representa as *produções* ou *regras de produção;*
- *S* representa o *símbolo inicial.*

As restrições para estes conjuntos são:

- Os conjuntos V_N e V_T não possuem elementos em comum, assim: $V_N \cap V_T = \varnothing$;
- O conjunto resultante da união de V_N e V_T é denotado por V, assim $V_N \cup V_T = V$;
- O conjunto de produções P possui regras da forma: $\alpha \rightarrow \beta$. Onde α é um símbolo de V^+ e β é um símbolo de V^*.
- O símbolo *S* é sempre um símbolo de V_N.

Aplicando-se as regras definidas em G, teremos uma *linguagem* gerada por G, denotada L(G). Assim, uma seqüência de símbolos pertence a L(G) se a seqüência possui somente termos terminais e pode ser derivada a partir do símbolo inicial *S* [21].

[32] No inglês, o texto é *case insensitive*.

Entre os vários tipos de gramática, as *Gramáticas Livre de Contexto* ou do *Tipo 2* são de relevância para este trabalho. Nestas gramáticas toda regra de produção em P obedece às seguintes restrições:

$\alpha \rightarrow \beta$; onde α é uma única variável e β é qualquer seqüência de caracteres, com exceção da seqüência vazia $\in$.

A $\rightarrow \beta$; fica sendo a regra de produção das gramáticas livre de contexto

O nome *livre de contexto* deriva desta última regra de produção que permite que **A** seja substituída por β independentemente do contexto no qual **A** aparece.

A forma normal de Chomsky para gramáticas livre de contexto determina que qualquer linguagem livre de contexto pode ser gerada por uma gramática na qual todas as produções são da forma [21]:

A→BC ou **A→a**; onde **A**, **B** e **C** são variáveis e **a** é um termo terminal.

O conjunto das regras de produção para as *NEJBs* seguem estas definições. No entanto, a notação adotada aqui é a notação BNF estendida, similar à utilizada na RFC nº 822 [35]. Algumas construções da BNF estendida facilitam o entendimento das regras de produção e são de grande relevância para este trabalho. Estas são reproduzidas a seguir com acréscimos para adequação às *NEJBs*.

Regra de Composição de regras de produção:

nome → definição

Esta regra determina que toda ocorrência de *nome* é substituída pelos elementos especificados em *definição*, onde *nome* é um único *símbolo não-terminal* e *definição* pode ser um composto de *símbolos não-terminais* e *terminais*. Esta definição está de acordo com as especificações das regras de produção para uma gramática livre de contexto [21].

Um exemplo de regras de produção neste formato:

DataAbrev → Dia "/" Mes "/" Ano

Dia → 0-31

Mes → 1-12

Ano → 1900-1997

Regra de Alternativa:

regra1 / regra2

A barra "/" representa um "ou lógico", assim "primeiro" / "segundo" especifica que tanto primeiro quanto segundo são aceitos.

Regra de Agrupamento:

(regra3 regra4)

Elementos em parênteses são tratados como uma seqüência, como partes de um mesmo grupo. Por exemplo:

("ab" ("c"/"d") "ef") permite as seqüências: "abcef" e "abdef"

Regra de Repetição:

<n><m>regra5*

onde:

<n> é opcional e representa o número mínimo de ocorrências de *regra5*.

<m> também é opcional, representa o número máximo de ocorrências de *regra5*.

Na ausência de *<n>* e *<m>*, a regra permite 0 ou infinitas ocorrências de *regra5*, permitindo qualquer número de iterações.

Por exemplo: 1*("abc") permite seqüências: "abc", "abcabc", …

2*3("abc") permite seqüências: "abcabc" e "abcabcabc".

Regra Opcional:

[regra6]

Esta regra determina que o conteúdo entre chaves é opcional, podendo ocorrer ou não. Trata-se de uma representação de: *1regra6, no mínimo zero e no máximo 1 ocorrência.

Regra de Repetição Específica:
n regra7

Esta regra especifica um número exato de repetições da regra. Por exemplo:

2(“ab”) 3(“c”) permite somente a seqüência: “ababccc”.

Regra Comentário:
; comentário

Permite que comentários sejam colocados sempre à direita de uma regra de produção. Por exemplo:

Start → Item1 / Item2 ; regra inicial das produções

4.3.2 Aplicação para o Conjunto das *NEJBs*

A definição de uma gramática para o conjunto das *NEJBs* exige algumas definições relação ao *alfabeto* utilizado pela gramática. O *alfabeto* ou *vocabulário* é o conjunto de vocábulos definidos pela língua portuguesa. Este conjunto, apesar de possuir um grande número de elementos, é finito.

As siglas e abreviaturas devem ser controladas de algum modo. Uma sugestão seria a criação de um banco de dados oficial com a grafia correta e o significado adotado. Outra solução seria a adoção de uma ou mais definições como um padrão a ser seguido. Seria necessária uma custódia de um órgão (governamental ou não) que se responsabilizasse pela manutenção e atualização desta informação. Hoje, temos conhecimento de diversas agências nacionais como: CONAMA, IBAMA, entre outras. A existência de um banco de dados oficial ajudaria na identificação destas agências, de seus responsáveis e inclusive dos seus períodos de existência.

Os *símbolos não-terminais* para a gramática de *NEJBs* é um conjunto extenso e está listado no glossário que encontra-se no **Apêndice A**.

4.4 Regras de Produção para as *NEJBs*

Como um primeiro passo no estabelecimento da definição formal foi definido um conjunto de regras que se aplicam a qualquer texto escrito em português do Brasil. Na **Seção 4.4.2** são apresentadas as regras de produção para o conjunto das *NEJBs*.

4.4.1 Regras de Produção para o Português do Brasil

As regras apresentadas na **Tabela 5** definem o conjunto de caracteres válidos quando compondo ou analisando um texto em português. As regras foram baseadas no vocabulário definido no Formulário Ortográfico aprovado pela Academia Brasileira de Letras em 1943 e regulamentado pela Lei nº 5.765. Acréscimos foram feitos para incluir:

- caracteres especiais utilizados no âmbito das *NEJBs*, por exemplo o símbolo "§";
- termos específicos do processamento de texto, como por exemplo o *carriage return* e o *linefeed*.

De acordo com a definição aqui adotada, equações, gráficos, tabelas e similares serão tratados como elementos externos ao texto da norma, e assim sendo, não serão processados ou verificados dentro do escopo deste trabalho.

REGRAS DE PRODUÇÃO			COMENTÁRIOS
Ctl	→	< caracteres de controle ASCII, mais o DEL>	
Cr	→	<"retorno de carro" do ASCII, *carriage return*>	
Lf	→	<nova linha do ASCII, *linefeed*>	
CarEsp	→	<espaço do ASCII, *space*>	
Tab	→	<tab horizontal do ASCII, *horizontal-tab*>	
Crlf	→	Ctl Lf	
Esp	→	1*(CarEsp / Tab)	; espaço genericamente
EspLn	→	Esp Crlf	; *Esp* com nova linha
Pt	→	"."	; caracter ponto final
2p	→	":"	; caracter dois pontos
Vg	→	","	; caracter vírgula
Pv	→	";"	; caracter ponto-e-vírgula
Hf	→	"-" / "–"	
Br	→	"\" / "/"	; todas as barras
Figura, tabela,	→	<externos>	; elementos fora do escopo

REGRAS DE PRODUÇÃO			COMENTÁRIOS
equação, ...			destas regras são considerados externos
LminOrd	→	“a” / “b” / “c” / “d” / “e” / “f” / “g”/ “h” / “i” / “j” / “l” / “m” / “n” / “o” / “p” / “q” / “r” / “s” / “t” / “u” / “v” / “x” / “z”	; letra minúscula brasileira ordinária, 23 letras.
LminEsp	→	“ç” / “ü” / “à” / “ã”/ “õ”	
LminAcent	→	“á” / “é” / “í” / “ó” / “ú” / “â” / “ê” / “î” / “ô” / “û”	
LetraMin	→	LminOrd / LminEsp / LminAcent	
LmaiOrd	→	“A” / “B” / “C” / “D” / “E” / “F” / “G”/ “H” / “I” / “J” / “L” / “M” / “N” / “O” / “P” / “Q” / “R” / “S” / “T” / “U” / “V” / “X” / “Z”	; 23 letras do alfabeto, não considera acentuação.
LmaiEsp	→	“Ç” / “Ü” / “À” / “Ã” / “Õ”	
LmaiAcent	→	“Á” / “É” / “Í” / “Ó” / “Ú” / “Â” / “Ê” / “Ô” / “Ã” / “Õ”	
LetraMai	→	LmaiOrd / LmaiEsp / LmaiAcent	; qualquer letra maiúscula
LetraEsp	→	LminEsp / LmaiEsp	; qualquer letra especial
LetraOrd	→	LminOrd / LmaiOrd	; qualquer letra ordinária
LetraAcent	→	LminAcent / LmaiAcent	; qualquer letra acentuada
Letra	→	LetraMai / LetraMin / LetraAcent	; qualquer letra
Separador	→	“**-**”	; hífen - ASCII 30
PontFinal	→	“**.**” / “**!**” / “**?**” / “**:**”/ “**...**”	; pontuação de final de frase
PontGen	→	“**,**” / “**;**” / “**–**” / PontFinal	; pontuação genérica
PontItem	→	“*****” / “**§**”	
Delimit	→	<“> / “**[**“ / “**]**” / “**(**“ / “**)**” / “’”	; delimitadores de texto
DigArab	→	“0” / “1” / “2” / “3” / “4” / “5” / “6” / “7” / “8” / “9”	
DigRom	→	“M” / “D” / “C” / “L” / “X” / “V” / “I”	; dígitos romanos
Número	→	NumArab / NumRom / NumOrd	
NumInt	→	(NumInt [pt] 3DigArab) / 1*3DigArab	
NumArab	→	NumInt / NumFrac	
NumRom	→	[*M] [CM / 0*1CD / 0*1D3*C] [XC / 0*1XL / 0*1L3*X] [IX / 0*1IV / 0*1V3*I]	
NumOrd	→	NumInt (“ º “ / “°”)	
NumFrac	→	NumInt “,” 1*2DigArab	; no formato 1.123,45
NomeOrd	→	1*(Letra [Separador/Esp])	; nomes com certo significado
NomeEsp	→	Sigla / 1*(Letra [Separador/Esp])	; nome com significado especial
Nome	→	NomeOrd / NomeEsp	; glossariados ou não
Abrev	→	1*Letra[pt]	
Porcent	→	NumArab [Esp] “%” [Esp Descr]	; porcentagem
Sigla	→	1*(Letra / Número)	; qualquer Sigla
Cifra	→	Sigla “$” [Esp] NumArab	
Palav	→	1*(Letra [Separador])	
Vocab	→	Palav / Nome / Número	; qualquer vocábulo
Descr	→	“**(**“ 1*Vocab “**)**”	; descrição entre parênteses
Expr	→	(<”>/ “’”) 1*Vocab (<”>/ “’”)	; expressão entre aspas
Frase	→	1*(Vocab / Descr / Expr [esp]) [PontGen]	
ParagGen	→	1*Frase EspLn	

Tabela 5 Regras para a definição de texto brasileiro, baseado na língua portuguesa

4.4.2 Regras de Produção para o conjunto das *NEJBs*

As regras de produção para o conjunto das *NEJBs* são apresentadas na **Tabela 6** a seguir. A notação adotada foi a BNF que é uma notação que permite gerar regras que estabelecem a sintática de cada elemento pertencente a um conjunto e além disso, permite definir restrições sobre a hierarquia entre esses elementos.

As regras de produção, parte da definição de uma gramática, foram descritas para o conjunto das *NEJBs* com a intenção de servir como:

- guia na criação e escrita de novas normas;
- base para a construção de um processo de análise das normas pertencentes ao conjunto das *NEJBs*.

A preocupação concentra-se no aspecto estrutural, hierárquico e morfológico das normas. O significado dos elementos não pode ser representado através desta definição formal.

