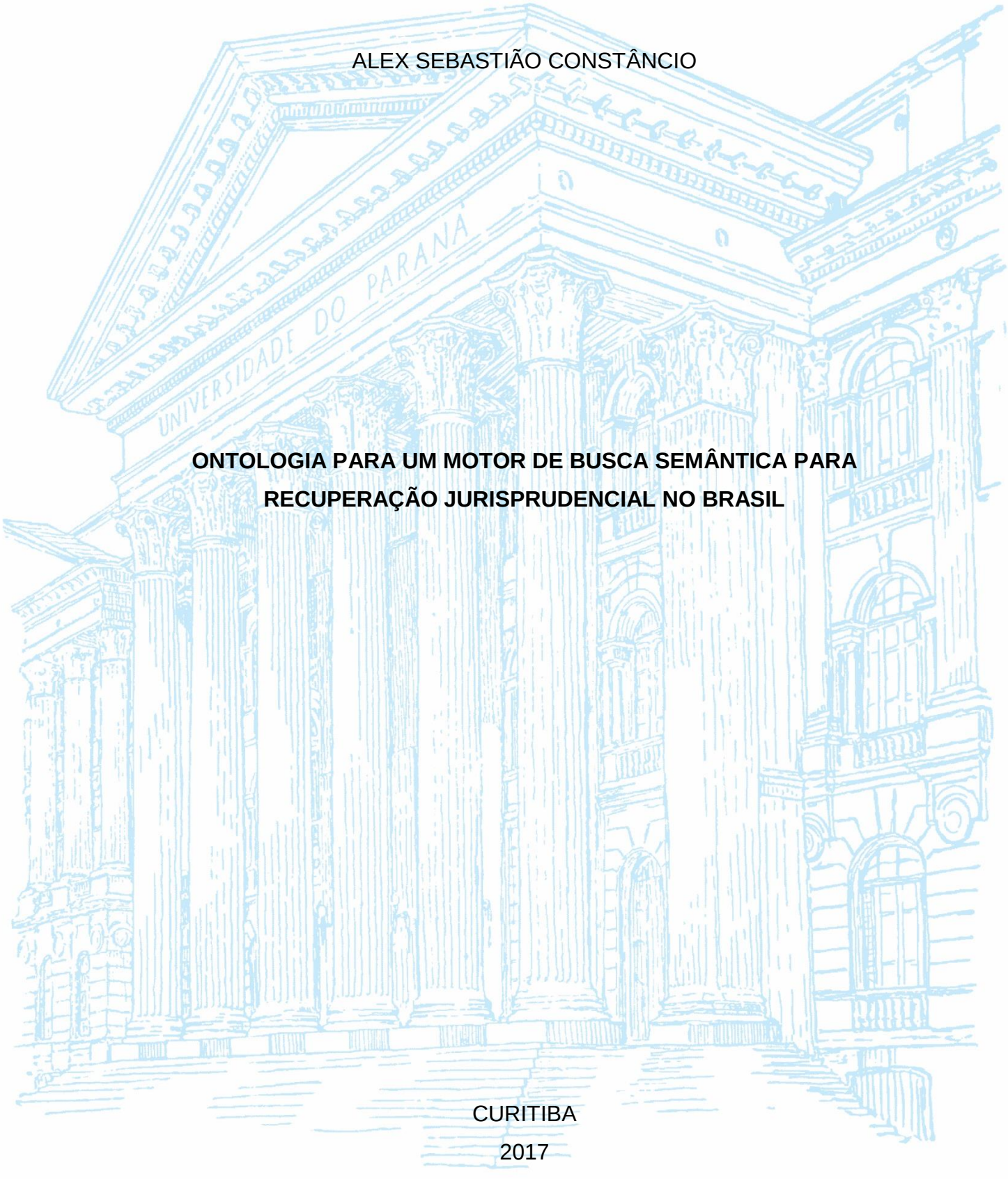
UNIVERSIDADE FEDERAL DO PARANÁ

ALEX SEBASTIÃO CONSTÂNCIO



ONTOLOGIA PARA UM MOTOR DE BUSCA SEMÂNTICA PARA RECUPERAÇÃO JURISPRUDENCIAL NO BRASIL

CURITIBA

2017

ALEX SEBASTIÃO CONSTÂNCIO

ONTOLOGIA PARA UM MOTOR DE BUSCA SEMÂNTICA PARA RECUPERAÇÃO JURISPRUDENCIAL NO BRASIL

Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Ciência, Gestão e Tecnologia da Informação, Área de Concentração: Informação, Tecnologia e Gestão do Setor de Ciências Sociais Aplicadas da Universidade Federal do Paraná como requisito parcial à obtenção de título de Mestre em Ciência, Gestão e Tecnologia da Informação.

Orientadora: Prof.ª. Drª. Deborah Ribeiro Carvalho

CURITIBA

2017

UNIVERSIDADE FEDERAL DO PARANÁ. SISTEMA DE BIBLIOTECAS.
CATALOGAÇÃO NA FONTE

Constâncio, Alex Sebastião
Ontologia para motor de busca semântica para recuperação jurisprudencial no Brasil / Alex Sebastião Constâncio. - 2017.
202 f.
Orientadora: Deborah Ribeiro Carvalho .
Dissertação (Mestrado) – Universidade Federal do Paraná. Programa de Pós- Graduação em Ciência, Gestão e Tecnologia da Informação, do Setor de Ciências Sociais Aplicadas.
Defesa: Curitiba, 2017

1. Web semântica. 2. Semântica (Direito). 3.Jurisprudência - Recuperação da informação. 4. Ontologia. I. Deborah Ribeiro Carvalho II. Universidade Federal do Paraná. Setor de Ciências Sociais Aplicadas. Programa de Pós- Graduação em Ciência. Gestão e Tecnologia da Informação. III. Título.

CDD 025.0634

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DO PARANÁ
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
Setor CIÊNCIAS SOCIAIS APLICADAS
Programa de Pós-Graduação CIÊNCIA, GESTÃO E TECNOLOGIA DA INFORMAÇÃO

TERMO DE APROVAÇÃO

Os membros da Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em CIÊNCIA, GESTÃO E TECNOLOGIA DA INFORMAÇÃO da Universidade Federal do Paraná foram convocados para realizar a arguição da Dissertação de Mestrado de ALEX SEBASTIÃO CONSTÂNCIO intitulada: Ontologia para um motor de busca conceitual para recuperação jurisprudencial no Brasil, após terem inquirido o aluno e realizado a avaliação do trabalho, são de parecer pela sua APROVAÇÃO.

CURITIBA, 08 de Fevereiro de 2017.

DEBORAH RIBEIRO CARVALHO
Presidente da Banca Examinadora (UFPR)

HELENA DE FÁTIMA NUNES SILVA
Avaliador Interno (UFPR)

CÉSAR ANTONIO SERBENA
Avaliador Externo ()

DEDICATÓRIA

Dedico esta dissertação a meus pais, Adelino e Arlete, que colocaram a minha educação e meus valores acima de todas as suas prioridades e à minha esposa Denise, pelo estímulo e paciência constantes que soube demonstrar durante as muitas horas em que estive me dedicando a esta conquista.

AGRADECIMENTOS

Primeiramente a Deus, pois tudo o que foi possível, o foi por Sua vontade e infinita misericórdia.

À minha família que por vezes recebeu menos atenção do que merecia para que eu pudesse exercer a dedicação necessária a esta realização.

À minha orientadora, Prof.ª. Drª. Deborah Ribeiro Carvalho, que acreditou no potencial da minha proposta, empreendendo seus melhores esforços e compartilhando seus melhores valores no sentido de vê-la finalmente realizada.

À Prof.ª. Drª. Helena de Fátima Nunes Silva e ao Prof.°. Dr. Cesar Antônio Serbena por seus prestimosos conselhos, sugestões e recomendações durante a banca de qualificação.

Aos demais professores do Programa de Pós-graduação em Ciência, Gestão e Tecnologia da Informação, assim como aos secretários Cesar e Emerson que, por seu empenho diário, tornam possível a existência deste pináculo do saber.

Ao meu chefe Antônio Rodrigues Barros, na Divisão de Sistemas de Informação da Universidade Federal do Paraná, por sua compreensão frente às exigências da rotina estudantil e aos colegas da mesma divisão por seu incentivo diário.

A Yeda, Liziane e Elias pela paciência e pela disposição em dividir seus conhecimentos tão duramente conquistados.

Aos respondentes do questionário de coleta de opinião, cuja disposição, boa vontade e ideias muito contribuíram para a valorização desta pesquisa.

A todos que de alguma forma, por atos ou omissões, palavras ou silêncio, contribuíram para que os meus esforços pudessem frutificar.

EPÍGRAFE

“Não temas, eu venci o mundo”
Jesus Cristo

“Não há problema que não possa ser solucionado pela paciência”
Francisco Cândido Xavier

“O teu trabalho é a oficina em que podes forjar a tua própria luz”
Emmanuel

“A ciência é, portanto, uma perversão de si mesma, a menos que tenha como fim último, melhorar a humanidade”
Nikola Tesla

“Descobrir consiste em olhar para o que todo mundo está vendo e pensar uma coisa diferente”
Roger Von Oech

“Hoje é um bom dia para tentar”
Quasimodo

“É preciso uma grande dose de coragem para enfrentar seus inimigos, mas muito mais para enfrentar os seus amigos”
Alvo Dumbledore

“Ao infinito e além”
Buzz Lightyear

RESUMO

Com o crescimento massivo do volume de documentos jurisprudenciais no Brasil, advogados e magistrados enfrentam a dificuldade de localizar informações relevantes em meio a milhões de decisões disponíveis na Internet. Os recursos hoje oferecidos para a satisfação de tal necessidade não mais correspondem aos anseios, motivando o descortinamento de novas abordagens, técnicas e ferramentas para alcançar melhores resultados na tarefa de recuperação jurisprudencial. Dentro destas iniciativas encontram-se os Motores de Busca Semântica, que fazem uso de ontologias para identificar os conceitos e ideias presentes nas decisões judiciais e assim minorar os efeitos nocivos do chamado Problema da Recuperação da Informação. Atendendo ao propósito de desenvolver uma ontologia para a construção de um Motor de Busca Semântica, foram executadas as seguintes etapas: a) conversão do Vocabulário Jurídico Controlado mantido pelo Superior Tribunal de Justiça do Brasil na ontologia Vocabulario-2016-06; b) combinação da ontologia Vocabulario-2016-06 com a ontologia JurisTJPR de Molinari, originando a ontologia OntoLegis; c) construção e avaliação de um Motor de Busca Semântica baseado na ontologia OntoLegis. O Motor de Busca Semântica experimental contruído foi comparado com o sistema de busca jurisprudencial do Tribunal de Justiça do Paraná e também colocado para experimentação pública por meio do sítio ConsultaJuris. Este sítio também disponibilizou um questionário para avaliação da opinião dos experimentadores. Os experimentos e respostas coletadas mostraram que o potencial para aplicação do Motor de Busca Semântica no domínio da recuperação jurisprudencial é real, mas que também ainda demanda aperfeiçoamentos. Resultaram destes esforços uma ontologia peso leve voltada a recuperação jurisprudencial composta po 10.210 classes e 12.595 rótulos, um Motor de Busca Semântica experimental baseado naquela ontologia e uma sistemática para conversão de parâmetros de busca em classes seletoras.

Palavras-chave: Web semântica, Recuperação da Informação, Direito

ABSTRACT

With the massive increase in the volume of jurisprudential documents in Brazil, lawyers and magistrates face the difficulty of finding relevant information amid millions of decisions available on the Internet. The features available nowadays to satisfy such a need no longer correspond to the desires, motivating the unfolding of new approaches, techniques and tools to seek to achieve better results in the task of recovering jurisprudence. Within these initiatives are the Semantic Search Engines, which make use of ontologies to identify the concepts and ideas present in the judicial decisions and thus mitigate the harmful effects of the so-called Information Retrieval Problem. Considering the purpose of developing an ontology for the construction of a Semantic Search Engine, the following steps were performed: a) conversion of the Controlled Legal Vocabulary maintained by the Superior Court of Justice of Brazil into the Vocabulario-2016-06 ontology; b) combining the Vocabulario-2016-06 ontology with the JurisTJPR ontology from Molinari, giving rise to the OntoLegis ontology; b) construction and evaluation of a Semantic Search Engine based on the OntoLegis ontology. The experimental Semantic Search engine was compared with the jurisprudential search system provided by the Tribunal de Justiça Paraná and also released for public experimentation through the ConsultaJuris website. Such website also offered a questionnaire to evaluate the opinion of experimenters. Experiments and the collected responses showed that the potential for Semantic Search Engine application in the area of jurisprudential retrieval is real and with noticeable gains, but also still requires improvements. Those efforts have resulted in a lightweight ontology for jurisprudential retrieval composed by 10,210 classes and 12,595 labels, an experimental Semantic Search Engine based on that ontology and a process for converting query parameters into selector classes.

Key-words: Semantic web, Information retrieval, Law

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE FIGURAS

LISTA DE EQUAÇÕES

LISTA DE SIGLAS

CBR	*Case-based Reasoning*
CNJ	Conselho Nacional de Justiça
IA	Inteligência Artificial
IJR	*Intelligent Jurisprudence Research*
ODP	*Ontology Design Pattern*
OEG	*Ontology Engineering Group*
OWL	*Ontology Web Language*
PRI	Problema da Recuperação da Informação
RDF	*Resource Definition Framework*
SKOS	*Simple Knowledge Organization System*
STJ	Superior Tribunal de Justiça
STF	Supremo Tribunal Federal
TJPR	Tribunal de Justiça do Paraná
TJRS	Tribunal de Justiça do Rio Grande do Sul
TJSC	Tribunal de Justiça de Santa Catarina
TJSP	Tribunal de Justiça de São Paulo
UML	*Unified Modeling Language*
UP	*Unified Process*
W3C	*World Wide Web Consortium*

SUMÁRIO

1 INTRODUÇÃO

A jurisprudência é a ocorrência de um **conjunto de decisões judiciais** consonantes produzidas em tribunais a respeito de casos semelhantes, aplicável a hipóteses similares (DINIZ, 2012), uniforme e constantemente (MAXIMILIANO, 2011). Portanto, trata-se de um conjunto de casos com características similares e que apresentam uma **sucessão de decisões harmônicas** (MAXIMILIANO, 2011; MOLINARI, 2011; DINIZ, 2012).

A jurisprudência evidencia que existe um entendimento comum ou preponderante a respeito de determinado conjunto de eventos, o que é usado para reforçar a convicção de um magistrado em favor de sua decisão a respeito de uma circunstância particular.

Por outro lado, os advogados também fazem uso de conteúdos jurisprudenciais para avaliar de que forma a justiça tem tratado determinado tipo de Litígio Processual e também para construir os argumentos em favor dos interesses de seus clientes.

Assim, tanto advogados quanto magistrados recorrem frequentemente aos conteúdos jurisprudenciais para o desempenho de suas atividades diárias e precisamente por este motivo é que os tribunais no Brasil fornecem ferramentas digitais para a busca e recuperação de documentos desta natureza.

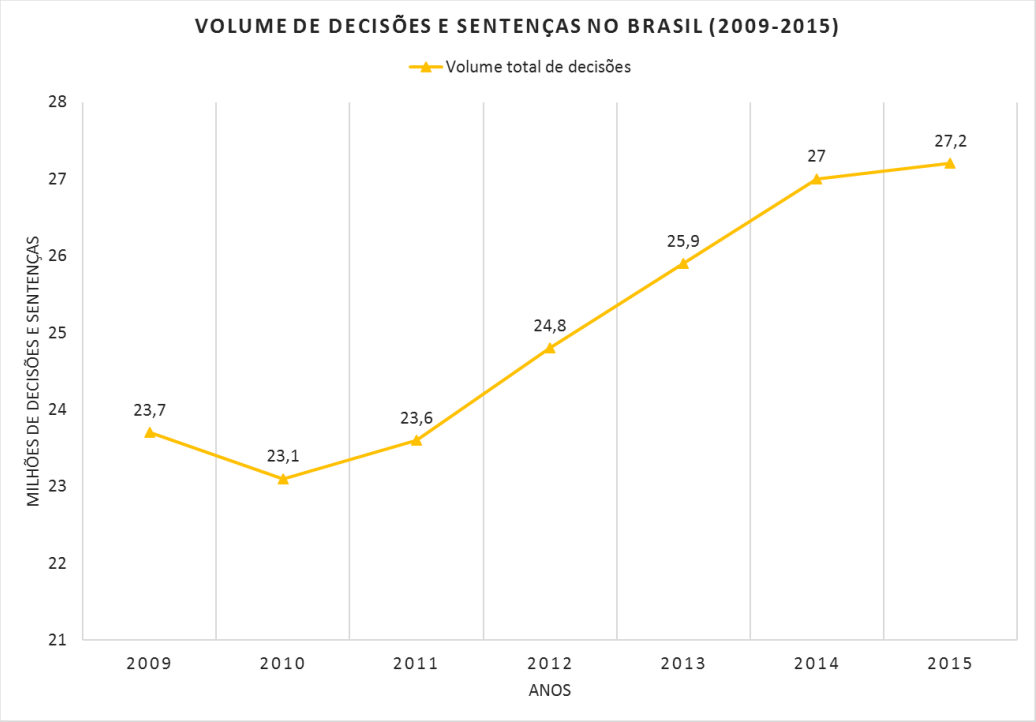
As decisões judiciais, em sua maioria, são de domínio público, estando à disposição para consulta por meio da Internet. Apenas as decisões a respeito dos casos classificados como segredo de justiça ficam inacessíveis à consulta geral, como por exemplo crimes envolvendo menores de idade, crimes de natureza sexual, investigação de organizações criminosas ou a quebra de sigilo de dados pessoais.

O volume de informação jurisprudencial brasileira publicado anualmente na rede mundial de computadores é massivo. De acordo com o relatório **Justiça em Números 2016**, publicado anualmente pelo **Conselho Nacional de Justiça** (CNJ) do Brasil, o volume de decisões judiciais e sentenças produzidas no ano de 2015, envolvendo todas as unidades de julgamento do Poder Judiciário, atingiu a marca de **27,2 milhões** (somando todos os três níveis de tribunais existentes), duzentas mil a mais que no ano anterior (BRASIL, 2016).

O relatório Justiça em Números é o resultado de um esforço do CNJ no sentido de coletar dados e elaborar estatísticas que permitam a produção de indicadores de desempenho do Poder Judiciário do Brasil, conhecido como iniciativa Q-Justiça (SERBENA, 2013).

Na Figura 1 está colocada uma série histórica extraída do relatório Justiça em Números 2016, onde é possível observar o volume e a taxa de crescimento anual da produção de decisões judiciais no Brasil, com valores expressos em milhões.

Figura 1 – Série histórica do volume de decisões e sentenças no Brasil no período de 2009 a 2015



Fonte: Adaptado de Relatório Justiça em Números 2016 (BRASIL, 2016)

Pelo gráfico na Figura 1 fica evidente que o volume total de decisões apresenta tendência de crescimento desde o ano de 2010.

Visto que a produção de decisões é da ordem de milhões ao ano, tem-se um volume de decisões armazenadas textualmente que atinge a magnitude de dezenas de milhões, dispersas por diversos tribunais. Cada um destes oferece seus próprios sistemas de busca e recuperação de decisões.

Sistemas de recuperação de informação digital (como é atualmente o caso dos documentos jurisprudenciais) são frequentemente chamados de **motores de busca**. Tais sistemas aceitam um conjunto de critérios de busca e retornam documentos que, sob algum aspecto, sejam correlatos com aqueles critérios (KISHORE et al., 2012; KAVITHA et al., 2015).

No entanto, os motores de busca disponibilizados no âmbito das jurisprudências não oferecem facilidades à altura do volume de conteúdo existente (MOLINARI, 2011), pois:

a) o volume de respostas é frequentemente maior do que o consultante consegue ou tem interesse em administrar (mais de 50);
b) os resultados não vêm ordenados por um critério de relevância semântico (maior relação com as ideias envolvidas) (MOENS, 2001; MOLINARI, 2011), o que é um problema, pois os usuários de sistemas de recuperação frequentemente tendem a selecionar os dez primeiros resultados (DRAGONI et al., 2012);
c) os resultados são obtidos a partir da presença de termos exatos, eventualmente fonetizados, não recuperando conteúdos que tenham significado equivalente, mas que empregam palavras diferentes (ADELMANN et al., 2013).

Como consequência, o resultado das consultas é extenso, contendo muitos documentos com fraca ou nenhuma relação com os interesses do consultante. Esta situação é conhecida como **Problema da Recuperação da Informação** ou simplesmente PRI (MOENS e DE BUSSER, 2002).

Diversas pesquisas apontam para a necessidade de se aprimorar a recuperação jurisprudencial por meio de processos de **busca conceitual**, justificando que neste domínio de documentos que tal modalidade de busca se faz mais relevante (MOENS e DE BUSSER, 2002; MOLINARI, 2011).

A busca conceitual pode oferecer aprimoramentos na indexação de documentos, que então podem ser utilizados para a proposta de melhores motores de busca, reduzindo assim o esforço e o tempo para a localização de um conteúdo específico.

A construção de um **Motor de Busca Semântica** é dependente da existência de uma **ontologia** (KASSIM e RAHMANY, 2009), artefato abstrato que identifica, estrutura, classifica, organiza e relaciona conceitos a respeito de um dado domínio de conhecimento.

Tecnologicamente falando, a ontologia constitui insumo fundamental para a construção de agentes de software automatizados que possam realizar trabalhos

baseados em conhecimento (KASSIM e RAHMANY, 2009; DRAGONI et al., 2012; KAVITHA et al., 2015), que poderiam contribuir para os resultados da iniciativa Q-Justiça (SERBENA, 2013). Esta, por sua vez, precisa ser construída sob critérios particulares para o propósito ao qual será empregada (GRUBER, 1995), o que leva à necessidade de uma **ontologia jurídica voltada à recuperação jurisprudencial**.

Nenhum órgão federal oferece uma ontologia com vistas a recuperação jurisprudencial semântica. O **Superior Tribunal de Justiça** (STJ) oferece em sua página[1] dedicada à jurisprudência, na Internet, um **Vocabulário Jurídico Controlado** que enumera, conceitua e relaciona termos e expressões jurídicos. Em situação análoga, o **Supremo Tribunal Federal** (STF) também oferece sua própria versão do mesmo tipo de recurso[2].

No entanto, nenhum dos dois oferece, efetivamente, uma **ontologia jurisprudencial oficial**, que possa dar subsídio à recuperação jurisprudencial semântica por meio de um **Motor de Busca Semântica**, que é o objeto de estudo desta pesquisa.

1.1 Problematização

Os motores de busca oferecidos pelos tribunais no Brasil fazem uso de uma tecnologia conhecida como *full-text search* (MOLINARI, 2011; BEPPLER e FERNANDES, 2012), que é um recurso frequentemente encontrado em sistemas gerenciadores de banco de dados. Este modelo de indexação e recuperação é baseado na presença de palavras no texto. Trata-se de um **Motor de Busca Estatístico** que somente é capaz de perceber a presença ou ausência de termos exatos em um documento e as suas quantidades de ocorrência. Este tipo de tecnologia de motor de busca é fortemente sensível ao PRI (MOENS e DE BUSSER, 2002).

A alternativa para lidar com o PRI no domínio das consultas jurisprudenciais parece ser um **Motor de Busca Conceitual** ou **Motor de Busca Semântica**, no qual os documentos são indexados não pela mera presença de termos, mas pelos **conceitos** que estes termos identificam no contexto do próprio conteúdo do documento.

---

[1] http://www.stj.jus.br/SCON/thesaurus, acesso em 17 jan. 2017
[2] http://www.stf.jus.br/portal/jurisprudencia/pesquisarVocabularioJuridico.asp, acesso em 17 jan. 2017

Como já colocado, Motores de Busca Semântica empregam um sistema de indexação e recuperação construídos sobre uma **ontologia** (KASSIM e RAHMANY, 2009; KAVITHA et al., 2015), que servem para mapear os conceitos de um domínio, estabelecendo uma base comum de entendimento destes conceitos e de relacionamentos entre estes. Portanto, para a construção de um Motor de Busca Semântica para o domínio das jurisprudências, é necessário se ter uma ontologia de conceitos jurisprudenciais.

Motores de Busca Semântica são objetos de estudo de uma área conhecida como **Web Semântica**, que é um esforço colaborativo liderado pelo **World Wide Web Consortium** (W3C) e que oferece e desenvolve tecnologias para a construção de conteúdo digital, disponibilizado na Internet, baseado em informação e conhecimento (KAVITHA et al., 2015).

A estratégia de indexação semântica aplicada aos aspectos peculiares da língua portuguesa do Brasil vem diretamente ao encontro do suporte informatizado para consultas jurisprudenciais que minorem os efeitos do PRI. No âmbito da consulta jurisprudencial, este é um interesse específico do Poder Judiciário do Brasil.

Assim, dentro do contexto apresentado, é possível compreender que um problema hoje enfrentado pelos usuários de sistemas de consultas à jurisprudência no Brasil é a ocorrência das consequências indesejáveis do PRI no contexto da recuperação de conteúdos jurisprudenciais.

Sob o ponto de vista desta pesquisa e em observação à alternativa tecnológica com maiores oportunidades para a mitigação dos problemas vivenciados pelos usuários de conteúdo jurisprudencial no Brasil, é possível enunciar como problema de pesquisa: **a não identificação dos conceitos presentes em documentos jurisprudenciais por meio dos Motores de Busca Jurisprudencial hoje em atividade no Brasil**.

Ressalta-se que existiram esforços isolados para a construção de ontologias jurídicas e jurisprudenciais no Brasil, sendo um exemplo a ontologia **JurisTJPR** (MOLINARI, 2011). Este esforço efetivamente resultou em uma ontologia jurídica, mas a mesma não tinha o objetivo de mapear a terminologia presente no texto dos documentos jurisprudenciais para conceitos, o que limita sua aplicabilidade em um motor de busca.

Relembrando, em um contexto tecnológico, as ontologias devem ser concebidas e construídas com um propósito particular em vista, para efetivamente se

mostrarem úteis àquele propósito (GRUBER, 1995). Portanto, o insumo indispensável para a construção de um Motor de Busca Semântica (ontologia), com vistas a este fim específico, ainda é um artefato indisponível.

A iniciativa de conceber uma ontologia jurídica exige a colaboração de profissionais da área do Direito dedicados a modelar o conhecimento deste domínio na forma de construções ontológicas.

Se uma alternativa possível seria a de reunir uma equipe que aceitasse o desafio de construir tal ontologia e iniciar o esforço sem qualquer insumo preliminar além do conhecimento técnico, outra possível abordagem seria a de fazer uso de artefatos existentes e já validados por especialistas e estudiosos da área, que pudessem ser sistematicamente empregados para a concepção desta nova ontologia.

Tais artefatos poderiam ser outras ontologias jurídicas e também os tesauros jurídicos. Ambos os artefatos existem e estão disponíveis, a saber, a ontologia JurisTJPR de Molinari (a ser discutida na seção 2.4) e o Vocabulário Jurídico Controlado (Tesauro Jurídico) do Superior Tribunal de Justiça do Brasil.

Frente ao problema em questão e à indisponibilidade de uma ontologia jurisprudencial própria voltada a atender a um Motor de Busca Semântica e Jurisprudencial para o Brasil, a proposta desta pesquisa é a de elaborar e avaliar tal ontologia objetivando a futura construção daquele Motor de Busca.

Da circunstância identificada surge a questão de pesquisa: **quais conceitos precisam compor uma ontologia para que esta suporte um Motor de Busca Semântica que possa reduzir o Problema da Recuperação da Informação no âmbito da recuperação jurisprudencial no Brasil?**

Todos os encaminhamentos de pesquisa e seus esforços consequentes foram delineados com o propósito específico de responder à questão de pesquisa.

1.2 Objetivos

Face ao panorama de consulta jurisprudencial no Brasil, que é caracterizado por motores de busca baseados em *full-text search* e, portanto, que sofrem com as consequências trazidas pelo Problema da Recuperação da Informação, é possível estabelecer o seguinte **objetivo geral** para esta pesquisa: **elaborar uma ontologia que suporte um Motor de Busca Semântica que possa reduzir o Problema da Recuperação da Informação no domínio da consulta jurisprudencial no Brasil**.

Assim, derivados do objetivo geral e em conjugação com o propósito de aproveitar artefatos já existentes e validados, foram definidos os seguintes objetivos específicos:

a) elaborar uma ontologia preliminar a partir do Tesauro Jurídico do STJ;
b) elaborar uma ontologia para a recuperação jurisprudencial pela combinação da ontologia preliminar com a ontologia JurisTJPR proposta por Molinari (2011);
c) construir e avaliar um Motor de Busca Semântica experimental que utilize a ontologia proposta.

Os objetivos específicos estabelecem etapas, representam marcos dentro da pesquisa e contribuem para o atingimento do objetivo geral, além de cooperarem para a estratégia de aproveitamento de esforço já realizado e validado.

1.3 Justificativa

Considerando-se que o já mencionado Problema da Recuperação da Informação torna-se progressivamente mais agudo à medida que o volume de documentos de natureza jurisprudencial aumenta e que as informações oficiais providas pelo Relatório Justiça em Números 2016 indicam clara tendência de crescimento, é evidente que a área de recuperação jurisprudencial somente sofrerá com as limitações particulares da tecnologia empregada pelos motores de busca *full-text search*.

Consonantemente, a presente pesquisa justifica-se pelo fato de o Problema da Recuperação da Informação, no âmbito da Recuperação Jurisprudencial no Brasil, ser ainda experimentado pelos profissionais que dependem deste recurso para o eficaz desempenho de suas atribuições.

1.4 Contribuições

A realização da corrente pesquisa oferece algumas contribuições de diferentes naturezas, colocadas e pormenorizadas nas seções a seguir.

