USO DE ONTOLOGIAS PARA A EXTRAÇÃO DE INFORMAÇÕES EM ATOS JURÍDICOS EM UMA INSTITUIÇÃO PÚBLICA

USE OF ONTOLOGIES FOR THE AUTOMATIC INFORMATION EXTRACTION IN LEGAL ACTS IN A STATE INSTITUTION

Eduardo Jaime Quirós Batres, Mestre
Central de Processamento – Universidade Federal de Viçosa (UFV)
dojai@ufv.br
Alcione de Paiva Oliveira, Doutor
Departamento de Informática – Universidade Federal de Viçosa (UFV)
alcionepaiva@yahoo.com
Bruno Ventorim Gabrielli
Departamento de Informática – Universidade Federal de Viçosa (UFV)
Vinci Pegoretti Amorim
Departamento de Informática – Universidade Federal de Viçosa (UFV)
Alexandra Moreira, Mestre
Grupo de Sistemas de Apoio à Decisão – Universidade Federal de Viçosa (UFV)
xandramoreira@yahoo.com.br

RESUMO

Com a expansão da Internet e a disponibilidade das informações em geral, surge um crescente anseio por parte de cidadãos e organizações de terem à sua disposição não só informações que dizem respeito a terceiros, mas também as informações a seu respeito ou que diretamente os afetem. Dentro deste contexto incluem-se as normas em geral e mais especificamente os atos emanados do serviço público. Este trabalho apresenta uma ferramenta automatizada, utilizando técnicas de extração automática de informações, com o intuito de extrair as principais informações contidas nos atos administrativos da Universidade Federal de Viçosa (UFV), visando a facilitar a utilização ampla dessas informações que, por serem de natureza pública, expandem seu interesse além das fronteiras do órgão emissor. Para isto se fez necessária a extração e estruturação das informações contidas nos mais diversos documentos eletrônicos dispersos pelos órgãos emissores. A ferramenta faz uso de uma ontologia construída especificamente para este propósito, possibilitando a geração de uma base de conhecimento cujo conteúdo reflete os campos obrigatórios e necessários para caracterizar um ato administrativo.

PALAVRAS-CHAVE: Extração de informação. Ontologia. Tesauros. Atos jurídicos.

1 INTRODUÇÃO

Com o crescimento da Internet e a disponibilidade das informações em geral surge um crescente anseio por parte dos cidadãos e organizações de ter à sua disposição não só informações que dizem respeito a terceiros mas também as informações a seu respeito ou que diretamente os afetem. Dentro deste contexto incluem-se as normas em geral e mais especificamente os atos emanados do serviço público.

Neste trabalho procura-se mostrar as vantagens e possibilidades da extração automática de informações específicas contidas numa norma, assim como os possíveis benefícios da estruturação de documentos jurídicos visando à sua posterior padronização evitando com isto a presença de termos ambíguos e auxiliando na padronização da sua estrutura gramatical. É apresentada uma ferramenta automatizada capaz de extrair campos específicos, importantes na análise de um ato administrativo, povoar uma base de dados e ainda alertar quando um campo, considerado importante ou indispensável ao documento, não possa ser identificado.

Desta forma pode-se melhorar a recuperação das informações contidas no documento, facilitando sua análise, tanto no sentido jurídico quanto administrativo (freqüência de certos tipos de atos, indivíduos envolvidos, assuntos etc.). Além disso, o uso de uma ontologia específica permite a definição de um domínio no qual será possível classificar o documento jurídico possibilitando a melhoria no processo de extração de informações e o intercâmbio do conhecimento obtido através do processo. O domínio selecionado é o dos atos administrativos da Universidade Federal de Viçosa (UFV).

Este artigo está estruturado da seguinte forma: a próxima seção descreve a tarefa da extração de informações; a seção 3 descreve viabilidade do uso de ontologias no âmbito do direito; a secção 4 aborda as características dos documentos jurídicos; a seção 5 apresenta os resultados e discussões; e a seção 6 apresenta as conclusões do trabalho.

2 EXTRAÇÃO DE INFORMAÇÕES

Inicialmente, a motivação principal na Extração de Informações (EI) era a de povoar automaticamente bases de dados (Jacobs e Rau, 1993). Entretanto, sistemas de extração de informações também permitem melhorar o desempenho de sistemas de recuperação de informações através da integração e sintetização da informação, evitando desta forma a ocorrência de redundâncias em textos que tratam do mesmo assunto.

Segundo Wives e Loh (1999) a EI foi uma evolução natural da área de recuperação de informações. Cada dia que passa, mais e mais informações surgem nos meios eletrônicos e uma porcentagem significativa das mesmas é composta de informações que, de alguma forma, podem ser estruturadas ou inter-relacionadas. Assim, ao invés de encontrar textos que contenham informações e permitir ao usuário procurar o que lhe interessa, esta nova área passou a se preocupar em encontrar as informações dentro dos textos e tratá-las de forma a

apresentar algum tipo de conhecimento novo e útil para o usuário. A idéia é aproveitar todo o conhecimento humano que há em textos escritos e, mesmo que tal conhecimento novo não seja resposta direta às indagações do usuário, processá-lo sob a premissa de que ele deve contribuir na satisfação das necessidades de informação do mesmo.