REGRAS DE PRODUÇÃO				COMENTÁRIOS
1.	**Norma**	→	(Cabeçalho Corpo)	; regra inicial
2.	Vocab	→	Palav / Nome / Número / Termo	; redefinição da regra anterior
3.	AbrevN	→	"Nº" / "N.º" / "N." / "N.°" / "N°." / "N°"	; variações de nº.
4.	PrepDE	→	"de"	; preposição DE
5.	DataExt	→	Dia Esp PrepDE Esp Mês Esp PrepDE Esp Ano	; data por extenso
6.	Dia	→	1*2DigArab <numeral arábico de 1 a 31>	; no mínimo 1DigArab
7.	Mês	→	"Janeiro" / "Fevereiro" / "Março" / "Abril" / "Maio" / "Junho" / "Julho" / "Agosto" / "Setembro" / "Outubro" / "Novembro" / "Dezembro"	; sem distinção entre maiúsculas e minúsculas
8.	Ano	→	1DigArab [pt] 3DigArab	; por ex.: 1.996 ou 1987
9.	Única	→	"Único" / "Única"	; ex. "Capítulo Único"
10.	Cabeçalho	→	NormaLabel [Apelido] [DispInic] [EscopoN]	; normas jurídicas
11.	NormaLabel	→	(TipoJur Esp AbrevN [Esp] NormaID [Esp] [hf/vg] [Esp] PrepDE Esp DataExt EspLn) / TipoConst	; verdadeiro para todos casos observados.
12.	TipoJur	→	"Lei" / "Lei Delegada" / "Decreto" / "Decreto-Lei" / "Emenda Constitucional" / "Medida Provisória" /TipoConstGen / TipoJurEsp	; indiferente para maiúsculas ou minúsculas.
13.	TipoJurEsp	→	(("Portaria" / "Instrução" ["Normativa"]) / "Resolução" / "Ordem de Serviço") Instituição	; *tipojur* que requer uma instituição após descrição
14.	TipoJurGen	→	<quaisquer outros tipos de normas jurídicas>	; para expansão futura
15.	TipoConst	→	"Constituição da República Federativa do Brasil"	; não requer numeração.
16.	TipoConstGen	→	<quaisquer outros tipos de normas constitucionais>	; para expansão futura
17.	Apelido	→	Descr / Frase EspLn	; qualquer texto

REGRAS DE PRODUÇÃO				COMENTÁRIOS
18.	DispInic	→	Descr / (Frase [pt]) EspLn	; disposições iniciais
19.	EscopoN	→	Respons [EspLn] Frase [2p] EspLn	; texto e return
20.	Corpo	→	1*(Seção / Artigo)	
21.	Seção	→	Livro / Título / Capítulo / SecaoC / SubSec	
22.	Livro	→	"Livro" SeçãoLabel 1*Título	; requer no mínimo 1 Título
23.	Título	→	"Título" SeçãoLabel 1*(Capítulo / OutraSec / Artigo)	; caso comum requer Capítulo
24.	Capítulo	→	"Capítulo" SeçãoLabel 1*(OutraSec/Artigo)	; caso comum possuir Artigo
25.	OutraSec	→	(SeçãoC / SeçãoS)	
26.	SeçãoC	→	"Seção" SeçãoLabel 1*(SubSec / Artigo)	; seção comum
27.	SubSec	→	"Sub-Seção" SeçãoLabel 1*Artigo	; sub-seção
28.	SeçãoS	→	["Seção"] SeçãoLabel 1*Artigo	; ocorre Seção sem Label
29.	Componente	→	1*(Artigo / Parágrafo / ItensNum / ItensLetra / ItensCom / Anexo / DispFin)	
30.	Artigo	→	ArtigoLabel Caput [Parags / ItensNum / ItensLetra / ItensCom / *Anexo]	
31.	Caput	→	Frase [pt/2p] EspLn	
32.	Parags	→	Parag *(EspLn Parag)	
33.	ItensLetra	→	ItemLetra *(pv EspLn ItemLetra) pt	
34.	ItensNum	→	ItemNum *(pv EspLn ItemNum) pt	
35.	ItensCom	→	ItemCom *(pv EspLn ItemCom) pt	
36.	Anexo	→	AnexoLabel <conteúdo externo>	; referência ao arquivo
37.	Parag	→	ParagLabel ParagGen [ItensNum / ItensLetra / ItensCom / *Anexo]	
38.	ItemNum	→	ItemNumLabel ParagGen [ItensLetra / ItensCom / *Anexo]	
39.	ItemLetra	→	ItemLetraLabel ParagGen [ItensCom / *Anexo]	
40.	ItemCom	→	ItemComLabel ParagGen *Anexo	; item comum
41.	DispFin	→	[NomeOrd vg DataExt [pv] ParagGen [pt]] *(NomeOrd [EspLn])	; disposições finais: data e nomes
42.	SeçãoLabel	→	[[Esp] SeçãoID [EspLn] [hf]] SecTit EspLn	
43.	ParagLabel	→	(("§" [Esp] ParagID [".”] ["º"/"°"]) / ParagUn) [Esp] [hf] [Esp]	; **§ X.º** ou parágrafo único
44.	ArtigoLabel	→	("Art." Esp ArtigoID [.] ["º"/"°"] [Esp] [hf] [Esp])	; a partir de ID 10 sem **º**.
45.	ItemLetraLabel	→	ItemLetraID (")" / pt)	; por exemplo: **a)** ou **a.**
46.	ItemNumLabel	→	ItemNumID [Esp] [hf] [Esp]	
47.	ItemComLabel	→	[1*DigArab]	; pode ser vazio
48.	AnexoLabel	→	[]"Anexo" [AnexoID] []	; HTML reference
49.	NormalID	→	[*3DigArab [pt]] 3DigArab	; formato 999.999
50.	SeçãoID	→	NumRom / Único	; um número romano
51.	ArtigoID	→	1*DigArab / NumInt	; no mínimo 1DigArab
52.	ParagID	→	1*DigArab / NumInt	; sem máximo
53.	ItemLetraID	→	LetraMinOrd	; só uma letra
54.	ItemNumID	→	1*DigArab / NumInt / NumRom	
55.	AnexoID	→	NumInt / NumRom	
56.	SecTit	→	Frase	; título da seção
57.	ParagUn	→	"Parágrafo Único"	; diversas grafias
58.	Instituição	→	Palavra *(br / hf / Esp Palavra)	; no mínimo 1 palavra

REGRAS DE PRODUÇÃO		COMENTÁRIOS
59.	Respons → 1*(Frase / Instituição) [PontFinal EspLn]	; identificado por palavras-chave

Tabela 6 Regras para a definição formal de normas estatutárias jurídicas brasileiras[33]

A partir da definição formal das *NEJBs*, esta passa a ser adotada como um mecanismo formalizador e padronizador das normas deste conjunto. A ferramenta *WebifyLaw*, discutida no **Capítulo 5**, adota esta definição como base para a sua implementação.

Assim, além da restrição conceitual apresentada na **Seção 4.1.1**, normas são consideradas parte do conjunto de *NEJBs* para o escopo deste projeto, se adicionalmente seguem as regras que foram apresentadas nesta seção e que serão apresentadas na seção seguinte.

4.4.3 Regras de Validação

As regras de produção apresentadas na seção anterior especificam detalhes da estruturação de cada componente de uma norma. No entanto, para que uma norma possa ser validada, é necessário que alguns elementos tenham seus conteúdos verificados. Estes elementos correspondem àqueles que não possuem uma etiqueta padrão que é o caso específico de *apelido*, *disposições iniciais* e *escopo normativo*.

Estas regras de validação foram levantadas através da análise exaustiva de normas. O conjunto de regras aqui apresentado tem a intenção de servir como um guia para estas restrições. Trata-se de um conjunto aberto para que novas ocorrências observadas possam ser acrescentadas. Para o escopo deste trabalho as regras de validação são as seguintes:

- Na categoria de **apelido** definida pela regra 17, a frase deve ser iniciada pelas palavras: "lei" ou "código" para ser reconhecida como o apelido da norma. Por exemplo:
    - "Lei da Introdução"
    - "Código Florestal"

[33] Apesar de não fazer parte da especificação da notação utilizada, as regras foram numeradas para facilitar a futura referência às mesmas.

- "Código Civil"

- Para a categoria de **disposições preliminares** definida pela regra 18, o conjunto de palavras que iniciam a frase é composto por:

Acrescenta	Cria	Estabelece	Limita	Restabelece
Adota	Dá	Estatui	Mantém	Retifica
Altera	Declara	Estende	Modifica	Revoga
Aprova	Decreta	Exclui	Permite	Simplifica
Assegura	Define	Expede	Promulga	Suspende
Atribui	Determina	Extingue	Reajusta	Torna
Autoriza	Disciplina	Faculta	Reconhece	Transfere
Baixa	Dispensa	Fixa	Regula	Veda
Concede	Dispõe	Institui	Regulamenta	
Consolida	Eleva	Isenta	Releva	

- O **escopo normativo** definido pela regra 19 é identificado pela ocorrência das seguintes seqüências de palavras no início da frase:

A ministra	O conselho
A presidenta	O diretor
Considerando	O ministro
Decreta:	O presidente
Faço saber	O vice-presidente
Havendo	Resolve:
O chefe	

Um segmento de uma norma só pode ser identificado como um dos três tipos de segmentos acima mencionados – *apelido*, *disposições preliminares* e *escopo normativo* –se as restrições apresentadas forem seguidas.

4.5 Regras de Produção para as Citações

As regras de produção para as citações têm um escopo diferente das regras de produção para *NEJBs* pois analisam o <u>conteúdo</u> de um segmento de norma. Essas regras devem ser aplicadas somente aos segmentos identificados como *disposições iniciais*, *escopo normativo* e *componente*, de acordo com as regras 18, 19 e 29 da **Tabela 6**, respectivamente. Isto porque somente estes segmentos possuem um texto que pode conter uma citação.

REGRAS DE PRODUÇÃO			COMENTÁRIOS
1.	ag0	→ <agências nacionais, estaduais…>	; ex: CONAMA, IBGE, IBAMA, IBDF.
2.	ag1	→ “Federal” / “Estadual”	;classe de agência
3.	na	→ NumInt <número arábico inteiro>	; mapeamento para a regra na **Tabela 5**.
4.	nr	→ NumRom <número romano>	; idem à regra 3.
5.	dm	→ Mês <mês no calendário gregoriano>	; idem à regra 3.
6.	lt	→ LminOrd	; idem à regra 3.
7.	apelido	→ “Cód”(“igo”/”.”) ” Civil”/ “Cód”(“igo”/”.”) “ Florestal” / “Lei d”(“e”/”a”) “ Introdução”/ “Cód”(“igo”/”.”) “ de Processo Civil”/ “Cód”(“igo”/”.”) “ Comercial”	; apelidos das normas registrados (somente os mais utilizados).
8.	ss0	→ “Constituição” [“ Federal”]	; classe “ss”: substantivo no singular
9.	ss1	→ Apelido / ”Decreto”[-]”Lei” / ”Lei” [-]”Delegada” / ”Lei”/ ”Decreto Legislativo” / ”Decreto” [ag1] / ”Resolução” [[d(a/o)] ag0]/ ”Instrução” [”Normativa”] ([“ “/ ”/” /(“ d“(“a“/“o“) “ “)]ag0)/ ”Portaria” [”Normativa”]([“ “/ ”/” /(“ d“(“a“/“o“))]ag0)/Ordem de Serviço([“ “/ ”/” /(“ d“(“a“/“o“)“ “)]ag0) / Medida Provisória	
10.	ss2	→ “Cap“(“ítulo“/“.“) / “Seção“ / “Título“ / “Livro“ / “Parte“ / “Disposições“	
11.	ss3	→ ”Artigo” [“ único”]/”Art.” [“ único”]	
12.	ss4	→ “Parágrafo”/ ”Par.”/ ”§ ”/ ”Item”/ ”Inciso”	
13.	ss5	→ “Alínea“ / “Letra“	
14.	sp1	→ “Leis”/”Decretos”/”Portarias”	; classe “sp”: substantivo no plural
15.	sp2	→ “Cap”(“ítulos”/”s.”)	
16.	sp3	→ “Art”(“igos”/”s.”)	; art. ou artigos
17.	sp4	→ “Parágrafos“ / “§§“ / “Itens“ / “Incisos“	
18.	sp5	→ “Alíneas” / “Letras”	
19.	sx	→ “Anexo”	
20.	citação	→ “Presente “ ss1	; início das regras de citação
21.	citação	→ [ss4 “ “ na/nr[“,”][“ d(a/o)”]] ss3 “ “ na [,]“ “ (“d“/“n“)(“e“/“a“/“o“) ss1 (“ “/”/”) na ”, de “ na “ de “ dm “ de “ na	
22.	citação	→ [ss4 “ “ (na/nr)[,][“ d“ (“a“/“o“)]] ss3 “ “ na [,] “ “ (“d“/“n“)(“e“/“a“/“o“) “ “ ss1(“ “ / ”/”) na “, de “ dm de na	
23.	citação	→ [ss4 “ “ (na/nr)[,][“ d“ (“a“/“o“)]] ss3 na [,]“ “ (“d“/“n“)(“e“/“a“/“o“) “ “ ss1(“ “/ ”/”) na “, de “ na	
24.	citação	→ [ss4 “ “ (na/nr)[,][“ d“ (“a“/“o“)]] ss3 na [,] (“d“/“n“)(“e“/“a“/“o“) ss1(/”/”) na	
25.	citação	→ [ss4 “ “ (na/nr)[,][“ d“ (“a“/“o“) “ ”]] ss3 na [,] (“d“/“n“)(“e“/“a“/“o“) “ ” (ss0/ss1)	
26.	citação	→ ss4 “ “ (na/nr) [,] [“ d“ (“a“/“o“) “ “] ss3 “ “ na	
27.	citação	→ ss4 “ “ (na/nr) [“ deste “ ss3]	
28.	citação	→ ss4 [,] “ d“ (“a“/“o“) “ “ ss3 “ “ na [,] (“d“/“n“)(“a“/“o“) “ “ ss1(/”/”) na “, de “ na “ de “ dm “ de “ na	
29.	citação	→ ss4 [,] “ d“ (“a“/“o“) “ “ ss3 “ “ na [,] “ “ (“d“/“n“)(“a“/“o“) ss1(“ “/ ”/”) na “, de “ dm “ de “ na	
30.	citação	→ ss4[,] “ d“(“a“/“o“) “ “ ss3 “ “ na[,] “ “ (“d“/“n“)(“a“/“o“) ss1(“ “ / ”/”) na [“, de na“]	
31.	citação	→ ss4[,] “ d“ (“a“/“o“) “ “ ss3 “ “ na[,] (“d“/“n“)(“a“/“o“) “ “ ss0/ss1	
32.	citação	→ ss4 “ ” (na/nr) [,] “ d“ (“a“/“o“) “ “ ss3 “ “ na	
33.	citação	→ ss4 “ “ (na/nr) [,] [“ d“ (“a“/“o“)] “ “ ss3 “ anterior“	
34.	citação	→ ss4 “ “ (na/nr) [“ deste “ss3]	
35.	citação	→ [ss4 na“, d“ (“a“/“o“)“ “] ss3 “ “ na [,] “ “ (n/d)est(e/a) “ “ ss1[(“ ”/ ”/”) na [,] “ de “ na “ de “ dm “ de “ na]	
36.	citação	→ [ss4 “ “ (na/nr) [,] “ d“ (“a“/“o“)] ss3 “ “ na [,] “ “ (n/d)est(e/a) “ “ ss1(“ “	