1.4.1 Para a Ciência

Muitas pesquisas vêm sendo realizadas no âmbito da recuperação de informação semântica por estudiosos de todo o mundo. Este volume pode ser constatado em uma pesquisa no Google Acadêmico. No mês de julho de 2016, uma consulta pela expressão *"semantic web"* produziu um resultado de mais de 18 mil respostas, unicamente para artigos publicados a partir do ano de 2015.

Tais estudos focalizam em seu maior volume os conteúdos existentes em língua inglesa. Ocorre que estudiosos de outras origens realizaram avaliações apontando que em diversos casos a eficácia das técnicas é dependente de aspectos linguísticos particulares (AFONSO, 2013).

Visando avaliar as tendências de pesquisas científicas a respeito dos temas relacionados com o problema em estudo, procedeu-se com uma pesquisa bibliográfica integrativa que, dentro do período de 2000 a 2016, buscou por publicações científicas em espectro mundial para a identificação de esforços sobre a temática de recuperação jurisprudencial.

A localização dos artigos, teses e dissertações se deu por meio de repetidas consultas ao portal **EBSCOHost**, **Portal de Periódicos Capes** e por meio do **Google Acadêmico**, todos metamotores de busca. Diversas iterações de consultas foram necessárias porque os Motores de Busca utilizados também estão sujeitos ao PRI.

Os resultados das consultas, por diversos critérios, produziram em uma lista de documentos potenciais que foram avaliados um a um por meio de seus títulos e resumo, o que veio a resultar em um extrato menor. Os documentos presentes neste foram então avaliados por meio da leitura integral do seu conteúdo, de onde um conjunto ainda mais restrito foi atingido.

Finalmente, uma avaliação das referências existentes nos documentos lidos em detalhe levou à procura por algumas delas, o que veio a configurar o conjunto que é apresentado resumidamente a seguir.

As consultas conduzidas no portal **EBSCOHost** incluíram unicamente fontes do período de 2008 a 2016 e que foram avaliadas por pares. Uma gama de resultados foi obtida, muitos deles repetidos em resposta aos diversos critérios. Tais resultados estão apresentados na **Erro! Fonte de referência não encontrada.**, na qual em cada l inha se pode ver a sentença de busca utilizada e o volume de fontes retornadas.

Tabela 1 - Resultados retornados pelo EBSCOHost para temas envolvendo motores de busca e jurisprudência

Sentença de consulta	**Encontrados**
“search engine” AND mining	348
(jurisprudence OR precedent) AND ("information retrieval")	34
(jurisprudence OR precedent) AND ("information retrieval") AND mining	3
Crime AND mining	127
Legal AND mining	433
Legal AND text mining	28
Total	937

Fonte: O Autor (2017)

De forma análoga, a partir do **Portal de Periódicos Capes**, no período de 2008 a 2016, apenas artigos avaliados por pares, diversos resultados foram obtidos, muitos deles repetidos em resposta aos diversos critérios. O resultado das consultas está apresentado na Tabela 2.

Tabela 2 - Resultados retornados pelo Periódicos Capes para temas envolvendo motores de busca e jurisprudência

Sentença de consulta	**Encontrados**
“search engine” AND mining	539
(jurisprudence OR precedent) AND ("information retrieval")	20
(jurisprudência OR precedente) AND ("recuperação da informação")	2
(jurisprudence OR precedent) AND ("information retrieval") AND mining	0
(jurisprudência OR precedente) AND ("recuperação da informação") AND mineração	0
Crime AND mining	338
Legal AND mining	3583
Total	4482

Fonte: O Autor (2017)

A conclusão deste levantamento é de que existe pouca pesquisa realizada no campo da recuperação semântica de conteúdo jurisprudencial, o que sugere uma oportunidade para a exploração da conjugação destes dois temas.

A exploração das técnicas busca semântica (ou conceitual) especificamente aplicadas ao português do Brasil ainda oferece um volume pequeno de publicações, representando também terreno para avaliações e propostas vinculadas especificamente à língua portuguesa brasileira.

1.4.2 Para a Sociedade

A consulta jurisprudencial faz parte do dia a dia de advogados e magistrados, tendo impacto na qualidade, velocidade e custo do julgamento dos processos judiciais.

No entanto, é suprida unicamente pelas técnicas de indexação tradicionais, baseadas em *full-text search*, com expressões exatas e operadores booleanos (MOLINARI, 2011); nenhum tipo de esforço envolvendo a análise de conteúdos assistida por técnicas de indexação conceitual está em uso (MOLINARI, 2011; BEPPLER e FERNANDES, 2012), o que fornece um volume ainda muito grande de resultados para o consumo dos consultantes.

Um novo sistema de indexação e recuperação de documentos jurisprudenciais, baseado em semântica, que reduza os esforços e custos tanto para advogados quanto para juízes, potencializará decisões mais consistentes, obtidas a custo e tempo menores, o que resultará em benefício para toda a sociedade, já que os tempos envolvidos hoje na realidade brasileira são considerados insatisfatórios.

1.4.3 Para o Poder Judiciário

A pesquisa em questão está alinhada a interesses do Conselho Nacional de Justiça do Brasil que visa aprimorar o processo de justiça brasileiro por meio da utilização de suporte tecnológico, nos esforços hoje denominados E-Justiça e Q-Justiça (SERBENA, 2013).

Sendo a ontologia o ponto de partida para a construção de agentes de software autônomos com operação baseada em conhecimento (KAVITHA et al., 2015), a disponibilidade de tal artefato passaria a viabilizar uma série de iniciativas tecnológicas para aprimorar a coleta de dados necessária para as iniciativas E-Justiça e Q-Justiça.

1.4.4 Para o Programa de Pós-graduação em Ciência, Gestão e Tecnologia da Informação

Uma nova ontologia jurídica que viabilize um sistema de indexação e recuperação de conteúdo jurisprudencial representa um esforço na recuperação de informação, que é um dos objetos de estudo do Programa de Pós-graduação em Ciência, Gestão e Tecnologia da Informação (PPGCGTI), estando alinhada aos interesses deste.

A pesquisa também oportuniza a exploração conjugada de áreas da ciência até então pouco estudadas interdisciplinarmente pelo PPGCGTI, a saber, a Ciência da Informação, a Tecnologia da Informação e o Direito.

1.4.5 Para o pesquisador

O pesquisador tem interesse pessoal em temas relacionados a análise de dados, *big data*, Mineração de Dados, Mineração de Texto e Aprendizado de Máquina e Recuperação de Informação. Todos estes têm o potencial de ganhar com a existência de uma ontologia e com as tecnologias da Web Semântica.

O escopo da pesquisa sobre a problemática da consulta jurisprudencial parece oferecer um bom campo de estudos para a exploração destes interesses com resultados para a ciência, sociedade e programa de pós-graduação.

A partir da problemática e contexto apresentados confrontados pela questão e objetivos da pesquisa, as próximas seções apresentam o referencial teórico que explora e delimita os estudos existentes a respeitos dos temas em questão, os encaminhamentos metodológicos articulados para a elaboração de uma resposta às demandas observadas, os resultados atingidos pela consecução de esforços sistemáticos com a respectiva análise e um conjunto de considerações finais a respeito destes, acompanhadas de sugestões para trabalhos futuros.

2 REFERENCIAL TEÓRICO

A presente dissertação discorre a respeito das problemáticas envolvidas na indexação e recuperação de conteúdo jurisprudencial produzido no Brasil, focalizando especificamente a recuperação semântica ou conceitual dos conteúdos jurisprudenciais.

Um motor de busca voltado para a recuperação jurisprudencial apresenta o desafio de precisar fazer uso de uma indexação conceitual em alternativa aos motores *full-text search*, de natureza estatística e que são basicamente baseados na presença de termos (MOENS e DE BUSSER, 2002). Trata-se, portanto, de um **Motor de Busca Conceitual** ou **Motor de Busca Semântica**, que objetiva relacionar palavras e expressões a conceitos e tais conceitos aos documentos.

O tema é objeto de pesquisas ao redor de todo o globo e a solução, quando constituída na forma de um modelo de **Motor de Busca Semântica**, torna-se objeto de estudos da **Web Semântica** e requer uma revisão bibliográfica ampla a respeito de diversos temas de estudo, de forma a identificar tanto caminhos a seguir quando outros a evitar.

Estes estudos abrangem tanto os aspectos genéricos das diversas tecnologias aplicáveis ao problema proposto, quanto esforços de pesquisadores sobre este mesmo tema, com foco específico na análise de suas características, virtudes, restrições e conclusões. Tais conteúdos e esforços estão condensados e apresentados nas seções que integram este referencial teórico.

2.1 Recuperação jurisprudencial

A recuperação jurisprudencial é uma atividade recorrente ao longo das atividades de advogados e magistrados. Estes fazem uso de decisões já tomadas em tribunais para fundamentar seus próprios argumentos.

As seções seguintes apresentam o que são as jurisprudências e os documentos desta natureza, qual sua importância e quais os fatores de importância em sua recuperação.

2.1.1 Conteúdos Jurisprudenciais

A jurisprudência é o conjunto de decisões judiciais uniformes realizadas por juízes e magistrados quando estes avaliam um processo judicial (WEBER, 1999;

MOLINARI, 2011). Este conjunto de decisões consonantes a respeito de temas ou disputas similares estabelece um entendimento e interpretação comuns a respeito de determinado conflito entre partes.

Em países que utilizam o sistema de justiça baseado no modelo *common law* (como nos Estados Unidos e Austrália), as jurisprudências apresentam força de lei (são entendidos como norma jurídica), pois uma vez que uma decisão é tomada em favor de uma determinada ideia ou argumento, esta tende fortemente a ser repetida (MOENS e DE BUSSER, 2002).

No entanto, mesmo em países que implementam o modelo *civil law*, como é o caso do Brasil, as jurisprudências (que representam uma fonte não formal de Direito, sem aplicação obrigatória) são utilizadas tanto para reforçar a argumentação construídas por advogados (estabelecendo diretrizes a se tomar ou evitar no sentido de atingir o sucesso) como a convicção de julgamento dos magistrados (WEBER, 1999; BEPPLER e FERNANDES, 2012). A partir da promulgação da Constituição de 1998, a jurisprudência adquiriu caráter vinculante para decisões do Supremo Tribunal Federal, elevando sua importância como fonte normativa (THEODORO JÚNIOR, 2015).

A sentença é o resultado de um ato decisório do juiz, que encerra o processo judicial, na qual foi proferida sua decisão a respeito de acatar ou não a solicitação da parte titular do interesse de conflito, ou seja, da parte autora do processo. É também entendida como senteça a extinção do processo quando a parte autora declara sua perda de interesse pelo andamento do processo. As sentenças podem ser **terminativas**, quando não avaliam o mérito da causa, ou **definitivas**, quando decidem total ou parcialmente sobre o mérito da causa (THEODORO JÚNIOR, 2015).

Pelos motivos apresentados, as jurisprudências no Brasil são públicas (exceção dos processos classificados como segredo de justiça) e, frequentemente, pesquisadas pelos profissionais do Direito. As jurisprudências podem também ser acessadas livremente pelos cidadãos. No entanto, este tipo de documento é constituído de uma terminologia própria que por vezes é considerada difícil para o leitor leigo, tornando seu conteúdo praticamente ininteligível (CHEN et al., 2013).

2.1.2 Recuperação jurisprudencial

Assim como textos científicos, o volume de documentos de jurisprudência produzido anualmente no Brasil ultrapassa qualquer capacidade de análise humana

(vide Figura 1) e requer assistência computacional para efeitos de seleção criteriosa (BUENO et al., 1999; FERAUCHE e DE ALMEIDA, 2011; MOLINARI, 2011; KAVITHA et al., 2015).

Desde há alguns anos, todo o conteúdo jurisprudencial produzido no Brasil é diretamente disponibilizado na Internet (BUENO et al., 1999). Neste contexto, ferramentas têm sido desenvolvidas na intenção de promover algum tipo de facilidade de filtragem que possa direcionar a atenção dos consultantes para os conteúdos de seu interesse.

Um grande fator de dificuldade no âmbito deste tipo de recuperação de documentos é que muitos dos conceitos importantes para um consultante não estão textualmente presentes no documento. Sua escrita indireta requer que o leitor realize uma **inferência** para a percepção dos mesmos (MOENS e DE BUSSER, 2002).

Apesar de não existir um formato padronizado para a construção de um documento jurisprudencial, existem convenções que são amplamente adotadas, resultando em um conteúdo semiestruturado e semipadronizado, com seções específicas e campos de dados que podem ser utilizados para facilitar a identificação de informações e conceitos relevantes (BUENO et al., 1999; WEBER, 1999). Ainda que esta estrutura efetivamente possa oferecer facilitadores para sua indexação e uma posterior recuperação mais precisa do conteúdo jurisprudencial, tal recurso não necessariamente aproveitado pelos atuais sistemas de busca.

2.2 Motores de busca

Sistemas de recuperação de informação digital (como é o caso dos documentos jurisprudenciais) são frequentemente chamados de **motores de busca** (do inglês, *Search Engine*) (KISHORE et al., 2012; KAVITHA et al., 2015), que nada mais são do que sistemas computacionais especializados em localizar e recuperar informações (codificadas em diferentes tipos de mídia). Tais sistemas aceitam um conjunto de critérios de busca (palavras, frases e operadores) e retornam os documentos que, sob algum aspecto, sejam correlatos com aqueles critérios (KISHORE et al., 2012; KAVITHA et al., 2015).

Estas ferramentas, no entanto, ainda se encontram no estágio de Motores de Busca do tipo *full-text search* (BUENO et al., 1999; MOLINARI, 2011; BEPPLER e FERNANDES, 2012), acumulando todas as suas limitações, notadamente o **Problema da Recuperação da Informação** (MOENS e DE BUSSER, 2002) que é

caracterizado pela incidência de alta taxa de **revogação** e baixa taxa de **precisão** (LANCASTER, 2004).

Aprofundamentos com relação a estes conceitos estão colocados nas próximas seções.

2.2.1 Recuperação de textos

A recuperação de textos é uma especialização do serviço de **Recuperação da Informação**, no qual um conjunto de termos ou critérios de busca é fornecido ao motor de busca que os utiliza para localizar documentos que melhor os representem (MOENS, 2001; KISHORE et al., 2012).

Algumas métricas foram propostas para medir o grau de eficiência de um motor de busca, sendo que as mais comuns são a **revocação** (do inglês ***recall***) e a **precisão** (do inglês ***precision***) (WEBER, 1999; MOENS, 2001; LANCASTER, 2004; KISHORE et al., 2012).

A precisão mede a taxa de relevância geral do resultado retornado por um motor de busca. Outra forma de definir precisão seria a medida da capacidade total de recuperar textos relevantes. A precisão $\mathbf{P}$, para uma dada consulta $\mathbf{Q}$, é dada pela Equação 1.

$$\mathbf{P}(\mathbf{Q}) = \frac{\boldsymbol{contagem\ de\ resultados\ relevantes}}{\boldsymbol{contagem\ de\ resultados\ totais}} \quad \textbf{(1)}$$

Em um sistema ideal, $\mathbf{P(Q)}$ deve valer 1, pois neste caso, todos os resultados trazidos são relevantes ou úteis para o consultante (LANCASTER, 2004).

Já a revocação mede a taxa de acerto dos resultados relevantes, ou seja, quanto dos resultados que efetivamente são do interesse do consultante foram identificados. Alternativamente, revocação poderia ser definida como a capacidade de evitar a recuperação de textos irrelevantes. A revocação $\mathbf{R}$, de uma dada consulta $\mathbf{Q}$, é fornecida pela Equação 2.

$$\mathbf{R}(\mathbf{Q}) = \frac{\boldsymbol{contagem\ de\ resultados\ relevantes\ retornados}}{\boldsymbol{contagem\ total\ de\ resultados\ relevantes}} \quad \textbf{(2)}$$

Novamente, a medida ideal de $\mathbf{R(Q)}$ é 1, pois neste caso todos os resultados relevantes foram retornado pelo motor de busca (LANCASTER, 2004).

Um motor de busca será tanto mais eficiente em matéria de recuperação quanto a precisão e a revocação dos seus resultados se aproximarem de 1 (LANCASTER, 2004).

Frequentemente, motores de busca empregam estratégias de indexação estatísticas ou probabilísticas, baseados na frequência de ocorrência de palavras. Os sistemas que operam segundo tal estratégia são geralmente conhecidos como sistemas *full-text search* (MOENS, 2001).

Neste processo, o que ocorre é que o sistema procura identificar documentos que apresentem as palavras fornecidas como parâmetros de busca, geralmente ordenando os resultados por um critério de relevância estatístico.

Um dos critérios de relevância estatísticos mais utilizado é conhecido como **tf-idf** (ROBERTSON, 2004; WEISS et al., 2010), sigla para "***Term-Frequency-Inverse Document Frequency***".

A formulação desta métrica está apresenta da na Equação 3.

$$\mathbf{tf-idf(i,j) = tf(t_i, d_j) \,.\, idf(t_j)} \qquad \mathbf{(3)}$$

O valor de **tf-idf(i, j)** do termo **i** ($\mathbf{t_i}$) com relação ao documento **j** ($\mathbf{d_j}$) é na verdade dado pelo produto de outras duas métricas, $\mathbf{tf(t_i, d_j)}$ e $\mathbf{idf(t_j)}$. O primeiro fator significa "***Term Frequency***", ou "*Frequência do termo*" e constitui-se basicamente da contagem simples de ocorrência do termo $\mathbf{t_i}$ no documento $\mathbf{d_j}$.

Por exemplo, em um *corpus* (coleção) de **125** documentos, o documento $\mathbf{d_j}$ é formado por **170** palavras, sendo que dentro deste o termo $\mathbf{t_i}$ é a palavra "**casa**" que ocorre três vezes, então o $\mathbf{tf(t_i, d_j)}$ valerá exatamente **3**.

Já o segundo fator significa "***Inverse Document Frequency***" ou "**Frequência Inversa de Documento**" é uma métrica informacional (SPARCK JONES, 1972) e é dado pela Equação 4

$$\mathbf{idf(t_i) = log\left(\frac{N}{n(t_i)}\right)} \qquad \mathbf{(4)}$$

Neste caso, $\boldsymbol{N}$ corresponde a quantidade total de documentos que formam o *corpus* (**125** para o exemplo anterior) e $\mathbf{n(t_i)}$ corresponde a quantidade de documentos do *corpus* que apresentam ao menos uma ocorrência do termo $\boldsymbol{t_i}$ ("**casa**" no caso do

exemplo anterior). Assumindo que o termo “**casa**” ocorra em 67 dos 125 documentos, então o valor do **tf-idf(“casa”, d_j)** é dado por:

$$\mathbf{tf-idf}(\text{"casa"}, d_j) = \mathbf{tf}(\text{"casa"}, d_j) \cdot \log\left(\frac{125}{67}\right) = 3 \cdot \log(1,8657) =$$
$$\mathbf{tf-idf}(\text{"casa"}, d_j) = 0,8125$$

Frequentemente a métrica **tf-idf** participa do processo de recuperação de documentos com o objetivo de valorar a importância dos termos ou expressões indexadas, de forma a influenciar na priorização dos diversos documentos recuperados (ROBERTSON, 2004) ou como peso de atributos em mineração de texto (WEISS et al., 2010).

Tal se dá por que o valor desta métrica aumenta à medida que um dado termo ocorre mais raramente dentro do *corpus*. Neste sentido, o termo em questão tem maior poder **discriminante** e assim é considerado mais caracterizante e **mais relevante**.

Apesar da fama e confiabilidade de que a métrica **tf-idf** goza em atividades de recuperação de informação textual (ROBERTSON, 2004), quando o espaço de busca é composto por um volume expressivo de documentos (centenas de milhares ou mais), sistemas baseados em recuperação estatística frequentemente não correspondem aos anseios de seu usuário.

No contexto onde o volume de documentos no espaço de busca atinge a magnitude de milhões (que é o caso da recuperação jurisprudencial, segundo Figura 1), a mera contagem de palavras isoladas não mais fornece suficiente grau de especificidade nos resultados (MOENS, 2001; MOENS e DE BUSSER, 2002; KAVITHA et al., 2015).

2.2.2 Recuperação de informação legal

A recuperação de informação de natureza legal constitui um difícil desafio tecnológico, visto que o formalismo legal absoluto ainda não foi atingido (SCHWEIGHOFER e WINIWARTER, 1994). Como consequência, os conteúdos textuais de natureza legal ainda sofrem das ambiguidades e imprecisões próprias da linguagem natural (WEBER, 1999), mesmo que com frequente utilização de expressões particulares e padronizadas.

Outro aspecto dificultador desta categoria particular de sistema de recuperação de informação é de que sua precisão não está vinculada à identificação de termos ou expressões, mas sim à identificação de **conceitos** (MOENS, 2001). Um

sistema de recuperação estatístico (*full-text search*), puramente baseado na presença de termos se mostra insuficientemente preciso neste domínio particular de documentos (KAVITHA et al., 2015).

Diante da situação, é necessária a utilização de estratégias adicionais para indexar e recuperar tais documentos, efetivamente vinculando o conteúdo textual aos conceitos de interesse de quem realiza a consulta.

Uma estratégia comumente empregada é a utilização de um tesauro de termos legais (SCHWEIGHOFER e WINIWARTER, 1994). Assim, o usuário fornece um conjunto de palavras e estas remetem a diversas outras, aumentando a abrangência da consulta (WEBER, 1999). Este processo é conhecido como **expansão de consulta** ou, do inglês, *query expansion* (DRAGONI et al., 2012).

2.2.3 Problema da Recuperação da Informação

É comum que em motores de busca do tipo *full-text search*, muitos dos documentos que seriam considerados relevantes para o usuário sejam excluídos da lista de respostas, forçando-o a alterar seus critérios de consulta e repetir o processo por diversas vezes. Esta situação é conhecida como o **Problema da Recuperação da Informação** (PRI) (MOENS, 2001).

A forma mais simples para identificar a ocorrência do PRI é pela observação do comportamento do usuário de um motor de busca. Por exemplo, quando um consultante fornece critérios de busca e recebe diversas respostas, algumas pertinentes e muitas não aderentes ao seu interesse, repetirá a busca variando os critérios de busca. Assim, o consultante repete o processo com diversas configurações de parâmetros e anota os resultados até que se sinta satisfeito.

Sob o ponto de vista da **Ciência da Informação**, no campo da Recuperação da Informação, tal circunstância é caracterizada pela ocorrência de alta **revocação** e baixa **precisão** (LANCASTER, 2004). Quando $\mathbf{R(Q)} = \mathbf{P(Q)} = \mathbf{1}$ tem-se a solução do Problema da Recuperação da informação, pois para uma dada consulta $\mathbf{Q}$ submetida ao motor de busca, todos os resultados relevantes foram retornados ($\mathbf{R(Q)} = \mathbf{1}$) e absolutamente mais nenhum outro ($\mathbf{P(Q)} = \mathbf{1}$).

Se por um lado é importante que o motor de busca identifique mais documentos de interesse para seu usuário e menos documentos sem interesse, é também importante que a indexação seja automática, pois além do volume de conteúdo ser tamanho a ponto de ultrapassar a capacidade humana e qualquer

orçamento para tal, a indexação intelectual (executada por humanos) normalmente deixa de considerar significados menos comuns (SCHWEIGHOFER e WINIWARTER, 1994).

O que se procura em sistemas que sofrem com os efeitos do PRI é localizar documentos que discutam sobre **ideias importantes** identificadas pelas palavras fornecidas, e não simplesmente pela presença daquelas exatas palavras. Ou seja, o que o usuário deseja, muitas vezes, não é uma pesquisa por palavras, mas por temas e/ou **conceitos**.

Portanto, um caminho para a solução para o Problema da Recuperação da Informação parece ser um sistema que seja capaz de identificar ideias e conceitos presentes no conteúdo textual e construir um sistema de indexação documental com base nestes conceitos.

Correntemente, o maior problema na construção de tais sistemas encontra-se na identificação dos conceitos presentes no conteúdo textual. É neste contexto que técnicas e tecnologias providas pela **Web Semântica** ganham importância.

2.2.4 Recuperação semântica de textos

Motores de Busca Conceitual são objetos de estudo da **Web Semântica** (HEFLIN, 2007; DRAGONI et al., 2012; KAVITHA et al., 2015), que fornece tecnologias voltadas à recuperação semântica (ou conceitual) de textos (KASSIM e RAHMANY, 2009). Nesta modalidade de recuperação de informação, os conteúdos de interesse são caracterizados por conceitos e não por palavras (ou termos). Com base em uma indexação conceitual, o que o Motor de Busca faz é converter os parâmetros de busca fornecidos em conceitos e então localizar os documentos que discutem tais conceitos.

A Web Semântica é um esforço organizado pelo **World Wide Web Consortium** (W3C), fundamentado em tecnologias abertas tais como *Resource Description Framework* (RDF), *Ontology Web Language* (OWL), *eXtensible Mark-up Language* (XML), *Simple Knowledge Organization System* (SKOS) e *Universal Resource Identifier* (URI) (KAVITHA et al., 2015).

No contexto da Web Semântica, a informação deve receber significado específico em uma forma tal que soluções automatizadas possam operar de forma mais “inteligente” (HEFLIN, 2007; MARTÍNEZ-GONZÁLEZ e ALVITE-DÍEZ, 2014). Inteligência aqui denota a capacidade de operação que agentes de software passarão

a demonstrar, tendendo a satisfazer mais proximamente às demandas por informação.

Motores de busca que fazem uso de recuperação conceitual ou semântica são chamados de **Motores de Busca Semântica** e adotam técnicas de indexação e representação de conceitos. Estes conceitos, segundo a Web Semântica, são representados computacionalmente e organizados por um artefato abstrato chamado de **ontologia**.

As seções a seguir abordarão os conceitos fundamentais que constituem as tecnologias básicas que dão suporte à recuperação conceitual ou semântica de conteúdo informacional.

2.3 Ontologias

Uma forma de conceituar ontologia foi proposta por Thomas Gruber como sendo “uma especificação explícita de uma conceituação” (GRUBER, 1995, p. 908) em seu artigo versando sobre sistemas baseados em conhecimento, onde sugeria o uso de ontologias na especificação de software. Parreiras (2012) complementa esta definição para indicar que tal especificação deve ser formal.

A conceituação em questão engloba os objetos, conceitos e demais entidades que existem em um domínio de interesse e, muito importante, as relações entre estes. Assim, a ontologia pode ser entendida como uma abstração da realidade, compreendendo os aspectos relevantes desta para o propósito em vista.

Ontologia é objeto de estudos da filosofia, que origina o termo, e lá pode ser entendida como “a ciência do que é, dos tipos e estruturas dos objetos, propriedades, eventos, processos e relações em todas as áreas da realidade” (KASSIM e RAHMANY, 2009). É o estudo filosófico da natureza da existência e por isso promove a identificação e **categorização dos conceitos** em estudo (PARREIRAS, 2012).

Já em termos computacionais, ontologia é um artefato abstrato que armazena e organiza os conceitos relevantes para algum objetivo e assim, representa o conhecimento em termos de categorias e suas propriedades, relações entre estas e também em termos de um vocabulário próprio do domínio (DRAGONI et al., 2012; GRIFFO et al., 2015).

Por meio da formalização trazida pela ontologia, um sistema computacional pode não apenas organizar e representar o conhecimento a respeito de um domínio de interesse, mas também vincular definições textuais, satisfazendo as necessidades

de homogeneidade e consistência conceitual tanto do lado humano quanto computacional do sistema (GRUBER, 1995; KAVITHA et al., 2015).

Ontologias podem ser classificadas quanto a sua especificidade (RAMOS JÚNIOR, 2008; GRIFFO et al., 2015), havendo quatro diferentes tipos de ontologia, a saber:

a) **ontologia fundacional** (*fundational ontology*): define um conjunto de categorias ontológicas livres de domínio;
b) **ontologia de núcleo** (*core ontology*): define um conjunto de conceitos a respeito de um campo de estudos que ainda é suficientemente genérica para representar vários domínios de conhecimento;
c) **ontologia de domínio** (*domain ontology*): define um conjunto de conceitos específicos para um domínio do conhecimento, dando a seus termos conotações peculiares;
d) **ontologia peso leve** (*lightweight ontology*): declara e organiza um conjunto de termos e expressões pertencentes a um dado domínio, servindo basicamente a sistemas de recuperação de informação (GRIFFO *et al.*, 2015).

É possível enumerar alguns motivos para a construção de uma ontologia, tais como (KAVITHA et al., 2015):

a) permitir que uma máquina faça uso de conhecimento em alguma aplicação;
b) dar a diversas máquinas a capacidade de compartilhar o conhecimento;
c) auxiliar na compreensão dos conceitos de um determinado domínio;
d) estabelecer uma base consensual de conceitos a respeito de determinado tema.

Em matéria de sistemas de recuperação de informação, as ontologias cumprem o papel de enriquecer as informações vinculadas aos conteúdos, elevando a eficácia das consultas (DRAGONI et al., 2012; KAVITHA et al., 2015).

A Web Semântica fornece tecnologias para a construção e utilização de ontologias em agentes de software. As seções seguintes discorrem a respeito destas tecnologias.

2.3.1 Resource Description Framework

Os dois primeiros padrões da Web Semântica recomendados pelo W3C foram o *Resource Description Framework* (ou simplesmente RDF) e o RDF Schema (HEFLIN, 2007). O RDF é um modelo de dados que permite construir uma rede semântica que armazena declarações a respeito de um domínio.

A Figura 2 apresenta um exemplo de três declarações de descrevem um aspecto de um dado domínio de estudo. São três frases que registram um aspecto da realidade daquele domínio e assim capturam uma parte do conhecimento relativo ao mesmo.