Nas palavras de Wives e Loh (1999) "As aplicações de sistemas de descoberta de conhecimento em textos são inúmeras. Qualquer domínio que utilize intensivamente textos não-estruturados (documentos escritos), tais como as áreas jurídicas, policiais, cartórios e órgãos de registros, empresas em geral, etc., podem beneficiar-se destes sistemas."

Do ponto de vista das técnicas utilizadas, a EI pode ser vista como qualquer método que filtre informações de um grande volume de texto. Grishman (1997) estreita a definição para "a identificação de instâncias de uma classe particular de eventos ou relacionamentos num texto em linguagem natural, e a extração de argumentos relevantes do evento ou do relacionamento". Daí Grishman conclui que a extração de informações envolve a criação de uma representação estruturada da informação selecionada e extraída do texto.

Tipicamente, a extração de informação envolve a identificação de padrões que representam um contexto chave dentro do texto. Além disso, a EI utiliza um conjunto de filtros que, junto com os padrões, irão representar, de forma estruturada, a informação contida em cada texto, possibilitando a atualização de uma base de dados ou a melhora de uma recuperação de informações posterior. (Jacobs e Rau,1993)

Da mesma forma, já num domínio jurídico, especificamente o dos atos administrativos da UFV, um sistema de extração de informações estará, por exemplo, interessado em extrair as seguintes instâncias: número do ato, órgão expedidor, autoridade competente da qual emana o ato, destinatários do ato, determinação (o que deve ser feito), motivação (o que levou à elaboração do ato), etc.

Daí, por mais fácil que seja identificar as instâncias verifica-se que o entendimento da linguagem natural é crucial a qualquer sistema cujo objetivo seja a extração de informações. Isto porque a identificação da informação desejada só é possível através do reconhecimento do papel conceitual (relações semânticas que são definidas pelo papel que o item desempenha no contexto.) que a identifica (Riloff, 1999). Alerta similar é feito por Voorhees (1999) quando, ao abordar as dificuldades da análise de documentos em linguagem natural, mostra que mesmo as técnicas de análise denominadas não lingüísticas, implicitamente, exploram o conhecimento lingüístico existente.

Ainda segundo Riloff (1999) a maioria dos sistemas de EI utiliza algum padrão para identificar as informações potencialmente relevantes. Por exemplo, no domínio do ato jurídico

seriam utilizados os termos “nomear” e “nomeio” como padrões indicativos a quem se destina a norma (ato). Em geral os padrões de extração podem ser vistos como um *frame* que é ativado por uma expressão lingüística que leva a extrair as frases circundantes.

Uma forma de capturar e registrar a semântica dos elementos terminológicos de uma determinada especialidade é por meio da criação de uma estrutura que contenha os termos usados no domínio, bem como suas relações. Esta estrutura vem sendo denominada pelos pesquisadores da área de representação do conhecimento de ontologia terminológica. (Sowa, 2001).

3 ONTOLOGIAS

O conceito de ontologia surgiu, originalmente, na área de Filosofia como o estudo “do que existe”. Na área de Inteligência Artificial o termo ontologia possui uma conotação diferente, sendo usado como um termo geral para denotar sistemas conceituais utilizados como veículos promotores do compartilhamento e reutilização do conhecimento. Adicionalmente também são usadas na integração inteligente da informação, recuperação e extração de informação, sistemas de informação cooperativos, extração de informações, comércio eletrônico e gestão do conhecimento. O poder das ontologias está no fato de que elas fornecem um entendimento comum, padronizado e compartilhado de um domínio, conhecimento este que pode ser comunicado ou compartilhado entre pessoas e sistemas aplicativos (Horrocks *et al*., 2000). Assim ela é capaz de fornecer um vocabulário de termos e relações com os quais se pode modelar o domínio. Moreira (2003) apresenta uma análise detalhada sobre o conceito de ontologias no âmbito da ciência da computação.

As ontologias podem permitir uma ampliação da extração de informação ao fornecer um sistema conceitual expresso por um conjunto de termos e suas relações que permitem, a partir de um determinado termo, a localização de termos mais amplos ou mais genéricos, sinônimos, termos oposto e termos associados em geral.