REGRAS DE PRODUÇÃO			COMENTÁRIOS
		na [,]“ de “ na “ de “ dm “ de “ na)	
37.	citação	→ ss1(“ “ / ”/”) na[,][“ de “ na]“ de “ dm “ de “ na [“ e seu “ sx “ “ nr]	
38.	citação	→ ss1(“ “ / ”/”) na[[,][“ de “dm] “ de “ na]	
39.	citação	→ sp3 “ “ na *(“, “na) “ e “ na[,]“ d“ (“a“/“o“) “ “ ss1(“ “/”/”) na [[,][de na] de dm de na]	
40.	citação	→ sp3 “ “ na*(, na) e na[,] d(a/o) ss1(/”/”)na	
41.	citação	→ sp3 “ “ na *(, na) e na[,] (“d“/“n“)“est“(“e“/“a“) “ “ ss1	
42.	citação	→ ss5 “ “ lt“, “ [ss4 “ “] nr	
43.	citação	→ ss3 “ “ na, sp5 lt *(, lt) e lt d[oa] ss1 na[, de na de dm de na]	
44.	citação	→ sp5 “ “ lt*(, lt) e lt do ss3 na da ss1(/”/”)na[, de na de dm de na]	
45.	citação	→ sp5 “ “ lt “ a “ lt “ do “ ss3 na	
46.	citação	→ sp5 “ “ lt “ a “ lt “ do “ ss4 “ anterior“	
47.	citação	→ sp5 “ “ lt *(“, “lt) “ e “ lt “ do “ss4 “ anterior“	
48.	citação	→ ss3 “ “ na[,] “ do mesmo “ ss1	
49.	citação	→ ss3 “ “ na[,] “ “ ss5 “ “ lt “ d“(“a“/“o“) “ “ (ss0/ss1)[(“ “/ ”/”)na]	
50.	citação	→ ss3 “ “ na “ e “ sp4	
51.	citação	→ ss3 “ “ na, “ “ nr	
52.	citação	→ ss3 “ “ na “, “ ss4 “ “ nr “, d“ (“a“/“o“) ss1[(“ “ / ”/”) na [“, de “ na “ de “ dm “ de “ na]]	
53.	citação	→ ss3 “ “ na “,“ ss4 “ “ (na/nr) [“, d“ (“a“/“o“) “ “ (ss0/ss1)]	
54.	citação	→ ss3 na[“;“/“,“]ss4 único[,](ss5 lt,)? d(a/o) ss1(/”/”)na, de na de dm de na	
55.	citação	→ ss3 “ “ na[“;“/“,“] “ “ ss4 “ único [[,][ss5 lt,] “d“ (“a“/“o“) “ “ ss1(“ “/ ”/”) na]	
56.	citação	→ ss4 “ anterior, “ sp5 lt “ a “ lt	
57.	citação	→ (ss3/ss4) “ anterior“	
58.	citação	→ ss4 “ “ na[,]“ do mesmo “ ss3	
59.	citação	→ sp4 “ “ na(, na)* “ e “ na “ do “ ss3 “ “ na (“n“/“d“)“est“(“a/e“) “ “ ss1	
60.	citação	→ sp4 “ “ na(, na)* “ e “ na “ do “ ss3 “ “ na “ d“ (“o“/“a“) “ “ ss[01] [(“ ”/ ”/”)na[“, de “ na “ de “ dm “ de “ na]]	
61.	citação	→ ss3 “ “ na[,] “ “ sp4 “ “ (na/nr) *(“,“(na/nr)) “ e “ (na/nr) [[,] “ da “ ss0]	
62.	citação	→ sp4 “ “ n(a/r) *(“, “n(a/r))* “ e “ (na/nr) “dest“(“e“/“a“) “ “ (ss3/ss4)	
63.	citação	→ sp4 “ “ (na/nr) *(, na/nr) “ e “ (na/nr) d(a/o)s? ss2[do ss1(na [, de na de dm de na])]	
64.	citação	→ sp4 “ “ (na/nr) *(“, “ na/nr) “ e “ (na/nr)[“ do “ (ss3/ss4) “ na “ [“d“ (“o“/“a“) “ “ (ss0/ss1)]]	
65.	citação	→ sp4 “ “ nr “ a “ nr “dest“ (“e“/“a“) (ss3/ss4)	
66.	citação	→ sp4 “ “ nr “ a “ nr [“ do “ (ss3/ss4) na [d[oa] (ss0/ss1)]]	
67.	citação	→ sp4 “ anteriores“	
68.	citação	→ ss2/ss3 “ “ na	
69.	citação	→ ss2 “ “ (na/nr) “ “ (“n“/“d“)“est“ (“e“/“a“) “ “ (ss1/ss2)	
70.	citação	→ [\“][Cc]aput[\“]“ deste “ ss3	
71.	citação	→ [\“][Cc]aput[\“][“ do “ ss3 “ “ na]	
72.	citação	→ sx “ “ nr[“ d“ (“a“/“o“) “ “ ss1(“ “/ ”/”) na [“, de “na]]	
73.	citação	→ sx “ a este “ ss1	
74.	citação	→ [“n“/“d“](“e“/“E“)“st“ (“e“/“a“) (ss1/ss2/ss3/ss4)	
75.	citação	→ “referid”(“o“/“a“) “ “ (ss1/ss2/ss3/ss4)	
76.	citação	→ sp1 “ “ 1*(na[,][“ de “ na “ de “ dm “ de “ na])[,] “ e “ na[“, de “ na “ de “ dm “ de “ na]	
77.	citação	→ sp1 “ “ na[[,] “ de “ na “ de “ dm “ de “ na][,] e na[, de na de dm de na]	

Tabela 7 Regras de Produção para as Citações em *NEJBs*

Os casos analisados para a composição das regras de produção para citações foram escolhidos aleatoriamente dentre as ocorrências de citações nos textos das normas. Não é intenção propor um conjunto fechado para tais regras. A composição das citações ainda assume uma forma livre e o número de casos de citações é muito grande. O propósito é que este conjunto de regras se mantenha aberto para a inserção de novos casos conforme estes são reconhecidos.

4.6 Modelo de Dados para *NEJBs*

O modelo de dados para *NEJBs* foi definido para trabalhar em conjunto com a definição formal apresentada na seção anterior. A definição formal especifica como identificar segmentos de uma norma e o modelo de dados é o responsável pelo armazenamento dos segmentos de modo a ser compatível com a definição formal. O modelo de Dados para *NEJBs* foi especificado para atingir os seguintes objetivos:

- Servir de modelo para as aplicações envolvendo *NEJBs*;
- Servir como suporte para a aplicação da definição formal especificada na seção anterior, fornecendo a modelagem para armazenar os segmentos identificados;
- Criar um modelo de alto-nível estabelecendo as entidades relevantes para o conjunto de *NEJBs* e determinando os principais relacionamentos entre elas;
- Permitir a uniformização do armazenamento de normas pertencentes ao conjunto *NEJBs*, facilitando o intercâmbio de dados;
- Permitir a recuperação de elementos citados ou de elementos específicos;

4.6.1 Modelo Entidade Relacionamento Estendido

Para a representação do modelo de dados para *NEJBs* foi utilizado o modelo Entidade-Relacionamento Estendido[34] – mE-Rx [14], largamente aplicado em SGBDs. O mE-Rx é um modelo conceitual que representa as informações em alto nível sem a preocupação de implementação. Possui uma simbologia particular onde as principais construções são:

CONSTRUÇÕES	DEFINIÇÃO	REPRESENTAÇÃO
ENTIDADE	"É uma coisa do mundo real com existência independente" [Elmasri, p. 40].	
ATRIBUTOS	São propriedades das entidades ou dos relacionamentos. Podem ser mapeados como: 1. atributos, 2. atributos-chave; ou 3. multivalorados, Apresentam uma variação na representação, conforme exibido ao lado.	1. 2. 3.
RELACIONAMENTOS	Relacionamentos são criados para mapear as diversas interações entre entidades. Com relação ao grau: grau 2 são binários, grau 3 ternários e assim por diante. Sua cardinalidade pode ser: 1:1, 1:n ou m:n o que determina relacionamentos de 1 para 1, 1 para muitos e muitos para muitos respectivamente.	

Tabela 8 Construções utilizadas no modelo Entidade Relacionamento Estendido[35]

4.6.2 Esquema Conceitual do Modelo de Dados para *NEJBs*

O modelo de dados para *NEJBs* apresentado a seguir na **Figura 8**, está centralizado na entidade Norma Estatutária Jurídica, ou abreviando *NEJ*. Esta entidade representa de modo generalizado uma norma completa e de modo especializado um segmento de norma.

[34] Do original *EER Model – Enhanced Entity-Relationship Model* [Elmasri].
[35] Extraído de [Elmasri] p. 57.

De acordo com o modelo, a *NEJB* é composta da união de 2 outras entidades: cabeçalho e corpo, como mostra em detalhe, a **Figura 7**. Na prática, uma norma será composta de somente um cabeçalho e diversos outros elementos que irão compor o corpo.

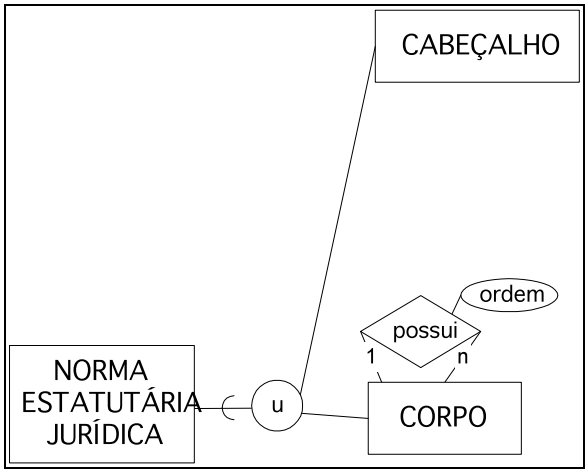


Figura 7 Composição de uma *NEJB*

A estruturação hierárquica do corpo de uma norma está representada no relacionamento possui que através de seu atributo ordem determina a ordem entre os diversos segmentos que possuem o mesmo pai.