Figura 2 – Três frases a respeito de um domínio que podem ser representadas por triplas RDF

```
Juiz julga processo
Juiz se chama João Sem-nome
Juiz é uma pessoa
```

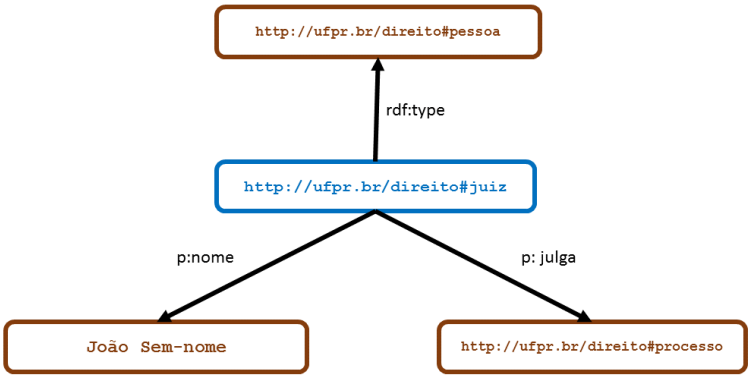
Fonte: o Autor (2017)

Cada uma destas frases está estruturada nos três componentes que constituem o que se costuma chamar de **tripla RDF** (HEFLIN, 2007): o **sujeito**, o **predicado** e o **objeto**.

Nos três casos, o **sujeito** é “Juiz”, pois é a respeito deste que se está registrando conhecimento. O sujeito se relaciona com três **objetos** distintos, respectivamente, “processo”, “João Sem-nome” e “pessoa”. Estes três relacionamentos são estabelecidos pelos três **predicados** intermediários, respectivamente, “julga”, “se chama” e “é uma”.

Tais triplas podem ser organizadas, estruturadas e graficamente representadas na forma de uma rede semântica, ou grafo semântico, e nesta forma correspondem ao mostrado na Figura 3.

Figura 3 - Rede semântica correspondente às três frases a respeito de um domínio



Fonte: o Autor (2017)

Dentro das regras do RDF, o sujeito sempre é encarado como um *resource*, ou seja, um recurso na concepção do W3C, e neste sentido, sempre identificado por um URI (*Universal Resource Identifier*). URI é uma tecnologia recomendada pelo W3C para a identificação não-ambígua de qualquer tipo de objeto.

No caso do exemplo, o sujeito “Juiz” é traduzido para o URI “http://ufpr.br/direito#juiz”, que deve ser entendido como um identificador totalmente qualificado e contextualizado. Neste caso, o nome “juiz” está definido dentro do contexto “http://ufpr.br/direito”, podendo ser diferenciado de outros contextos como, por exemplo, “http://cbf.com.br/membros”, que poderia identificar e atribuir outro significado para “juiz” (“http://cbf.com.br/membros#juiz”).

No caso desta rede existem três diferentes objetos, pois assim também o era nas três declarações anteriores: “http://ufpr.br/direito#pessoa”, “http://ufpr.br/direito#processo” e “João Sem-nome”. Objetos podem ser tanto *resources* (neste caso, sempre identificados por URIs) ou valores literais, neste caso exemplificado unicamente por “João Sem-nome”.

Os arcos que unem os nodos representam os predicados e **sempre partem do nodo sujeito** para o nodo objeto. O exemplo da Figura 3 mostra três diferentes predicados:

a) “p:nome” é um predicado utilizado para identificar uma propriedade do sujeito, neste caso, seu nome, que é um valor literal (“João Sem-nome”);
b) “p:julga” é um predicado que descreve algum tipo relação de domínio com o objeto, descrito como sendo um recurso e identificado por um URI (“http://ufpr.br/direito#processo”);
c) “rdf:type” é um predicado que descreve uma relação primitiva do RDF, neste caso para demonstrar que o sujeito é um tipo particular de um outro objeto (“http://ufpr.br/direito#pessoa”).

As três declarações originais descrevem aspectos de um domínio e a rede semântica é uma representação destes fatos estabelecidos formalmente, em um modelo **passível de processamento computacional**. Por este motivo pode ser objeto de operação de algoritmos de inferência para fins de construção de agentes automatizados de software aplicáveis a variados fins, incluindo Motores de Busca.

Finalmente, existe uma forma de armazenar tal rede em um arquivo digital conhecido como **documento RDF**. Um exemplo deste documento, equivalendo à rede mostrada na Figura 3 está colocado na Figura 4.

Figura 4 - Documento RDF que armazena as tripas RDF correspondentes às três frases a respeito de um domínio

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns:p="http://ufpr.br/direito#">
  <rdf:Description rdf:about="http://ufpr.br/direito#juiz">
    <p:julga rdf:resource="http://ufpr.br/direito#processo"/>
    <p:nome>João Sem-nome</p:nome>
    <rdf:type rdf:resource="http://ufpr.br/direito#pessoa"/>
  </rdf:Description>
</rdf:RDF>
```

Fonte: o Autor (2017)

Trata-se também de um documento XML, outro padrão recomendado pelo W3C. Ressalta-se que a exploração dos detalhes da especificação XML do W3C está além do escopo deste referencial teórico, pois tais documentos têm propósitos gerais de **estruturação e compartilhamento** de dados. Minúcias relativas a tal recomendação estão livremente disponíveis em página própria do W3C[3].

Um documento RDF é precisamente um documento XML constituído de *tags* específicas. O termo ***tag*** vem do inglês e é a forma padrão de nomear um identificador de seção em um documento XML (e, portanto, RDF). Cada seção (ou parte) de um documento XML é chamada de **elemento**.

Um elemento é delimitado por uma *tag* de início (por exemplo, `<rdf:Description>`) e outra de fim (por exemplo, `</rdf:Description>`).

Um elemento do tipo `<rdf:Description>` declara o **sujeito** do conjunto de triplas (“http://ufpr.br/direito#juiz”) que são complementadas pelos sub-elementos. Estes declaram simultaneamente os **predicados** (`<p:julga>`, `<p:nome>`, `<rdf:type>`) e os **objetos** (“http://ufpr.br/direito#processo”, “João Sem-nome” e “http://ufpr.br/direito#pessoa”) que estão vinculados ao sujeito.

Sempre existe um elemento raiz, dentro do qual todos os outros estão contidos (`<rdf:RDF>`) e nele encontra-se a definição do *namespace* XML “p”, que é utilizado como abreviatura (ou *alias*) para uma **definição do domínio** do conhecimento em vias de captura (“http://ufpr.br/direito”), também codificado na forma de URI. Em XML, o propósito da definição de um *alias* para um *namespace* é tornar a

[3] https://www.w3.org/XML, acesso em 17 jan. 2017

escrita e a leitura do documento mais fácil, confortável e menos suscetível a erros. O *alias* permite o uso de um prefixo curto que representa um URI longo.

Nesta forma, é possível armazenar e compartilhar conhecimento codificado a respeito de um domínio particular. A compatibilidade com o formato XML simplifica a adoção do padrão, por ser também uma recomendação W3C largamente aceita e utilizada (HEFLIN, 2007).

Associada à especificação RDF está o RDF Schema, que define diversos tipos de dados primitivos aplicáveis a definições RDF. Um exemplo é “rdf:type”, já visto no exemplo. Alguns outros incluem:

a) “rdfs:Class” e “rdf:Property”, para definição de entidades e propriedades;
b) “rdfs:subClassOf”, “rdfs:subPropertyOf”, “rdfs:domain” e “rdfs:range”, para definição de relacionamento entre entidades;
c) “rdfs:label” e “rdfs:comment”, para documentação.

No entanto, mesmo fornecendo seu vocabulário de primitivas, o RDF Schema demonstrou não ter expressividade semântica suficiente para descrever os conceitos de um domínio na complexidade que os mesmos se apresentam e com capacidade computável desejável (HEFLIN, 2007). Tais limitações levaram o W3C a desenvolver uma evolução que preenchesse esta lacuna, que veio a ser o padrão *Ontology Web Language* (OWL), voltado para a construção de ontologias.

2.3.2 *Ontology Web Language*

A recomendação W3C para a construção de ontologias se dá pela especificação *Ontology Web Language* (OWL), que em 2004 passou a figurar como a principal recomendação W3C para aplicações da Web Semântica (HEFLIN, 2007).

OWL é uma extensão de RDF, de maneira que preserva seus conceitos e os enriquece com outros. O que OWL faz é aumentar o repertório de construções primitivas de RDF, permitindo a construção de taxonomias (estruturas hierárquicas) de entidades e suas propriedades. Estas entidades identificam conceitos e são chamadas de **classes**.

A Figura 5 apresenta uma nova versão, agora codificada em OWL da Figura 2, que codifica as triplas RDF do domínio de exemplo. Neste caso, o documento OWL foi enriquecido com definições próprias deste padrão.

Figura 5 - Documento OWL correspondente ao documento RDF com as três triplas RDF