Quanto à viabilidade do uso de ontologias no âmbito do direito destaca-se o estudo apresentado por Valente (1995) que aborda o uso de ontologias em IA e Leis. De acordo com Valente, para se construir uma ontologia sobre uma lei é necessário adotar-se uma perspectiva baseada numa visualização sistêmica que distinga o sistema do ambiente e que determine como estes dois elementos interagem. Ressalta que, ao contrário de outros domínios, como da

mecânica e da medicina, no domínio das leis não é fácil distinguir qual é o sistema e qual o meio ambiente. Isto ocorre porque há dois sistemas que podem ser representados:

- **o sistema legal** que é uma organização legal composta por vários sub-organismos, e nesta perspectiva o modelo do sistema legal conteria informações sobre a estrutura do sistema legal em termos de suas agências, estrutura interna etc.;
- **o domínio da aplicação legal** que é o domínio da sociedade que o sistema legal pretende regular. Como um todo, as leis controlam o comportamento social dos cidadãos, entretanto normas, e.g., atos administrativos, específicos regulam subsistemas da sociedade identificados por objetos, ações etc.

Valente (1995) justifica o uso, de uma ontologia específica para uma lei (norma) pelo seguinte:

1. *"primeiramente porque é inevitável: qualquer sistema baseado em conhecimento legal terá algum comprometimento ontológico que determinará o que pode ou não pode ser feito. Daí, havendo necessidade de se fazer estes comprometimentos é melhor que seja de forma clara e concisa através de uma ontologia;*
2. *ontologias podem ser utilizadas para produzir linguagens especialistas ou para representar formalismos, portanto é possível construir interpretadores que, mesmo que ineficientes, resolvam o problema da análise da linguagem especialista;*
3. *em terceiro lugar, e provavelmente mais importante, ontologias podem ser utilizadas para implementar estratégias de "dividir e conquistar". A ontologia define o domínio em termos de categorias primitivas de conhecimento (entidades) e seus relacionamentos. Assim, a ontologia divide o mundo em pedaços que podem ser solucionados separadamente. Ou seja, uma ontologia define qual é o problema básico que deve ser solucionado;*
4. *há uma grande potencialidade de que ontologias na área jurídica comecem a se interconectar e desta forma levar a uma visão global das potencialidades do uso de ontologias para explicar todo um fenômeno jurídico, ao contrário do que ocorre até o momento (1995) onde a maioria das definições tem surgido para solucionar problemas específicos (locais do pesquisador);*
5. *finalmente, uma ontologia define que tipos de conclusões, garantias e cadeia de argumentos são usuais ou válidos no domínio, e isto pode ser um fator importante se o raciocínio jurídico for visto como o produtor e avaliador de argumentos legais."*

Em geral, as ontologias são representadas por meio de linguagens formais, dentre elas pode-se citar: *Ontolingua* (Gruber, 1993); Farquhar *et al*. ,1995; Farquhar *et al* ,1997), que é

baseada em *Knowledge Interchange Format* (KIF) (Genesereth e Fikes, 1992) e *Frame Logic* (Kifer *et al*. ,1995). Nesta pesquisa foi utilizada a *Ontology Inference Layer* (OIL) (Horrocks *et al*., 2000). Klein *et al*. (2001) destacam as potencialidades do desenho do projeto OIL como:

- prover a maioria das primitivas de modelagem disponíveis nas ontologias orientadas nos modelos de quadros e lógica de descrições;
- possui uma semântica simples, limpa e bem definida; e
- suporte automático para um mecanismo automático de inferência (e.g., consistência das classes e verificação de classificação).

Para facilitar a definição de ontologias utilizando o OIL (Bechhofer, 2003) desenvolveu em Java um editor denominado OilEd. Embora ainda em desenvolvimento, trata-se de uma poderosa ferramenta de elaboração e testes de uma ontologia construída utilizando o formalismo do OIL. A ferramenta permite a análise da definição utilizando um mecanismo de inferência (FaCT) e também permite exportar a ontologia em diversos formatos aptos a serem utilizados em outros ambientes e aplicações

Algumas ontologias definem os conceitos informalmente e não por meio de axiomas e definições, capturando a semântica dos elementos pela posição de seus respectivos termos em uma estrutura de relações denominada de sistemas de conceitos. Tais ontologias são denominadas por Sowa (2001) de ontologias terminológicas. Segundo Sowa, a diferença entre uma ontologia terminológica e uma formal é de grau: a medida que mais axiomas são acrescentados a uma ontologia terminológica, ela pode evoluir para uma ontologia formal ou axiomatizada. Dentre as ontologias terminológicas destaca-se a classe dos tesauros, utilizado amplamente na área da ciência da informação. Craven (2002), define tesauros como ferramentas para o controle de vocabulário que através do uso de indexadores e buscadores sobre quais termos usar, pode melhorar significativamente a qualidade da extração de informações. Segundo Moreira (2003), uma ontologia como vista pela ciência da computação é um sistema de conceitos, da mesma forma que os tesauros, e como tal pertence ao nível epistemológico e não ao ontológico. A diferença em relação aos tesauros pode ocorrer em termos de linguagem, de nível de formalização e de propósitos. Neste sentido pode ser adequado que, no âmbito da ciência da computação, os tesauros sejam enquadrados como ontologias. Esta pesquisa fez uso de um tesauros, justamente com o propósito de ampliar a capacidade de extração de informação da ferramenta. De agora em diante, serão usados os termos tesauros e ontologia segundo a interpretação de Sowa.