Tendo o enfoque na generalização, a *NEJ* possui uma classificação em diversos tipos de normas. A entidade Outras foi representada para permitir a expansão futura no caso de surgirem outros tipos de normas.

Partindo para o enfoque na especialização, a entidade corpo, seguindo a definição formal, é classificada em seção, componente ou anexo, cada um destes representados nas entidades respectivas.

No modelo de dados também está previsto o controle sobre a emissão das normas pelos órgãos formuladores de normas através da entidade Órgão Formulador e do relacionamento emite.

Várias entidades não foram detalhadas propositadamente para não impor restrições neste modelo. O modelo de dados deve ser detalhado de acordo com cada aplicação e então mapeado para tabelas de dados de acordo com o SGBD escolhido.

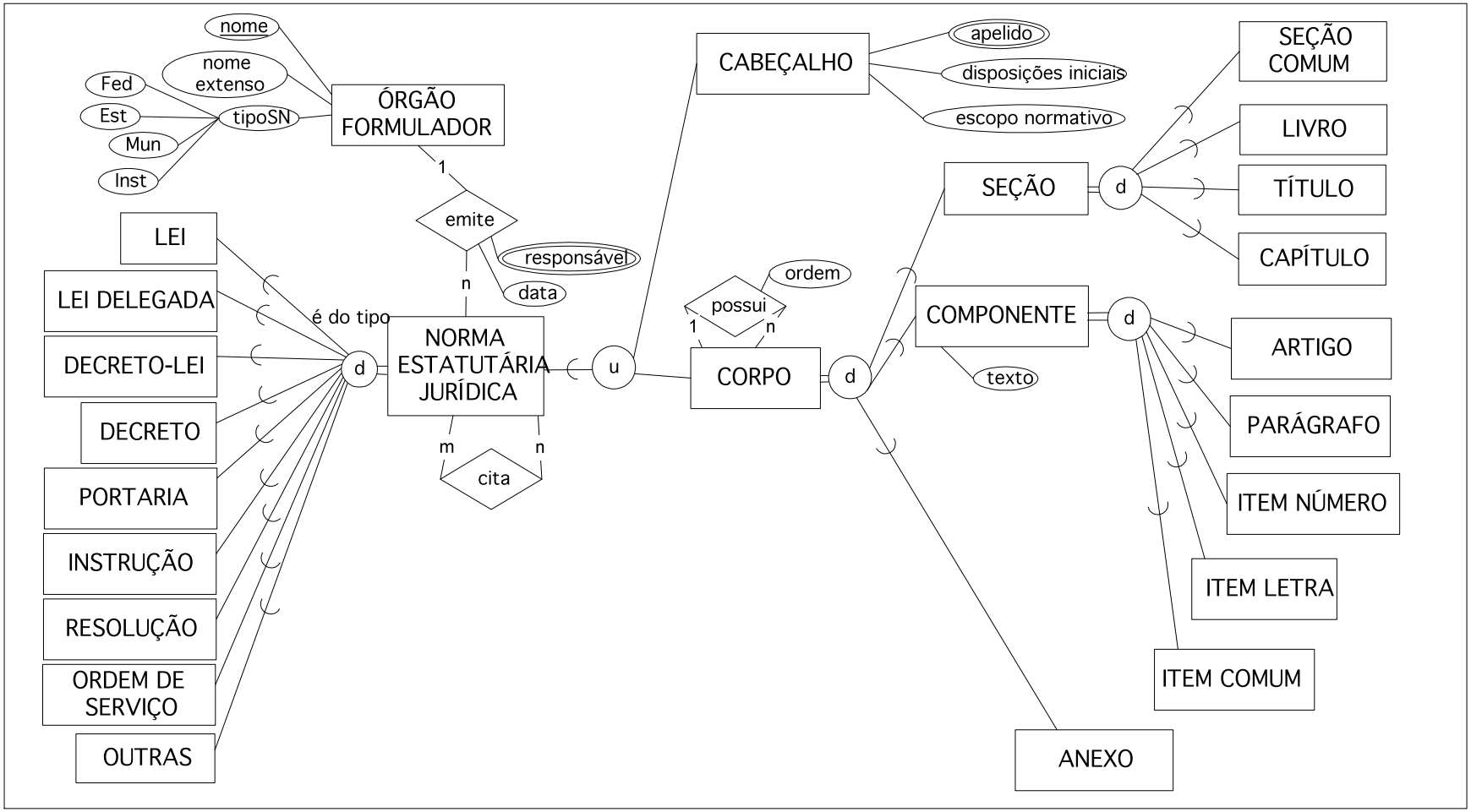


Figura 8 Modelo de Dados para *NEJBs*

4.6.3 Citações no Modelo de Dados para *NEJBs*

Neste modelo, as citações estão representadas como relacionamentos entre os segmentos de uma norma. Inicialmente definido como um relacionamento muitos para muitos (m:n), este foi detalhado na **Figura 9** para refletir a seguinte situação:

- Um segmento de norma comumente contém uma ou mais citações compostas. Alguns exemplos de citação composta encontram-se na **Tabela 9** abaixo:

CITAÇÃO COMPOSTA	CITAÇÕES UNÍVOCAS
"nos casos **dos arts. 106 e 107"**	1. Lei nº 3017, Artigo 106 2. Lei nº 3017, Artigo 107
"no caso do **art. 219, I, II e III**"	1. Lei nº 3.017, Artigo 219, item I 2. Lei nº 3.017, Artigo 219, item II 3. Lei nº 3.017, Artigo 219, item III
"disposto nos **arts. 148 a 150**"	1. Lei nº 5.869, Artigo 148 2. Lei nº 5.869, Artigo 149 3. Lei nº 5.869, Artigo 150

Tabela 9 Exemplos de citações compostas

- As citações compostas devem ser separadas em várias citações unívocas referenciando um único segmento de uma norma, para que os relacionamentos possam ser armazenados em um banco de dados.

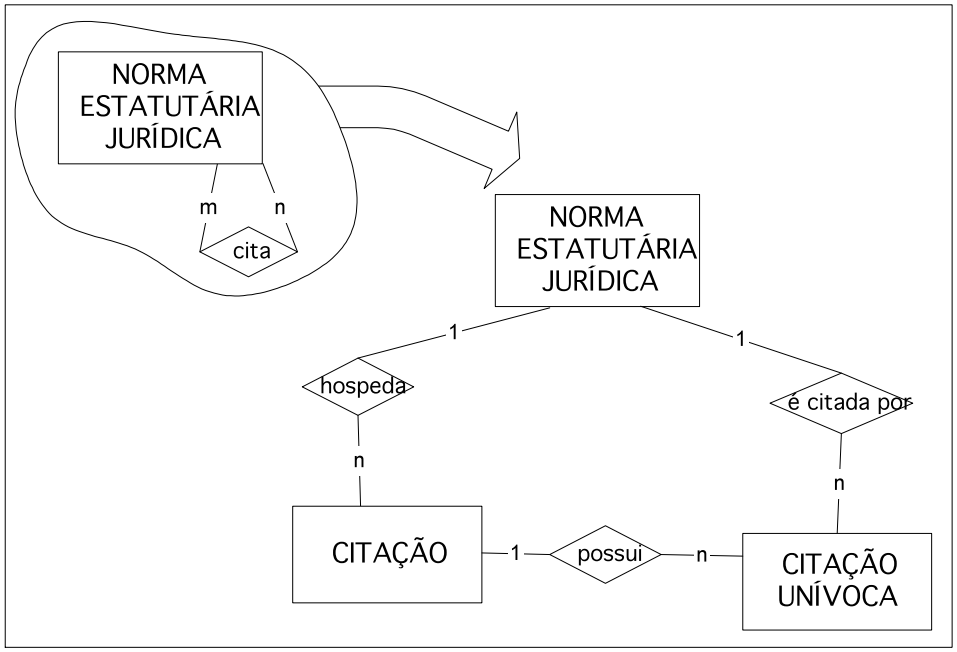


Figura 9 Esquema conceitual para a representação de citações

Ao separar citações compostas em diversas citações unívocas pode-se identificar todos os segmentos que uma dada citação referencia, estabelecendo as relações entre estes no modelo de dados. O relacionamento <u>é citada por</u> permite, por exemplo, dado um segmento de norma, que se recupere todos os segmentos que fazem referência a ele. Este é um relacionamento importante no conjunto de normas principalmente para poder levantar normas que foram modificadas ou revogadas por normas mais novas.

5 A FERRAMENTA WEBIFYLAW

A ferramenta *WebifyLaw* implementa o processo de conversão proposto no **Capítulo 3**, aplicado ao conjunto de *NEJBs* definido no **Capítulo 4**. Do conjunto de *NEJBs* foram selecionadas 45 normas para teste dentre as quais encontram-se a Constituição Federal de 1988, o Código Civil e o Código de Processo Civil.

A ferramenta *WebifyLaw*, como resultado, gera automaticamente hiperdocumentos codificados na linguagem HTML para serem disponibilizados facilmente na WWW da Internet. Os hiperdocumentos são gerados a partir de documentos no padrão STDNORM definido na **Seção 3.3**. Adicionalmente, a ferramenta *WebifyLaw* automaticamente produz dois índices para os hiperdocumentos gerados: um índice por categoria e um índice cronológico.

Para o armazenamento de dados, a ferramenta *WebifyLaw* utiliza o modelo de dados proposto na **Seção 4.6** e simula as tabelas em arquivos do tipo texto.

5.1 Implementação

A ferramenta *WebifyLaw* foi implementada para comprovar que a *conversão automática* pode ser efetivamente executada através do *processo de conversão* proposto neste trabalho. Sua implementação não prevê uma interface gráfica com o usuário. A ferramenta foi implementada em módulos com várias funções que devem ser executadas diretamente a partir da linha de comando.

Opções de configuração são passadas para a ferramenta *WebifyLaw* através da inclusão de um arquivo na linha de comando. A implementação prevê o uso de um arquivo de configuração padrão, chamado default.cfg, e neste caso dispensa a inclusão na linha de comando. O arquivo de configuração determina o diretório raiz e vários diretórios auxiliares onde a ferramenta *WebifyLaw* procura por informação ou armazena os resultados do processamento. A **Tabela 10** abaixo mostra as etiquetas padrão a serem utilizadas no arquivo de configuração.

DEFINIÇÃO	ETIQUETA	DIRETÓRIO[36]
Diretório raiz, a partir do qual são criados os outros diretórios	DirRaiz	d:\WebifyLaw\
Diretório para os hiperdocumentos gerados automaticamente.	DirHTxt	hipertxt\
Diretório para as tabelas de dados	DirBD	bdados\
Diretório de trabalho, para armazenamento de arquivos temporários, se assim for necessário.	DirTrab	system\
Diretório para outras opções de configuração	DirCfg	config\
Diretório para os componentes externos às normas.	DirCExt	compext\
Diretório para a localização dos índices gerados automaticamente.	DirIndex	index\

[36] Os diretórios aqui apresentados são somente ilustrativos.

DEFINIÇÃO	ETIQUETA	DIRETÓRIO[36]
Diretório para as mensagens geradas durante a execução dos módulos.	DirMsg	log\

Tabela 10 Opções de configuração

O arquivo de configuração deve ser escrito mantendo-se uma etiqueta em cada linha e utilizando o caracter dois pontos – ":", para a separação entre etiqueta e valor.

A ferramenta *WebifyLaw*, por possuir características de processamento em *batch mode*, gera um arquivo com mensagens das ações executadas durante o processamento. Este arquivo possui a extensão .log e é nomeado de acordo com a data da execução da ferramenta.

5.1.1 Linguagem de Programação

Para a implementação da ferramenta *WebifyLaw* foi escolhida a linguagem de programação Perl [43] pelas razões mencionadas abaixo:

- Facilidade de manipulação de textos, permite analisar e converter grande volume de informação rapidamente.
- Suporte ao uso de expressões regulares.
- Portabilidade.
- Facilidade de criação de programas para a Web.
- Distribuição gratuita através de licença artística.

A ferramenta foi desenvolvida utilizando o compilador/interpretador Perl para Windows 95® na sua versão 5.004_02. O código desenvolvido totalizou 2.674 linhas sem comentários.