```
<!DOCTYPE rdf:RDF [
  <!ENTITY owl "http://www.w3.org/2002/07/owl#">]>
  <rdf:RDF xmlns:owl ="http://www.w3.org/2002/07/owl#"
           xmlns:rdf ="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
           xmlns:rdfs="”http://www.w3.org/2000/01/rdf-schema#”
           xmlns:p = “http://ufpr.br/direito#”>
  <owl:Ontology rdf:about="">
    <rdfs:label>Ontologia Exemplo</rdfs:label>
    <rdfs:comment>Exemplo de ontologia de domínio</rdfs:comment>
  </owl:Ontology>
  <owl:Class rdf:ID=”Processo” />
  <owl:Class rdf:ID=”Pessoa” />
  <owl:DatatypeProperty rdf:ID=”cpf” />
  <owl:Class rdf:ID=”Juiz” />
    <rdfs:subClassOf rdf:resource=”#Pessoa” />
  </owl:Class>
  <owl:ObjectProperty rdf:ID=”julga”>
    <rdf:type rdf:resource=”&p;TransitiveProperty” />
    <rdfs:domain rdf:resource="#Processo" />
  </owl:ObjectProperty>
  <owl:AllDifferent>
    <owl:distinctMembers rdf:parseType=”Collection”>
      <p:Juiz about=”#João Sem-nome”/>
    </owl:distinctMembers>
  </owl:AllDifferent>
</rdf:RDF>
```

Fonte: o Autor (2017)

A raiz do documento OWL é a mesma de um documento RDF, pois na prática, o primeiro é um caso particular do segundo. O elemento `<owl:Ontology>` evidencia que se trata de um documento OWL e não apenas RDF.

Em RDF, um objeto anotado com o predicado “rdf:type” estabelece uma relação do tipo generalização-especialização, identificando implicitamente duas categorias de objetos, sendo uma mais abrangente que a outra (HEFLIN, 2007). Em OWL existe uma construção própria para a construção de taxonomias, a classe, identificada por `<owl:Class>`.

Classes OWL representam conjuntos ou categorias de recursos e podem ser organizadas de forma a constituir uma hierarquia, ou seja, uma taxonomia. Classes mais próximas da raiz da taxonomia identificam categorias mais genéricas e as mais próximas das folhas, categorias mais específicas.

No exemplo da Figura 5, a classe “Juiz” é uma subclasse de “Pessoa” (como estava implicitamente colocado no documento RDF), pois naquele domínio todos os juízes são pessoas.

É uma convenção que os nomes de classes em OWL sejam iniciados por letras maiúsculas, seguindo o modelo *Pascal case* para nomes com múltiplas

palavras. Por exemplo, uma classe que identificasse o *homo sapiens* teria o nome “HomoSapiens” (HEFLIN, 2007).

Em OWL uma classe pode ter dois tipos de propriedade. Propriedades de dado servem para caracterizar algum aspecto da classe e podem reter dados de tipos primitivos, como valores numéricos ou textuais. Já as propriedades de objeto são usadas para relacionar uma classe à outra.

No documento OWL em exemplo, foi definida uma propriedade de dados chamada “cpf”. Esta propriedade não existe no exemplo RDF e foi colocada apenas para demonstrar o caso de uma propriedade de dado.

Existe também a propriedade de objeto “julga”, que representa o predicado que une os recursos “Juiz” e “Processo”, expressando a declaração original “Juiz julga processo”.

Como já colocado, o predicado “rdf:type” é representado pela construção OWL “rdfs:subClassOf”.

O último predicado existente tinha o propósito de identificar um juiz específico por seu nome e em RDF era representado por “p:nome”. Em OWL existem construções específicos para a instanciação de classes.

O último elemento do documento (`<owl:AllDifferent>`) é um exemplo de conjunto de instâncias de classe, ou seja, de definição de indivíduos particulares que têm um nome para diferenciá-los. O único indivíduo do exemplo é uma instância identificada da classe “Juiz” e corresponde ao uso do predicado “p:nome” no RDF ou a declaração “Juiz se chama João Sem-nome”.

A especificação OWL, até então na versão 2, é composta por diversas construções que visam dar expressividade inclusive para domínios com relações complexas (HEFLIN, 2007). Não é escopo deste documento esgotar toda a definição da OWL. Esta pode ser encontrada livremente na página do W3C[4].

2.3.3 Engenharia de ontologias

A construção de uma ontologia é um processo de projeto de um artefato, ou seja, as decisões que levam a incluir ou não um determinado conceito em uma ontologia são decisões de projeto (GRUBER, 1995).

[4] https://www.w3.org/OWL, acesso em 17 jan. 2017

As decisões de projeto devem estar relacionadas com os propósitos da ontologia e não com a concepção abstrata e naturalista de uma verdade absoluta (GRUBER, 1995). Em outras palavras, quando se constrói uma ontologia, deve-se ter em vista de que forma e para que fins a mesma será utilizada (PRESSUTI e GANGEMI, 2008).

O processo de desenvolvimento de ontologias já pode ser considerado amadurecido (SIMPERL *et al.*, 2010), contando com diversas metodologias disponíveis, voltadas especificamente para o que se chama de Engenharia de ontologias (GÓMEZ-PÉREZ e SUÁREZ-FIGUEROA, 2009).

Alguns exemplos de metodologias para a construção de ontologias incluem (GÓMEZ-PÉREZ e SUÁREZ-FIGUEROA, 2009; GUERRA, 2012):

a) **TOVE** (*TOronto Vitual Enterprise*): metodologia baseada na lógica de predicados de primeira ordem, que é utilizada para representar atividades, estados, tempos, recursos e custo;
b) **ONIONS** (*ONtologic Integration os Naive Sources*): dá ênfase a ontologias de domínio e fornece diversos métodos para coletar informações de variadas fontes e integrá-las na ontologia;
c) **METHONTOLOGY**: está entre as mais conhecidas e utilizadas (SIMPERL *et al.*, 2010), foi influenciada por processos oriundos tanto da engenharia de software quanto da engenharia do conhecimento e apresenta uma estrutura de ciclo de vida claramente definida, com tarefas e atividades alocadas em fases específicas;
d) **UPON** (*Unified Process for ONtologies*): metodologia que faz uso da *Unified Modeling Language* (UML), com aspectos do *Unified Process* (UP), sendo baseada em casos de uso e com ciclo de vida iterativo;
e) **OTK** (*On-To-Knowledge*): metodologia desenvolvida pela cooperação de diversas entidades europeias com vista a criação de ontologias que serão aplicadas em sistemas de gestão do conhecimento;
f) **DILIGENT**: metodologia voltada a interação de ontologistas em um processo colaborativo e possivelmente distribuído.

Destas, a METHONTOLOGY, a OTK e a DILIGENT foram as que realmente deram ao processo de construção de ontologias o formalismo e a dimensão da engenharia (NEON, 2010).

Todas estas metodologias foram elaboradas imaginando uma circunstância na qual a ontologia seria criada sem nenhum tipo de insumo ontológico (do inglês, *from the scratch*), partindo totalmente do zero (GÓMEZ-PÉREZ e SUÁREZ-FIGUEROA, 2009; NEON, 2010).

Diferentemente desta presunção, foi elaborada a metodologia NeOn, que se define como uma **metodologia baseada em cenários**, com vistas a construção de redes de ontologias e com foco não apenas em construir novas ontologias, mas também na reutilização e adaptação de ontologias e outros recursos ontológicos e não ontológicos já existentes (GÓMEZ-PÉREZ e SUÁREZ-FIGUEROA, 2009; NEON, 2010).

Adicionalmente, a NeOn oferece suporte disponível em suas páginas na Internet, que conta com um livro *online* que detalha todas as suas etapas, tarefas e cenários (GÓMEZ-PÉREZ *et al.*, 2008; NEON, 2010).

Como um dos objetivos declarados na seção 1.2 é o de construir uma nova ontologia jurídica com base na ontologia **JurisTJPR** de Molinari (2011) e no tesauro jurídico do STJ (dois recursos cobertos), a metodologia NeOn foi a selecionada para dirigir os esforços desta pesquisa.

2.3.4 A metodologia NeOn

Diferentemente de outras metodologias, a NeOn não se caracteriza pela existência de um ciclo de vida predefinido e rígido. Ao invés disso, sugere atividades e caminhos aplicáveis em um dos nove cenários que identifica (NEON, 2010). O argumento é de que estes nove cenários cobrem a maioria das situações típicas vivenciadas durante a construção de ontologias, como a reengenharia de ontologias existentes, alinhamento, modularização, localização e culturização (tradução para outras línguas e culturas) e integração com outros recursos, tais como folksonomias e tesauros.

Os nove cenários previstos pela NeOn (GÓMEZ-PÉREZ e SUÁREZ-FIGUEROA, 2009; NEON, 2010) são:

a) **cenário 1** (da especificação à implementação): a nova ontologia é construída sem a preexistência de qualquer recurso ontológico ou não ontológico;

b) **cenário 2** (reutilização e reengenharia de recursos não ontológicos): diretrizes específicas são fornecidas para realizar a construção de ontologias a partir de recursos não ontológicos existentes;
c) **cenário 3** (reutilização de recursos ontológicos): recursos ontológicos existentes são utilizados para a construção de redes de ontologias;
d) **cenário 4** (reutilização e reengenharia de recursos ontológicos): recursos ontológicos existentes são reutilizados e reelaborados para satisfazer novas demandas;
e) **cenário 5** (reutilização e combinação de recursos ontológicos): diversos recursos ontológicos relativos ao mesmo domínio são combinados para a geração de uma nova ontologia;
f) **cenário 6** (reutilização, combinação e reengenharia de recursos ontológicos): similar ao cenário 5, mas com o adicional de efetuar reengenharias nos recursos existentes antes da combinação para construção da nova ontologia;
g) **cenário 7** (reutilização de Padrões de Projeto para Ontologias): diretrizes para a aplicação de Padrões de Projeto para Ontologias;
h) **cenário 8** (reestruturação de recursos ontológicos): suporte para operações de transformação estrutural, como extensão, poda, reorganização de ramos e modularização em ontologias existentes;
i) **cenário 9** (localização – tradução – de recursos ontológicos): adaptação de ontologias para outras línguas e culturas, produzindo ontologias multiculturais.

A metodologia NeOn não apresenta uma instrução específica e claramente delineada para o processo de combinação de duas ou mais ontologias. Em seu cenário 5, a metodologia limita-se a sugerir o uso do **NeOn Toolkit** para a realização desta tarefa.

Outra orientação de uso da NeOn é enquadrar os objetivos e insumos disponíveis em um ou mais cenários previstos e aplicar técnicas, diretrizes, e ferramentas providas pela metodologia em um dos dois modelos de ciclo de vida suportados (GÓMEZ-PÉREZ e SUÁREZ-FIGUEROA, 2009), a saber, desenvolvimento em modelo de cascata (do inglês, *waterfall*) ou modelo iterativo-incremental.

2.3.5 Tesauros e o *Simple Knowledge Organization System*

O termo tesauro tem origem no latim, *thesaurus*, e sua etimologia remete ao conceito de tesouro, neste caso, o tesouro do conhecimento expresso pelas palavras (SILVA, 2013). A expressão se popularizou com a publicação do livro "*Thesaurus of English Words and Phrases*" de Peter Mark Roget, em 1852 (SILVA, 2013). O tesauro de Roget já estava organizado em função de conceitos e não palavras (PEREIRA e BUFREM, 2005).

Tesauros são vocabulários controlados que reúnem e conceituam termos próprios de um domínio do conhecimento, demonstrando utilidade em mecanismos de recuperação de informação baseada em conceitos (MARTÍNEZ-GONZÁLEZ e ALVITE-DÍEZ, 2014). Tais termos são denominados de descritores (SLYPE *et al.*, 1991).

Assim como as ontologias, os tesauros explicam termos e estabelecem relações entre eles (SLYPE *et al.*, 1991), que os aproxima, em estrutura, às ontologias (DRAGONI *et al.*, 2012; SILVA, 2013). Frequentemente, há três tipos de relações presentes em um tesauro:

a) **hierarquia**: descreve uma relação de generalização e especialização entre termos, identificando qual tem significado mais abrangente e qual mais restrito;
b) **equivalência**: estabelece a sinonímia entre termos, identificando termos com significados iguais;
c) **associação** (ou correlação): estabelece que dois termos estão relacionados de uma forma que não é nem hierárquica e nem de equivalência.

São estas relações que permitem enriquecer a percepção semântica dos conteúdos textuais. Os relacionamentos semânticos (equivalência e associação) e taxonômicos (hierarquia) presentes nos tesauros estabelecem sua estrutura e têm uma correspondência direta ou muito próxima com a estrutura de uma ontologia peso leve (MARTÍNEZ-GONZÁLEZ e ALVITE-DÍEZ, 2014).

Devido a sua importância, os tesauros apresentam normas internacionais para sua elaboração (CLARKE e ZENG, 2012), sendo a ISO 2788 de 1986 (ISO 2788:1986 *Documentation – Guidelines for the Establishment and Development of*

Monolingual Thesauri) e a ISO 25964 de 2011, que é dividida em duas partes, a primeira de 2011 (ISO 25964-1:2011 *Information and Documentation – Thesauri and Interoperability with other Vocabularies – Part 1: Thesauri for Information Retrieval*) e a segunda de 2013 (ISO 25964-2:2013 *Information and Documentation – Thesauri and Interoperability with other Vocabularies – Part 2: Interoperability with Other Vocabularies*).

Apesar de estar fundamentado em termos, as relações existentes entre os mesmos conferem aos tesauros uma natureza também conceitual (CLARKE e ZENG, 2012), o que fica transparente no uso intercambiado dos termos “conceito” e “termo” presente nas normas ISO (CLARKE e ZENG, 2012).

O W3C também incluiu o suporte a tesauros em suas tecnologias voltadas para a Web Semântica[5], na forma do *Simple Knowledge Organization System* (SKOS), que opera em conjunto com o RDF (MARTÍNEZ-GONZÁLEZ e ALVITE-DÍEZ, 2014). Assim como a OWL, o SKOS é uma tecnologia estendida do RDF.

2.4 Pesquisas relativas à recuperação jurisprudencial

A construção de sistema de recuperação de informação a partir de um *corpus* (coleção) de documentos legais, e notadamente de um *corpus* de documentos jurisprudenciais, vem sendo objeto de pesquisas há quase duas décadas.

Os primeiros experimentos de interesse foram realizados em 1991, na forma de um sistema chamado NEUROLEX (BOCHEREAU *et al.*, 1991) baseado em **Inteligência Artificial** (IA), mais precisamente uma **Rede Neural Artificial** (RNA) de múltiplos níveis.

O NEUROLEX tinha o objetivo de simular o processo de tomada de decisão de um juiz para, dessa forma, gerar um conjunto de regras de decisão em linguagem natural que serviria para dar suporte a decisões legais por parte da prefeitura de Paris.

As jurisprudências analisadas serviram para treinar a rede, de forma que as regras que determinam a legalidade ou ilegalidade de uma decisão municipal pudessem ser identificadas. O domínio daquelas decisões limitou-se a questões de trânsito.

O modelo de treinamento do NEUROLEX se baseava em um conjunto de quatro parâmetros que poderiam assumir diversos valores. Estes valores eram

[5] https://www.w3.org/2001/sw/wiki/SKOS, acesso em 17 jan. 2017

expressões textuais diretamente retiradas das decisões e serviam como palavras chave que designavam circunstâncias do caso. A rede neural era treinada para identificar as combinações destas expressões textuais e encontrava associações entre as mesmas e o fato de o caso ter sido ou não considerado abuso de poder por parte da prefeitura.

Tratava-se de um modelo fortemente dependente do uso padronizado de certas expressões em certas partes do texto. A violação desta regra poderia levar a incorreta identificação de valores para os parâmetros envolvidos, ameaçando a confiabilidade dos resultados.

Neste mesmo ano, Dick (1991) apresentou um estudo a respeito da necessidade de reimaginar Motores de Busca, para dar a estes a capacidade de consulta conceitual, necessária para se atingir melhores resultados na recuperação de documentos legais.

Dentro desta proposta, a autora discute a importância da representação do conhecimento e dos conceitos para que se possa realizar uma verdadeira recuperação por conceitos. Os **grafos conceituais** foram defendidos como uma ferramenta matematicamente bem fundada, eficiente para a representação de conhecimento oriundo de textos escritos em linguagem natural.

Já no Brasil, Weber (1999) apresentou uma proposta de indexação de documentos jurisprudenciais por meio da técnica de **IA** chamada ***Case-based Reasoning*** (CBR). Em sua tese de doutorado, a autora cunhou a expressão ***Intelligent Jurisprudence Research*** (IJR) para denominar seu processo de recuperação de jurisprudências.

A pesquisa se justificou por que os métodos de consulta à jurisprudência existentes na época (*full-text search* presente nos SGBDs) não faziam uso de conhecimento do domínio jurídico, o que resultava em respostas numerosas e irrelevantes, que constitui o **Problema da Recuperação da Informação** (MOENS, 2001).

A técnica de CBR parte do princípio que problemas similares apresentam soluções similares, o que foi defendido como apropriado para o âmbito de consultas jurisprudenciais. Soluções fundamentadas na técnica de CBR são exclusivamente baseadas em conhecimento do domínio (*knowledge-based indexing*), que era precisamente a inovação pretendida pela pesquisa.

Um aplicativo chamado PRUDENTIA foi construído e experimentado. Este sistema construía um conjunto de índices que descreviam cada um dos documentos do *corpus* e depois comparava os critérios de consulta submetidos a tais índices, para selecionar os documentos mais apropriados. Os índices eram vetores de pares nome-valor, definidos por especialistas em jurisprudência.

Os valores dos índices eram obtidos diretamente dos textos por meio de um processo denominado ***template mining***, que procurava por padrões textuais específicos, sistemática profundamente dependente da estrutura dos documentos indexados.

O sistema PRUDENTIA foi experimentado em um *corpus* de casos jurisprudenciais obtidos no Tribunal de Justiça de Santa Catarina (TJSC). A autora descreveu os resultados como superiores aos fornecidos por processos de indexação textual estatística, mas não apresentou nenhum estudo para demonstrar tal observação.

No mesmo ano, Bueno *et al.* (1999) apresentaram uma nova alternativa ao processo de recuperação de documentos jurisprudenciais brasileiros também pelo uso de **CBR**. Os autores descreveram seu processo como “Recuperação Inteligente de Jurisprudências”. A pesquisa se justificou pelos mesmos motivos apresentados por WEBER (1999).

Um sistema denominado JurisConsulto foi construído e fez uso de um vocabulário controlado de termos e um tesauro próprio da área jurídica para evitar que critérios de busca não indexados (terminologia leiga) fossem submetidos.

O modelo de indexação e recuperação era precisamente o mesmo do PRUDENTIA. Os pares nome-valor também foram definidos por especialistas em jurisprudência.

Para extração dos valores característicos dos índices definidos, o JurisConsulto fez uso da estrutura semipadronizada dos documentos e executava um processo de reconhecimento por três critérios: local de ocorrência no texto, inferência baseada em palavras-chave do Código Penal Brasileiro e terminologia específica, baseada em um tesauro jurídico.

O processo de recuperação foi descrito como baseado em um cálculo da similaridade entre os índices de um documento e os critérios de consulta, levando a um índice de relevância do documento frente àqueles critérios.

O protótipo do JurisConsulto foi experimentado sobre casos jurisprudenciais obtidos no Tribunal de Justiça de Santa Catarina. Os resultados foram comparados com o PRUDENTIA, onde os autores advogaram ter obtido melhores resultados sem, no entanto, apresentarem qualquer tipo de estudo ou estatística para suportar tal conclusão, assim como quais critérios de comparação foram utilizados.

Em seu artigo, Moens (2001) faz um apanhado geral a respeito das técnicas mais modernas, frequentes e promissoras da época, quando aplicadas ao que definiu como **Problema da Recuperação da Informação**.

A autora vincula o PRI ao contexto da recuperação de informação jurisprudencial, quando muitos dos documentos desejados não são retornados, enquanto que outros indesejados o são. Assim, o usuário precisa repetir sua consulta inúmeras vezes, realizando pequenas alterações nos critérios de consulta e vasculhando as respostas à busca de novos itens de interesse.

Moens argumentou que havia pouca pesquisa na área e que este domínio é particularmente complexo, pois os resultados desejados não estão vinculados meramente às palavras presentes nos textos, mas sim aos conceitos que estas expressam. Em função da natureza particular do conteúdo dos textos legais e de seu uso, o PRI mostra-se mais agudo neste contexto.

Várias das técnicas foram descritas na forma de subproblemas, como a recuperação propriamente dita, aprendizado e representação de conhecimento legal, classificação de conteúdos legais, extração de informação, agrupamento de documentos e resumo automático de textos.

No ano seguinte, em seu artigo, Moens e De Busser (2002) abordaram novamente o tema de recuperação de informação e descreveram de forma genérica o que seria um sistema voltado a este fim, então especializando a definição para a recuperação de documentos jurisprudenciais.

À esta época os autores estavam vinculados a um projeto chamado MOSAIC, ainda em andamento, que visava estabelecer um modelo geral para a recuperação de jurisprudências. Este modelo e seu protótipo eram uma clara resposta ao PRI, evidenciando a fraca aplicabilidade de consultas *full-text search*, existente nos SGBDs.

O artigo apresenta os vários aspectos que contribuem para uma decisão judicial (*Facts*, *Factors*, *Issues* e *Theories*), definindo-os como o conteúdo mais

importante de um documento jurisprudencial. Estes aspectos nem sempre estão textualmente expressos nos documentos de decisão, cabendo ao leitor inferi-los.

Seguindo tendências já aplicadas anteriormente, os autores discutiram o potencial do CBR no escopo da recuperação jurisprudencial, mas apontaram como desafiador o processo de extrair as representações dos casos automaticamente a partir do texto das decisões.

Também discorreram a respeito de técnicas que podem ser úteis para extrair do texto os conceitos relevantes, implicitamente presentes em uma decisão judicial: *topic segmentation* (particionar o conteúdo em parágrafos, sentenças e passagens e buscar nestes segmentos os termos importantes), *concept identification*, (identificação de temas mais abstratos a partir de termos específicos, sendo que tais servirão para classificar as decisões e compor o índice de consulta) e *rhetorical structure identification* (emprega técnicas de reconhecimento de linguagem natural para identificar os termos centrais e periféricos existentes na estrutura textual).

Os autores deixaram claro que o sucesso no aprimoramento dos resultados de recuperação jurisprudencial requer a extração de conceitos existentes no texto e que as consultas precisam utilizar os critérios de seleção para identificar tais conceitos.

O artigo de Ferauche e De Almeida (2011) apresenta os resultados de técnicas de Mineração de Texto na construção de classificadores automáticos. O objetivo deste estudo era comparar a qualidade de classificação automática frente à classificação manual, já em execução, a partir de uma ontologia existente.

A base de treinamento foi construída a partir de ementas jurisprudenciais (decisões jurídicas) retiradas de revistas jurídicas e já classificadas por processo intelectual, executado por especialistas do próprio tribunal.

Os documentos foram submetidos ao PRETEXT II para gerar seus vetores de atributos e redução da dimensionalidade do vetor de características. Os resultados foram convertidos para arquivos ARFF, para uso no software de domínio público Waikato Environment for Knowledge Analysis (Weka)[6].

Um conjunto de ementas já classificadas, mas diferentes das utilizadas para treinamento foi submetido para avaliar a qualidade dos classificadores.

[6] Disponível livremente em http://www.cs.waikato.ac.nz/ml/weka, acesso em 17 jan. 2017

Não houve taxas de acerto expressivas, de onde os autores concluíram que a abordagem puramente estatística dos algoritmos foi incapaz de identificar as categorias jurídicas corretas em relação à ontologia utilizada na classificação intelectual.

Uma outra estratégica foi apresentada por Molinari em sua dissertação de mestrado, (MOLINARI, 2011), na qual argumenta que os recursos de consulta jurisprudencial existentes no Brasil são tecnologicamente insuficientes. Por ocasião desta situação, o autor justificou sua pesquisa como uma proposta de solução, apresentando uma ferramenta capaz de superar as deficiências existentes.

O autor também apresentou como situação de problema a necessidade de classificação intelectual das jurisprudências. Neste sentido, sua pesquisa apresenta um método para classificação automática. Unindo a classificação automática a uma ontologia específica, denominada **JurisTJPR**, a dissertação procura apresentar um Motor de Busca baseado em conceitos fundamentado nas tecnologias da Web Semântica e com isso ter uma ferramenta de consulta jurisprudencial mais útil.

A ontologia JurisTJPR modela o conhecimento e as regras que norteiam o ambiente onde ocorrem as decisões. Para tal, foram declaradas e relacionadas **220** classes. Este conhecimento modelado foi persistido na forma de um documento OWL chamado **JurisTJPR.owl**.

O processo de Mineração de Texto foi aplicado para identificar em um *corpus* de ementas jurisprudenciais providas pelo Tribunal de Justiça do Paraná (TJPR), quais destas se relacionam com quais conceitos da ontologia construída, produzindo assim um sistema de indexação por conceitos, o que é uma clara resposta às colocações de Moens e De Busser (2002).

A seção de resultados informa alguns percentuais elevados (89% e 91%) na precisão das consultas, mas o autor informou que não fora possível realizar nenhuma comparação com o sistema então existente no TJPR para estabelecer estatísticas que validassem seus esforços.

Em seu artigo, Beppler e Fernandes (2012) descrevem um sistema experimental para Mineração de Texto de bases jurisprudenciais do TJSC. Os autores justificaram sua pesquisa por fatores de limitação da tecnologia vigente que remetem, uma vez mais, ao PRI.

Basicamente todas as fases de um processo de Mineração de Texto foram utilizadas até se chegar a uma representação tabular do *corpus* em utilização e então executar um algoritmo de extração de **Regras de Associação**.

As regras, da forma **SE x ENTÃO y**, correspondem à saída do sistema e são submetidas ao uso de especialistas da área do direito do próprio Tribunal. Os autores utilizaram a implementação do algoritmo **Apriori** existente no Weka.

O artigo não relata resultados ou avaliações a respeito da qualidade das regras obtidas e nem de que forma as mesmas são utilizadas pelos especialistas, mas reiteram que a qualidade das mesmas poderia ser elevada se o sistema fosse aprimorado para identificar **termos compostos**, muito comuns na área do direito. Eles complementam a observação apontando o uso de um **tesauro especializado** para viabilizar esta melhoria.

Em seu artigo, Chen *et al.* (2013) descreveram um processo para utilização de **Mineração de Texto** para dar suporte a consultas jurisprudenciais da Justiça de Taiwan.

Diferentemente, no entanto, de outras aparentemente existentes soluções, os autores descreveram um processo para permitir tais consultas **a toda a população**. Os autores argumentaram que a Mineração de Texto tem sido largamente utilizada em uma variedade de problemas, mas raramente sobre bases de documentos legais e quando o são, fornecem resultado apenas para profissionais do Direito.

Os autores definem **termo legal** (do vocabulário próprio dos especialistas do direito e que efetivamente é utilizado na construção dos textos legais) e **termo popular** (palavras e expressões utilizadas por leigos, que remetem aos mesmos conceitos dos termos legais).

O artigo descreve um Motor de Busca baseado em Mineração de Texto. Os documentos são organizados em tópicos por um algoritmo de agrupamento (*clustering*) e as buscas são expressas em termos populares, convertidos em termos legais, que efetivamente correspondente às características discriminantes dos documentos agrupados.

Os autores evidenciam que em geral os processos de seleção de características normalmente elegem substantivos para compor o vetor representante, mas no caso de textos legais, outras categorias sintáticas (verbos e adjetivos) também se mostram importantes. Outro fator de aprimoramento dos vetores de características

foi a inclusão de termos relevantes extraídos das leis que normalmente são citadas nos textos jurisprudenciais.

Finalmente, os termos legais que foram identificados como os mais caracterizantes foram submetidos à apreciação de especialistas e a partir de suas opiniões, foram ajustados.

Os autores fizeram uso de uma API da Google que é capaz de calcular um fator de similaridade (***Normalized Google Distance***) entre dois termos, mapeando termos populares para legais.