4 CARACTERÍSTICAS DOS DOCUMENTOS JURÍDICOS

Como no restante do Poder Executivo, numa instituição de ensino superior existem também normas que o poder originário permite emanar (delegação) das autoridades constituídas da instituição. Portanto dos conselhos universitários tem-se as resoluções, do Reitor as portarias, das diretorias e chefias de departamento os atos. Todas essas normas estão regulamentadas, de alguma forma, não apenas quanto ao seu conteúdo, como também quanto a sua forma. No que diz respeito à forma não há rígidos padrões quanto ao formato do documento e, embora existam padrões a serem seguidos quanto ao conteúdo, nem sempre isto ocorre.

Seja qual for a norma, ocorre um problema inerente à atividade humana que é a dificuldade de interpretação causada pelo uso da linguagem natural. A linguagem natural, quando utilizada para gerar um ordenamento que, em principio, deve ser rígido e único, muitas vezes resulta em normas com interpretação dúbia, determinações com duplo sentido ou contraditórias, textos contendo lacunas, incertezas, ambigüidades, obscuridade etc.

Um problema muito comum em qualquer ordenamento jurídico é o das antinomias (contradição entre duas leis), que podem ser causadas justamente pelo uso inadequado da linguagem natural. Na realidade, o direito não tolera antinomias, entretanto ocorre, regularmente, o conflito de normas. Os casos mais comuns segundo Bobbio (1995) são: uma norma manda fazer uma coisa e outra proíbe; uma norma manda fazer e outra permite não fazer; e uma norma proíbe fazer e outra permite fazer. Existem diversos critérios para resolver o problema de conflito de normas.

Problemas de conflitos de leis poderiam ser evitados, ainda que não na sua totalidade, pelo uso de ferramentas de recuperação ou de extração de informações que permitam analisar o conteúdo do ordenamento em vigor e de algoritmos que fizessem pré-análise dos critérios de resolução dos conflitos.

4.1 O Problema

Além dos problemas descritos, existem outros problemas que ocorrem no serviço público, seja no Brasil ou no exterior, que também dificultam a organização do conhecimento gerado pelo Estado. A seguir são descritos alguns destes problemas, em particular aqueles do serviço público no Brasil:

- não há normas que regulem o uso da linguagem natural ou a forma detalhada de se apresentar os documentos jurídicos, exceção feita às leis;

- a norma se preocupa apenas em determinar o conteúdo exigido mas não a forma de apresentação. Os padrões de localização de determinado conteúdo surgem das normas da linguagem natural, do bom senso e do uso e costumes legislativo, processual e administrativo;
- como no serviço público tudo tem que ser documentado e não há padrões de classificação da importância deste ou daquele documento (todos são considerados de "igual valor"), a tendência natural dos indivíduos, devido a características psicológicas humanas, é desconsiderar a importância dos documentos não diretamente relacionados consigo mesmo ou com sua função;
- a rotatividade natural dos dirigentes faz com que muitas informações sejam perdidas, tanto no sentido estrito como prático.

Focalizando os atos administrativos gerados dentro do âmbito da UFV, encontramos os seguintes problemas específicos:

- falta de padronização entre os atos de órgão e outro, o que também ocorre com outros documentos de valor jurídico significativo;
- cada órgão arquiva seus próprios atos ocorrendo perdas constantes dos documentos eletrônicos devido a vírus de computador, falta de cuidados, treinamento inadequado de quem manipula, etc.;
- havendo mudança nos dirigentes ocorrem mudança na forma de elaboração dos documentos, método de arquivamento, etc.;
- mesmos os atos remetidos para publicação sofrem das conseqüências da falta de copias de segurança;
- os métodos de extração da informação contida nos documentos é visual e pode depender de várias pessoas tendo em vista que os documentos não se encontraram concentrados num único órgão.

Do acima exposto considerou-se viável e necessário trabalhar numa ferramenta cujo objetivo a seguir descreve-se.

4.2 Os Objetivos

O objetivo da pesquisa aqui descrita foi produzir uma ferramenta computacional para analisar o conjunto de normas emitidas pela UFV, determinando padrões que permitissem extrair informações relevantes de um conjunto semi-estruturado de informações. Essas

informações seriam armazenadas em uma base de dados estruturada, de modo a facilitar consultas, ou exportadas na forma de uma estrutura conceitual representando o domínio (tesauros ou ontologia terminológica), com o objetivo de comunicar as informações extraídas.