5.1.2 Módulos e Bibliotecas

A ferramenta *WebifyLaw* é composta por três módulos: *webifySegm*, *webifyLink* e *webifyIndex* que implementam, respectivamente, a segmentação e hierarquização, a inferência de *links* e a geração de hiperdocumentos e a geração de índices.

As funções estão divididas entre seis bibliotecas diferentes:

- **libConfig**: Biblioteca de configuração, inclui entre suas funções a leitura do arquivo de configuração e mecanismos para determinar os caminhos para a localização dos arquivos de entrada e diretórios de saída. Faz a configuração de variáveis globais utilizadas por todos os módulos.
- **libStdIO**: Biblioteca de funções de entrada e saída. Possui funções implementando diversos mecanismos de importar e exportar a informação. Em particular, implementa a função de leitura do padrão STDNORM, os filtros *filtroTXT* e *filtroETEXT*, e as funções que simulam importação e exportação de uma base de dados.
- **libNormas**: Implementa funções para a segmentação de normas e identificação e análise de tipos de normas e segmentos. Inclui também funções para a hierarquização de segmentos.
- **libLink**: Biblioteca de inferência de *links*. Implementa funções para a extração, análise e estabelecimento de *links* entre os segmentos. Incorpora as funções para o *parse* de *tokens* e análise léxica da citação, além das funções para endereçamento na forma normal de endereçamento de segmentos.
- **libHTML**: Biblioteca de geração de hiperdocumentos. Implementa funções para a geração de código em linguagem HTML na criação de hiperdocumentos. As

funções nesta biblioteca geram automaticamente páginas HTML com codificação diferente de acordo com o tipo do segmento.

- **libIndex**: Biblioteca da criação de índices. Contém funções para a geração de índices por categoria e cronológico em páginas codificadas em HTML.

5.2 A Geração de Hiperdocumentos

Como resultados do processo de conversão, a ferramenta *WebifyLaw* gera automaticamente a informação para o banco de dados e hiperdocumentos a partir dos dados dos segmentos. Os dados das 45 normas selecionadas para teste resultaram em um total de 10.660 segmentos.

Nesta implementação, cada hiperdocumento gerado contém o texto completo de uma norma. Assim, cada norma é armazenada em um único arquivo HTML. Esta é uma opção particular desta implementação. Outra opção seria gerar um hiperdocumento para cada seção de uma norma, por exemplo.

Para cada norma armazenada no banco de dados, a geração de hiperdocumentos realiza consultas para recuperar todos os segmentos que compõem aquela norma. Os segmentos são, então, codificados. O hiperdocumento resultante é armazenado em um arquivo HTML de nomenclatura controlada.

A etiqueta de nome do arquivo é resultante da combinação do tipo da norma e do seu identificador. O tipo da norma é representado por no máximo duas letras de acordo com a **Tabela 11** a seguir.

SIGLA	TIPO DA NORMA
CF	Constituição Federal
D	Decreto
DL	Decreto-Lei
L	Lei
LD	Lei Delegada
LM	Lei Municipal
R	Resolução
RN	Resolução Normativa
P	Portaria
PN	Portaria Normativa
IN	Instrução Normativa

Tabela 11 Siglas para os tipos de normas

O número identificador da norma segue a sigla sem a inclusão de espaços. Exemplos de algumas etiquetas geradas são: D1207.html, DL719.html, L3017.html, D1752.html.

No caso de a norma possuir uma instituição associada, a etiqueta compõe-se, adicionalmente, da sigla da instituição e de dois dígitos identificando o ano da norma, no seguinte formato:

*siglaTipoDaNorma_siglaInstituição*NúmeroIdentificador-Ano

Exemplos de etiquetas para este caso são: R_CONAMA10-94.html, P_IBAMA29-96.html, P_IBDF17-88.html.

Os hiperdocumentos foram codificados de modo coerente com seu significado estrutural, utilizando os recursos da linguagem HTML para diferenciar os diversos tipos de segmentos. O **Quadro 3** a seguir mostra um exemplo de um hiperdocumento gerado automaticamente pela ferramenta *WebifyLaw*.

DECRETO Nº 1.207, DE 1º DE AGOSTO DE 1994.

Acrescenta inciso ao artigo 14 do Decreto nº 96.036, de 12 de maio de 1988.

O PRESIDENTE DA REPÚBLICA, no uso da atribuição que lhe confere o artigo 84, inciso IV, da Constituição, e tendo em vista o disposto na Lei nº 7.646, de 18 de dezembro de 1987,
DECRETA:

Artigo 1º - Fica acrescido ao artigo 14 do Decreto nº 96.036, de 12 de maio de 1988, o inciso IV, com a seguinte redação:

"Art. 14. .....
.....
IV - reconhecíveis como exclusiva ou principalmente destinados à utilização em unidades digitais de processamento da Posição 8471.91.0100, da Nomenclatura Brasileira de Mercadorias - NBM (microcomputadores e estações de trabalho)."

Artigo 2º - Este Decreto entra em vigor na data de sua publicação.

ITAMAR FRANCO
José Israel Vargas

Informações geradas automaticamente em *2 de Outubro de 1997* e mantidas por:
Alessandra Dorante

Quadro 3 Exemplo de hiperdocumento gerado automaticamente

Nos hiperdocumentos gerados pela ferramenta *WebifyLaw* foi mantido um único estilo de formatação coerente para cada tipo de segmento. Esta característica facilita a identificação visual dos tipos de segmentos. Exemplificando, os cabeçalhos de seções utilizam as construções próprias para *headings* da linguagem HTML (<H1>..</H1>, <H2>…</H2>, e assim sucessivamente), as etiquetas dos *artigos* são sempre grafadas utilizando o estilo *strong* da linguagem HTML (…). Segmentos como *escopo normativo*, *disposições iniciais* e *finais* são sempre grafados utilizando o estilo enfático (…).

Nos hiperdocumentos que possuem seções internas, a ferramenta *WebifyLaw* inclui um índice interno para facilitar a navegação, como mostra o **Quadro 4**. Para possibilitar a referenciação pelas citações cada segmento do documento tem seu endereço gerado na forma normal de endereçamento de segmentos apresentada na **Seção 3.6.2**. O resultado do endereçamento é utilizado na geração das etiquetas para âncoras em linguagem HTML – etiquetas ….

PORTARIA IBAMA N° 29, DE 26 DE ABRIL DE 1996

Índice

Capítulo I - DA REPOSIÇÃO FLORESTAL OBRIGATÓRIA , DO PLANO INTEGRADO FLORESTAL E DA ASSOCIAÇÃO FLORESTAL

Seção I - DA REPOSIÇÃO FLORESTAL
Seção II - DO PLANO INTEGRADO FLORESTAL - PIF
Seção III - DA ASSOCIAÇÃO FLORESTAL

Capítulo II - DAS VISTORIAS E LAUDOS

Capítulo III - DAS PENALIDADES

Capítulo IV - DAS DISPOSIÇÕES GERAIS E TRANSITÓRIAS

…

Quadro 4 Exemplo de um trecho de hiperdocumento com índice interno

O **Quadro 5** mostra o código fonte em linguagem HTML com endereçamento de segmentos para o documento apresentado no **Quadro 3**.

```
<html><head><title>DECRETO Nº 1.207, DE 1º DE AGOSTO DE 1994.</title></head><body>
<h1>DECRETO Nº 1.207, DE 1º DE AGOSTO DE 1994.</h1>
<hr>
<p><em><a name="DI1">
Acrescenta inciso ao <a href="msg.html">artigo 14 do Decreto nº 96.036, de 12 de maio de 1988</a>.
</a></em>
<p><strong><a name="EN2">
O PRESIDENTE DA REPÚBLICA, no uso da atribuição que lhe confere o <a href="CF.html#T4C2S2A84IR6">artigo 84, incisos IV, da Constituição</a>, e tendo em vista o disposto na <a href="msg.html">Lei nº 7.646, de 18 de dezembro de 1987</a>,</a></strong>
<p><strong><a name="EN3">
DECRETA:
</a></strong>
<p><strong><a name="A1">Artigo 1º</a></strong>
 - Fica acrescido ao <a href="msg.html">artigo 14 do Decreto nº 96.036, de 12 de maio de 1988</a>, o <a href="#A1IR4">inciso IV</a>, com a seguinte redação:
<br><em><a name="A1CM1">
"Art. 14. ..................................................................................
..................................................................................
</a></em>
<br><em><a name="A1CM2">
IV - reconhecíveis como exclusiva ou principalmente destinados à utilização em unidades digitais de processamento da Posição 8471.91.0100, da Nomenclatura Brasileira de Mercadorias - NBM (microcomputadores e estações de trabalho)."
</a></em>
<p><strong><a name="A2">Artigo 2º</a></strong>
 <a href="D1207.html">Este Decreto</a> entra em vigor na data de sua publicação.
<p><em><a name="A2DF1">ITAMAR FRANCO</a></em>
<p><em><a name="A2DF2">José Israel Vargas</a></em>
<hr>
<font size=-1>Informações geradas automaticamente em 26 de Outubro de 1997 e mantidas por:
<blockquote><author><a href="mailto:alessand@www.ifqsc.sc.usp.br">Alessandra Dorante</a></author>
</blockquote>
</font>
</body></html>
```

Quadro 5 Código fonte gerado automaticamente em linguagem HTML

5.3 Uso da Ferramenta *WebifyLaw*

O processo de conversão automática é realizado com o uso da ferramenta *WebifyLaw* seguindo os passos abaixo:

1. Para cada documento no padrão STDNORM executar o módulo webifySegm, utilizando a seguinte linha de comando:

```
$> perl webifySegm.pl doc1STDNORM default.cfg
```

2. Executar a inferência de *links* sobre os dados processados no passo anterior e a geração dos hiperdocumentos, com a seguinte linha de comando:

```
$> perl webifyLink.pl default.cfg
```

3. Gerar os índices:

```
$> perl webifyIndex.pl default.cfg
```

Após a execução destes três passos que fazem a conversão automática é necessário fazer a disponibilização dos resultados. Esta pode ser feita simplesmente através do mecanismo de cópia. O diretório escolhido como diretório raiz do sistema deve ser copiado para um *web site*. Os *links* foram gerados com endereçamento relativo para evitar a invalidação com a eventual mudança da localização física das páginas.

5.4 Resultados

Os resultados da aplicação do processo de conversão automática, utilizando a ferramenta *WebifyLaw*, estão organizados em três **Seções**: **5.4.1**, **5.4.2** e **5.4.3**, que respectivamente, analisam os resultados da extração das citações, da geração de *links* e apresentam os resultados da conversão em um *site* na WWW.

Finalmente, a **Seção 6.2** aborda sugestões para trabalhos futuros.

5.4.1 Resultados da Extração das Citações

O resultado da extração das citações é muito importante para o processo de inferência de *links*. Somente citações extraídas são passíveis de serem convertidas em *links*. Para avaliar o resultado da extração das citações é necessário ter uma fonte de comparação. Para a ferramenta *WebifyLaw* a avaliação da extração de *links* foi feita através da comparação entre citações marcadas manualmente e citações extraídas automaticamente. No conjunto de 45 normas foram processados 1.565 segmentos marcados. O resultado da extração está na **Tabela 12**:

Segmentos com citações corretas:	Iguais à marcação manual	1.295	82.5%
Segmentos com citações extras:	Não marcadas manualmente	87	5.5%
Segmentos com citações não coincidentes:	Diferentes da marcação manual	183	12%
		1.565	

Tabela 12 Resultados da extração das citações por segmento[37]

Nos 183 segmentos em que a marcação não coincidiu com a marcação, existem diferentes casos, que são agrupados em:

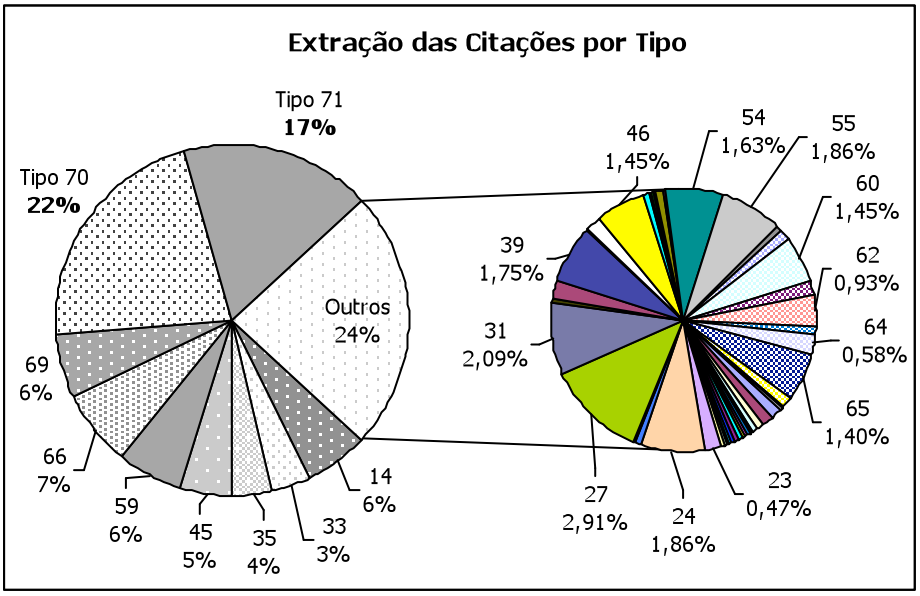
Citação não reconhecida	Correspondem a novos casos a serem incorporados na definição formal de citações e interpretados pela ferramenta *WebifyLaw*.
Uso de apelido de norma	A citação utiliza o apelido para fazer referência à norma. O uso de apelidos não é suportado pela ferramenta *WebifyLaw* nesta versão.
Erro na grafia sintática	Erros de grafia sintática na citação, como por exemplo o uso trocado de plural e singular ou grafia incorreta de palavras que desempenham função chave no reconhecimento. A ocorrência desses casos pode ser facilmente eliminada pelo uso de corretores ortográficos.
Marcação manual não coincidente	A marcação manual não foi executada corretamente: ou incorpora elementos que não fazem parte da citação, ou não abrange a citação completa. Em ambos casos provoca erro na comparação.