Assim, um vetor de características é gerado a partir da consulta fornecida em terminologia leiga e este é então comparado com os vetores que caracterizam os documentos jurisprudenciais. Para cada documento indexado, é realizado o **cálculo da distância de cosseno**. Distâncias inferiores a um limiar elegem o documento como participante da resposta e também determinam seu grau de relevância.

Para testar o método, os autores selecionaram notícias policiais de um provedor de notícias confiável e utilizaram uma amostra aleatória destas para servirem como critérios de consulta. De cada notícia foram retirados os termos populares, convertidos em termos legais e então utilizados para a seleção dos documentos.

Para avaliar os resultados, os autores criaram uma enquete que visava medir, a partir de profissionais e estudantes da área de Justiça, qual o grau de satisfação no uso do novo sistema de indexação, quando comparado com o sistema correntemente à disposição.

Após a tabulação dos resultados, um **teste estatístico T** foi utilizado para comparar as respostas e a conclusão estatística é de que a solução proposta retornou melhores resultados na consulta jurisprudencial.

As tecnologias e pesquisas apresentadas neste referencial teórico fornecem embasamento conceitual para compreender as estratégias metodológicas encaminhadas dentro do propósito de conceber uma ontologia especialmente voltada à construção de um Motor de Busca Semântica. Tais encaminhamentos são apresentados e pormenorizados nas próximas seções.

3 ENCAMINHAMENTOS METODOLÓGICOS

Esta seção apresenta os encaminhamentos metodológicos que orientaram a pesquisa e os processos aplicados para o atingimento do objetivo geral já apresentado.

O produto de todas as tratativas e processos encaminhados metodologicamente por esta pesquisa é uma **ontologia peso leve** (vide seção 2.3) denominada **OntoLegis**. Trata-se de uma ontologia voltada especificamente à recuperação jurisprudencial semântica, com vistas a aplicação na construção de um Motor de Busca Semântica.

Esta seção também classifica a pesquisa em questão dentro dos diversos critérios acadêmicos comumente aplicados.

3.1 Caracterização da pesquisa

Os diversos critérios de classificação desta pesquisa (GIL, 2002) estão colocados a seguir.

a) propósito: **pesquisa descritiva** que objetiva descortinar os fatores característicos gerais que norteiam as consultas jurisprudenciais realizadas pelos profissionais do Direito que delas dependem e também **pesquisa explicativa**, que visa avaliar as características necessárias a uma ontologia jurisprudencial para que a mesma possa dar subsídios para a construção de Motores de Busca Semântica no domínio da recuperação jurisprudencial no Brasil;

b) natureza dos dados: **pesquisa qualitativa** no que se refere aos dois principais insumos para a confecção da nova ontologia (tesauro jurídico provido pelo Superior Tribunal de Justiça do Brasil e ontologia Juris TJPR de Molinari); a mesma classificação se aplica às entrevistas feitas a dois advogados consultores entrevistados; **pesquisa quantitativa** no que se refere à avaliação de opinião e experiência de um Motor de Busca Semântica experimental chamado de ConsultaJuris (detalhado na seção 4.5), assim como ao cálculo de estatísticas descritivas utilizadas para análise objetiva da ontologia e sua cobertura de um *corpus* de decisões de teste;

c) obtenção dos dados: quanto aos insumos de entrada para submissão à metodologia NEON, os dados foram obtidos diretamente de seus detentores (Superior Tribunal de Justiça e pesquisador Molinari); no tocante à avaliação quantitativa de opiniões, um instrumento de quatro questões foi incluído junto à ferramenta ConsultaJuris, disponibilizada para experimentação pública na Internet;
d) natureza: **pesquisa aplicada**, pois gera conhecimento de aplicação para resposta a situações de problema identificadas no contexto de estudo explorado.

Entende-se que estes aspectos são compatíveis com o objetivo da pesquisa, com a resolução do problema de pesquisa declarado e com a resposta à questão de pesquisa, todos enunciados na seção 1 desta dissertação.

3.2 Procedimentos

As subseções a seguir detalham cada um dos objetivos específicos estabelecidos na seção 1, assim como as fontes bibliográficas que ofereceram base referencial e teórica para o atingimento de cada um dos mesmos.

3.2.1 Organização geral da pesquisa

Os encaminhamentos descritos nesta subseção foram estabelecidos e organizados visando o atingimento dos objetivos específicos declarados na seção 1.2 e estão sinteticamente apresentados no Quadro 1.

Quadro 1 – Objetivos específicos frente aos encaminhamentos metodológicos realizados e suas motivações

O que	Como	Por quê?
Elaborar uma ontologia preliminar a partir do Tesauro Jurídico do STJ	a) Construção de aplicativo para converter documento de tesauro em arquivo de ontologia; b) Aplicação do cenário 2 da Metodologia NeOn; c) Adaptação da ontologia resultante às necessidades particulares de um Motor de Busca.	a) Tesauro oferece um vocabulário com mais de 13.000 descritores organizados taxonomicamente por profissionais especialistas na área; b) Metodologia NeOn oferece orientações específicas

		para reaproveitamento de recursos não ontológicos; c) Com o tesauro convertido em ontologia, o artefato resultante poderá ser combinado com a ontologia JurisTJPR.
Elaborar uma ontologia para a recuperação jurisprudencial pela combinação da ontologia preliminar com a ontologia JurisTJPR	a) Identificação de conceitos comuns nas duas ontologias; b) Compatibilização de convenções de nomenclatura; c) Utilização do recurso de combinação de ontologias do software de domínio público Protégé 5.	a) A ontologia resultante contemplará tanto os conceitos de vocabulário derivados do tesauro quanto os conceitos que descrevem o ambiente e os processos relacionados produção e consumo de decisões judiciais; b) O Protégé 5 oferece uma funcionalidade pronta para a combinação de ontologias.
Construir e avaliar um Motor de Busca Semântica experimental que utilize a ontologia final elaborada	a) Construção de um aplicativo para recuperar decisões do TJPR; b) Submeter um protótipo de Motor de Busca Semântica a dois advogados consultores; c) Submeter um protótipo de Motor de Busca Semântica a comunidade de usuários; d) Avaliar quantitativamente as características da ontologia frente ao *corpus* de teste.	a) Ter um *corpus* para avaliação da ontologia e do Motor de Busca Semântica; b) As estatísticas resultantes da aplicação da ontologia **OntoLegis** sobre uma amostra significativa de decisões permitirá avaliar seu potencial de identificação de conceitos e formar uma expectativa de sucesso de um Motor de Busca Semântica baseado na mesma; c) A percepção do uso de um Motor de Busca experimental pode fornecer subsídios para correções e aprimoramentos.

Fonte: o Autor (2017)

Tais objetivos foram abordados a partir de um processo articulado e dirigido para os seus atingimentos. Na Figura 6 está colocado um diagrama que resume todos os encaminhamentos executados ao longo da pesquisa, em organização cronológica, que culminaram na elaboração da ontologia **OntoLegis**.

Figura 6 – Etapas do fluxo geral de procedimentos metodológicos realizados ao longo da pesquisa

Fonte: o Autor (2017)

Nesta pesquisa, optou-se por construir a nova ontologia **OntoLegis** pelo aproveitamento de artefatos já existentes, pois estes já estão validados por seus respectivos elaboradores.

A primeira etapa foi dedicada ao tratamento preparatório necessário do **Tesauro Jurídico** do **Superior Tribunal de Justiça**, que é o resultado do esforço de especialistas da **Secretaria de Jurisprudência** daquele tribunal na construção de um

vocabulário controlado abrangente e baseado nos textos das próprias decisões (seção 4.1).

A segunda etapa foi dedicada a converter o Tesauro Jurídico que, por sua organização, estrutura e conteúdo, estava em condição de servir como ponto de partida para uma ontologia peso leve intermediária (RAMOS JÚNIOR, 2008), (seção 4.2).

Na terceira etapa se procedeu com a combinação da ontologia peso leve com a ontologia de domínio **JurisTJPR** de Molinari (2011), que é o resultado da pesquisa daquele autor e que relata que a mesma foi desenvolvida e validada por especialistas da área de jurisprudência do Tribunal de Justiça do Paraná (MOLINARI, 2011). Desta combinação surgiu a ontologia **OntoLegis (**seção 4.3).

Na quarta etapa foi constituído um *corpus* de documentos jurisprudenciais reais obtido a partir do sítio do **Tribunal de Justiça do Paraná**, visando a execução de experimentos para avaliação da ontologia produzida. Os documentos que integram este *corpus* foram obtidos por meio de um aplicativo que visita o sítio do TJPR, realiza consultas da mesma forma que seria feita por um visitante humano, recupera os documentos, interpreta seu formato particular e os armazena localmente (seção 4.5.1).

A quinta etapa consistiu em construir um Motor de Busca Semântica que empregasse a ontologia produzida para operação sobre o *corpus* constituído, realizando assim os experimentos de recuperação baseada em indexação semântica (seção 4.5).

A sexta etapa consistiu em submeter o Motor de Busca Semântica construído a diversas estratégias de avaliação:

a) realizar entrevistas com dois consultores em Direito para apresentar o Motor de Busca Semântica construído e colher suas impressões a respeito do mesmo;
b) experimentar cenários de busca e comparar seus resultados com um motor de busca fundamentando na tecnologia *full-text search*;
c) realizar medições estatísticas para mensurar o grau de cobertura e indexabilidade da ontologia frente a um volume de documentos;
d) submeter o Motor de Busca Semântica a experimentação de um grupo de profissionais da área e colher suas impressões por meio de um questionário.

Especificamente no caso da última forma de avaliação (experimentação por profissionais), a estratégia foi publicar o motor de busca na Internet para facilitar sua experimentação e disponibilizar em conjunto o questionário para coleta de opiniões.

Por tratar-se de uma coleta de opinião, é entendido que tal experimento tem natureza qualitativa. Assim, optou-se por um sistema de amostragem não-probabilística, de tal forma que a amostragem foi dirigida.

O experimento foi noticiado em um grupo de interesse chamado **GEDEL** (Grupo de Estudos em Direito Eletrônico), formado por 273 membros, todos profissionais do Direito atuantes como advogados, juizes e desembargadores. A comunicação feita apresentou brevemente o projeto de pesquisa e convidou seus usuários a experimentarem o motor de busca (denominado de ConsultaJuris) e a responder a um questionário que este oferece.

Além do GEDEL, houve também disseminação pessoal entre pessoas conhecidas que atuam como advogados e, portanto, fazem uso frequente do recurso de recuperação jurisprudencial.

O questionário é um instrumento de coleta de opiniões de quatro perguntas que visa medir, ainda que superficialmente, a impressão deixada pelo recurso de busca semântica. O instrumento em questão foi elaborado para ser intencionalmente curto com o objetivo de estimular ao máximo a adesão de respondentes. A página do ConsultaJuris com mesmo está integralmente colocada no **Apêndice I**.

As questões presentes no questionário foram:

a) **Q1:** “Qual a sua avaliação quanto aos resultados obtidos nas consultas no ConsultaJuris?”, a ser respondida com uma das seguintes opções: “Péssimo”, “Ruim”, “Indiferente”, “Bom” e “Ótimo”;
b) **Q2:** “Quanto aos resultados retornados, quando comparado ao mecanismo de busca do TJPR, você considera o ConsultaJuris?” , a ser respondida com uma das seguintes opções: “Péssimo”, “Ruim”, “Indiferente”, “Bom” e “Ótimo”;
c) **Q3:** “Você retornaria para realizar novas pesquisas no ConsultaJuris?” , a ser respondida com uma das seguintes opções: “Sim” e “Não”;
d) **Q4:** “Espaço para outras análises, sugestões de melhorias, etc.”, que é uma questão aberta.

A proposta original de aplicação do questionário seria de uma única etapa, ou seja, o questionário seria disponibilizado por um tempo, as respostas seriam coletadas e então avaliadas. No entanto, a primeira das respostas ofereceu informação suficientemente importante para promover uma melhoria no motor de busca.

Desta forma, houve duas diferentes versões do motor de busca com duas coletas distintas de opinião. O primeiro período de coleta está compreendido entre 04/01/2017 e 06/01/2017 e o segundo entre 07/01/2017 e 14/01/2017.

O primeiro período ofereceu respostas que influenciaram critérios de valoração das classes seletoras e permitiriam o aperfeiçoamento de um dos processos realizados pelo motor de busca.

Detalhamentos destas avaliações estão apresentados na seção 4.8.1.

3.2.2 Modelo de ciclo de vida empregado na Engenharia de Ontologia

Considerando que a nova ontologia **OntoLegis** deveria ser produzida por meio da reutilização e combinação de artefatos já existentes (ontologia JurisTJPR e tesauro do STJ), optou-se por adotar a metodologia **NeOn**. As condições e objetivos desta pesquisa foram determinantes para a seleção de dois dos nove cenários fornecidos pela metodologia NeOn, a saber:

a) **cenário 2**: reutilização e reengenharia de recursos não ontológicos (tesauro do STJ);
b) **cenário 5**: reutilização e combinação de recursos ontológicos (ontologia Vocabulario-2016-06, proveniente do tesauro do STJ, e ontologia JurisTJPR).

3.2.3 Encaminhamento dos objetivos

Conforme a seção 1.2, o primeiro dos objetivos específicos (objetivo específico (a)) está enunciado da seguinte forma: **elaborar uma ontologia preliminar a partir do Tesauro Jurídico do STJ.**

O atingimento deste objetivo envolveu a transformação do Tesauro Jurídico do STJ (entendido como um **recurso não-ontológico**) em uma ontologia peso leve, realizada por uma série de processos preparatórios para compatibilizar o artefato origem (tesauro) com os propósitos do artefato destino (ontologia).

O segundo objetivo específico está enunciado da seguinte forma: **elaborar uma ontologia para a recuperação jurisprudencial pela combinação da ontologia preliminar com a ontologia JurisTJPR, proposta por Molinari (2011).**

O atingimento deste objetivo foi possível a partir de um esforço de compatibilização entre as duas ontologias envolvidas, ambas existentes na forma de documentos OWL.

O terceiro objetivo específico está enunciado da seguinte forma: **construir e avaliar um Motor de Busca Semântica experimental que utilize a ontologia final elaborada.**

A construção do Motor de Busca Semântica contribui para a avaliação da ontologia final dentro de seu propósito primário e é um dos critérios divisados para este fim. Além desta estratégia, outras duas também foram empregadas:

a) cálculo de métricas da estatística descritiva (totalização, média e desvio padrão), que objetivou principalmente aferir o grau de cobertura que a ontologia **OntoLegis** apresenta em relação a um *corpus* de documentos jurisprudenciais reais;
b) apresentação e experimentação por dois advogados consultores e coleta de suas impressões e sugestões.

Ao operar sistematicamente para o atingimento de cada um dos objetivos específicos que colaboram para o atingimento do objetivo geral desta pesquisa, entende-se que os esforços necessários para a atingimento deste último foram também realizados.

3.3 Alinhamento conceitual

De forma a evidenciar a contribuição que o referencial teórico incluído na seção 2 forneceu para embasar os encaminhamentos metodológicos previstos para a realização da pesquisa, foi elaborado o Quadro 2 no qual estão relacionados os autores principais, o tema de suporte e a atividade de procedimento metodológico realizada.

Quadro 2 - Referencial teórico que dá suporte aos temas de encaminhamentos metodológicos

Referência	Tema	Encaminhamento
Heflin, 2007	RDF e OWL	Compreensão dos conceitos e formatos de documentos para armazenagem de ontologias
Gómez-Pérez e Suárez-Figueroa, 2009 Neon, 2010	Engenharia de ontologias e NEON	Elaboração sistematizada de ontologias
Silva, 2013	Tesauros	Interpretação do Tesauro Jurídico do STJ
Molinari, 2011	Ontologia JurisTJPR de Molinari	Combinação com Tesauro para produção de nova ontologia
Gómez-Pérez e Suárez-Figueroa, 2009 Neon, 2010	Avaliação de ontologias	Avaliação do grau de cobertura da ontologia produzida

Fonte: o Autor (2017)

Diversas outras fontes contribuíram para dar um entendimento apropriado do campo de estudos, das dificuldades inerentes ao mesmo e estratégias de abordagem disponíveis. As mesmas não estão listadas no quadro anterior, mas foram citadas ao longo do texto e condensadas na forma de referências bibliográficas.

3.4 Ferramentas empregadas

Algumas ferramentas foram utilizadas para a consecução de etapas da pesquisa. Tais ferramentas foram selecionadas por familiariedade do autor e por causa de seus recursos de produtividade. As mesmas foram utilizadas em diversos momentos, todos em contribuição ao atingimento dos objetivos específicos já apresentados.

As ferramentas adotadas foram as seguintes:

a) **Microsoft© Word 2016**, foi utilizado para carregar o documento que continha o Tesauro Jurídico do STJ e imprimi-lo em uma impressora virtual PDF (seção 4.1.1);
b) **Foxit© PDF Reader**[7] foi utilizado durante os processos de conversão do Tesauro do STJ texto plano (seção 4.1.1);

[7] Livremente disponível no endereço https://www.foxitsoftware.com/pt-br/downloads, acesso em 17 jan. 2017

c) **Notepad++** versão 5[8] foi utilizado durante os processos de simplificação da versão em texto plano do Tesauro do STJ, pois fornece a facilidade de substituição de textos com **expressões regulares** (espécie de sintaxe simplificada para a especificação de padrões de partículas de texto) e remoção de linhas vazias (seção 4.1.1);

d) **Embarcadero Delphi XE3**, foi utilizado na construção do utilitário para conversão do Tesauro do STJ em sua forma de texto plano para uma ontologia peso leve na forma de um documento OWL (seção 4.2);

e) **Protégé** versão 5.0[9] foi utilizado para combinar a ontologia peso leve intermediária e a ontologia de domínio **JurisTJPR** para gerar a nova ontologia **OntoLegis** (seção 4.3);

f) **Microsoft VisualStudio 2016 Community[10]** foi utilizado na construção de um sítio da Internet que disponibiliza o Motor de Busca Semântica e o questionário para coleta de opiniões a respeito do mesmo, utilizando linguagem **C# 5**, os *frameworks* de desenvolvimento **ASP.NET MVC** versão **4.5.1** e **EntityFramework** versão **6.0** (seção 4.5);

g) **Microsoft SQL Server 2012 Express[11]** foi utilizado para abrigar os documentos do *corpus* indexados semanticamente e outras estruturas de operação do Motor de Busca Semântica (seção 4.5).

A aplicação articulada das ferramentas apresentadas, ordenada pelo fluxo de encaminhamentos anteriormente proposto culminou com a concepção de uma nova ontologia peso leve batizada de **OntoLegis**. Os processos, etapas e preparativos necessários para a consecução de todos estes passos estão descritos na seção 4, a seguir.

[8] Livremente disponível no endereço https://notepad-plus-plus.org, acesso em 17 jan. 2017
[9] Livremente disponível no endereço http://protege.stanford.edu/products.php#desktop-protege, acesso em 17 jan. 2017
[10] Livremente disponível no endereço https://www.visualstudio.com/downloads, acesso em 17 jan. 2017
[11] Livremente disponível no endereço https://www.microsoft.com/en-US/download/details.aspx?id=29062, acesso em 17 jan. 2017

4 RESULTADOS E ANÁLISES

Esta seção descreve as etapas realizadas que tornaram possível o atingimento do objetivo geral apresentado na seção 1, na forma de uma **ontologia peso leve** (GRIFFO *et al.*, 2015) denominada **OntoLegis**, voltada especificamente à recuperação jurisprudencial semântica, com vistas à aplicação na construção de um Motor de Busca Semântica.

As seções a seguir detalham os encaminhamentos levados a efeito para o atingimento dos objetivos específicos desta pesquisa.

4.1 Preparação do Tesauro Jurídico do Supremo Tribunal Federal

O **Vocabulário Jurídico Controlado** do **Superior Tribunal de Justiça** do Brasil (STJ), aqui referido como **Tesauro do STJ**, é mantido pela Secretaria de Jurisprudência daquele órgão e pode ser livremente consultado no sítio eletrônico da Internet[12]. Nesta mesma página encontra-se um *link* para outra com instruções para interpretação dos termos e forma de construção do tesauro.

O tesauro foi obtido por meio de pedido formal encaminhado por e-mail[13] e constitui-se de um documento do Microsoft© Word™ (arquivo digital no formato ".doc") chamado de **Tesauro 06.2016.doc**, correspondendo a um arquivo de **16,3 megabytes** de tamanho, composto de **4525 páginas** e **346.503** palavras, contadas pelo próprio processador de textos.

Trata-se de um documento voltado à leitura de pessoas e não ao processamento de máquina, pois apresenta cabeçalhos e rodapés informativos, assemelhando-o a um relatório.

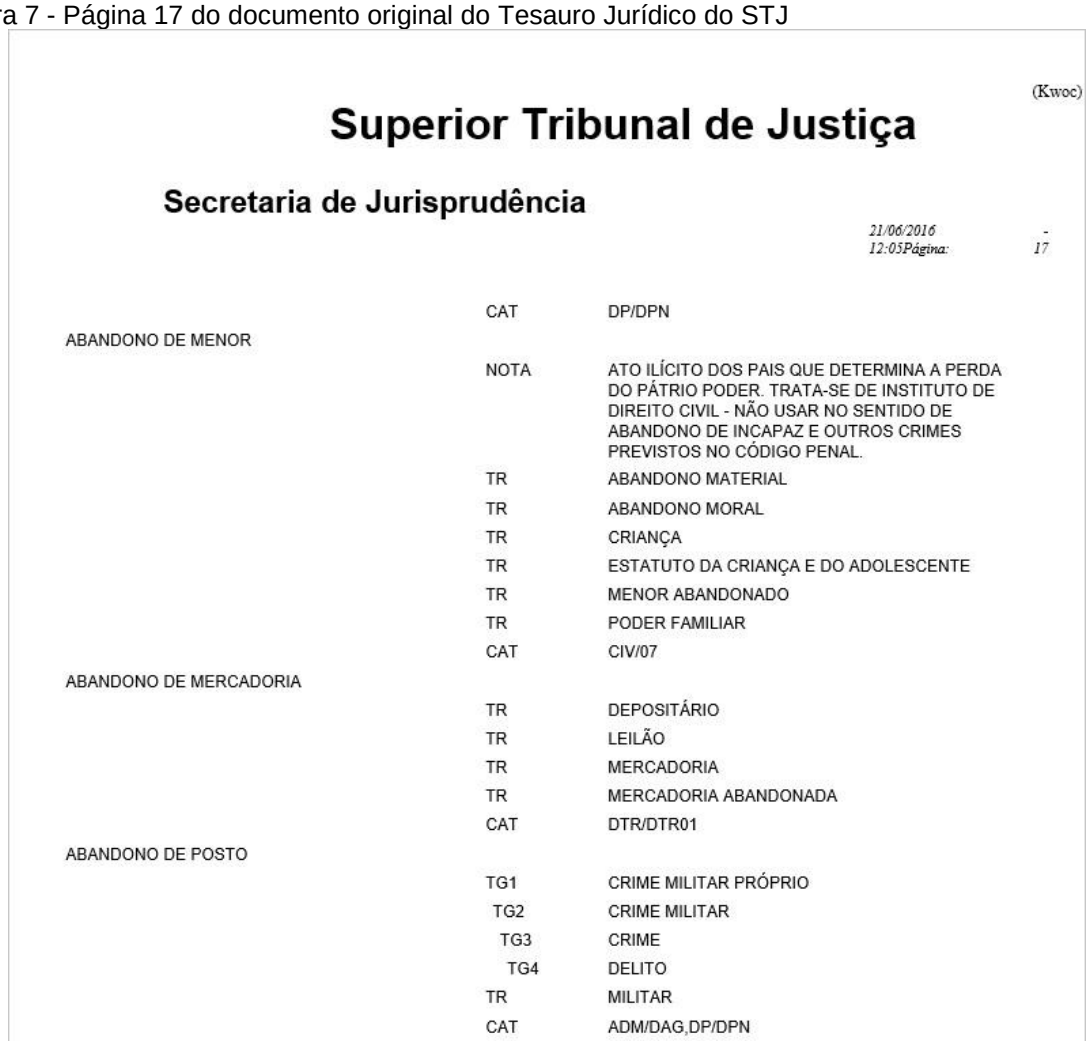
Por não se tratar de um documento em formato amigável ao consumo de agentes automatizados, como seria o caso de um documento XML, seu aproveitamento em alguma aplicação desta natureza requer preparativos semi-automatizados que tirem proveito de sua formatação para extrair o conhecimento de cada uma de suas definições.

Na Figura 7 pode ser visto o exemplo de uma página do documento digital que abriga o tesauro.

---

[12] http://www.stj.jus.br/SCON/thesaurus, acesso em 17 jan. 2017
[13] secretaria.jurisprudencia@stj.jus.br

Figura 7 - Página 17 do documento original do Tesauro Jurídico do STJ



(Kwoc)

Superior Tribunal de Justiça

Secretaria de Jurisprudência

21/06/2016 12:05Página: - *17*

	CAT	DP/DPN
ABANDONO DE MENOR		
	NOTA	ATO ILÍCITO DOS PAIS QUE DETERMINA A PERDA DO PÁTRIO PODER. TRATA-SE DE INSTITUTO DE DIREITO CIVIL - NÃO USAR NO SENTIDO DE ABANDONO DE INCAPAZ E OUTROS CRIMES PREVISTOS NO CÓDIGO PENAL.
	TR	ABANDONO MATERIAL
	TR	ABANDONO MORAL
	TR	CRIANÇA
	TR	ESTATUTO DA CRIANÇA E DO ADOLESCENTE
	TR	MENOR ABANDONADO
	TR	PODER FAMILIAR
	CAT	CIV/07
ABANDONO DE MERCADORIA		
	TR	DEPOSITÁRIO
	TR	LEILÃO
	TR	MERCADORIA
	TR	MERCADORIA ABANDONADA
	CAT	DTR/DTR01
ABANDONO DE POSTO		
	TG1	CRIME MILITAR PRÓPRIO
	TG2	CRIME MILITAR
	TG3	CRIME
	TG4	DELITO
	TR	MILITAR
	CAT	ADM/DAG,DP/DPN

Fonte: o Autor (2017)

Ao topo, em letras negritadas e fonte maior, se pode observar um cabeçalho próprio de documentos informativos ou relatórios. Nas linhas abaixo existem três **descritores** (“**ABANDONO DE MENOR**”, “**ABANDONO DE MERCADORIA**” e “**ABANDONO DE POSTO**”).

A página do próprio sítio eletrônico do STJ[14] informa que um **descritor** é um termo simples ou composto autorizados pelo Tesauro para representar conceitos. Estes descritores fazem uso da terminologia usual dos ministros e é selecionada dos próprios acórdãos presentes na base de jurisprudência do STJ. Já o **não-descritor** é um termo com conceituação equivalente a um descritor, mas cujo uso não é autorizado, servindo apenas para identificar sinonímia.

[14] Acessível pelo endereço http://www.stj.jus.br/SCON/thesaurus/ajuda_thes.jsp, acesso em 17 jan. 2017

Para cada descritor existem termos e expressões relacionadas, categorizadas por meio do uso de prefixos padronizados. A mesma página eletrônica apresenta os significados destes prefixos, colocados a seguir:

a) **TR** que significa Termo Relacionado (ou **associação**) identifica termos e expressões que guardam algum tipo de relação não específica com o descritor, considerada suficientemente importante para ser evidenciada; por exemplo, o descritor "**ABANDONO DE MENOR**" está relacionado com o termo "**CRIANÇA**";
b) **TGx**: que significa Termo Genérico (ou **relação hierárquica**) identifica um termo ou expressão que apresenta um significado mais amplo que abrange o próprio descritor; por exemplo, o descritor "**ABANDONO DE POSTO**" é entendido como um caso particular do descritor "**CRIME MILITAR PRÓPRIO**" (TG1), que por sua vez é um caso particular de "**CRIME MILITAR**" (TG2), caso particular de "**CRIME**" (TG3) que finalmente é um caso particular de "**DELITO**" (TG4); o numeral ao lado do prefixo identifica o grau de distância hierárquica de um descritor em relação a outro;
c) **NOTA** identifica um texto explicativo em linguagem natural que descreve o significado do descritor; por exemplo, o descritor "**ABANDONO DE MENOR**" apresenta uma nota;
d) **TEx** que significa Termo Específico (ou **relação hierárquica**) identifica um termo ou expressão que apresenta um significado mais específico que o descritor; por exemplo o descritor "**DELITO**" (presente na página 1343 do documento) apresenta como termos específicos os descritores "**CONTRAVENÇÃO PENAL**" (TE1) e "**CRIME**" (TE1), ambos com diversos outros termos específicos (TE2) e assim sucessivamente; a exemplo dos termos genéricos, os numerais também identificam a distância hierárquica entre os termos e o descritor;
e) **USE** que significa Equivalência (ou sinônimo), identifica um termo ou expressão que tem significado equivalente ao descritor, mas que é considerado como **não-descritor**; por exemplo, o descritor "**AÇÃO TRABALHISTA**" (presente na página 48) apresenta como termo equivalente "**RECLAMAÇÃO TRABALHISTA**";

f) **UP** que significa Uso Proibido (ou sinônimo), identifica um termo ou expressão que tem significado equivalente ao descritor, mas que é considerado um não-descritor (exatamente como é o caso de “**USE**”); por exemplo, o descritor “**PRINCÍPIO DA ADSTRIÇÃO**” (encontrado na página 1685 do documento) apresenta como um sinônimo proibido a expressão “**PRINCÍPIO DA CONGRUÊNCIA**”;

g) **CAT** que significa Categoria, enumera um conjunto de siglas de documentos legais que guardam relação com o descritor; por exemplo o descritor “**PRINCÍPIO DA ALTERNATIVIDADE**” (encontrado na página 3369 do documento) apresenta a categoria “**DP/DP02**”, ou seja, Código de Direito Penal de 2002.

A página de informações a respeito do tesauro o classifica como polierárquico, ou seja, é possível que um dado descritor seja um termo específico de **mais de um descritor**. Um exemplo encontra-se na página 2127 do documento, onde está apresentado o descritor “**TÍTULO DE CRÉDITO**”, que inclui o **Termo Específico** (TE1) “**CDB**”. No entanto, na página 143 também se encontra “**CDB**” figurando como **Termo Específico** (TE1) do descritor “**APLICAÇÃO FINANCEIRA**”. Ainda, na página 337, encontra-se o termo “**CDB**” sub o rótulo “**TE2**” (Termo Específico) em relação ao termo “**CERTIFICADO DE DEPÓSITO**”.

Este tesauro estabelece uma taxonomia de termos, identificando sinônimos e termos relacionados, o que o habilita a servir como ponto de partida para a construção de uma **ontologia peso leve** (RAMOS JÚNIOR, 2008).

Uma vez que o conteúdo do documento em seu formato original não foi elaborado para consumo de máquina (não está organizado em nenhum padrão de documento voltado ao uso computacional), foi necessário realizar um processamento preparatório de natureza semi-automatizada, descrito nas próximas seções.

4.1.1 Conversão do documento Word para documento de texto plano

A estratégia para conciliar o formato de texto plano com o leiaute do conteúdo do documento (importante para sua correta interpretação) foi alcançada pelo uso conjunto do **Microsoft Word 2016**, **Foxit© PDF Reader** e **Notepad++**. O processo de conversão foi:

a) Abrir o documento **Tesauro 06.2016.doc** no Microsoft Word;
b) Comandar sua impressão na impressora PDF do Foxit PDF Reader, gerando o arquivo **Tesauro 06.2016.pdf**;
c) Abrir o documento PDF recém gerado no Foxit PDF Reader;
d) Comandar seu salvamento na forma de um documento texto, produzindo o arquivo **Tesauro 06.2016.txt**.

Na Figura 8 está apresentada a página 17 do documento em formato de texto plano, gerado pela utilização do Foxit PDF Reader.

Figura 8 - Página 17 do Tesauro Jurídico do STJ após impressão em formato de texto plano

Fonte: o Autor (2017)

O processo produziu um arquivo em codificação UTF-8 e tamanho de 13,5 megabytes.

Com o emprego dos recursos do Notepad++ foi possível simplificar o documento em texto plano por meio dos recursos de remoção de linhas vazias e substituição por expressões regulares. As expressões regulares foram utilizadas para remover os cabeçalhos e redimensionar as endentações, de tal forma que o conteúdo passou a assumir a conformação visível na Figura 9.

Figura 9 - Página 17 do Tesauro Jurídico do STJ em formato de texto plano após simplificações com uso do Notepad++