Adicionalmente foi utilizado um tesauros que, como ontologia terminológica, auxiliou na identificação das instâncias que representam um contexto chave dentro do texto, sendo este contexto um arcabouço que indica onde se encontra a informação a ser extraída. Portanto, o conceito de ontologia é encontrado em dois momentos na ferramenta: no processo de extração de informação, fornecendo a estruturação conceitual do domínio dos atos com vistas a ampliação do poder de reconhecimento da ferramenta; e no processo de exportação, de modo a prover uma descrição dos conceitos e relações extraídas. O tesauros foi construído a partir de uma análise dos termos encontrados nos atos, contando-se com o auxílio de um especialista da área de direito.

A ontologia, em conjunto com a estruturação dos dados também responde aos três pontos apresentados por Buneman (1997), a saber: a fonte de dados poderá ser tratada como base de dados quando necessário; a ontologia fornece o "formato muito flexível", sugerido por Buneman, permitindo o intercâmbio das informações com outras bases de dados; e mesmo quando utilizada a base de dados, através da ontologia pode ser gerada a informação semi-estruturada para propósitos de navegação.

4.3 A Fonte de Informações

O conjunto de informações semi-estruturadas é composto pelos atos emitidos no âmbito da UFV e publicados no *Campus Oficial* através da Imprensa Universitária. Nesta publicação também podem ser encontradas as Portarias provindas do Magnífico Reitor, as Resoluções dos Conselhos Universitários, os Editais e resumos de legislação federal pertinentes a instituições federais. A escolha dos atos ao invés das portarias, resoluções e/ou editais deveu-se aos seguintes fatores:

- o número total de atos publicados é bem superior à soma das demais normas publicadas no *Campus Oficial*, isto ocorre porque os atos podem emanar dos seguintes órgãos: 6 Pró-Reitorias, 4 centros, 27 departamentos, 5 diretorias e outros órgãos como a Escola Técnica, o Colégio Universitário etc.;
- a redução do conjunto de normas possíveis de serem analisadas apenas aos atos, evita que características muito especificas das outras normas possa influenciar negativamente a análise do conjunto ou evitar a extração de informações relevantes;

- apesar de se tratar de uma espécie das possíveis normas que podem emanar de uma instituição de ensino federal, os atos tratam dos mais diversos assuntos sendo, portanto, ricos em informações e possibilidades de extração das mesmas. Esta riqueza de possibilidades traz consigo maiores dificuldades na elaboração dos algoritmos de extração, mas enriquece a demonstração da viabilidade de estruturação das informações neles contidos;
- tanto os atos como a portaria possuem certa padronização o que torna viável a elaboração de protótipo da ferramenta no curto espaço de tempo de seis meses a um ano. Entretanto, como nas portarias a pessoa que elabora o documento eletrônico, salvo raros casos, sempre era a mesma, as portarias também foram eliminadas do universo a ser pesquisado por não representarem um desafio, ao contrário dos atos que são elaborados por cada seção de expediente dos departamentos, centros e diretorias, daí resultando distintos padrões de escrita em linguagem natural que certamente enriquecerão o trabalho.

4.4 O Método

Valente (1995), ao tratar do uso de ontologias em Inteligência Artificial e Leis, discorre que é necessário primeiro adotar uma perspectiva ao visualizar o fenômeno legal. Esta perspectiva pode ser baseada numa visualização sistêmica que pode começar pela delimitação do sistema e do ambiente e de como estes dois interagem.

Inicialmente foi definida a ontologia que descreve o domínio. Como ferramenta de especificação da Ontologia foi utilizada a OIL (*Ontology Inference Layer*) (Horrocks *et al*., 2000) que é uma proposta de representação (baseada na Web) da camada de inferência para ontologias. Em Oil é possível utilizar tanto as primitivas típicas das linguagens *frame-based* como a capacidade inferencial da lógica de descrições.

Para elaborar a definição da ontologia foi utilizada a ferramenta OilEd (*OIL Editor*) que permite gerar a ontologia no padrão DAML+OIL. O DAML (*DARPA Agent Markup Language*) é uma extensão do XML e do RDF que possui um rico conjunto de construtores para permitir a criação de ontologia e simultaneamente permite exportar a ontologia em uma linguagem de marcações, de tal forma que possa ser lido e interpretado por sistemas computacionais. Desta forma o poder das ontologias, que é descrever objetos e suas relações com outros objetos (o que implica em descrever conhecimento), é potencializado pelo padrão DAML que passa a ser o meio de divulgação e compartilhamento deste conhecimento. Obtida a definição formal, esta foi utilizada para desenvolver a ferramenta de extração de informações.

A figura 1 mostra os módulos que compõem a ferramenta assim como os passos envolvidos no processo de extração das informações. Antes de iniciar o processo de extração de informação o ato é pré-processado, sendo que suas unidades léxicas são identificadas e marcadas. Após essa etapa é iniciado o processo de extração de informação com o auxílio de tesauros. O método de análise é simples, uma vez que se baseia em um conjunto de padrões que descrevem a estrutura do ato. O poder de extração do método é ampliado pelo tesauros. Em seguida as informações extraídas são estruturadas e inseridas em uma base de dados relacional. Finalmente, as informações podem ser exportadas na forma de ontologias, por meio do padrão DAML+OIL.