Tabela 13 Casos de extração de citação não coincidentes

Nesses 1.565 segmentos foram marcadas manualmente 1.428 citações, e foram reconhecidas automaticamente 1.703 citações, numa diferença de 275 citações "descobertas" pela ferramenta *WebifyLaw*.

Para a extração das citações foram analisados 75 tipos diferentes de citações, baseados na definição formal das citações para as *NEJBs*, apresentada na **Seção 4.5**. Através da análise apresentada no **Quadro 6**, fica claro que a maioria das citações varia pouco com relação ao seu tipo, concentrando-se nos dois tipos básicos seguintes:

> ***Tipo 70***: citações são do tipo "*artigo 1º*", "*artigo 186*", que seguem a forma geral de *artigo* seguido de *identificador*.
>
> ***Tipo 71***: reconhece citações do tipo: "*neste decreto*", "*desta Constituição*", "*este artigo*", "*nesta seção*", "*deste parágrafo*".



Quadro 6 Estatísticas da extração de citações por tipo de citação

[37] Considere-se que cada segmento pode conter mais de uma citação.

5.4.2 Resultados da Geração de *Links*

A geração automática de *links* baseia-se nas marcações feitas pela extração das citações. Analisa o conteúdo das citações entre as marcações <cit>…</cit>. Como resultado desta análise gera o endereço do segmento citado na forma normal de endereçamento de segmentos apresentada na **Seção 3.6.2**.

Antes de gerar as âncoras em linguagem HTML referenciando o segmento, este é procurado entre os segmentos presentes na base de dados. Se o segmento é encontrado a etiqueta HTML é estabelecida fazendo referência ao endereço do segmento. Caso contrário a âncora é estabelecida direcionada para um arquivo padrão contendo uma mensagem.

No caso da existência de citações compostas, o *link* foi gerado para somente um segmento, que é o segmento de menor nível na hierarquia, conforme mostram os exemplos de caso a seguir.

Os resultados da geração dos *links* são analisados a partir dos casos encontrados no conjunto de *NEJBs* selecionados. Os exemplos de casos abaixo procuram relatar a variedade de casos e sua complexidade. Para cada caso é sempre apresentado primeiro o texto original conforme figurado na norma, seguido do trecho gerado automaticamente pela ferramenta *WebifyLaw*, em linguagem HTML.

Caso 1 – Citação com Referências Explícitas

Neste caso, todos os segmentos da citação possuem menção explícita ao seu tipo e identificação. É o caso que apresentou maior volume de citações. Os exemplos a seguir mostram algumas das variações possíveis dentro desse caso.

" V. improbidade administrativa, nos termos do art. 37, § 4º."[38]

Neste exemplo a citação não menciona outra norma. A citação refere-se a um segmento que pertence à mesma norma do segmento hospedeiro. Nestes casos, a citação é chamada "interna". Apesar de ser uma citação interna, para identificar o segmento citado é necessário "resolver o endereço". Isto deve ser feito para todos os tipos de citação. A partir do "endereço abreviado" mencionado na citação é gerado o endereço completo do segmento citado. A etiqueta href abaixo mostra o endereço completo.

" improbidade administrativa, nos termos do ****art. 37, § 4º****."

"Artigo 270 - Este Código regula o processo de conhecimento (Livro I), de execução (Livro II), cautelar (Livro III) e os procedimentos especiais (Livro IV)."[39]

Este exemplo é similar à citação anterior, com a diferença que este trecho apresenta diversas citações referenciando segmentos classificados como seções. No caso em particular fazem referência à seção do tipo livro. Casos similares incluem referências a títulos e capítulos.

"Artigo 270
 - Este Código regula o processo de conhecimento (****Livro I****), de execução (****Livro II****), cautelar (****Livro III****) e os procedimentos especiais (****Livro IV****)."

"Artigo 1º - O inciso VI do art. 2º do Decreto nº 1.791, de 15 de janeiro de 1996, passa a ..."[40]

Uma variação da citação com referências explícitas mencionando uma norma "externa". A citação possui uma referência explícita a uma norma diferente da norma do segmento hospedeiro.

"Artigo 1º
 - O ****inciso VI do art. 2º do Decreto nº 1.791, de 15 de janeiro de 1996****, passa a ..."

[38] Artigo 15, inciso V da Constituição.
[39] Lei nº 5.869, artigo 270.
[40] Decreto nº 1.965, artigo 1º.

Caso 2 – Citação Composta

A citação composta referencia mais de um segmento de menor nível na hierarquia da citação. A âncora gerada para este tipo de citação refere-se somente ao primeiro elemento da lista. Esta é uma opção da implementação adotada na ferramenta *WebifyLaw*. Os outros elementos da lista podem ser obtidos pela proximidade com o elemento citado, sendo exibidos para o usuário na mesma página. Os casos de citações compostas são abordados nos exemplos abaixo.

"Artigo 10 - Ficam revogadas as disposições em contrário especialmente as alíneas "n" e "o" do artigo 2º da Resolução/CONAMA/004/85."[41]

O primeiro caso de citação composta é a citação de múltiplos segmentos que pertencem a um mesmo e único tipo. É chamada de citação do tipo lista. A lista contém um número variável de elementos, sem um limite fixo. Os elementos são separados por vírgula quando são mais do que dois.

"Artigo 10 - Ficam revogadas as dos posições em contrário especialmente as ****alíneas "n" e "o" do artigo 2º da Resolução/CONAMA/004/85****."

"Artigo 207 - É nulo e de nenhum efeito, quanto aos contraentes e aos filhos, o casamento contraído com infração de qualquer dos ns. I a VIII do art. 183."[42]

Neste caso de citação composta, os segmentos são apresentados na forma de um intervalo, caracterizado pela presença dos limites inferior e superior.

"Artigo 207
- É nulo e de nenhum efeito, quanto aos contraentes e aos filhos, o casamento contraído com infração de qualquer dos ****ns. I a VIII do art. 183****."

Caso 3 – Citação com Referência de Contexto

Na citação com referência de contexto, o endereço da âncora é determinado com base no endereço do segmento hospedeiro, utilizando assim o contexto em que a citação aparece para determinar seu endereço. Este tipo de citação é chamada de citação modificada pela

[41] Artigo 10, da Resolução CONAMA nº 10 , de 1993.
[42] Artigo 207 da Lei nº 3.017.

presença de um vocábulo, chamado de modificador. O modificador e o contexto da citação são elementos-chave na identificação do segmento citado.

"§3º A indicação de que trata o parágrafo anterior será feita no prazo de trinta dias..."[43]

Neste exemplo, é usado o modificador "anterior". Outras variações deste modificador podem ser "antecedente" e "precedente". Na citação acima, o segmento modificado é o segmento citado. Seu endereço é determinado subtraindo-se um do nível correspondente do endereço do hospedeiro.

"§3º
A indicação de que trata o ****parágrafo anterior**** será feita no prazo de trinta dias ...".

"§5º O representante de que trata o inciso IV deste artigo será indicado pelo Ministro ..."[44]

O modificador nesta citação é o modificador de segmento atual. Algumas variações deste tipo de modificador são: "deste", "desta", "neste", "nesta", "este", "esta", "esse". Na geração do endereço da citação, o nível modificado é copiado do segmento hospedeiro.

"§5º
O representante de que trata o ****inciso IV deste artigo**** será indicado pelo Ministro..."

"I. a afirmação do autor, ou a certidão do oficial, quanto às circunstâncias previstas nos nºs. I e II do artigo antecedente;"[45]

Esta citação apresenta uma segunda variação do uso do modificador tipo anterior, na qual o segmento modificado não é o segmento citado em última instância.

a afirmação do autor, ou a certidão do oficial, quanto às circunstâncias previstas nos ****nºs. I e II do artigo antecedente****;

Caso 4 – Vários Casos de Citação

Os diversos casos identificados anteriormente aparecem, muitas vezes, combinados em uma mesma citação. Alguns destes casos são apresentados a seguir.

[43] Artigo 3º, §3º, Decreto nº 1.752.
[44] Artigo 3º, §5º, Decreto nº 1.752.
[45] Artigo 232, inciso I da Lei nº 5.869.

"O VICE-PRESIDENTE DA REPÚBLICA, no exercício do cargo de Presidente da República, usando das atribuições que lhe confere o art. 84, incisos IV e VI, da Constituição, e tendo…"[46]

Este é um caso que apresenta citação composta com referência à norma externa.

"O VICE-PRESIDENTE DA REPÚBLICA, no exercício do cargo de Presidente da República, usando das atribuições que lhe confere o ****art. 84, incisos IV e VI, da Constituição****, e tendo…"

"§4º Nas causas de pequeno valor, … , atendidas as normas das alíneas "a", "b", e "c" do parágrafo anterior."[47]

Neste exemplo, a citação é uma combinação de uma citação composta e de uma referência de contexto.

"§4º Nas causas de pequeno valor, … , atendidas as normas das ****alíneas "a", "b", e "c" do parágrafo anterior****."

"§3º Os partidos políticos reger-se-ão pelo disposto, no que lhes for aplicável, nos arts. 17 a 22 deste Código e …"[48]

Esta variação é uma combinação do uso do modificador com uma lista de segmentos.

"§3º Os partidos políticos reger-se-ão pelo disposto, no que lhes for aplicável, nos ****arts. 17 a 22 deste Código**** e …"

Existem alguns casos com certa complexidade que para a geração automática é necessário adaptar o mecanismo de extração e estabelecimento de *links*. Estes casos não foram tratados aqui por representarem uma pequena porcentagem do total das citações, como mostra o **Quadro 6**. No entanto, o tratamento desses casos é necessário para se alcançar uma completude no reconhecimento de citações. Alguns desses casos são:

"estabelecido no *art. 5º deste Decreto*, conforme seu enquadramento nas condições especificadas no **referido artigo**, entre a documentação …"[49]

A citação acima em negrito é um caso de citação que para interpretação é necessário fazer referência a outra citação, normalmente a citação antecedente (apresentada em itálico). A identificação desta citação só é possível com o auxílio da interpretação da antecedente.

[46] Decreto nº 1.752.
[47] Artigo 20, §4º da Lei nº 5.869.
[48] Lei nº 3.017, artigo 16, §3º.
[49] Artigo 1º, Decreto nº 1.070.

"...ou de pacto antenupcial (arts. 178, Par. 9°, I, c, 274, 289, I, e 311)"[50]
"... contado o prazo do dia em tiverem ciência do casamento (arts. 180, III, 183, XI, 209 e 213)."[51]

As citações acima são complexas na sua estruturação. Apresentam identificadores de segmentos de diversos níveis, sem a referência explícita dos tipos. Por exemplo os segmentos I e c aparecem sem a identificação do tipo, que seria inciso e alínea respectivamente. Esta característica dificulta sua análise e identificação automática.

Estes casos mais complexos para serem identificados corretamente necessitam de um tratamento especial.