```
ABANDONO DE LAVOURA
    TG1 CRIME CONTRA A ECONOMIA POPULAR
    TG2 CRIME
    TG3 DELITO
    TR  LAVOURA
    CAT DP/DPN
ABANDONO DE MENOR
    NOTA    ATO ILÍCITO DOS PAIS QUE DETERMINA A PERDA
    DO PÁTRIO PODER. TRATA-SE DE INSTITUTO DE
    DIREITO CIVIL - NÃO USAR NO SENTIDO DE
    ABANDONO DE INCAPAZ E OUTROS CRIMES
    PREVISTOS NO CÓDIGO PENAL.
    TR  ABANDONO MATERIAL
    TR  ABANDONO MORAL
    TR  CRIANÇA
    TR  ESTATUTO DA CRIANÇA E DO ADOLESCENTE
    TR  MENOR ABANDONADO
    TR  PODER FAMILIAR
    CAT CIV/07
ABANDONO DE MERCADORIA
    TR  DEPOSITÁRIO
    TR  LEILÃO
    TR  MERCADORIA
    TR  MERCADORIA ABANDONADA
    CAT DTR/DTR01
ABANDONO DE POSTO
    TG1 CRIME MILITAR PRÓPRIO
    TG2 CRIME MILITAR
    TG3 CRIME
    TG4 DELITO
    TR  MILITAR
    CAT ADM/DAG,DP/DPN
ABANDONO DE RECÉM-NASCIDO
    TG1 PERICLITAÇÃO DA VIDA E DA SAÚDE
    TG2 CRIME CONTRA A PESSOA
    TG3 CRIME
    TG4 DELITO
    TR  EXPOSIÇÃO DE RECÉM-NASCIDO
    TR  RECÉM-NASCIDO
    CAT DP/DPA
ABANDONO DO CARGO
    USE ABANDONO DE CARGO
```

Fonte: o Autor (2017)

O descritor **“ABANDONO DE MENOR”** exemplifica um caso onde a aplicação de expressões regulares foi insuficiente para a completa solução da reformatação do documento. Neste caso, existe um prefixo **“NOTA”** seguido de um texto longo fragmentado em diversas linhas. Para casos como este foi confeccionada uma rotina, implementada em linguagem Delphi (a escolha pela linguagem se deu por familiariedade do autor), para acondicionar todo o texto em uma única linha. O pseudocódigo de tal rotina pode ser visto no **Apêndice H**.

A situação ocorreu com outros **6.746** rótulos. Um exemplo da mesma página após a integração de linhas está apresentado na Figura 10.

Figura 10 - Página 17 do Tesauro Jurídico do STJ em formato de texto plano após integração de linhas

```
677 ABANDONO DE LAVOURA
678     TG1 CRIME CONTRA A ECONOMIA POPULAR
679     TG2 CRIME
680     TG3 DELITO
681     TR  LAVOURA
682     CAT DP/DPN
683 ABANDONO DE MENOR
684     NOTA    ATO ILÍCITO DOS PAIS QUE DETERMINA A PERDA DO PÁTRIO PODER. TRATA-SE DE INSTITUTO DE DIREITO CIVIL - NÃO USAR NO SENTIDO
685     TR  ABANDONO MATERIAL
686     TR  ABANDONO MORAL
687     TR  CRIANÇA
688     TR  ESTATUTO DA CRIANÇA E DO ADOLESCENTE
689     TR  MENOR ABANDONADO
690     TR  PODER FAMILIAR
691     CAT CIV/07
692 ABANDONO DE MERCADORIA
693     TR  DEPOSITÁRIO
694     TR  LEILÃO
695     TR  MERCADORIA
696     TR  MERCADORIA ABANDONADA
697     CAT DTR/DTR01
698 ABANDONO DE POSTO
699     TG1 CRIME MILITAR PRÓPRIO
700     TG2 CRIME MILITAR
701     TG3 CRIME
702     TG4 DELITO
703     TR  MILITAR
704     CAT ADM/DAG,DP/DPN
705 ABANDONO DE RECÉM-NASCIDO
706     TG1 PERICLITAÇÃO DA VIDA E DA SAÚDE
707     TG2 CRIME CONTRA A PESSOA
708     TG3 CRIME
709     TG4 DELITO
710     TR  EXPOSIÇÃO DE RECÉM-NASCIDO
711     TR  RECÉM-NASCIDO
712     CAT DP/DPA
713 ABANDONO DO CARGO
714     USE ABANDONO DE CARGO
715     CAT ADM/DAG,DP/DPL
```

Fonte: o Autor (2017)

Para efeitos de simplificação do algoritmo responsável pela carga do tesauro, as seguintes convenções foram mantidas na estrutura do documento:

a) descritores ocupam uma linha exclusiva e aparecem já no início da mesma;
b) prefixos de detalhes dos termos (“NOTA”, “TR”, “TEx”, “TGx”, “USE”, “UP” e “CAT”) são iniciados como um caractere de tabulação, para diferenciá-los de um termo;
c) os prefixos de detalhes são seguidos por um segundo caractere de tabulação que atua como separador dos seus objetos;
d) os objetos dos rótulos de detalhes ocupam uma única linha.

Neste novo formato, o documento, agora renomeado para **Tesauro-2016-06.txt**, está contido em um arquivo de **2,31 megabytes**, distribuídos em **92.892 linhas** de texto. Foram contados **13.336** diferentes descritores de termos de vocabulário presentes no documento.

As estratégias para a administração destes casos e sua compatibilização com o propósito de derivar uma ontologia estão colocados na próxima seção.

4.1.2 Remoção de descritores sem contribuição informacional

Descritores que não apresentavam contribuição informacional para os propósitos da ontologia foram descartados.

Por exemplo, o tesauro inclui diversos descritores que correspondem a anos, sendo que estes não apresentam quaisquer rótulos a não ser "**CAT**". Alguns destes estão apresentados na Figura 11.

Figura 11 – Exemplo de descritores do Tesauro Jurídico do STJ sem contribuição informacional

```
1902
       CAT    STJ/ANO
1908
       CAT    STJ/ANO
1916
       CAT    STJ/ANO
1932
       CAT    STJ/ANO
1933
       CAT    STJ/ANO
```

Fonte: o Autor (2017)

Ao todo, **139** descritores sem contribuição informacional foram descartados, resultando em **13.197** descritores remanescentes em condições de contribuir para a construção de uma ontologia.

4.1.3 Tratamento de termos de tesauro com complementos

Na Figura 12 é possível ver um exemplo de um descritor com complemento. O descritor é o termo **"ACRE"**, enquanto que o que foi chamado de complemento está entre parênteses, correspondendo à palavra **"AC"**.

Figura 12 – Exemplo de descritor do Tesauro Jurídico do STJ com complemento entre parênteses

```
ACRE (AC)
       UP     ESTADO DO ACRE
       CAT    STJ/UF
```

Fonte: o Autor (2017)

Foram identificados **264** descritores que apresentavam complementos, com diversos usos diferentes para estes, não sendo possível a aplicação de uma convenção única para lidar com todos eles. Por exemplo, em alguns casos tratava-se simplesmente de uma sigla, enquanto que em outros o objetivo era oferecer desambiguação em relação a outro descritor igual.

O aplicativo de conversão de tesauro em ontologia precisou contar com implementações particulares para aplicar as diversas estratégias diferentes de tratamento de complementos.

Em um caso a presença do complemento entre parênteses não oferecia nenhum diferencial particular, apenas identificava um descritor que já estava presente nas relações como um termo sinônimo. Neste caso, tal complemento foi simplesmente removido sem oferecer qualquer perda semântica. Na Figura 13 está mostrado o descritor em questão. Notar que o complemento **"ANELL"** apresentava grafia incorreta, já que visava identificar o descritor **"ANEEL"**.

Figura 13 – Exemplo de descritor do Tesauro Jurídico do STJ cujo complemento foi removido

```
AGÊNCIA NACIONAL DE ENERGIA ELÉTRICA (ANELL)
       UP     ANEEL
       TR     DNAEE
       TR     ENERGIA ELÉTRICA
       TR     MEDIDOR
       TR     SUBESTAÇÃO DE ENERGIA ELÉTRICA
       TR     TARIFA DE    USO DO SISTEMA DE TRANSMISSÃO DE ENERGIA ELÉTRICA (TUST)
       TR     TARIFA DE USO DO SISTEMA DE DISTRIBUIÇÃO DE ENERGIA ELÉTRICA (TUSD)
       CAT    ADM/DAB
```

Fonte: o Autor (2017)

Alguns outros casos também foram solucionados com a simples remoção do complemento, pois identificavam um descritor (**"CRIME"**) já listado como termo genérico. Um exemplo foi **"SUBSTÂNCIA NOCIVA À SAÚDE"**, exemplificada na Figura 14.

Figura 14 – Exemplo de descritor do Tesauro Jurídico do STJ que apresenta o complemento "CRIME"

```
SUBSTÂNCIA NOCIVA À SAÚDE (CRIME)
       TG1    CRIME CONTRA A SAÚDE PÚBLICA
       TG2    CRIME CONTRA A INCOLUMIDADE PÚBLICA
       TG3    CRIME
       TG4    DELITO
       TR     ENVENENAMENTO
       TR     SUBSTÂNCIA TÓXICA
       TR     VENENO
       CAT    DP/DPH
```

Fonte: o Autor (2017)

Houve também o caso particular do complemento "**CONTRAVENÇÃO**", que tem um comportamento bem semelhante ao de "**CRIME**", já discutido. No entanto, o complemento "**CRIME**" remete diretamente a um descritor existente no tesauro, o que não é o caso de "**CONTRAVENÇÃO**". Na Figura 15 está exemplificada uma ocorrência deste caso.

Figura 15 – Exemplo de descritor do Tesauro Jurídico do STJ que apresenta o complemento "CONTRAVENÇÃO”

```
JOGO DE AZAR (CONTRAVENÇÃO)
       TG1    CONTRAVENÇÃO RELATIVA À POLÍCIA DE COSTUMES
       TG2    CONTRAVENÇÃO PENAL
       TG3    DELITO
       TR     BINGO
       TR     CAÇA-NÍQUEIS
       TR     JOGO DO BICHO
       CAT    DP/DPB,DP/DPG,DP/DPZ
```

Fonte: o Autor (2017)

O que diferencia os dois casos é que neste o complemento (“**CONTRAVENÇÃO**”) não identifica textualmente o seu descritor correspondente (“**CONTRAVENÇÃO PENAL**”). No entanto, em todos os casos, o uso do complemento foi feito em um termo que já tinha uma relação taxonômica, quer seja direta quer seja indireta, com o descritor “**CONTRAVENÇÃO PENAL**”, de forma que, a exemplo de “**CRIME**”, não houve perda semântica com sua remoção.

Outras ocorrências se deram pelo uso de complementos para identificar siglas. Na Figura 16 está apresentado um caso como exemplo.

Figura 16 – Exemplo de descritor do Tesauro Jurídico do STJ com duplicação, diferenciado por sigla

```
AGÊNCIA NACIONAL DE SAÚDE SUPLEMENTAR
       USE    AGÊNCIA NACIONAL DE SAÚDE SUPLEMENTAR (ANS)
       CAT    ADM/DAB,DC/DCTH
AGÊNCIA NACIONAL DE SAÚDE SUPLEMENTAR (ANS)
       UP     AGÊNCIA NACIONAL DE SAÚDE SUPLEMENTAR
       UP     AGÊNCIA NACIONAL DE SAÚDE (ANS)
       UP     AGÊNCIA NACIONAL DE SAÚDE
       UP     ANS
       TR     ASSISTÊNCIA MÉDICA
...diversos TR
       TR     TUNEP
       CAT    ADM/DAB,DC/DCTH
```

Fonte: o Autor (2017)

Nestes casos, assim como em outros, existem dois descritores idênticos, diferenciados apenas pela presença do complemento. O tratamento foi combinar os dois descritores, gerando um único em sua forma **sem complemento**, eventualmente adicionando a sigla como um sinônimo. Desta forma, o complemento foi preservado e constará como uma forma alternativa de identificação do conceito, além de o tesauro diminuir de um termo.

Neste caso particular, o descritor sem sigla referenciava o seu correspondente com sigla. Tal referência foi removida, visto que o descritor com sigla passou a inexistir no tesauro.

Outros casos como este ocorreram em **179** descritores, totalizando **180** descritores que foram entendidos como duplicados, diferenciados unicamente pela presença do complemento contendo uma sigla.

A lista completa de todos os descritores com complementos pode ser encontrada no **Apêndice A**.

Nem todos os complementos puderam ser removidos, pois seu uso previa a desambiguação do termo. Por tratar-se de informação com valor semântico, os **84** complementos remanescentes tiveram que ser tratados no nível de ontologia, o que está discutido na seção seguinte.

4.2 Conversão do tesauro na ontologia Vocabulário-2016-06

A conversão de um tesauro em ontologia já conta com um processo que foi padronizado na forma de um *design pattern* (Neon, 2010), presente no sítio **http://ontologydesignpatterns.org**.

Há dois *design patterns* voltados para a conversão de um tesauro em uma ontologia. Um caso é aplicável quando o tesauro se encontra no que foi chamado de **formato baseado em registro** (*record-based data model*) e o outro quando o tesauro se encontra no **formato relacional** (*relational-based data model*). O tesauro do STJ atende ao primeiro formato (baseado em registros).

O *design pattern* nada mais é do que um modelo de convenções que determinam de que forma os vários constituintes do tesauro são mapeados para construções de uma **ontologia peso leve** (RAMOS JÚNIOR, 2008). É um conjunto de convenções que oferece um modelo padronizado de operação, fazendo com que os critérios utilizados no processo sejam sempre os mesmos.

Dentro do que determina o *design pattern*, o processo realizado sobre o tesauro, no formato mais compacto que foi produzido, seguiu as seguintes convenções:

a) os descritores são utilizados para gerar os nomes das classes que os representam dentro da ontologia, sob a notação *pascal-case*, que constitui uma convenção de boa prática em OWL (HEFLIN, 2007);
b) para cada uma das classes produzidas a partir dos descritores do tesauro é produzida uma instância com o mesmo nome; esta recomendação não estava presente no *design pattern*, mas é necessária por que as relações

em OWL não podem ocorrer entre classes, apenas entre instâncias (também chamadas de indivíduos) da mesma;

c) os descritores são mapeados para anotações do tipo **"rdfs:label"**, em sua versão em minúsculo, estarão nesta forma para efeitos de indexação e recuperação; a partir deste momento, os descritores passam a ser chamados de **rótulos**.

d) descritores que não apresentam **Termos Genéricos** (**TG1**) produzem classes OWL que não apresentam superclasses;

e) descritores que apresentam **Termo Genérico de nível 1** (prefixo "**TG1**") produzem classes OWL que têm como superclasse a classe correspondente ao TG1;

f) descritores que apresentam **Termos Específicos** (prefixos "**TEx**") foram ignorados por que as construções OWL modelam relações com superclasses (mais genéricas) e não subclasses (mais específicas); no entanto, não existe perda semântica alguma em tal sistemática;

g) **Termos Relacionados** a descritores (prefixo "**TR**") produzem classes (que poderão ou não apresentar superclasses) e são vinculadas à instância de classe do descritor por meio de uma propriedade de objeto (forma como a OWL modela uma relação entre instâncias de classes) denominada "**relacionadoCom**"; o *design pattern* recomenda uma relação chamada "**relatedClass**", mas como a relação é na prática entre instâncias, considerou-se o nome sugerido como inadequado;

h) **Termos Equivalentes** (sob os prefixos "**UP**" e "**USE**") podem ou não produzir classes; quando o termo equivalente identifica um descritor do tesauro (que produzirá uma classe) então esta relação é modelada com o relacionamento "**owl:equivalentClass**"; por outro lado, quando o termo equivalente corresponde a um não-descritor, este é representado por uma anotação do tipo **"rdfs:label"** na instância de classe do descritor;

i) os prefixos "**NOTA**" e "**CAT**" do tesauro não contribuem para o propósito da ontologia e por isso foram ignorados.

O vínculo entre dois termos (termo relacionado) é resolvido pela instanciação da relação "**relacionadoCom**", como informado no item "g". Tal construção não é nativa na OWL e é representada por meio de uma propriedade de objeto

(“**ObjectProperty**”) declarada no início do documento OWL. Na Figura 17 pode ser vista a declaração desta propriedade.

Figura 17 – Declaração da propriedade de objeto "relacionadoCom" no documento Vocabularario-2016-06.owl

```
<owl:ObjectProperty rdf:about="relacionadoCom">
   <rdfs:domain rdf:resource="http://www.w3.org/2002/07/owl#Thing"/>
   <rdfs:range rdf:resource="http://www.w3.org/2002/07/owl#Thing"/>
</owl:ObjectProperty>
```

Fonte: o Autor (2017)

Para a realização do processo de aplicação do *design pattern* foi construído um aplicativo em linguagem Delphi que tem como entrada o tesauro e como saída um documento OWL, chamado de **Vocabulario-2016-06.owl**. O processamento deste aplicativo foi precisamente a aplicação das regras do *design pattern*.

Para uma melhor compreensão do processo, tome-se, por exemplo, os dois descritores presentes na Figura 18.

Figura 18 – Fragmento do Tesauro Jurídico do STJ a ser convertido em ontologia no formato OWL

```
ABANDONO COLETIVO DE TRABALHO
       TG1    CRIME CONTRA A ORGANIZAÇÃO DO TRABALHO
       TG2    CRIME
       TG3    DELITO
       TR     PARALISAÇÃO DE TRABALHO
       TR     SERVIÇO DE INTERESSE COLETIVO
       CAT    DP/DPD
ABANDONO DA CAUSA
       UP     ABANDONO DO PROCESSO
       UP     ABANDONO DE CAUSA
       TR     AUSÊNCIA INJUSTIFICADA
       TR     DESÍDIA
       TR     PEREMPÇÃO
       TR     PROCESSO JUDICIAL
       CAT    CPC/CPC06,CPC/CPC08
```

Fonte: o Autor (2017)

Após o processamento realizado, os descritores e seus respectivos prefixos produziram as classes OWL mostradas na Figura 19. Cores foram utilizadas para facilitar a identificação dos mapeamentos entre os dois tipos de artefatos.

Os nomes das classes correspondem diretamente aos descritores, sob a convenção *pascal-case*. Por exemplo, o descritor “**ABANDONO DE CAUSA**” produziu uma classe cujo identificador (“**rdf:ID**”) é “**AbandonoDeCausa**”, assim como sua instância.

No entanto, para efeitos de indexação e recuperação, não é o nome da classe que tem utilidade. A forma textual original do descritor, que é o que vai figurar no texto dos documentos (e também nos critérios de busca usados pelo motor de busca) foram preservadas na forma de um **rótulo**, armazenado pelo elemento **`<rdfs:label/>`**, em

letras minúsculas, pois assim favorecem o processamento durante a execução da indexação e da busca.

Figura 19 – Fragmento de ontologia gerada a partir do Tesauro Jurídico do STJ

```
<owl:Class rdf:ID="#AbandonoColetivoDeTrabalho">
     <rdfs:subClassOf rdf:resource="#CrimeContraAOrganizaçãoDoTrabalho"/>
     <rdfs:label>abandono coletivo de trabalho</rdfs:label>
</owl:Class>

<owl:NamedIndividual rdf:about="#AbandonoColetivoDeTrabalho">
     <rdf:type rdf:resource="#AbandonoColetivoDeTrabalho"/>
     <relacionadoCom rdf:resource="#ParalisaçãoDeTrabalho"/>
     <relacionadoCom rdf:resource="#ServiçoDeInteresseColetivo"/>
</owl:NamedIndividual>

<owl:Class rdf:ID="#AbandonoDaCausa">
     <rdfs:label>abandono da causa</rdfs:label>
     <rdfs:label>abandono de causa</rdfs:label>
     <rdfs:label>abandono do processo</rdfs:label>
</owl:Class>