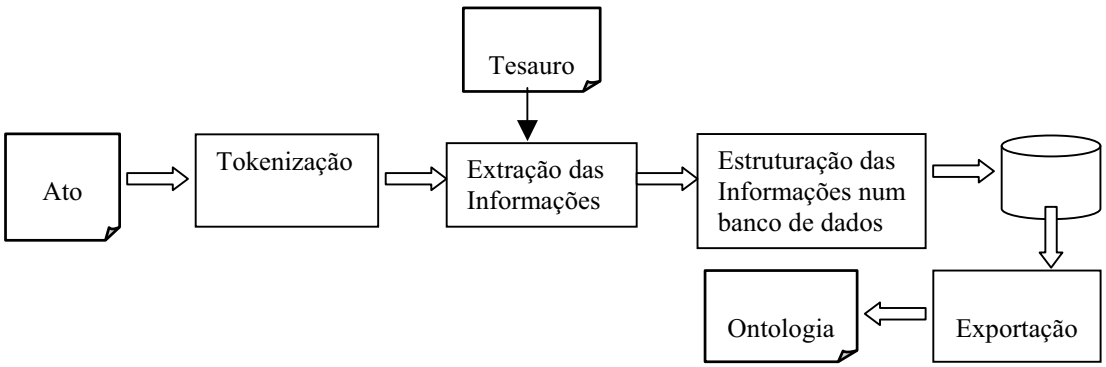

Figura 1: fluxo da extração de informações

Etapas tais como a eliminação de *stopwords*, radicalização comuns e outros sistemas de extração de informações não foram implementados, uma vez que poderiam alterar os nomes próprios.

5 RESULTADOS E DISCUSSÃO

O estudo de caso dividiu-se em duas etapas: a primeira foi o teste do sistema com o conjunto dos documentos disponíveis, a segunda consistiu em levar o sistema de extração de informações até um usuário que conhecesse da metodologia, sistemas, costumes, processo etc. da parte administrativa da instituição para que pudesse avaliar o sistema de modo prático, indicando suas falhas e potencialidades.

Durante a implementação do sistema foi utilizado um subconjunto dos atos disponíveis em formato digital. Na fase de implementação do sistema foram utilizados menos da metade dos atos disponíveis, escolhidos aleatoriamente, tendo em vista o número reduzido

de documentos eletrônicos disponíveis para implementação e testes da ferramenta. Considerou-se de bom senso não utilizar todos os documentos inicialmente para se ter alguns que pudessem ser usados no teste final do sistema.

5.1 Teste Inicial da Ferramenta

Os atos escolhidos, em número de 109, foram submetidos à ferramenta de extração com resultados aquém do desejado, uma vez que a recuperação total das informações dos atos correspondeu a apenas 47,8% dos documentos. As causas detectadas para esse baixo desempenho foram:

- 6,5% dos atos não recuperados apresentavam erros de português, e.g., "reoslve" ao invés de "resolve", "desgnar" ao invés de "designar" e outros similares;
- 19,6% das falhas de extração deveu-se a uma falha no projeto da ferramenta que foi elaborada, inicialmente, com nenhuma tolerância à ausência dos campos pré-definidos. O que não deveria ter acontecido porque a ferramenta visa extrair informações e caso não existam, não há como extraí-las, mesmo que juridicamente devessem estar presentes;
- 19,3% dos atos não tiveram suas informações extraídas devido a um problema na estrutura das frases, onde frases distintas com o mesmo significado não estavam sendo reconhecidas devido à sua ausência no tesauros;
- os restantes 6,8% foram devidos a problemas os mais diversos mas principalmente a diferenças estruturais nos atos bem distintas entre si.

Após os testes iniciais verificou-se a necessidade de melhoria da ferramenta com relação aos principais aspectos insatisfatórios, inclusive porque se tratavam de problemas incoerentes com a idéia da ferramenta. Primeiramente, a ferramenta destina-se à extração de informações, portanto deve ser tolerante com a falta de campos nos atos administrativos mesmo que estes sejam campos obrigatórios do ponto de vista administrativo. Com esta alteração mais 19,6% dos documentos passaram a ser recuperados. Em segundo lugar, a ferramenta utiliza uma ontologia terminológica justamente para definir o conhecimento dos termos jurídicos que são equivalentes, similares, contraditórios, etc. Entretanto uma das equivalências comuns não tinha sido adicionada ao tesauros inicial. Portanto, adicionou-se ao tesauros a definição de que as frases "para, sob a presidência do primeiro, constituírem" e "para constituírem, sob a presidência do primeiro" constituem unidades léxicas equivalentes,

conforme proposto em Fox *et al*. (1988). Com isto foi possível extrair corretamente as informações de mais 19,3% dos documentos.