Os exemplos acima caracterizam com fidelidade o estado atual da tecnologia empregada na ferramenta *WebifyLaw*.

5.4.3 Site: Projeto WebifyLaw

Os resultados da utilização da ferramenta *WebifyLaw* na conversão automática de *NEJBs* pode ser vistos em: http://www.ifqsc.sc.usp.br/hpp/alessandra/webify/webify_result.html. As páginas geradas automaticamente contêm uma referência textual explícita incluindo a data da geração.

[50] Lei nº 3.017, Artigo 233, inciso II.
[51] Artigo 178, §4º, inciso II da Lei nº 3.017.

6 CONCLUSÃO

O processo de conversão proposto neste trabalho provou ser eficiente e confiável para a conversão automática de textos estruturados. Além da geração dos hiperdocumentos, uma característica importante do processo é a geração automática de informação para alimentar bancos de dados. Assim, o principal resultado do processo de conversão são os segmentos da informação. Outras ferramentas podem a vir utilizá-los na geração de diversos "produtos", além dos hiperdocumentos, como por exemplo consultas dinâmicas a segmentos isolados, selecionando aqueles que tratam de um determinado assunto, ou cumprem um requisito específico; ou mesmo a inserção de anotações ou comentários esclarecedores sobre a informação.

Definindo-se o conjunto de documentos alvo, as maiores dificuldades concentram-se em estabelecer as definições formais da estruturação dos documentos e das citações.

No caso das citações, a maior dificuldade é gerar um conjunto de regras que identifique todos os possíveis casos de citações.

À parte de alguns casos de citação, o processo de conversão proposto neste trabalho mostrou que a conversão automática de textos estruturados pode ser implementada como um processo que pouco requer da interferência do usuário com alta confiabilidade nos resultados.

6.1 Contribuições do Trabalho

As contribuições deste trabalho podem ser separadas nas áreas de conversão automática e hipertexto, e na área de tratamento de documentos legais. Na área de conversão automática e hipertexto as principais contribuições foram:

1. Especificação do formato para entrada de dados, o padrão STDNORM, visando uniformizar a troca de informações e permitir contribuições externas de documentos ao processo de conversão de automática.

2. Aplicação da teoria de Gramáticas e Linguagens Formais no estabelecimento de uma definição formal de textos estruturados, permitindo que futuras definições formais sejam feitas conforme realizado neste trabalho.

3. Definição de um método para a segmentação de textos estruturados com base na especificação formal. Contribuição importante para a geração de informação para bases de dados e para constituir hiperdocumentos automaticamente.

4. Especificação da forma normal para o reconhecimento e a representação de citações. Contribuição muito importante para permitir que as relações explícitas entre

documentos sejam reconhecidas automaticamente e representadas de modo uniforme dentro de um conjunto de hiperdocumentos.

5. Definição do método para a extração e estabelecimento automático de citações com base na forma normal para citações, permitindo que relações explícitas presentes nos documentos seja convertida em relações entre segmentos na base de dados e *links* nos hiperdocumentos.

6. Um processo de conversão automática de textos estruturados para hiperdocumentos, através da utilização de técnicas, principalmente das áreas de Compiladores, Gramáticas e Linguagens Formais, sem a necessidade de interferência do usuário ou do uso de Processamento de Linguagem Natural.

Na área de análise de documentos legais, as principais contribuições foram:

1. Definição formal para a estruturação de *normas estatutárias jurídicas brasileiras*, uma adaptação da teoria de Gramáticas e Linguagens Formais para o domínio de documentos legais.

2. Especificação um modelo de dados para normas estatutárias jurídicas brasileiras utilizando o modelo entidade-relacionamento estendido como representação formal. O modelo de dados proposto pode ser utilizado como base na estruturação de outros aplicativos dentro do domínio legal.

6.2 Sugestões para Trabalho Futuro

A ferramenta *WebifyLaw* foi concebida como um mecanismo de teste para o processo de conversão. A sugestão para trabalho futuro é a criação de um sistema de conversão, tendo a ferramenta *WebifyLaw* como núcleo de desenvolvimento. Neste sistema de conversão algumas características desejáveis são:

- **Interface gráfica**: A execução dos módulos da ferramenta *WebifyLaw* via linha de comando é uma solução simples mas, é mais propícia a erros e força o usuário a conhecer os parâmetros adequados. A implementação de uma interface gráfica com mecanismos de ajuda viria a possibilitar que usuários leigos utilizassem o sistema.

- **Interação com o usuário confirmando inferências de *links***: A ferramenta *WebifyLaw* faz a inferência dos *links* entre os segmentos de normas de modo automático. O usuário não é consultado durante o processo. No entanto, em alguns casos seria benéfica a confirmação humana, principalmente casos complicados de citações que não são reconhecidos pela ferramenta.

- **Base de dados mais robusta**: A ferramenta *WebifyLaw* simula a base de dados utilizando arquivos texto, o que prejudica no desempenho da conversão, principalmente quando o volume de dados é grande. A solução para este problema é realizar o armazenamento de dados utilizando um SGBD. A linguagem Perl possui mecanismos de interfaceamento com bases de dados no padrão SQL, que facilitam na execução desta tarefa.

- **Mecanismo de busca por assunto**: A procura por assunto é um mecanismo muito utilizado. Apesar de este não estar presente na implementação da ferramenta, a pesquisa pode ser feita utilizando uma das ferramentas de pesquisa disponíveis na Internet (Altavista, Yahoo!, e similares). No entanto, as normas, em particular, exigem um mecanismo de busca mais elaborado, com suporte para sinônimos e antônimos. Esta é uma questão vinculada aos sistemas de informação, e é largamente abordada em sistemas para a construção de *thesaurus*. A incorporação da construção de um *thesaurus* no sistema de conversão permitiria a realização desta tarefa.

- **Hiperdocumentos dinâmicos**: Os hiperdocumentos são gerados estaticamente a partir dos dados na base de dados. Isto implica que a cada alteração na base de dados os hiperdocumentos devem ser gerados novamente e disponibilizados para uso. Esta limitação pode ser eliminada incorporando funções para fazer a geração dinâmica dos hiperdocumentos a partir das consultas remotas feitas pelos usuários via Web.

- **Execução remota do sistema via site na Web**: A incorporação desta característica vai permitir que a conversão automática seja executada independentemente da localização dos dados e dos programas fonte. Possibilita também, o atendimento a um grande número de usuários sem necessidade da distribuição do sistema.

Outra sugestão é a aplicação do processo de conversão proposto em outros conjuntos de textos estruturados.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] Aho, Alfred V.; Sethi, Ravi; Ullman, Jeffrey D.; *Compilers – Principles, Techniques and Tools*. Addison-Wesley Publishing, March 1986. ISBN 0-201-10088-6.

[2] Allan, James; Automatic Hypertext Construction. PhD Thesis, Cornell University, January 1995.

[3] Balasubramaniam, V.; "State of Art Review on Hypermedia Issues and Applications". Disponível em: http://www.csi.uottawa.ca/~duchier/misc/hypertext_review/

[4] Begoray, John A.; "An introduction to hypermedia issues, systems and application areas". *International Journal of Man–Machine Studies*. Vol. 33, 1990, pp. 121-147.

[5] Berners-Lee, T. ; Fielding, R. T. ; Nielsen, H. Frystyk; "Hypertext Transfer Protocol – HTTP/1.0". Seção 2.1 "Augmented BNF", March 12, 1995. HTTP Working Group, INTERNET-DRAFT. Disponível em: http://infturing.informatik.uni-halle.de/www/http/HTTP1.0-ID_10.html, download em 08 de outubro de 1997.

[6] Bernstein, Mark. “An Apprentice That Discovers Hypertext *Links*”. *Hypertext: Concepts, Systems and Applications – Proceedings of the European Conference on Hypertext, INRIA, France, November 1990*. Rizk, A.; Streitz, N.; André, J. (eds.); Cambridge University Press, New York, 1990, pp. 212-223. ISBN 0-521-40517-3.

[7] Bobbio, Norberto; *Teoria do Ordenamento Jurídico*. Edunb – Editora da Universidade de Brasília. Brasília – DF, 1994. ISBN 85-230-0276-6.

[8] Bush, Vannevar; “As we may think”. *Atlantic Monthly*. July 1945, pp. 101-108.

[9] Coffin, S.; *UNIX System V Release 4: The Complete Reference*. McGraw-Hill, Sydney, p. 292-309.

[10] Conklin, Jeff; “Hypertext: An Introduction and Survey”. *IEEE Computer.* September 1987, pp. 17-41.

[11] Cooke, Peter; “An Encyclopedia Publisher's Perspective”. *Interactive Multimedia*, Apple Computer Inc., Microsoft Press, 1988.

[12] Donnelly, Charles; Stallman, Richard; “Glossary for The YACC-compatible Parser Generator”. November 1995, Bison Version 1.25. Disponível em http://w4.lns.cornell.edu/public/COMP/info/bison/bison_14.html, download em 08 de outubro de 1997.

[13] Drakos, Nikos; “From Text to Hypertext: A Post-Hoc Rationalization of LaTeX2HTML”. Disponível em: http://cbl.leeds.ac.uk/nikos/doc/www94/www94.html.

[14] Elmasri, R.; Navathe, S. B.; *Fundamentals of Database Systems*. Benjamin/Cummings Publishing Co. Inc., California 1987. ISBN 0-8053-0145-3.

[15] Ferreira, Aurélio Buarque de Holanda; *Novo Dicionário da Língua Portuguesa*. Ed. Nova Fronteira, Rio de Janeiro, 1975.

[16] Furuta, Richard; Plaisant, Catherine; Shneiderman, Ben; “A Spectrum of Automatic Hypertext Constructions”. Hypermedia. Taylor Graham, 1989, v. 01, n. 02, pp. 179-195.

[17] Goffinet, Luc; Noirhomme-Fraiture, Monique; “Automatic Hypertext Link Generation based on Similarity Measures between Documents”. Disponível em http://www.info.fundp.ac.be/~lgo/Hypertext/semantic_*links*.html.

[18] Gray, Matthew; "Internet Statistics Growth and Usage of the Web and the Internet". MIT, disponível em: http://www.mit.edu/people/mkgray/net/index.html, visitado em 20 de outubro de 1997.

[19] Halasz, Frank G.; "Reflections on NoteCards : Seven Issues for the Next Generation of Hypermedia Systems". *Communications of the ACM*. July 1988.

[20] Halasz, Frank G.; Schwartz, Mayer; "The Dexter Hypertext Reference Model". *Communications of the ACM*. February 1994, vol. 37, n. 02, pp. 30-39.

[21] Hopcroft, John E.; Ullman, Jeffrey D.; *Formal Languages and their Relation to Automata*. Addison-Wesley Publishing Co. Massachussets, 1969. ISBN 0-201-02893-9.

[22] Horton, William K. "What Should and Should Not Go Online". *Designing & Writing Online Documentation*. John Wiley & Sons, New York, 1990, pp. 15-30.

[23] ISO 8859-1, *Information Processing – 8-Bit single-byte coded graphic character sets – Part 1: Latin Alphabet No. 1*. International Organization for Standardization, February, 1987.

[24] Joyce, Michael; "Notes Toward an Unwritten Non–Linear Electronic Text, 'The Ends of Print Culture' (a work in progress)". *Postmodern Culture*. Vol. 2(1), September 1991.

[25] Kahn, Paul; "Linking Together Books: Experiments in Adapting Published Material into Intermedia Documents". *Hypermedia*. Taylor Graham, 1989, v. 01, n. 02, pp. 111-145.

[26] Lévy, Pierre; *As Tecnologias da Inteligência*. Editora 34 Literatura S/C Ltda. Rio de Janeiro, 1993. ISBN 85-85940-15-2.

[27] Martin, Peter W.; "Digital Law: Some Speculations on the Future of Legal Information Technology". Maio, 1995. Disponível em: http://www.law.cornell.edu/papers/fut95fnl.html, *download* em 01 de Novembro de 1996.

[28] Martin, Peter W.; "Pre-Digital Law: How Prior Information Technologies Have Shaped Access to and The Nature of Law". Junho, 1995. Disponível em: http://www.droit.umontreal.ca/CRDP/Conferences/DAC/MARTIN/MARTIN.html, *download* em 01 de Novembro de 1996.

[29] Naur, Peter (ed.); "Revised Report on the Algorithmic Language ALGOL 60." Communications of the ACM, Vol. 3 No.5, pp. 299-314, May 1960.