<owl:NamedIndividual rdf:about="#AbandonoDaCausa">
     <rdf:type rdf:resource="#AbandonoDaCausa"/>
     <relacionadoCom rdf:resource="#AusênciaInjustificada"/>
     <relacionadoCom rdf:resource="#Desídia"/>
     <relacionadoCom rdf:resource="#Perempção"/>
     <relacionadoCom rdf:resource="#ProcessoJudicial"/>
</owl:NamedIndividual>
```

Fonte: o Autor (2017)

O descritor "**ABANDONO COLETIVO DE TRABALHO**" foi mapeado para a classe "**AbandonoColetivoDeTrabalho**". Como o descritor apresenta o **Termo Genérico** (TG1) "**CRIME CONTRA A ORGANIZAÇÃO DO TRABALHO**", este foi utilizado para identificar sua superclasse, que é representada pela construção OWL **<rdfs:subClassOf rdf:resource="#CrimeContraAOrganizaçãoDoTrabalho"/>**.

4.2.1 Descritores de taxas

Alguns descritores do tesauro identificam taxas, apresentando inclusive símbolo de percentual, como pode ser observado na Figura 20. Descritores com estes nomes não são válidos para uso em um documento OWL e por isso foram modificados.

Figura 20 – Exemplo de descritores do Tesauro Jurídico do STJ para a descrição de taxas

```
14,20%
       UP     QUATORZE VÍRGULA VINTE POR CENTO
       CAT    STJ/PER
14,36%
       NOTA   REFERE-SE AO PERCENTUAL DO IPC UTILIZADO NO MÊS DE FEVEREIRO 1986 ...
       CAT    STJ/PER
```

Fonte: o Autor (2017)

Por exemplo, o descritor “**14,20%**” foi incluído como classe na ontologia, pois apresenta um sinônimo, mas o nome da mesma não pôde incluir o símbolo de percentual e por isso foi denominada “**Percentual14_20**”, onde o sinal de pontuação vírgula foi substituído por um símbolo de sublinhado para satisfazer restrições de nomes do formato OWL.

O tesauro já inclui o descritor “**TAXA**” que apresenta alguns Termos Específicos. De forma a enriquecer as relações semânticas, a classe “Taxa” foi utilizada como ancestral de todas as classes de taxas específicas. Neste caso, uma relação de hierarquia que não estava originalmente presente no tesauro foi estabelecida.

4.2.2 Comentários à conversão de tesauro em ontologia

Todo o processo de conversão de um artefato para outro (de tesauro para ontologia) não chegou a ocupar dois minutos de processamento em um hardware equipado com um processador Intel® Core™ i7 modelo 960 de 3.2Ghz, equipado com 4 gigabytes de memória RAM.

Foram produzidas **13.197** classes a partir dos descritores do tesauro. No entanto, houve **84** descritores do tesauro que mantiveram seus complementos, pois a retirada destes implicava em desambiguação de natureza semântica e não léxica.

Assim, algumas classes mantiveram seus complementos entre parênteses, que foram deixados para ser tratados em nível de ontologia. Tal processo está descrito na próxima seção.

4.2.3 Desambiguação semântica da ontologia

Houve a necessidade de equipar o aplicativo com uma interface especial para realizar a desambiguação de classes derivadas de termos com complementos que existiam para este fim e que por isso não foram removidos do tesauro. Na Figura 21 é mostrado um exemplo deste caso.

Figura 21 – Exemplo de termos do Tesauro Jurídico do STJ desambiguados pelo uso do complemento "CRIME"

```
INCÊNDIO
       TR     DESABAMENTO
       TR     FENÔMENO DA NATUREZA
       TR     FOGO
       TR     INUNDAÇÃO
       TR     PERIGO DE DESABAMENTO
       TR     PERIGO DE DESABAMENTO (CONTRAVENÇÃO)
       TR     QUEIMA CONTROLADA
       TR     QUEIMADA
       CAT    DPV/DPV3,DAMB/AM08
INCÊNDIO (CRIME)
       TG1    CRIME DE PERIGO COMUM
       TG2    CRIME CONTRA A INCOLUMIDADE PÚBLICA
       TG3    CRIME
       TG4    DELITO
       TE1    INCÊNDIO CULPOSO
       TR     DEPÓSITO DE COMBUSTÍVEL
       TR     DEPÓSITO DE EXPLOSIVO
       TR     DEPÓSITO DE INFLAMÁVEL
       TR     INUNDAÇÃO (CRIME)
       TR     MATERIAL DE SALVAMENTO
       TR     PERIGO
       TR     SALVAMENTO
       CAT    DP/DP07,DP/DPH,DPP/DPP08
```

Fonte: o Autor (2017)

Há dois descritores para o termo "**INCÊNDIO**", cada um para um diferente sentido que foi identificado para as ocorrências da palavra "incêndio". A diferença semântica é evidente pois os termos genéricos e relacionados são bem distintos. Na primeira forma o termo está vinculado a diversos fenômenos naturais ou ao menos, não intencionais ("**DESABAMENTO**", "**FENÔMENO DA NATUREZA**" etc.), enquanto que na segunda forma, anotada com o complemento "**CRIME**", a taxonomia construída identifica sua natureza criminosa.

Visando solucionar casos como este, a parte automatizada do processo de conversão de tesauro em ontologia preservou o complemento, anotando a classe com o mesmo. No entanto, esta anotação não será compatível com o formato OWL que é a destinação final da ontologia.

Também foi considerado que um processo automatizado não seria eficaz para tratar todos os casos. Por este motivo, o aplicativo de conversão contou com uma interface que permitiu a renomeação das classes ou a mera remoção dos complementos, conforme o que foi considerado mais apropriado.

No caso da Figura 21, a segunda forma da classe foi renomeada para "**IncêndioCriminoso**". Desta maneira, passaram a existir duas classes identificadas pelo mesmo rótulo, que é o que sucede na realidade, demonstrando que a ontologia já passa a representar uma situação verificada do domínio do problema. Nesta nova

denominação, a classe não teve mais necessidade do complemento diferencial e este foi suprimido.

A opção de transformar o nome da classe usando a partícula "**Criminoso**" visa satisfazer uma boa prática da OWL que é a de dar nomes significativos às classes (HEFLIN, 2007). A mera concatenação de palavras, por exemplo "**IncêndioCrime**" atingiria o objetivo de gerar uma outra classe para solucionar a presença do complemento, mas foi considerada desinteressante diante das convenções de elegância da língua portuguesa no Brasil.

Assim como na realidade, a única forma de determinar se o termo "**incêndio**" está denotando uma classe ("**Incêndio**") ou outra ("**IncêndioCriminoso**") é pela análise do contexto do seu uso.

Caso similar se deu com os casos do complemento "**CONTRAVENÇÃO**" que foram remanescentes do tesauro. Na Figura 22 está mostrado um exemplo.

Figura 22 – Exemplo de termos do Tesauro Jurídico do STJ desambiguados pelo uso do complemento "CONTRAVENÇÃO"

```
EMISSÃO DE FUMAÇA
       TG1    POLUIÇÃO DO AR
       TG2    POLUIÇÃO
       TR     EMISSÃO DE GÁS
       TR     EMISSÃO DE VAPOR
       TR     ESCALA RINGELMANN
       CAT    ADM/DAJ,DAMB/AM01,DAMB/AM06,DAMB/AM08,DAMB/
EMISSÃO DE FUMAÇA (CONTRAVENÇÃO)
       TG1    CONTRAVENÇÃO REFERENTE À INCOLUMIDADE PÚBLICA
       TG2    CONTRAVENÇÃO PENAL
       TG3    DELITO
       TR     EMISSÃO DE GÁS (CONTRAVENÇÃO)
       TR     EMISSÃO DE VAPOR (CONTRAVENÇÃO)
       CAT    DP/DPZ
```

Fonte: o Autor (2017)

Nestes casos, optou-se por fazer uso da mesma convenção utilizada para "**CRIME**", ou seja, a desambiguação foi realizada pela renomeação da segunda forma (anotada com o complemento) para "**EmissãoDeFumaçaCriminosa**". A palavra "**Criminosa**" foi adotada neste caso uma vez que não existe na língua portuguesa um adjetivo que indique a condição de ser contravenção. Além disso, de acordo com Garcia (2015) a contravenção é também um crime:

> Não há diferença ontológica (de essência ou substancial) entre crime e contravenção, possuindo a mesma natureza. A distinção está em que as contravenções são condutas que apresentam menor gravidade, quando comparadas com os crimes. Por isso, a sanção penal das contravenções é menos severa do que a punição dos crimes. Cabe ao legislador determinar quais são os crimes e as contravenções, conforme a relevância dos interesses jurídicos para a sociedade. (GARCIA, 2015, p. 155)

No **Apêndice B** encontra-se a lista de todas as operações de renomeação de classes que foram realizadas na ontologia.

Com estas operações, uma ontologia peso leve com o vocabulário jurídico derivado a partir do tesauro do STJ passou a existir, chamando-se **Vocabulario-2016-06**, construída a partir da versão de **junho de 2016** do tesauro.

No entanto, a mesma precisou ser simplificada, para fins de uso com um motor de busca. Pormenores deste processo estão descritos na seção seguinte.

4.2.4 Simplificação da ontologia jurídica derivada

A derivação da ontologia jurídica a partir do tesauro do STJ produziu **3.915** classes que apresentavam nomes diferentes, mas que compartilhavam total ou parcialmente os **mesmos rótulos e relações**, sendo em muitos casos explicitamente sinônimas.

As classes foram organizadas a partir de seus rótulos comuns em **1.652** grupos. Na Figura 23 pode ser observado um exemplo de um destes grupos.

Figura 23 – Exemplo de grupo de classes que compartilham o mesmo rótulo

```
<owl:Class rdf:ID="CompensaçãoDeCréditoTributário">
     <rdfs:subClassOf rdf:resource="#ExtinçãoDoCréditoTributário"/>
     <rdfs:label>compensação de crédito tributário</rdfs:label>
     <rdfs:label>compensação de tributo</rdfs:label>
     <rdfs:label>compensação do crédito tributário</rdfs:label>
     <rdfs:label>compensação tributária</rdfs:label>
</owl:Class>

<owl:NamedIndividual rdf:about="#CompensaçãoDeCréditoTributário">
     <owl:type rdf:about="#CompensaçãoDeCréditoTributário">
     <relacionadoCom rdf:resource="#CréditoPresumido"/>
     <relacionadoCom rdf:resource="#CréditoRemanescente"/>
     <relacionadoCom rdf:resource="#CréditoVincendo"/>
     <relacionadoCom rdf:resource="#DébitoTributário"/>
     <relacionadoCom rdf:resource="#DeclaraçãoDeCompensação"/>
     <relacionadoCom rdf:resource="#EncontroDeContas"/>
     <relacionadoCom rdf:resource="#SubstitutoTributário"/>
</owl:Class>

<owl:Class rdf:ID="CompensaçãoDoCréditoTributário">
     <rdfs:label>compensação de crédito tributário</rdfs:label>
     <rdfs:label>compensação do crédito tributário</rdfs:label>
</owl:Class>

<owl:NamedIndividual rdf:ID="CompensaçãoDoCréditoTributário">
     <owl:type rdf:about="#CompensaçãoDeCréditoTributário">
</owl:Class>

<owl:Class rdf:ID="CompensaçãoDeTributo">
     <rdfs:label>compensação de crédito tributário</rdfs:label>
     <rdfs:label>compensação de tributo</rdfs:label>
</owl:Class>

<owl:NamedIndividual rdf:about="#CompensaçãoDeTributo">
     <owl:type rdf:about="#CompensaçãoDeTributo">
</owl:Class>

<owl:Class rdf:ID="CompensaçãoTributária">
     <rdfs:label>compensação de crédito tributário</rdfs:label>
     <rdfs:label>compensação tributária</rdfs:label>
</owl:Class>

<owl:NamedIndividual rdf:about="#CompensaçãoTributária">
     <owl:type rdf:about="#CompensaçãoTributária">
</owl:Class>
```

Fonte: o Autor (2017)

Apenas uma das classes está ricamente descrita, as outras representam apenas formas alternativas do mesmo conceito. Estas classes foram, então, combinadas em apenas uma que as representasse.

No caso do exemplo da Figura 23, a classe selecionada foi "**CompensaçãoDeCréditoTributário**". Todas as outras foram suprimidas, com eventuais correções de referências, de forma a manter a consistência semântica da ontologia. Na Figura 24 pode ser visto o resultado da combinação das quatro classes em sua representante.

Figura 24 - Exemplo de classe representante de grupo de classes agrupadas por causa do compartilhamento de rótulos

```
<owl:Class rdf:ID="CompensaçãoDeCréditoTributário">
     <rdfs:subClassOf rdf:resource="#ExtinçãoDoCréditoTributário"/>
     <rdfs:label>compensação de crédito tributário</rdfs:label>
     <rdfs:label>compensação de tributo</rdfs:label>
     <rdfs:label>compensação do crédito tributário</rdfs:label>
     <rdfs:label>compensação tributária</rdfs:label>
</owl:Class>

<owl:NamedIndividual rdf:about="#CompensaçãoDeCréditoTributário">
     <rdfs:type rdf:resource="#CompensaçãoDeCréditoTributário"/>
     <relacionadoCom rdf:resource="#CréditoPresumido"/>
     <relacionadoCom rdf:resource="#CréditoRemanescente"/>
     <relacionadoCom rdf:resource="#CréditoVincendo"/>
     <relacionadoCom rdf:resource="#DébitoTributário"/>
     <relacionadoCom rdf:resource="#DeclaraçãoDeCompensação"/>
     <relacionadoCom rdf:resource="#EncontroDeContas"/>
     <relacionadoCom rdf:resource="#SubstitutoTributário"/>
</owl:Class>
```

Fonte: o Autor (2017)

Para a consecução do processo, o aplicativo de conversão de tesauro em ontologia foi adaptado para oferecer uma funcionalidade própria para o fim de combinação de classes. Tratou-se de uma etapa manual visto que, em muitos casos, classes com rótulos idênticos eram de fato diferentes, representando na verdade o resultado de desambiguação semântica para termos homônimos.

Ao final da revisão de todos os grupos, a quantidade de classes presentes na ontologia diminuiu para conter **10.185** classes, destas persistindo **89** em **44** grupos de classes que de fato davam diferentes significados aos mesmos elementos léxicos.

Como apenas classes com equivalência de conceitos foram combinadas, a ontologia foi simplificada sem que tenha havido perda semântica ou léxica.

O Quadro 3 apresenta estatísticas que dão a dimensão do trabalho realizado.

Quadro 3 - Resumo das métricas das etapas da conversão do Tesauro Jurídico do STJ na ontologia Vocabulario-2016-06.owl

Artefato	Métricas
Arquivo Tesauro 06.2016.doc	16,3 megabytes 4.525 páginas 346.503 palavras 13.336 descritores
Arquivo Tesauro-2016-06.txt	2,31 megabytes 92.895 linhas 13.197 descritores
Arquivo Vocabulario-2016-06.owl	4.41 megabytes 10.185 classes

Fonte: o Autor (2017)

A conversão do tesauro em ontologia foi concluída, mas foi observado que a mesmo precisaria de mais alguns aprimoramentos para adaptar-se ainda melhor aos propósitos da pesquisa. Estes aprimoramentos estão descritos na próxima seção.

4.2.5 Complementação da ontologia

Para auxiliar na avaliação da ontologia **Vocabulario-2016-06** obtida a partir do tesauro do STJ foi construído um Motor de Busca Semântica simples (a ser pormenorizado em seção posterior) que permitiu avaliar suas características, virtudes e fraquezas.

Uma deficiência observada foi a falta de versões alternativas e plurais para diversas expressões presentes em documentos jurisprudenciais presentes em um *corpus* de teste que foi construído a partir de decisões recuperadas do **Tribunal de Justiça do Paraná**.

Na Figura 25 pode ser observado um exemplo desta situação, onde a classe "**DanoMoral**" apresenta apenas um rótulo ("dano moral").

Figura 25 – Exemplo de classe candidata ao enriquecimento de rótulos

```
<owl:Class rdf:ID="DanoMoral">
     <rdfs:subClassOf rdf:resource="#Dano"/>
     <rdfs:label>dano moral</rdfs:label>
</owl:Class>

<owl:NamedIndividual rdf:about="#DanoMoral">
     <rdfs:type rdf:resource="#DanoMoral"/>
     <relacionadoCom rdf:resource="#AbaloEmocional"/>
.
. demais propriedades
.
    <relacionadoCom rdf:resource="#TeoriaDaPerdaDaChance"/>
</owl:NamedIndividual>
```

Fonte: o Autor (2017)

Experimentos mostraram que outras formas de expressar esta mesma classe são utilizadas, mas tais alternativas textuais não estavam contidas no tesauro do STJ. Naturalmente, o tesauro e a ontologia obedecem a critérios diferentes de construção, pois objetivam aplicações também diferentes, o que explica a inexistência destas formas alternativas.

No sentido de superar esta limitação que tem implicação direta na eficácia da ontologia no contexto de um Motor de Busca, o aplicativo que converte o tesauro em ontologia foi aprimorado para importar um arquivo no qual este tipo de complementação é codificado. Para cada classe que se deseja complementar com rótulos, basta adicionar uma linha como a mostrada na Figura 26.

Figura 26 – Exemplo de complementação de rótulo da classe “DanoMoral” para incluir plural e outras formas alternativas

```
DanoMoral=danos morais;dano de ordem moral;danos de ordem moral
```

Fonte: o Autor (2017)

Neste exemplo, a classe “**DanoMoral**” recebeu três rótulos adicionais (“**danos morais**”, “**dano de ordem moral**” e “**danos de ordens morais**”), que passarão a ser utilizados para identificá-la quando ocorrerem em documentos indexados.

O nome mostrado à esquerda do sinal de igualdade (“**DanoMoral**”) identifica a classe, enquanto que a lista de rótulos adicionais figura à sua direta. Cada rótulo deve ser separado por um sinal de ponto-e-vírgula, visto que os rótulos podem apresentar múltiplas palavras separadas por espaço.

Na Figura 27 pode ser vista a nova declaração da classe, após a adição dos novos rótulos.

Figura 27 – Configuração da classe “DanoMoral” após o enriquecimento de rótulos para inclusão de plural e formas alternativas

```
<owl:Class rdf:ID="DanoMoral">
     <rdfs:subClassOf rdf:resource="#Dano"/>
     <rdfs:label>dano moral</rdfs:label>
     <rdfs:label>dano de ordem moral</rdfs:label>
     <rdfs:label>danos de ordem moral</rdfs:label>
     <rdfs:label>danos morais</rdfs:label>
</owl:Class>

<owl:NamedIndividual rdf:about="#DanoMoral">
     <rdfs:type rdf:resource="#DanoMoral"/>
     <relacionadoCom rdf:resource="#AbaloEmocional"/>
.
. demais propriedades
.
    <relacionadoCom rdf:resource="#TeoriaDaPerdaDaChance"/>
</owl:NamedIndividual>
```

Fonte: o Autor (2017)

No **Apêndice C** pode ser encontrada a lista completa de todas as adições de rótulos que foram realizadas com vistas a dar à ontologia melhores condições de identificação de conceitos.

Uma vez que a ontologia peso leve **Vocabulário-2016-06** fora atingida, construída a partir do tesauro mantido pelo Superior Tribunal de Justiça, passou a ser possível avançar para a última etapa do primeiro objetivo específico, que foi sua combinação com a ontologia **JurisTJPR** de Molinari (2011).

4.3 Combinação das ontologias

A combinação das ontologias **JurisTJPR** e **Vocabulario-2016-06** derivada a partir do tesauro do STJ visa o aproveitamento do conhecimento modelado nestes dois artefatos, originando uma terceira ontologia que foi denominada **OntoLegis**.

Apesar da recomendação direta da Metodologia NeOn para adoção do **NeOn Toolkit**, optou-se pelo uso do **Protégé 5.0** para a realização do processo, unicamente por questão de familiaridade com os conceitos e a interface gráfica da ferramenta.

Os conceitos em comum nas duas ontologias foram combinados e resultaram em uma única representação na ontologia final (**OntoLegis**), sendo que esta nova versão das classes inclui atributos e relações oriundo das duas ontologias originais, resultando assim uma terceira versão mais enriquecida.

No entanto, a combinação das ontologias não foi limitada à mera aplicação do recurso existente na ferramenta. Foi necessária uma adaptação e preparação para compatibilização entre os dois artefatos, descrita nas seções 4.3.1 e 4.3.2.

4.3.1 Adição de rótulos à ontologia JurisTJPR

Para efeitos de recuperação de informação, as classes da ontologia JurisTJPR não dispõem de uma representação textual que corresponda às expressões contidas nos documentos jurisprudenciais. Em outras palavras, as classes presentes na JurisTJPR não possuíam rótulos.

As duas ontologias foram construídas com critérios e processos bem diferentes, o que resultou em estruturas também muito distintas. Um atributo que evidencia este fato é a quantidade de classes constituídas em cada artefato. São **220** na **JurisTJPR** e **10.185** na **Vocabulario-2016-06**.

Em função do tamanho da ontologia JurisTJPR, optou-se por realizar a adição de rótulos (elementos "**rdfs:label**") manualmente. Basicamente, a anotação em questão foi adicionada uma a uma novamente fazendo uso do aplicativo **Notepad++**, pois o arquivo JurisTJPR.owl é um documento XML e a operação direta na forma de texto foi considerada mais produtiva do que utilizar a interface gráfica do **Protégé**.

A regra foi usar os caracteres maiúsculos do nome das classes como marcadores para espaços separadores (a ontologia JurisTJPR também fez uso da notação *pascal-case*), eventualmente acrescentando proposições quando necessário. No Quadro 4 estão mostrados alguns exemplos de nomes de classes da ontologia

JurisTJPR e a anotação textual respectiva que foi adicionada para promover sua integração à ontologia de vocabulário.

Quadro 4 – Exemplos de classes da ontologia JurisTJPR que receberam anotações textuais como preparativo para combinação com Vocabulario-2016-06

Nome da classe	Anotação textual
AcaoResolucao	ação de resolução ações de resolução
Agravado	agravada agravadas agravado agravados
ConflitoCompetencia	conflito de competência conflitos de competência
MandadoSeguranca	mandado de segurança mandados de segurança

Fonte: o Autor (2017)

Alguns casos permitiram a simples separação das palavras, outras precisaram receber sinais de acentuação. Algumas receberam pronomes. Finalmente, houve casos onde várias das adaptações ocorreram concomitantemente. Praticamente todas receberam formas nos dois gêneros.

Algumas classes presentes na ontologia JurisTJPR não precisaram receber rótulos, pois correspondiam em nome e sentido a outras existentes na ontologia **Vocabulario-2016-06**.

A partir desta preparação, o que se obteve foi uma versão da ontologia **JurisTJPR** anotada com expressões textuais que poderiam ser encontradas no texto de documentos jurisprudenciais e que, portanto, poderiam ser indexadas e emprestar mais informação semântica para efeitos de recuperação.

4.3.2 Identificação de classes em comum

Das 220 classes existentes na ontologia **JurisTJPR**, 50 também estavam presentes na ontologia **Vocabulario-2016-06**. Algumas destas estão listadas no Quadro 5. A totalidade destas classes pode ser encontrada no **Apêndice D**.

Quadro 5 – Exemplos de classes comuns existentes tanto na ontologia JurisTJPR quanto na Vocabulario-2016-06

Nome original	Nome convertido
AcaoNulidade	AçãoDeNulidade
Comarca	Comarca
ConflitoNegativo	ConflitoNegativoDeCompetência
HabeasCorpus	HabeasCorpus
Impetrado	Impetrado
MandadoSeguranca	MandadoDeSegurançaMinistro

Fonte: o Autor (2017)

A primeira coluna mostra o nome da classe como a mesma estava declarada originalmente na ontologia **JurisTJPR**, enquanto que na segunda coluna mostra o novo nome que veio a receber para garantir que o processo de combinação efetivamente fosse capaz de relacionar as duas classes. A renomeação das classes foi realizada utilizando facilidade existente no Protégé.

Em alguns casos, a classe existente na ontologia **JurisTJPR** foi renomeada para acompanhar o padrão de nomenclatura utilizado na ontologia **Vocabulario-2016-06**, onde ocorrem acentuação e uso de pronomes.

4.3.3 Operação de combinação de ontologias no Protégé

A combinação de duas ou mais ontologia no Protégé é chamada de *Ontology Merge* e envolve a carga isolada dos arquivos OWL contendo as ontologias e a comando de combinação destas.

O produto se encarrega em adicionar as classes e relacionamentos existentes em cada um dos artefatos envolvidos. O resultado final é uma única ontologia composta pelas construções semântica das várias ontologias envolvidas.

No entanto, cada um dos arquivos utilizados apresentava classes descritas em espaços de nomes (do inglês, *namespaces*) diferentes. A ontologia JurisTJPR faz uso do espaço de nome “**http://www.ahmolinari.com/jurisTJPR.owl**”, enquanto que na ontologia Vocabulario-2016-06, o mesmo se chamava “**http://br.ufpr.ppgcgti/mestrado/2016/vocabulario**”.

Em OWL, os *namespaces* figuram como prefixos para os nomes e normalmente se assemelham a endereços da *World Wide Web*, mas são apenas *strings*. Assim, quando uma classe é denominada Voto em casa ontologia, na verdade

se chama “**http://www.ahmolinari.com/jurisTJPR.owl#Voto**” na JurisTJPR e “**http://br.ufpr.ppgcgti/mestrado/2016/vocabulario#Voto**” na Vocabulario-2016-16.

Por apresentarem nomes diferentes, o processo de combinação de ontologias acaba por simplesmente gerar um novo documento OWL com duas taxonomias distintas.

A solução da situação é provida pelo próprio Protégé, que apresenta um recurso para renomeação em lote de classes, onde é possível trocar apenas o nome do *namespace*. Assim, este recurso foi utilizado para renomear todas as classes existentes para fazer uso do *namespace* “**http://br.ufpr.ppgcgti/mestrado/2016/ontoLegis**”.

Neste contexto, o Protégé se encarregou de identificar eventuais correspondências de nomes (como é o caso das classes “**Voto**”, existente em cada um dos *namespaces*) e integrá-las em uma única definição combinada.

Assim, com a renomeação em lote de classes, o processo de combinação foi concluído, tendo como resultado um documento OWL chamado de **OntoLegis-2016-06.owl**, contendo a nova ontologia peso leve **OntoLegis**. O que se atingiu foi uma nova ontologia composta por **10.210** classes, pronta para servir de entrada para um Motor de Busca experimental que possa atestar a eficácia das relações semânticas na recuperação de conteúdo jurisprudencial.

4.4 Avaliação da ontologia elaborada

Algumas métricas estatísticas foram calculadas com base em um *corpus* de **59.386** decisões recuperadas a partir do próprio sítio de Internet do **Tribunal de Justiça do Paraná** (TJPR). Estas decisões correspondem ao período de **05/02/2016** a **01/11/2016**. Mais detalhes a respeito da obtenção deste *corpus* estão disponíveis na seção 4.5 que discute o Motor de Busca Semântica.

A ontologia **OntoLegis** é constituída por **10.210** classes que totalizam **12.595** rótulos. Estes rótulos são procurados em cada um dos documentos do *corpus* e os que são encontrados passam a identificar classes que serão vinculadas aos mesmos.

A título de recordação, deve-se ter em mente que cada classe da ontologia corresponde a um conceito do domínio em estudo.

O processo de identificar as classes que estão presentes em um *corpus* de documento constitui o próprio processo de indexação semântica (ou conceitual), que é a primeira etapa para a construção de um sistema de recuperação semântica de

informação, ou seja, um Motor de Busca Semântica. Detalhes específicos do processo de indexação semântica estão apresentados em seção posterior.

Para efeitos de análise estatística foram construídos quatro diferentes ranques:

a) os 100 (cem) documentos com maior número de classes identificadas;
b) os 100 (cem) documentos com menor número de classes identificadas;
c) as 100 (cem) classes mais frequentes no *corpus*;
d) as 100 (cem) classes menos frequentes no *corpus*.

A segunda estatística permite descortinar um aspecto importante da avaliação de qualidade da ontologia: se a mesma é ou não capaz de cobrir a **totalidade** dos documentos sob indexação.

Algumas métricas de resumo estatístico foram calculadas e encontram-se no Quadro 6.

Quadro 6 – Estatísticas descritivas calculadas para avaliação das relações estabelecidas entre documento indexados e classes da ontologia

	Média	**Maior**	**Menor**	**Nulos**
Classes por documento	59,05	387	1	0
Frequência de classes	343,48	58.859	0	4.072

Fonte: o Autor (2017)

Na primeira linha do quadro estão colocadas as estatísticas que dizem respeito à quantidade de classes que foram identificadas nos documentos do *corpus*. Estas contagens são resultado direto do processo de indexação semântica.

Pela análise da primeira linha já se pode concluir que a ontologia OntoLegis apresenta um grau de cobertura de **100%** em relação ao corpus de **59.386** decisões obtidas a partir do **TJPR**. A coluna "**Nulos**" indica que não existem documentos sem a presença de ao menos uma classe.

A amplitude de classes vinculadas a documentos varia de **1** a **387**, ou seja, existem decisões onde exatamente uma classe foi identificada, enquanto que no outro extremo, ao menos um documento apresenta o volume de 387 classes identificadas.

Sete decisões apresentam exatamente uma classe identificada (processos número **1559783-6, 1556846-6, 1561644-5, 1388081-8, 1513630-4, 1436765-8, 803270-0**) sendo quatro com ocorrência da classe "**Relator**" e os restantes com uma

ocorrência de cada uma das classes “**PlanoDeCargosESalarios**”, “**EstadoDeSergipe**” e “**RecursoJudicial**”

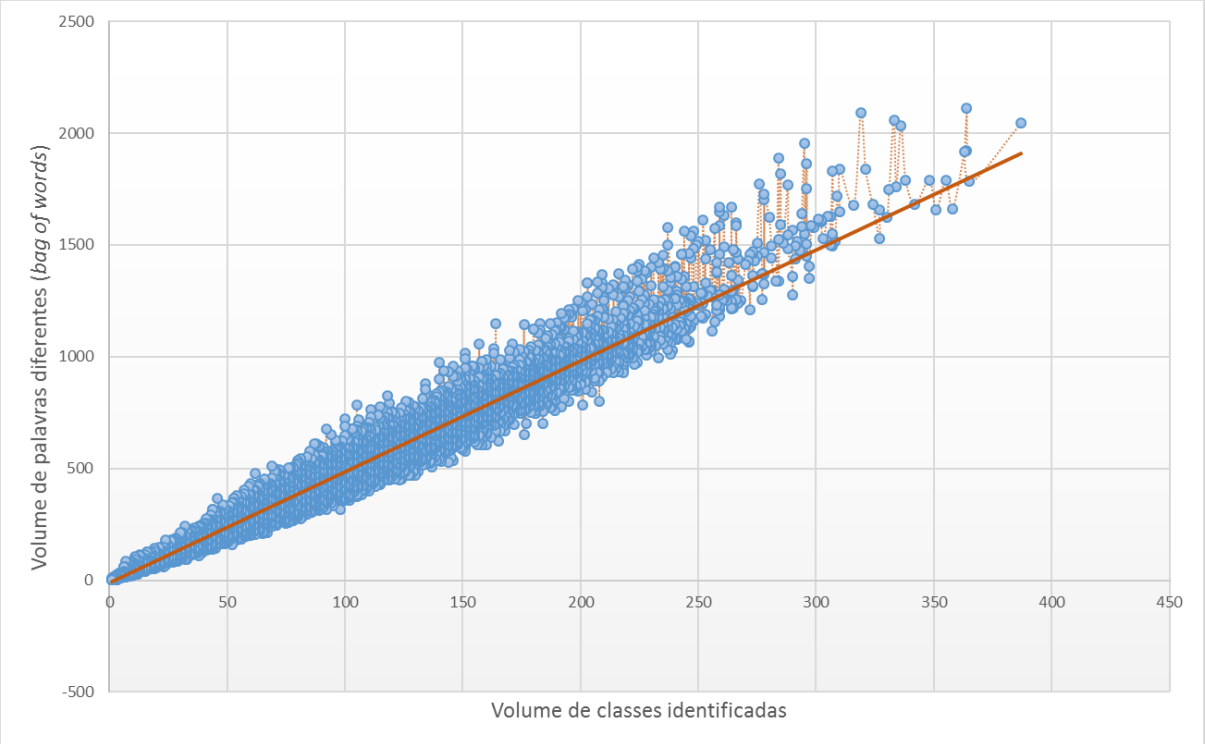
Foram identificados nove documentos vinculados a duas classes, sendo que a classe “**Relator**” foi identificada em cinco deles e “**Desembargador**” em três.

Para maiores detalhes, o **Apêndice E** apresenta a lista dos 100 documentos com a maior quantidade de classes vinculadas.

Em média, **59,05** classes foram identificadas em cada documento, mas o desvio padrão de **51,45** evidencia discrepâncias significativas na distribuição, fato reforçado pela amplitude de **387** na contagem de classes por documento.

A variação na contagem de classes é explicada pela variação significativa do **tamanho em palavras** de cada documento. Documentos mais longos apresentam mais conceitos e documentos mais curtos, menos conceitos. Na Figura 28 pode ser visto um diagrama de dispersão baseado em todo o *corpus*, onde a nuvem de valores e a reta de tendência demonstram a existência de uma correlação significativa.

Figura 28 - Diagrama de dispersão com a respectiva reta de tendência mostrando a relação direta entre volume de classes identificadas e volume de palavras presentes em um documento



Fonte: o Autor (2017)

Pela observação da Figura 28, é possível perceber que a reta de tendência (em cor alaranjada) acompanha a forma da nuvem formada pelos pontos que descrevem o volume de classes em relação ao volume de palavras únicas por documento (em cor azul).

A correlação existente entre volume de palavras de um documento e o volume de classes identificadas dentro do mesmo é corroborada pelo **Coeficiente de Correlação de Pearson**, que neste caso foi calculado em **0,9902**, ou seja, a dependência do volume de classes em relação ao volume de palavras é praticamente absoluta (**99,02%**). Por esta correlação, sempre que o documento aumentar em quantidade de palavras, haverá um aumento na quantidade de classes identificadas.

Na figura em questão foi considerado o volume de palavras distintas sem contar suas repetições, estrutura geralmente conhecida como *bag of words* (WEISS et al., 2010). Nestes casos também não foram removidas palavras como pronomes, que geralmente são interpretadas como desprovidas de valor informacional e conhecidas como *stop words* (WEISS et al., 2010), pois efetivamente ocorrem nos rótulos das classes e são consideradas para sua identificação.

Na segunda linha do Quadro 6 estão colocadas as estatísticas que dizem respeito à frequência de ocorrência das classes da ontologia, que corresponde a outra perspectiva de distribuição de classes e cobertura da ontologia em relação ao *corpus*.

O Quadro 7 apresenta as dez classes mais frequentemente identificadas no *corpus*. Na primeira coluna encontram-se os nomes das classes, na segunda sua frequência de ocorrência ou longo do *corpus* e na terceira e última, a métrica ***Inverse Document Frequency***, mais conhecido como **IDF** (ROBERTSON, 2004; WEISS et al., 2010).

Quadro 7 – Dez classes mais frequentemente encontrada nos documentos do *corpus*, com respectivo IDF

#	Classe	Frequência	IDF
01	EstadoDeSergipe	58859	0,00
02	Desembargador	38382	0,19
03	Artigo	37108	0,20
04	CódigoDeProcessoCivil	36906	0,21
05	OrganizaçãoSocial	36726	0,21
06	Autos	35970	0,22
07	RecursoJudicial	35636	0,22
08	Relator	34752	0,23
09	Prosseguimento	32681	0,26
10	DireitoEspecialDeSaque	32317	0,26

Fonte: o Autor (2017)

O quadro anterior revela um fato curioso: a classe com maior frequência de ocorrência é “**EstadoDeSergipe**”. O motivo é mais facilmente compreendido se os rótulos vinculados à classe em questão forem observados em detalhe. A definição dos rótulos da classe, em OWL, encontra-se no Quadro 8.

Quadro 8 - Rótulos da classe "EstadoDeSergipe" da ontologia OntoLegis, em um fragmento de documento OWL