Não se introduziu um dicionário para tentar melhorar a extração das informações nos documentos que possuíam erros de português porque, conforme será abordado em trabalhos futuros, estes problemas deveriam ser corrigidos durante a edição do texto. Isto porque a ferramenta pressupõe que o texto analisado já passou por uma correção ortográfica, seja durante a edição ou antes da alimentação ao sistema.

Uma vez efetuadas as adaptações descritas acima, obteve-se um índice de extração de informações satisfatório em 86,7% dos atos administrativos analisados, embora alguns campos passaram a não ser extraídos por não constarem do texto analisado.

5.2 Validação

O protótipo da ferramenta foi apresentado a uma secretária executiva da UFV, que já trabalhou como Chefe de Seção de Expediente num Departamento e que, portanto, conhece em profundidade a elaboração e trâmite dos atos administrativos. Acompanharam a apresentação outras duas secretárias.

Nessa apresentação a ferramenta foi muito bem recebida principalmente porque a metodologia atual de pesquisa por um documento é visual e manual. Tarefa esta extremamente árdua e que depende de um diálogo nem sempre exato.

Finalizados os testes, o sistema foi elogiado por todas as presentes pelas suas características gerais e facilidade de uso. Entretanto foram sugeridas duas modificações: que a ferramenta mostre na tela todos campos que conseguiu extrair, ou tantos campos quantos possíveis, já que isto ajudaria a distinguir a natureza do ato administrativo; e que, através do campo "número do ato", fosse possível recuperar o documento completo, ou seja, gerar um vínculo de todas as linhas retornadas da consulta para acesso direto ao documento que referencia.

Mesmo alertada sobre as limitações da ferramenta quanto à não extração de 100% das informações contidas nos documentos, a usuária comentou que era preferível recorrer ao arquivo físico duas de cada dez vezes ao invés de fazê-lo sempre. A usuária deixou claro que independentemente da implementação das sugestões acima descritas, a ferramenta seria de grande utilidade devido à dificuldade atual na recuperação dos documentos. Adicionalmente, dando mais uma sugestão, considerava que a ferramenta não deveria se limitar a extrair as informações dos atos, mas também das portarias do Reitor e no futuro de outros documentos da instituição.

6 CONCLUSÕES

Do ponto de vista deste trabalho, a ferramenta mais do que resolver ou amenizar um problema local, mostra que é possível através da aplicação de sistemas conceituais, como ontologias, obter soluções simples e funcionais que em muito podem auxiliar no dia a dia do serviço público.

Como a ferramenta conseguiu obter índices de extração de informações corretas superior a 86%, sua aplicabilidade à extração das informações contidas nos atos é viável. Levando-se em consideração os atuais problemas de desestruturação das informações, é de se esperar que seja possível no mínimo organizar e compartilhar 80% das informações originais. Entretanto o volume de documentos analisados foi pequeno, motivo pelo qual deve-se continuar testando o sistema, o que pode ser feito já em fase de produção tendo em vista os bons resultados obtidos.

Espera-se que, com o uso de ontologias, seja possível intercambiar as informações extraídas. Espera-se, adicionalmente, que este trabalho tenha conseguido mostrar a importância da definição de uma ontologia de domínio e que através desta seja possível elaborar ferramentas que procedam à extração de informações baseado na especificação ontológica, permitindo o compartilhamento do conhecimento.

Ainda é necessário melhorar a ferramenta em vários aspectos, principalmente no que diz respeito a ampliar a ontologia para incluir as portarias e, se possível, o domínio das resoluções dos conselhos universitários.

Também é necessário desenvolver *softwares* que auxiliem na elaboração das normas jurídicas. Esse *software* não seria muito complexo tendo em vista a ontologia previamente definida. O *software* também utilizaria o tesauros para indicar quais os termos que não devem aparecer num ato administrativo, quais termos são preferíveis a outros etc. A ferramenta poderia incluir um corretor ortográfico além de automaticamente povoar uma base de dados institucional já que saberia de antemão todas as instâncias que devem ser identificadas no texto e adicionadas à base de dados.

O *software* ideal para elaboração de leis precisa levar em consideração a hierarquia das leis, por exemplo, uma portaria não pode contrariar uma Lei Federal e esta não pode entrar em conflito com a Constituição Federal. Adicionalmente precisaria verificar a derrogação de leis, conflito entre a norma que está sendo elaborada e outra norma de igual hierarquia. Havendo ontologias bem definidas seria possível até verificar a legislação em vigor que permite que a norma seja elaborada, se há alguma que a proíba e assim por diante.

Para que isto seja possível é indispensável o uso de ontologias que definam claramente o domínio das normas e que existam padrões de armazenamento das normas legais que permitam o intercâmbio deste conhecimento. Este é um trabalho árduo que precisa da intervenção de indivíduos da área jurídica, lingüística, ciência da computação e ciência da informação, para que em conjunto possam ser definidas e elaboradas as ferramentas mais adequadas a cada necessidade.