[30] Network Wizards; "Internet Domain Survey". Disponível em http://www.nw.com/zone/WWW/report.html, visitado em 20 de outubro de 1997.

[31] Nielsen, Jacob; *Hypertext and Hypermedia*. Academic Press Inc., 1990. ISBN 0-12-518410-7.

[32] Raymond, Darrell R.; Tompa Frank WM.; "Hypertext and the Oxford English Dictionary". *Communications of the ACM*. July 1988, vol. 31, n. 7, pp. 871-879.

[33] Rearick, Thomas C.; "Automating the Conversion of Text Into Hypertext". *Hypertext/Hypermedia Handbook*. Berk, Emily; Devlin, Joseph; eds. Armadillo Associates, Inc. New York, 1991, pp. 113-142.

[34] RFC nº 1766 (*Request for Comment*), Tags for the Identification of Languages. Disponível em: http://ds.internic.net/rfc/rfc1766.txt.

[35] RFC nº 822 (*Request for Comment*), Standard for ARPA Internet Text Messages. Disponível em: http://ds.internic.net/rfc/rfc822.txt.

[36] Riner, Robert; "Automated Conversion". *Hypertext/Hypermedia Handbook*. Berk, Emily; Devlin, Joseph; eds. Armadillo Associates, Inc. 1991, pp. 95-111.

[37] Salton, Gerard; Allan, James; Buckley, Chris; "Automatic Structuring and Retrieval of Large Text Files". *Communications of the ACM*. February 1994, vol. 37, n. 2, pp. 97-108.

[38] Shneiderman, Ben; "Reflections on Authoring, Editing, and Managing Hypertext". *The Society of Text*. Barrett, Edward. ed. The MIT Press, 1991, pp. 115-131.

[39] Tapper, C.; "Citations as a tool for searching law by computer". In Niblett, B. (ed.); *Computer Science and Law*. 1980.

[40] Tittel, Ed; Gaither, Mark; Hassinger, Sebastian; Erwin, Mike; *World Wide Web com HTML & CGI: Bíblia do Programador*. Berkeley Brasil Editora, 1996. ISBN 85-7251-417-1.

[41] *vCard: The Electronic Business Card*. A versit Consortium Specification. Version 2.1, © 1996, Internet Mail Consortium. Disponível em: http://www.imc.org/pdi.

[42] Volpe Nunes, M.G.; Hasegawa, R.; *et. all*; "Desenvolvimento de um Revisor Gramatical para o Português Contemporâneo". *Relatório do Convênio ICMSC/USP - ITAUTEC/PHILCO.* Versão Julho de 1995. Documento de circulação restrita.

[43] Wall, Larry; Christiansen, Tom; Schwartz, Randal L.; *Programming Perl*. O'Reilly Associates, Inc, 1996. ISBN 1-56592-149-6.

[44] Wilson, Eve; "*Links* and Structures in Hypertexts Databases for Law". *Hypertext: Concepts, Systems and Applications – Proceedings of the European Conference on Hypertext, INRIA, France, November 1990*. Rizk, A.; Streitz, N.; André, J. (eds.); Cambridge University Press, New York, 1990, pp.195-223. ISBN 0-521-40517-3.

[45] Yankelovich, Nicole; Norman Meyrowitz; van Dam, Andries; "Reading and Writing the Electronic Book". *Hypermedia and Literary Studies*. Delany, Paul; Landow, George P. (eds.); MIT Press, Massachusetts, 1991; pp 53-79.

[46] Yankelovich, Nicole; Smith Karen E.; Garrett, L. Nancy and Meyrowitz, Norman; "Issues in Designing a Hypermedia Document System". *Interactive Multimedia*. Editors Sueann Ambron and Kristina Hooper, Microsoft Press, 1988.

APÊNDICE

Apêndice A GLOSSÁRIO PARA AS REGRAS DE PRODUÇÃO PARA O CONJUNTO DAS *NEJBs*

TERMO	DEFINIÇÃO
Norma	Uma entidade com identificação única dentro do conjunto de normas estatutárias jurídicas brasileiras, denominada somente *norma* por comodidade, composta necessariamente por um *cabeçalho* seguido de um *corpo*.
AbrevN	Diversos tipos de abreviação para "número" observados nas normas.
PrepDE	Diversos tipos de grafia para a preposição "de" observados nas normas.
DataExt	Formato para data por extenso, indicando a data da aprovação/divulgação da norma.

TERMO	DEFINIÇÃO
Dia	Um dia de acordo com o calendário gregoriano, variando de 1 a 31.
Mês	Um mês do calendário gregoriano.
Ano	Formato especificando a representação de um ano do calendário gregoriano.
Única	Diversos tipos de grafia para a indicação de uma seção ou parágrafo único.
Cabeçalho	Parte componente de uma norma que faz a identificação da norma unicamente dentro do seu contexto, no caso o sistema de normas a que pertence e ao conjunto que engloba normas brasileiras em geral. Esta definição de cabeçalho especifica a formatação através da identificação de seus sub-termos componentes, como a identificação do tipo *TipoJur*, do seu número *NormaID* e da sua data de publicação *DataExt*.
NormaLabel	Parte do *cabeçalho* que consiste somente da identificação da norma.
TipoJur	Estabelece determinados tipos de norma estatutárias jurídicas brasileiras.
TipoJurEsp	Tipos de normas estatutárias jurídicas considerados "especiais" por possuírem uma instituição à eles associada.
TipoJurGen	Quaisquer outros tipos de normas estatutárias jurídicas brasileiras que venham a surgir, e que sejam não-estatutárias são especificamente identificadas nestas regras.
TipoConst	Tipos de normas estatutárias jurídicas brasileiras especificamente presentes na Constituição Federal.
TipoConstGen	Quaisquer outros tipos de normas presentes na Constituição Federal, que não estão explicitamente identificadas na regra *TipoConst*.
Apelido	Nome que se dá em alternativa ao identificador oficial e comumente dado à uma *norma* para identificar, de modo geral em um escopo amplo, o assunto à que esta se refere. Por exemplo, o *código das águas*, *código florestal*.
DispInic	Parágrafo que segue a identificação formal da *norma* e que normalmente resume o assunto tratado. Também conhecidas como disposições preliminares.
EscopoN	Parágrafo que antecede o corpo da *norma* e que normalmente especifica quais as autoridades envolvidas na aprovação da *norma*, as autoridades envolvidas na execução da *norma* e as leis que permitem tais atribuições.
Corpo	Parte de uma *norma* que incorpora todo seu conteúdo, podendo ser uma *seção* ou um *componente*.
Seção	Qualquer seção de uma *norma*.
Livro	Um tipo de seção de uma norma com características específicas como a de possuir títulos como subseções.
Título	Um tipo de seção de uma norma, normalmente caracterizado por se tratar de um modo de organizar capítulos em temas generalizados.

TERMO	DEFINIÇÃO
Capítulo	Um tipo de seção de uma norma.
OutraSec	Um outro tipo de seção que não se encaixa em nenhum dos tipos especificados e tem como característica possuir a palavra "Seção" como identificadora do seu início ou possuir somente o título da seção sem nenhuma palavra-chave identificadora.
Componente	Qualquer componente de uma *norma* que não seja uma *seção*, identificada como uma unidade com significado semântico estabelecido.
Artigo	Artigo de uma *norma* que segue uma numeração constante a partir sempre do número 1, independente da existência de qualquer tipo de *seção*.
Caput	Texto inicial de um artigo, correspondente ao primeiro parágrafo do mesmo. Normalmente terminado por ponto no caso de final de frase ou por dois pontos (":") no caso do artigo possuir itens subordinados.
Parags	Conjunto de parágrafos pertencente a um outro componente de norma, como por exemplo um artigo.
ItensLetra	Conjunto de itens numerados por letras minúsculas do alfabeto, também conhecidos no meio jurídico como *alíneas*.
ItensNum	Conjunto de itens numerados por números romanos ou arábicos que normalmente estão subordinados a um outro componente, como por exemplo um *artigo*.
ItensCom	Conjunto de itens que possuem um símbolo, normalmente o asterisco ("*") ou o hífen ("-"), como elemento para itemização, ao invés de utilizar letras ou números como os itens anteriores.
Anexo	Componente de uma norma, correspondente a tabelas, gráficos, equações e similares.
Parag	Um parágrafo de uma *norma*, identificado unicamente dentro do componente de o contém (artigo ou corpo de norma).
ItemNum	Um item numérico pertencente à classe dos *ItensNum*, que pode ser identificado unicamente dentro do conjunto.
ItemLetra	Um item identificado por letras minúsculas do alfabeto, correspondente à alínea, que pode ser identificado unicamente dentro do conjunto dos *ItensLetra* ao qual pertence.
ItemCom	Um item único dentro de um conjunto de *ItensCom*, identificado pela ordem de apresentação.
DispFin	Disposições finais de uma *norma*, que normalmente contém o local e data de sua publicação e os nomes das pessoas responsáveis por ela.
LivLabel	Um rótulo que define a identificação de uma seção do tipo Livro.
TitLabel	Um rótulo que define a identificação de uma seção do tipo Título.
CapLabel	Um rótulo que define a identificação de uma seção do tipo Capítulo.
SecLabel	Um rótulo que define a identificação de uma seção sem tipificação específica.

TERMO	DEFINIÇÃO
ParagLabel	Um rótulo que identifica o início de um parágrafo.
ArtigoLabel	Um rótulo que identifica o início de um artigo.
ItemLetraLabel	Um rótulo com formatação específica que identifica um item letra.
ItemNumLabel	Um rótulo com formatação específica que identifica um item número.
ItemComLabel	Um rótulo com formatação específica que identifica um item comum.
AnexoLabel	Um rótulo que identifica o início de um anexo.
NormaID	Identificação numérica de uma norma.
SeçãoID	Identificação numérica de uma seção.
ArtigoID	Identificação numérica de um artigo.
ParagID	Identificação de um parágrafo.
ItemLetraID	Letra minúscula do alfabeto que identifica um item dentro de um conjunto.
ItemNumID	Número, normalmente em algarismos arábicos que identifica um item dentro de um conjunto.
TipoSec	Palavras-chave que identificam o início de uma seção do tipo *OutraSec*.
TipoLiv	Palavras-chave que identificam o início de uma seção do tipo *Livro*.
TipoTit	Palavras-chave que identificam o início de uma seção do tipo *Título*.
TipoCap	Palavras-chave que identificam o início de uma seção do tipo *Capítulo*.
TipoAnexo	Palavras-chave que identificam o início de um *Anexo*.
Citação	Uma citação qualquer presente em uma *norma*, fundamental para a identificação de relações entre diversas normas.
CitCodigo	A citação específica de um outro código ou sistema de normas dentro de uma norma. Esta citação pode ser composta pelo nome do código, identificação da norma citada dentro do código, do artigo, parágrafo, alínea e assim por diante.
CitNorma	Citação de uma norma pertencente ao mesmo sistema de normas da norma que a cita. Pode possuir referência a todos os componentes acima citados (artigo, parágrafo, alínea, etc.)
CitParag	Citação de unicamente um parágrafo dentro da mesma norma que o referencia.
CitArtigo	Citação de um artigo da mesma norma que o referencia.
SecTit	Parte da identificação de uma seção que representa seu título.
ParagUn	Diversos tipos de grafia para as palavras "parágrafo único" que identificam um parágrafo de um determinado artigo dentro de uma norma.

TERMO	DEFINIÇÃO
Instituição	Instituição vinculada ao estabelecimento de uma determinada norma. Em se tratando de normas jurídicas estatutárias brasileiras, tal instituição está restrita a representar instituições governamentais como o CONAMA, IBAMA, DIREN entre outras.
Respons	Conjunto de termos que especifica o "responsável" pela norma. Pode representar uma instituição, ou um cargo. Alguns exemplos são: "O Presidente da República", "O Chefe de Governo Provisório da República", "O Conselho Nacional do Meio Ambiente – CONAMA", "O Diretor de Recursos Naturais Renováveis", "O Presidente do Instituto". É identificado, principalmente, através do uso destas palavras-chave.
Abrev	Quaisquer abreviações utilizadas na língua portuguesa, como por exemplo: etc., ex., V.S.a., Ilmo., Sr., Sra., Prof., Prof.a, Prof.ª. Trata-se de um conjunto limitado, especificando quais são as abreviações válidas.

Tabela A Glossário dos termos definidos na Tabela 6.

IFSC-USP SERVIÇO DE BIBLIOTECA E INFORMAÇÃO