REFERÊNCIAS

BECHHOFER, Sean. **OilEd 3.5 manual**. Information Management -Group, Department of Computer Science, University of Manchester.
Disponível em: http://oiled.man.ac.uk/, Acesso em: 12 jan. 2003.

BOBBIO, Norberto. **Teoría General del Derecho**. Madrid : Debate, 1995. 278 p.

BUNEMAN, Peter. Semistructured Data. In: SYMPOSIUM ON PRINCIPLES OF DATABASE SYSTEMS (ACM), 16, 1997. **Tutorial in Proceedings**... [s.l. : s.n.], 1997. p. 5.

CRAVEN, Tim. **Tesauro construction – welcome to the introductory tutorial on tesauro construction**. Ontario, Canada Faculty of Information and Media Studies.
Disponível em: http://instruct.uwo.ca/gplis/677/thesaur/main00.htm, acesso em: 11 nov. 2002.

FARQUHAR, Adam et al. **Collaborative ontology construction for information integration.** Knowledge Systems Laboratory, Department of Computer Science, KSL-95-63, 1995. (Technical Report).

FARQUHAR, A., FIKES, R., RICE, J. The ontolingua server: a tool for collaborative ontology construction. **International Journal of Human-Computer Studies**, v. 46, p. 707-728, 1997.

FOX, E. et al. Building a large tesauro for information retrieval. In: CONFERENCE ON APPLIED NATURAL LANGUAGE, 2, 1988, Austin, Texas. **Proceedings…** Austin, Texas, 1988. p. 101-108.

GENESERETH, M. R., FIKES, R. E. **Knowledge Interchange Format** - Version 3.0, Reference Manual. Stanford : Computer Science Department, Stanford University, 1992.

GRISHMAN, Ralph. Information extraction; techniques and challenges. In: INTERNATIONAL SUMMER SCHOOL SCIE-97, 1997, New York. **Proceedings…** New York : Springer-Verlag, 1997. p. 10-27.

GRUBER, T. R. A translation approach to portable ontology specifications. **Knowledge Acquisition,** v. 5, p. 199-220, 1993.

HORROCKS, I. Et al. **The Ontology inference Layer OIL**. [s.l.] : Free University of Amsterdam, 2000. (Technical report).

JACOBS, Paul S., RAU, Lisa F. Innovations in text interpretation. **Artificial Intelligence,** [s.l.], 63, p. 143-191, 1993.

KIFER, M., LAUSEN, G., WU, J. Logical foundations of object orientated and frame-base languages. **ACM,** 42, 1995.

KLEIN, Michel et al. The relation between ontologies and XML schemas. **Electronic Articles in Computer and Information Science**, 6, 4, 2001.

MOREIRA, A. **Tesauros e Ontologias**: Estudo de Definições Presentes na Literatura das Áreas das Ciências da Computação e da Informação, Utilizando-se o Método Analítico-Sintético. Belo Horizonte: Escola de Ciência da Informação da UFMG, 2003. 153 p. (Dissertação, Mestrado em Ciência da Informação).

RILOFF, Ellen. **Information extraction as a stepping stone toward story understanding.** Montreal, Canada : MIT Press, 1999.

SOWA, John F. **Building, sharing and merging ontologies**.
Disponível na internet: <http://www.jfsowa.com/ontology/ontoshar.htm>. 2001, 21 p. Último acesso em : 14 mai. 2003.

VALENTE, A. **Legal Knowledge Engineering**; a modelling approach. The Netherlands : University of Amsterdam, IOS Press, 1995. 217 p.

VORHEES, Ellen M. Natural language processing and information retrieval. In: Pazienza, MT (Ed.). **Information Extraction: Towards Scalable Adaptable Systems**, New York: Springer, 1999. p. 32-48.

WIVES, Leandro Krug, LOH, Stanley. Tecnologias de descoberta de conhecimento em informações textuais; ênfase em agrupamento de informações. In: OFICINA DE INTELIGÊNCIA ARTIFICIAL (OIA) III, 1999, Pelotas (RS). **Proceedings…** Pelotas : EDUCAT, 1999. p. 28-48.

ABSTRACT
Juridical documents possess their own form of language that may be “obscure” for the lay. However, in general, juridical documents are rich in information that, if shared can aid other professionals and Justice related organizations. This includes the norms and acts emanated from the public institutions. This work presents an automated tool, using techniques of automatic extraction of information, with the aim of extracting the main information contained in the administrative acts of the Federal University of Viçosa (UFV), seeking to aid the access to that information, once it has a public nature. In order to accomplish the task the tool makes use of an ontology built specifically for this purpose, making possible the generation of a knowledge base whose content reflects the essential and necessary fields to characterize an administrative act.

KEYWORDS: Information extraction. Ontology. Thesaurus. Juridical documents.

Originais recebidos em 16/11/2004.