```
<rdf:Description rdf:about="#EstadoDeSergipe">
    <rdfs:label>sergipe</rdfs:label>
    <rdfs:label>estado de sergipe</rdfs:label>
    <rdfs:label>se</rdfs:label>
</rdf:Description>
```

Fonte: o Autor (2017)

Dentre os rótulos anotados para caracterizar lexicamente a classe “**EstadoDeSergipe**” está o rótulo “**se**” (destacado), ou seja, é a sigla da unidade federativa correspondente ao estado (“SE”). Esta sigla, no entanto, encontra-se presente no texto com a função de pronome reflexivo e não de substantivo. Em alguns casos estará na forma de ênclise (por exemplo, “intime-**se**”) e em outras, próclise (por exemplo “**se** fazem”). Como o Motor de Busca Semântica implementado para experimentação **não é capaz** de identificar a função sintática do termo “**se**” nos diversos contextos, a identificação errônea da classe ocorreu.

Este fenômeno se repetiu com a classe “**DireitoEspecialDeSaque**” (décima classe mais frequente), pois esta apresenta como rótulo a sigla “**des**”, que ocorre também na classe “**Desembargador**” (segunda classe mais frequente). Aqui é necessário um processo mais sofisticado de desambiguação, visto que em um caso se trata de uma sigla e em outro de uma abreviatura. As duas palavras apresentam semânticas bem distintas, a ser identificada unicamente com base no contexto de seu emprego.

A décima-primeira classe mais frequente é “**EstadoDoParaná**” (não presente no Quadro 7), explicada efetivamente pela sua elevada frequência, identificada pelos seus três rótulos (“**pr**”, “**paraná**” e “**estado do paraná**”). A frequência é alta porque o corpus foi constituído unicamente com documentos obtidos do **Tribunal de Justiça do Paraná**.

A última coluna do Quadro 7 apresenta o valor de uma métrica informacional chamada **IDF**, associada a cada classe. Como já comentado (vide seção 2.2.1), o IDF diminui com a quantidade de ocorrências (magnitude inversa), exprimindo assim sua menor capacidade de discriminação. Por este motivo, a classe mais frequente no

corpus (“**EstadoDeSergipe**”) apresenta of **IDF = 0,0**, enquanto que classe na décima posição do mesmo ranque apresenta um IDF maior (**IDF = 0,26**), exprimindo sua maior capacidade de discriminação.

O **IDF** foi utilizado no Motor de Busca Semântica com vistas a valorar a importância discriminante de cada classe, o que contribuiu para a ordenação do conjunto de resultados.

O **Apêndice F** apresenta uma lista com as 100 classes mais frequentemente identificadas no *corpus*.

A menor frequência é compartilhada por diversas classes, visto que **4.072** das classes da ontologia não foram referenciadas por nenhum documento do *corpus*. O **Apêndice G** apresenta uma lista com as 100 classes com menor frequência identificadas no *corpus*, mas que contam com ao menos uma referência.

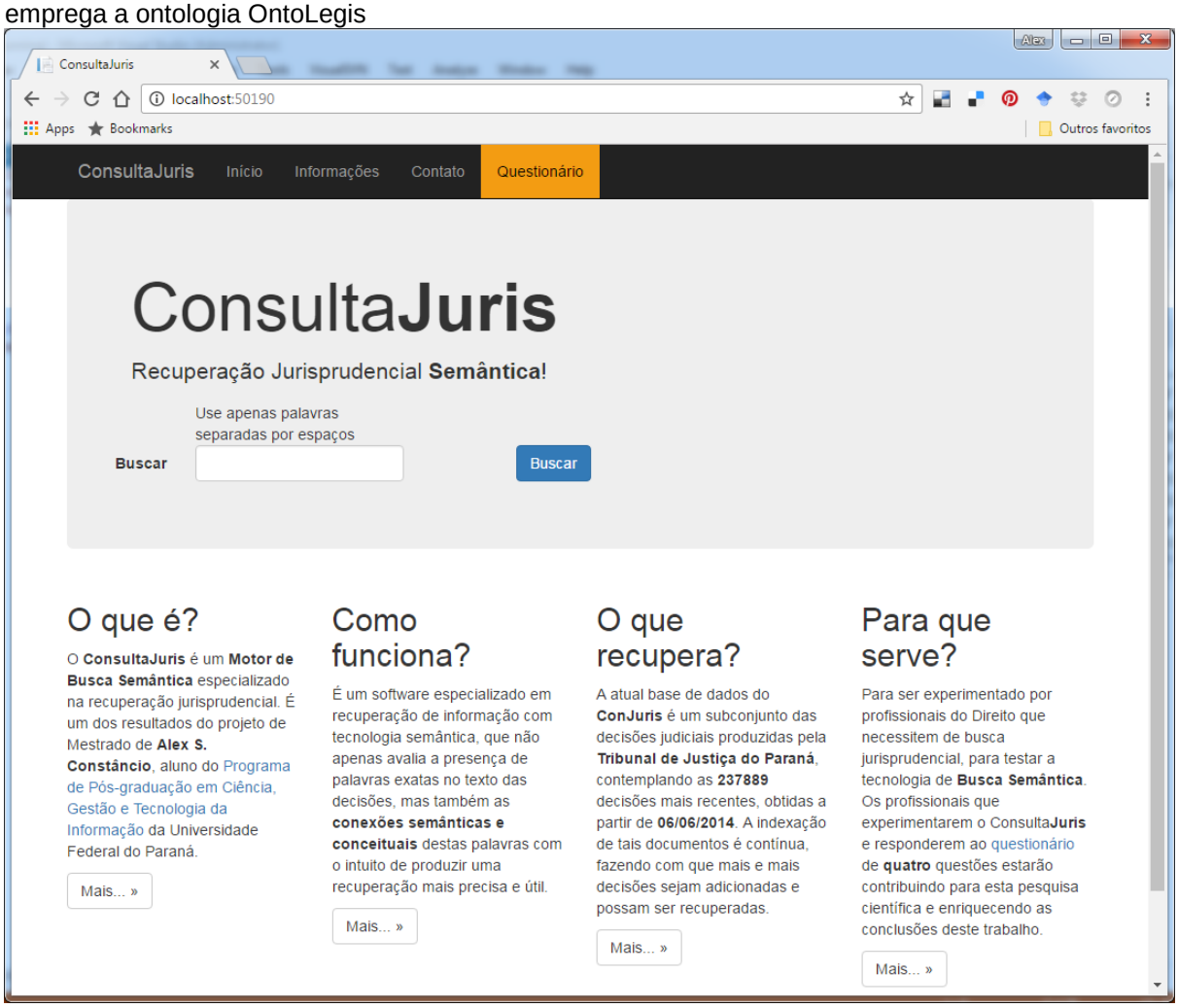
4.5 O Motor de Busca Semântica

Para efeitos de teste da ontologia **OntoLegis** foi construída uma versão experimental de um Motor de Busca Semântica seguindo os moldes operacionais dos motores de busca mais populares da internet, em suas versões mais simples, sem o uso de operadores booleanos.

Do ponto de vista de usuário, o Motor de Busca Semântica que emprega a ontologia **OntoLegis-06-2016** é um sítio da Internet[15]. A página principal deste sítio pode ser vista na Figura 29.

[15] http://www.consultajuris.com.br, acesso em 17 jan. 2017

Figura 29 - Página inicial do sítio ConsultaJuris, que implementa o Motor de Busca Semântica que emprega a ontologia OntoLegis

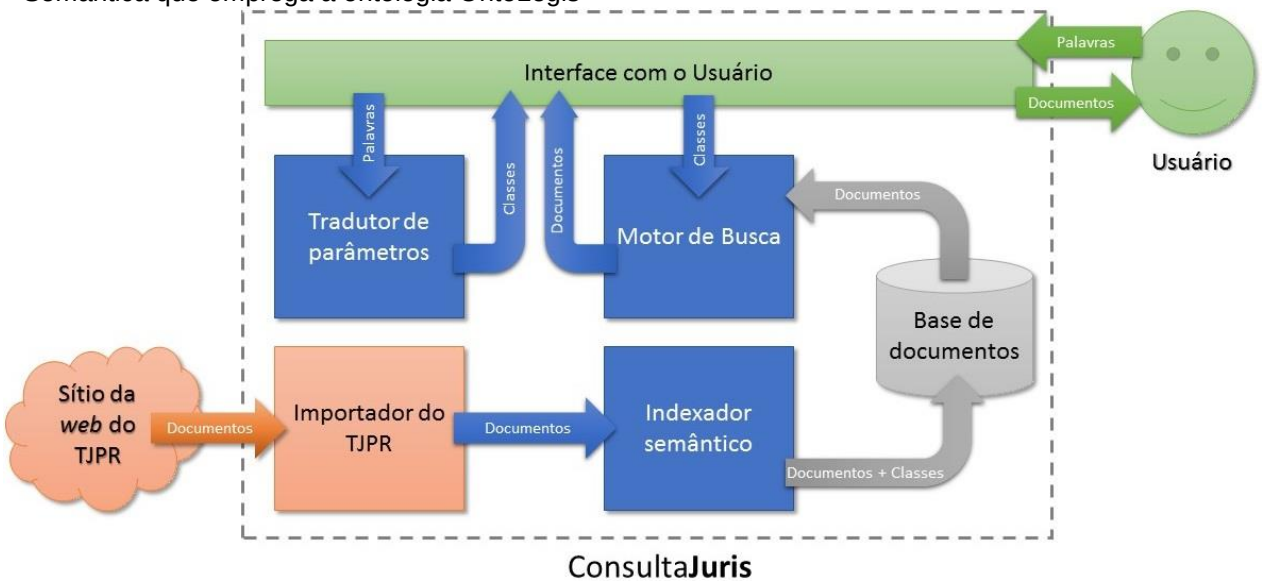


Fonte: o Autor (2017)

Neste motor de busca, o usuário fornece um conjunto de palavras que serão utilizadas como critérios de busca. Por trata-se de um motor de busca que opera sobre um *corpus* indexado a partir de representações de conceitos e não de palavras, o processo que se inicia no fornecimento dos parâmetros e é concluído com a apresentação de resultados obedece a alguns passos particulares que envolvem a tecnologia.

Na Figura 30 apresenta-se um diagrama de blocos que inclui todos os componentes que, articuladamente, colaboram para a operacionalidade do motor de busca dentro do paradigma de busca semântica.

Figura 30 – Diagrama de blocos do sítio ConsultaJuris.com.br, que implementa o Motor de Busca Semântica que emprega a ontologia OntoLegis



Fonte: o Autor (2017)

O sítio ConsultaJuris pode ser dividido em quatro componentes principais:

a) **Interface com o Usuário** é responsável por receber os parâmetros de busca e submetê-los ao motor de busca, cujos documentos resultantes são formatados e apresentados;
b) **Componentes Semânticos**, que empregam a ontologia OntoLegis para vários fins, como a tradução de parâmetros em classes, o motor de busca propriamente dito e o indexador semântico;
c) **Base de documentos**, que armazena os documentos do *corpus* e suas vinculações com as classes da ontologia;
d) **Importador do TJPR**, que é responsável por visitar o sítio do TJPR e realizar a importação de decisões, para sua posterior indexação, armazenagem e recuperação.

O usuário e o sítio do TJPR são elementos externos ao ConsultaJuris.

O Motor de Busca Semântica permitiu avaliar a ontologia e identificar oportunidades de melhoria, efeitos inesperados, deficiências, potenciais e a geração de estatísticas e métricas que deram condição de análises numéricas do comportamento da ontologia.

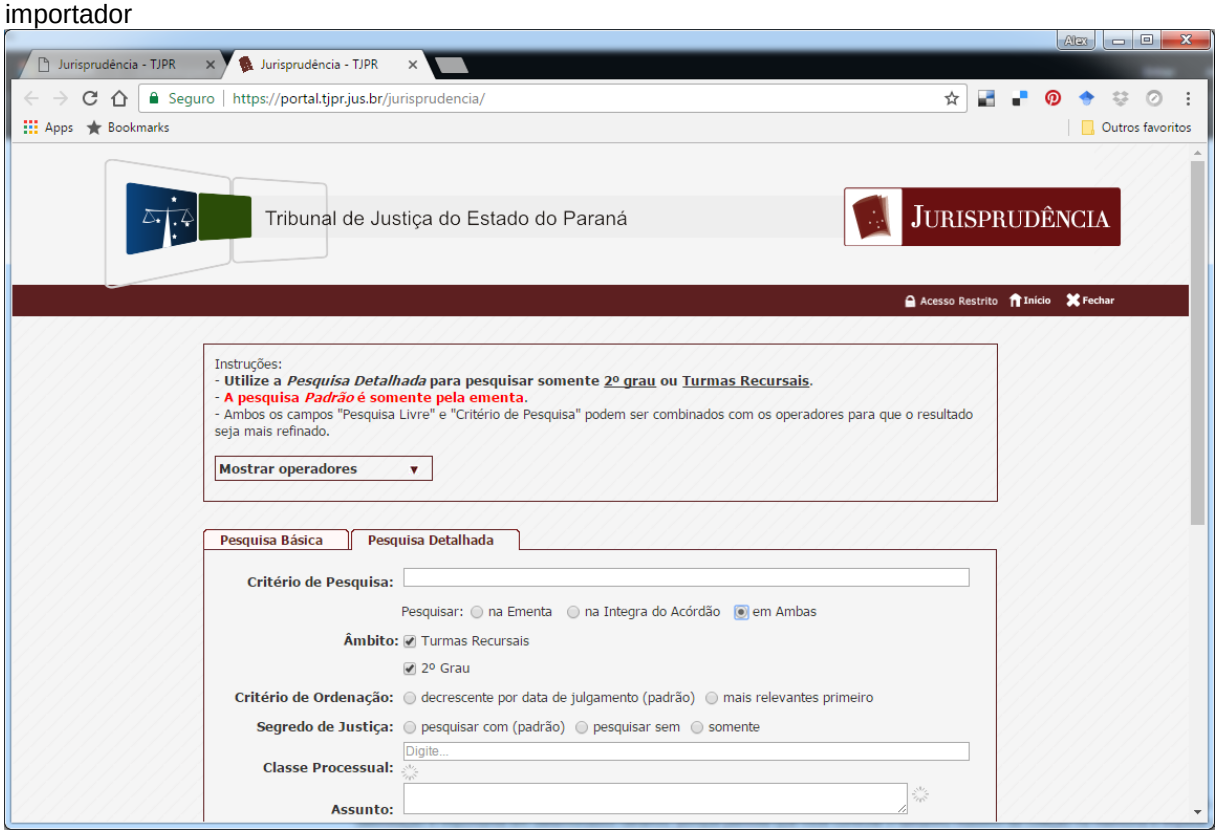
As seções a seguir apresentam os detalhes de cada uma das etapas realizadas durante a indexação semântica do *corpus* e a busca por documentos condicionada a este sistema de indexação.

4.5.1 Construção do *corpus* de teste

A construção do *corpus* de teste é uma parte importante do processo de validação, pois constituiu o conjunto de documentos indexados semanticamente com as classes da ontologia **OntoLegis** para posterior recuperação.

Para a construção de tal *corpus* foi necessária a produção de um aplicativo que imitasse o acesso à página de consulta a jurisprudência do TJPR[16] e dessa forma atuasse como um operador humano. Na Figura 31 pode ser vista uma imagem da página de consulta a jurisprudência do TJPR, já configurada para a realização da busca.

Figura 31 – Página de consulta a jurisprudência do TJPR, com a mesma configuração utilizada pelo importador

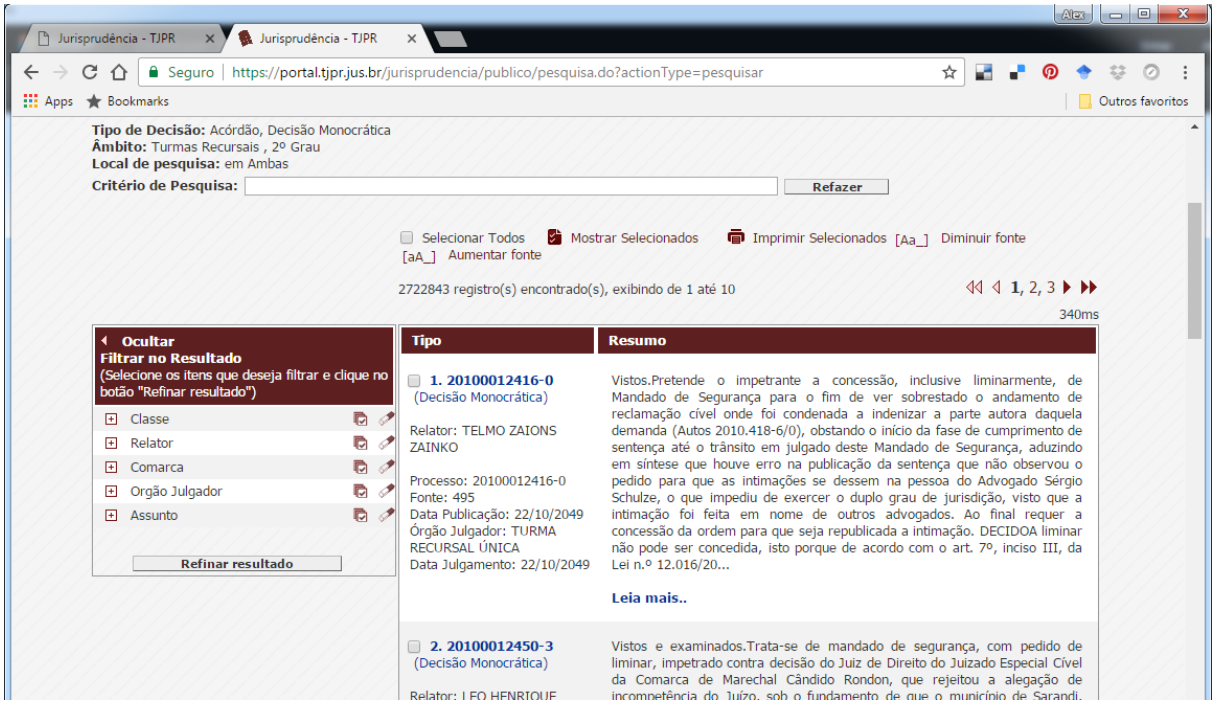


Fonte: Tribunal de Justiça do Paraná (2017)

Note-se que nenhum critério de busca (campo “Critérios de pesquisa”) é fornecido, ou seja, o objetivo é receber todos os documentos existentes. Esta busca, ao ser disparada, resulta no retorno da primeira de inúmeras páginas de resultados. Cada uma destas páginas comporta o resumo de dez decisões. A página de retorno para a consulta anterior é apresentada na Figura 32.

[16] https://portal.tjpr.jus.br/jurisprudencia, acesso em 17 jan. 2017

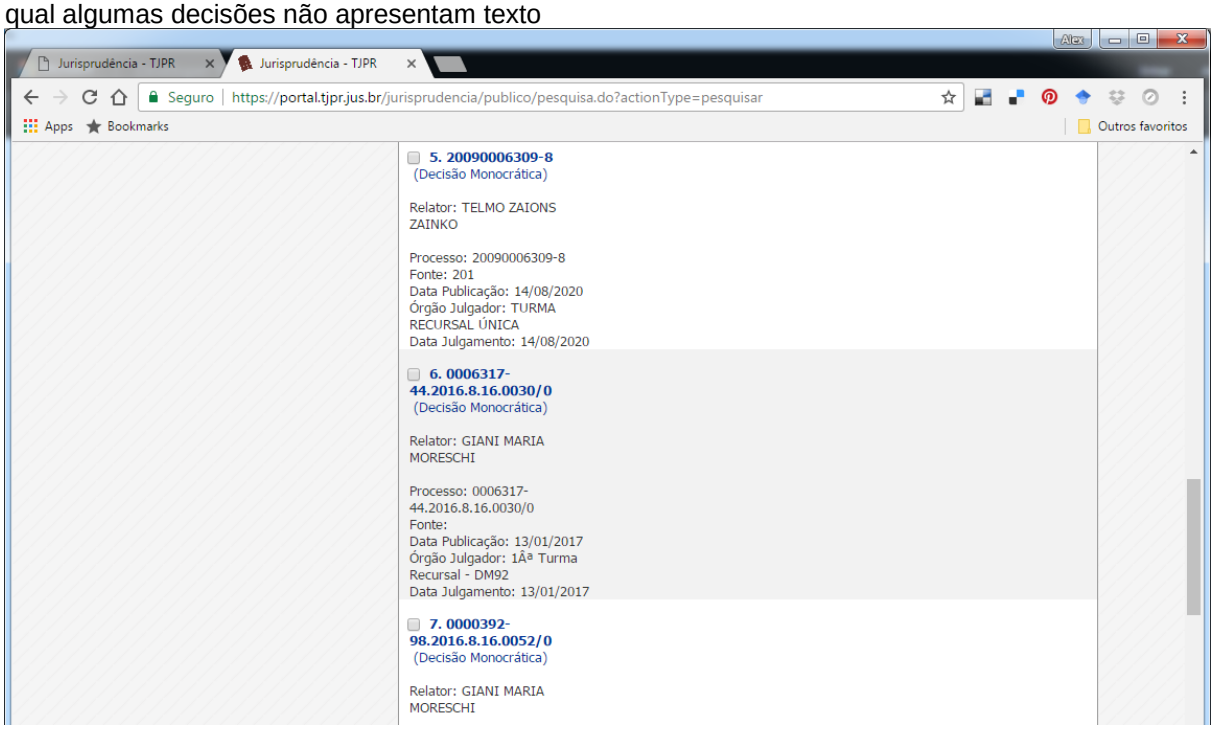
Figura 32 – Primeira página de resultados de uma consulta jurisprudencial sem critérios de busca no TJPR



Fonte: Tribunal de Justiça do Paraná (2017)

A página de resultados apresenta a quantidade total de documentos selecionados (neste caso, 2.722.843 decisões), mas apenas as dez primeiras são apresentadas. No entanto, ao se rolar a página para baixo é possível perceber que algumas decisões são desprovidas de conteúdo, como mostra a Figura 33.

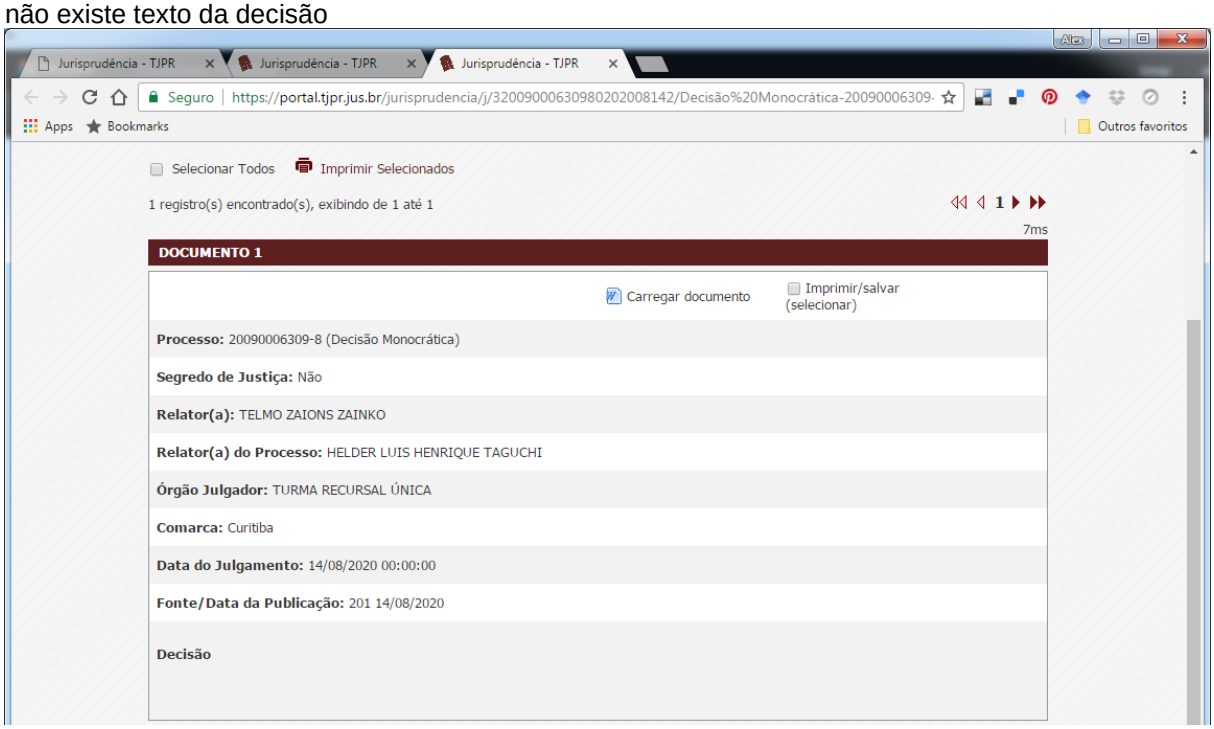
Figura 33 – Segmento final da primeira página de resultados da consulta jurisprudencial do TJPR, na qual algumas decisões não apresentam texto



Fonte: Tribunal de Justiça do Paraná (2017)

Nesta figura, quinta decisão retornada, correspondendo ao processo número 20090006309-8 não apresenta nenhum texto. Poder-se-ia imaginar que se trata de um segredo de justiça, mas ao se clicar no *hiperlink* sob o número do processo, recebe-se a página com o conteúdo integral da decisão, que é mostrada na Figura 34.

Figura 34 – Página de detalhes de decisão retornada pela consulta jurisprudencial do TJPR na qual não existe texto da decisão



Fonte: Tribunal de Justiça do Paraná (2017)

A Figura 34 apresenta detalhes a respeito da decisão, sendo um deles a negativa de que se trata de um segredo de justiça. O importador reconheceu e tratou situações assim, ignorando decisões que não contivessem texto indexável.

Assim, o importador submete uma consulta sem parâmetros, recebe a página de resultados (equivalente à apresentada na Figura 32), extrai os dados das decisões desta por meio da navegação para página com os detalhes de cada decisão (equivalente à mostrada na Figura 34) e avança para a próxima página, repetindo o processo.

Para cada página de resultados, o texto e alguns detalhes dos documentos (relator, órgão julgador, comarca e data de julgamento) são extraídos, indexados semanticamente e finalmente armazenados.

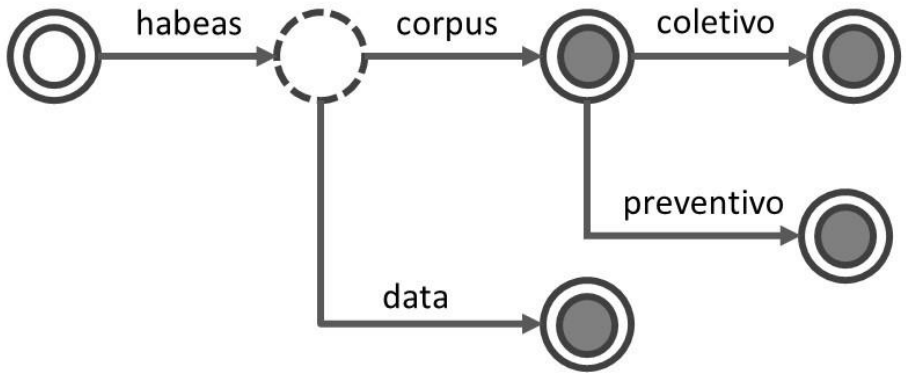
4.5.2 Carga da ontologia

A ontologia **OntoLegis** é composta por **10.210** classes identificadas a partir **12.595** rótulos. Existem rótulos compostos por uma única palavra (por exemplo, o

rótulo “**relator**” que identifica a classe “**Relator**”) assim como rótulos compostos por diversas palavras (por exemplo, o rótulo “**elemento subjetivo da obrigação**” que identifica a classe “**VínculoObrigacional**”).

Visando acelerar o processo de indexação, foi construído um autômato finito (também conhecido como máquina de estados) composto pelas palavras que constituem os diversos rótulos de cada classe presente na ontologia. O processo se assemelha a um analisador léxico de um compilador, mas opera no nível de palavra e não de caracteres. A Figura 35 apresenta um fragmento deste autômato.

Figura 35 - Fragmento de autômato finito utilizado pelo Motor de Busca Semântica durante o processo de indexação de documentos do *corpus* de teste



Fonte: o Autor (2017)

O autômato é um grafo dirigido no qual, neste caso, existem três tipos de nodo, sendo que cada um pode ou não identificar uma classe:

a) o nodo que contém um círculo é chamado **nodo inicial** e existe apenas para demarcar o início do reconhecimento;
b) o nodo vazio e com linhas tracejadas é chamado de **não-terminal** e indica que a última palavra reconhecida não forma um rótulo válido; para que o rótulo seja válido é necessário realizar mais um reconhecimento ou desconsiderar a última palavra lida;
c) o nodo com um círculo preenchido é chamado de **terminal** e indica que um rótulo foi reconhecido.

Os arcos que unem estes nodos representam as palavras que constituem os rótulos das classes. O fragmento de autômato apresentado na Figura 35 é capaz de reconhecer as sentenças “habeas corpus”, “habeas data”, “habeas corpus coletivo” e “habeas corpus preventivo”. A palavra “habeas” não é reconhecida como válida.

O processo de identificar uma classe significa basicamente consumir uma palavra do texto do documento, transitar para o nodo a partir do arco identificado por aquela palavra, consumir a próxima palavra, novamente transitar para o próximo nodo com base no arco identificado pela palavra atual e repetir o processo enquanto for possível (enquanto existirem arcos aferentes que correspondem a última palavra obtida do texto).

No momento em que a próxima transição não possa mais ocorrer avalia-se se o nodo identifica (nodo terminal) ou não (nodo não-terminal) um rótulo de uma classe. Em caso positivo, está identificada uma classe a partir de um de seus rótulos e o processo é reiniciado utilizando-se a última palavra processada. Em caso negativo, inicia-se o descarte das últimas palavras até o eventual atingimento de um nodo terminal. Se este último não existir, então conclui-se que o segmento de palavras até então lido não identifica um rótulo.

Quando a lista de palavras é inteiramente percorrida, tem-se a lista de classes identificadas para o documento em questão, sendo estas as que estarão diretamente vinculadas ao mesmo.

O autômato finito é uma ferramenta clássica utilizada por inúmeros subdomínios da ciência da computação e mostrou-se eficiente e robusta dentro do contexto da indexação semântica.

Com a ontologia carregada e o autômato finito construído, passa-se ao processo de indexar semanticamente o *corpus* de documentos.

4.5.3 Indexação semântica

A fase de indexação é responsável por atuar sobre cada um dos **59.386** documentos do *corpus*, identificando os conceitos (classes da ontologia) a partir dos rótulos que os caracterizam. Cada documento do *corpus* é processado individualmente.

A primeira etapa é a segmentação do texto em uma sequência de palavras isoladas. Neste caso, a segmentação assumiu sua forma mais simples, onde espaços em branco e caracteres não alfanuméricos foram entendidos como separadores e ignorados. As palavras são introduzidas nesta lista de palavras já com todas as letras convertidas para minúsculo, visto que os rótulos das classes da ontologia também estão armazenados seguindo esta convenção.

A partir de uma lista simples onde cada item é uma das palavras formadoras do texto original inicia-se o processo de identificar quais classes estão referenciadas a partir de seus rótulos utilizando o autômato finito construído a partir da ontologia.

É possível que um mesmo rótulo identifique mais de uma classe. Nestes casos, um mesmo estado terminal do autômato estará vinculado a mais de uma classe e ambas serão efetivamente vinculadas ao documento, o que constitui uma ambiguidade semântica e por estre motivo, uma imprecisão do processo de indexação.

A indexação na forma como está implementada **não é capaz** de solucionar o problema da ambiguidade, pois cada classe é identificada unicamente por seus correspondentes léxicos (rótulos), sem considerar o contexto linguístico ou sintático onde fora empregado. A resolução de tal ambiguidade ultrapassa os objetivos desta pesquisa.

Na atual configuração da ontologia **OntoLegis** existem **116** classes (**0,0011%**) que apresentam ambiguidade léxica por compartilharem rótulos entre si.

Com todos os documentos do *corpus* indexados semanticamente, já se tem o núcleo do Motor de Busca Semântica em condições de operar. A próxima etapa é mapear os parâmetros de busca fornecidos para classe da ontologia.

4.5.4 Tradução dos parâmetros de busca

No modelo que foi empregado para testes do Motor de Busca Semântica implementado, a funcionalidade foi a mais simples, sem a possibilidade de operadores booleanos ou simbologia especializada.

O usuário simplesmente fornece palavras que expressam as ideias ou conceitos nos quais está interessado e o sistema responderá com todos os documentos que estiverem relacionados a estes conceitos. Trata-se de uma interface simples e de uso muito fácil, não requerendo absolutamente nenhum conhecimento particular de uso do Motor de Busca.

Do usuário não é exigido que forneça palavras que identifiquem precisamente as classes da ontologia, este trabalho pertence ao Motor de Busca que procurará identificar as classes a partir das palavras fornecidas.

Para efeitos de consulta, pronomes, artigos e preposições são entendidos como *stopwords*, ou seja, sua presença entre os parâmetros de busca é ignorada. Por exemplo, se as palavras “**prestação**”, “**de**” e “**serviço**” são fornecidas como

parâmetros de busca, apenas “**prestação**” e “**serviço**” irão efetivamente figurar durante o processo de recuperação. O Quadro 9 apresenta a lista de *stopwords* consideradas nesta implementação do motor de busca.

Quadro 9 – Lista de *stopwords* utilizadas pelo Motor de Busca Semântica durante a etapa de tradução de parâmetros de busca

a à ao aos as às com da das de do dos e em na no o os sem

Fonte: o Autor (2017)

Por exemplo, um usuário poderia fornecer como critérios para sua busca as palavras “**prestação**”, “**serviço**”, “**contrato**” e “**multa**”, O Motor de Busca procurará identificar quais classes podem estar vinculadas a estas palavras. Na implementação atual tanto ordem quanto local das palavras são desprezados, ou seja, se as mesmas palavras figurarem em outra ordem (por exemplo, “**prestação**”, “**multa**”, “**contrato**” e “**serviço**”) as mesmas classes serão identificadas.

As classes que forem eleitas como as melhores representantes dos parâmetros de busca são chamadas de **classes seletoras**.

O processo de identificação funciona com base em taxa de semelhança entre parâmetros e rótulos de classes. As palavras fornecidas como parâmetros de busca são utilizadas para identificar as classes seletoras a partir de seus rótulos. Para efeitos de otimização, foi construído um dicionário que mapeia cada palavra individual para todas as classes que a apresentam em ao menos um de seus rótulos.

Assim, cada palavra sugere um conjunto inicial de classes seletoras. Após a recuperação de todas as classes que cada palavra pode identificar é feita uma contagem para cada uma destas, estabelecendo a quantidade de palavras que foram responsáveis por sua identificação. Por exemplo, as palavras “**prestação**” e “**serviço**” identificam as classes “**Prestação**”, “**Serviço**” e “**PrestaçãoDeServiço**”. Identificam também as classes “**PrestaçãoDeServiçosAComunidade**”, “**PrestaçãoPaga**”, “**Anuidade**” (por conta de seu rótulo “**prestação anual**”), “**ServiçoNocivo**” e outras 127 classes, totalizando 134 classes ao todo, pois todas estas apresentam ocorrências de ao menos uma das duas palavras em seus rótulos.

Para todas as quatro palavras utilizadas como exemplo de parâmetros (“**prestação**”, “**serviço**”, “**contrato**” e “**multa**”), a quantidade inicial de classes seletoras identificadas é de **197**.

A questão aqui é identificar quais destas classes apresentam maior probabilidade de corresponder ao interesse do usuário e descartar as outras, visando

assim reduzir a abrangência da busca e evitar que a mesma enverede por temas semanticamente irrelevantes, o que reduziria a eficiência do Motor de Busca (DRAGONI et al., 2012).

Para tal procedeu-se com um cálculo de relevância das classes identificadas **dentro do contexto dos parâmetros** fornecidos (não confundir com relevância no resultado da busca ou relevância dentro do *corpus*). Esta relevância será utilizada como critério de preservação de algumas classes seletoras e exclusão de outras.

Para o cálculo da relevância, algumas métricas intermediárias foram especialmente desenvolvidas com vistas a não apenas selecionar as classes mais prováveis a partir dos parâmetros de busca fornecidos, como também para valorar mais aquelas classes que utilizam mais destes parâmetros para sua seleção.

Para efeitos de convenção dentro desta discussão e exemplos que se seguem, deve-se entender:

a) **O** como sendo a ontologia **OntoLegis**, contendo **10.210** classes;
b) **D** como sendo o *corpus* de documentos coletados, formado por **59.386** decisões;
c) **P** como sendo o conjunto de todas as palavras fornecidas como parâmetros de busca, ou seja, **P = {"prestação", "serviço", "contrato", "multa"}**.

Inicia-se o cálculo da relevância a partir do cálculo do **grau de certeza individual** de cada parâmetro. Esta medida é calculada pela métrica **CertezaI(p, O)** formulada na Equação 5.

$$CertezaI(p,O) = \begin{cases} 1 + \frac{1}{|ClassesCom(O,p)|} \; quando \; |ClassesCom(p,O)| > 0 \\ 0 \; quando \; |ClassesCom(p,O)| = 0 \end{cases} \quad \textbf{(5)}$$

A certeza que uma palavra **p** fornece para identificar corretamente uma classe de **O** é a soma de 1 à razão entre 1 e a contagem de classes identificadas pela palavra **p**, calculada como o módulo do conjunto de classes retornado pela função **ClassesCom(p,O)**. Esta função simplesmente identifica quais classes apresentam algum rótulo com ao menos uma ocorrência de **p** e as retorna na forma de um conjunto.

O valor 1 é somado a cada razão por que esta métrica será utilizada posteriormente em um produtório e com valores maiores que 1 promoverá uma potencialização das certezas individuais quando consideradas em conjunto.

Por exemplo, dentro do *corpus* utilizado e na atual configuração da ontologia **OntoLegis**, para a palavra "**prestação**", o grau de certeza é dado pela equação 6.

$$CertezaI(\text{"prestação"}, O) = 1 + \frac{1}{|ClassesCom(\text{"prestação"}, O)|} = 1 + \frac{1}{34} = 1{,}0294 \quad \textbf{(6)}$$

Esta certeza é calculada para cada uma das palavras fornecidas como parâmetro. No Quadro 10 podem ser vistos os valores desta métrica para cada uma das quatro palavras utilizadas como exemplo.

Quadro 10 – Exemplos do cálculo de CertezaI(p) para diferentes palavras

Palavra p	ClassesCom(p,O)	CertezaI(p,O)
prestação	34	1,0294
serviço	103	1,0097
contrato	53	1,0187
multa	13	1,0769

Fonte: o Autor (2017)

A primeira coluna apresenta exemplos de palavras **p** submetidas como parâmetros para uma busca.

A segunda coluna apresenta o resultado da função **ClassesCom(p, O)**, que retorna a quantidade de classes existentes na ontologia **O** (neste caso, ontologia **OntoLegis**) que contém em ao menos um de seus rótulos uma ocorrência da palavra em questão.

A terceira coluna apresenta o cálculo da função **CertezaI(p, O)** que mede o grau de certeza que a respectiva palavra oferece na seleção de classes da ontologia em questão. Neste caso, a palavra "multa" oferece maior grau de certeza, pois existem apenas **13** classes com esta palavra em seus rótulos, enquanto "serviço" oferece menos certeza, pois existem **103** classes com esta palavra em seus rótulos.

O somatório de todas as classes identificadas a partir de cada uma das palavras (**34**, **103**, **53** e **13**) resulta em **203**, mas existem repetições de classes nestes conjuntos. O total de **classes distintas**, para este exemplo, é de **197**, ou seja, estas quatro palavras identificam total ou parcialmente, 197 classes da ontologia.

Em seguida é calculada a taxa de cobertura que todo o conjunto de parâmetros apresenta sobre os rótulos das 197 classes que foram identificadas. Esta cobertura serve para mensurar a correspondência existente entre os parâmetros e um rótulo de uma classe. Em outras palavras, a cobertura fornece um fator que mede quanto dos rótulos de uma classe está sendo mapeado pelos parâmetros de busca. Sua formulação está colocada na Equação 7.

$$Cobertura(P, c) = \max_{r_i \in R_c} \left(\frac{|ContidosEm(P, r_i)|}{|Significativas(r_i)|} \right) \quad \textbf{(7)}$$

Neste caso, $\mathbf{P}$ representa o conjunto de parâmetros, **c** representa uma classe dentre as 197 seletoras inicialmente identificadas e $\mathbf{r_i}$ representa cada um dos rótulos existentes em $\mathbf{R_c}$, que é o conjunto dos rótulos vinculados à classe **c**.

A função **Significativas(r_i)** retorna o conjunto das palavras individuais **significativas** (não correspondentes às *stopwords* vistas no Quadro 9) do rótulo, $\mathbf{r_i}$ enquanto que **ContidosEm(P, r_i)** representa o conjunto de palavras existentes em $\mathbf{P}$ que também existem no rótulo $\mathbf{r_i}$.

Por exemplo, seja ***c*** a classe "**AçãoDePrestaçãoDeContas**", para o seu rótulo "**ação de prestação de contas**", **ContidosEm(P, "ação de prestação de contas")** retorna o conjunto **{"prestação"}**, logo seu módulo vale **1** e **Significativas("ação de prestação de contas")** retorna **{"ação", "prestação", "contas"}**, portanto seu módulo vale **3**. Neste caso, **Cobertura(P, c)** = $\frac{1}{3}$ = 0,3333. Quanto mais palavras dos parâmetros forem encontradas nos rótulos, maior a cobertura dos parâmetros sobre estes rótulos.

Assim, calcula-se **Cobertura(P, c)** para cada um dos parâmetros e para cada uma das classes seletoras (**197** para o exemplo), o que permite ter o grau de cobertura dos parâmetros em relação a cada classe.

Para efeitos de exemplificação, tome-se a classe "**AçãoDePrestaçãoDeContas**" e seu único rótulo "**ação de prestação de contas**". Existe um único parâmetro de $\mathbf{P}$ ("**prestação**") presente no rótulo, composto por três palavras significativas, logo **Cobertura(P, "AçãoDePrestaçãoDeContas")** = 1/3 = **0,3333**.

Agora tome-se a classe “**PrestaçãoDeServiço**” e um de seus rótulos, “**prestação de serviço**”. Neste caso existem duas palavras de **P** (“**prestação**” e “**serviço**”) presentes no rótulo que é formado por duas palavras significativas, logo a **Cobertura(P, “PrestaçãoDeServiço”)** = 2/2 = 1.

Finalmente, tome-se a classe “**Serviço**” e um de seus rótulos “**serviço**”. Neste caso existe uma palavra de **P** (“**serviço**”) presente no rótulo, que é formado por também uma palavra. Logo a **Cobertura(P, “Serviço”)** = 1/1 = 1.

No Quadro 11 estão colocados os três valores calculados para efeitos de comparação.

Quadro 11 – Exemplos do cálculo de Cobertura(P, “Serviço”) para três classes existentes na ontologia OntoLegis

Classe c	Palavras em comum	Cobertura(P, c)
AçãoDePrestaçãoDeContas	`prestação`	0,3333
PrestaçãoDeServiço	`prestação` `serviço`	1,0000
Serviço	`serviço`	1,0000

Fonte: o Autor (2017)

A primeira coluna apresenta classes da ontologia, enquanto que a segunda apresenta palavras que estas classes apresentam em seus rótulos. Com base na contagem destas é possível calcular a cobertura, mostrada na terceira coluna.

A classe “PrestaçãoDeServiço” apresenta cobertura 1, pois todas as palavras significativas de ao menos um de seus rótulos (palavras “prestação” e “serviço” no rótulo “prestação de serviço”) encontram intersecção com os parâmetros de busca. O mesmo se dá com a classe “Serviço” (palavra “serviço” no rótulo “de serviço”).

Como estes exemplos demonstram, se apenas as coberturas forem consideradas, rótulos mais curtos (neste exemplo “serviço”) serão considerados tão relevantes quanto rótulos mais longos (neste exemplo “prestação de serviço”), o que viola a proposta de valorar mais as classes seletoras com mais parâmetros identificados em seus rótulos. Deseja-se considerar mais importante buscar pela classe “**PrestaçãoDeServiço**” do que pelas classes “**Prestação**” e “**Serviço**”. É para corrigir este desvio que existe o cálculo das certezas.

As certezas individuais são necessárias apenas como participantes do cálculo dos pesos das classes, que vão valorar mais a **combinação de palavras** do que as

palavras individuais. O peso que uma classe tem diante dos parâmetros de busca é dada pela Equação 8.

$$Peso(P, O, c) = \prod_{p_i \in P} CertezaI(p_i, O) \mid CertezaI(p_i, O) > 0, c \in O \quad \textbf{(8)}$$

Neste caso, $\boldsymbol{c}$ corresponde a uma das 197 classes selecionadas de $\mathbf{O}$ e $\mathbf{p_i}$ é um parâmetro existente em $\boldsymbol{P}$. Por tratar-se de um produtório, apenas fatores diferentes de zero devem figurar, daí a restrição de que apenas certezas maiores que zero devam ser consideradas.

A título de exemplificação, no Quadro 12 estão colocados os cálculos do peso para as classes "**AçãoDePrestaçãoDeContas**", "**Prestação**", "**PrestaçãoDeServiço**" e "**Serviço**".

Quadro 12 - Exemplos do cálculo de Peso(P,O,c)

Classe c	Peso(P,O,c)
AçãoDePrestaçãoDeContas	CertezaI("prestação") = 1,0294
Prestação	CertezaI("prestação") = 1,0294
PrestaçãoDeServiço	CertezaI("prestação").CertezaI("serviço") = 1,0294 . 1,0097 = 1,0394
Serviço	CertezaI("serviço") = 1,0097

Fonte: o Autor (2017)

O quadro anterior permite apreciar a valoração mais alta conferida ao peso da classe "**PrestaçãoDeServiço**" (1,0394), pois a esta foi avaliada como tendo uma maior correspondência entre seus rótulos e os parâmetros fornecidos.

Com estes cálculos intermediários realizados, é possível agora calcular a relevância de uma classe frente aos parâmetros fornecidos. Esta relevância vai estabelecer o critério para determinar quais das 197 classes identificadas pelos parâmetros em $\boldsymbol{P}$ permanecem na busca e quais serão excluídas.

A formulação da relevância de uma classe dentre as selecionadas frente aos parâmetros de busca fornecidos está colocada na Equação 9.

$$\mathrm{RelevanciaP(c, O, P)} = \mathrm{Peso(P, O, c)}\ \mathrm{Cobertura(P, c)} \quad \textbf{(9)}$$

A relevância conjuga peso e cobertura, de forma a avaliar simultaneamente o grau de correspondência dos parâmetros aos rótulos das classes (cobertura) e a combinação de parâmetros individuais (peso). O Quadro 13 mostra a lista com as 10

classes seletoras mais relevantes dentre as 197 selecionadas da ontologia **O** para os parâmetros em **P**.

Quadro 13 – As 10 classes mais relevantes para busca de documentos diante dos parâmetros “multa prestação serviço contrato”

#	Classe	Cobertura	Peso	RelevanciaP
01	Multa	1,0000	1,0769	1,0769
02	PrestaçãoDeServiço	1,0000	1,0394	1,0394
03	Prestação	1,0000	1,0294	1,0294
04	Contrato	1,0000	1,0189	1,0189
05	Serviço	1,0000	1,0097	1,0097
06	PrestaçãoDeSeriviçosÀComunidade	0,6667	1,0394	0,6929
07	MultaSubstitutiva	0,5000	1,0769	0,5385
08	Astreintes	0,5000	1,0769	0,5385
09	PenaDeMulta	0,5000	1,0769	0,5385
10	ViolaçãoDeContratoDeVendaÀPrestação	0,5000	1,0488	0,5244

Fonte: o Autor (2017)

Os cálculos realizados até o momento permitem valorar correspondências descobertas entre os parâmetros de busca e as classes disponíveis na ontologia, de forma a construir um conjunto inicial de classes candidatas a seletoras para a execução da recuperação de documentos, mas este conjunto precisa ser melhorado pois:

a) trata-se de um número expressivo de classes (197) o que implica em um conjunto expressivo de conceitos, que por sua vez aumentaria demasiadamente a amplitude da busca;
b) as classes foram mapeadas pela presença de palavras e nenhum critério semântico tomou lugar até o momento, o que significa não aproveitar a riqueza do conhecimento armazenado na ontologia para procurar estreitar o espaço de busca.

Assim, visando endereçar as duas situações, procedeu-se com uma segunda etapa de processamento dos parâmetros.

Inicialmente é realizada uma busca dentre as classes (197 para o exemplo corrente) para identificar a presença de classes relacionadas **verticalmente** na taxonomia, ou seja, se existem classes que guardam relação de superclasse e subclasse entre si.

A identificação da relação vertical é dada pelo cálculo da distância semântica entre duas classes. Se duas classes estão no mesmo ramo taxonômico, a distância semântica é dada pela contagem de arcos que as separam (RESNIK, 1995). É a forma mais simples de avaliação de similaridade semântica, mas é suficiente para os propósitos em vista.

Por exemplo, dentre as **197** classes foi identificada uma relação entre as classes "**ServiçoDeComunicação**" e "**ServiçoDeTelecomunicação**", onde a primeira é uma superclasse **direta** da segunda (distância semântica igual a **1**). Em tal situação a relevância de cada uma das classes é comparada. Se a superclasse (neste caso "**ServiçoDeComunicação**") tem relevância igual ou inferior a relevância de sua subclasse ("**ServiçoDeComunicação**"), será marcada para não ser expandida (processo a tomar lugar posteriormente), pois já existe uma subclasse mais relevante selecionada e que guarda algum grau de correspondência com os parâmetros de busca. Subclasses são classes mais específicas e por isso carregam mais informação, o que é interessante para incluir nos resultados de um Motor de Busca.

Se, por outro lado, a superclasse mostrar-se mais relevante que sua subclasse, ambas são mantidas sem qualquer cuidado particular.

Ainda dentro da intenção de reduzir o conjunto de classes a participarem da busca, outro processo é remover do conjunto de classes candidatas todas aquelas que apresentam cobertura inferior a 1 e que não apresentam relação semântica com as que apresentam cobertura igual a 1. Não ter relação semântica significa não participar como subclasse ou superclasse do mesmo ramo taxonômico.

Esta estratégia é uma aplicação direta da informação trazida pela estrutura taxonômica e permite uma filtragem de natureza semântica, o que vem ao encontro dos propósitos de uso da ontologia. Para o exemplo corrente, apenas cinco das 197 classes permaneceram no conjunto de classes candidatas ("**Multa**", "**PrestaçãoDeServiço**", "**Prestação**", "**Serviço**" e "**Contrato**").

Como última estratégia, procedeu-se com o que se chama expansão de consulta (do inglês, *query expansion*), onde a ideia é aumentar **criteriosamente** a amplitude da busca para procurar localizar mais documentos relacionados e ocorre somente nas classes remanescentes da filtragem semântica anterior.

Neste caso, expandir a consulta significa tão somente incluir no conjunto de classes todas as subclasses de todas as classes ainda presentes, excetuando-se as marcadas para não expansão.

Após todos estes processos, o que se tem é um conjunto de classes seletoras entendidas como as que o usuário da consulta tem interesse e que melhor caracterizam os documentos em vias de recuperação.

No caso de exemplo, das **197** classes inicialmente selecionadas a partir dos quatro parâmetros, apenas cinco classes restaram como seletoras ("**Multa**", "**PrestaçãoDeServiço**", "**Prestação**", "**Serviço**" e "**Contrato**"). O processo de expansão não conseguiu encontrar mais nenhuma classe com suficiente relevância para incluir.

De posse da lista de classes seletoras pelas quais se buscará os documentos, a próxima etapa é precisamente a recuperação destes.

4.5.5 Recuperação dos documentos

Uma vez que os documentos do *corpus* tenham sido indexados e que os parâmetros de busca tenham avaliados para a identificação das classes seletoras correspondentes, a recuperação se faz por um processo simples.

Cada uma das classes da ontologia foi vinculada a um conjunto de documentos que as referencia por meio de seus rótulos, logo é suficiente que as classes seletoras sejam utilizadas para a criação de uma relação de documentos.

É um processo direto, sem a necessidade de cálculos ou heurísticas. O documento será considerado apto a participar dos resultados se referenciar ao menos uma das classes seletoras.

No entanto, os resultados trazidos precisam ser mostrados em uma ordem de relevância semântica, pois esta ordem interessa ao usuário, que espera que documentos mais relacionados aos parâmetros de busca sejam mostrados antes.

Para atingir este objetivo é que existe um processo de ordenação dos resultados da busca.

4.5.6 Ordenação dos resultados

Como já colocado, a ordenação de resultados em um motor de busca guarda relação direta com a percepção de qualidade deste, pois os usuários em geral limitam-se a avaliar os dez primeiros resultados trazidos.

A recuperação produz uma lista de documentos e suas vinculações às classes que os identificaram. Alguns documentos poderão conter vínculos com mais de uma

classe seletora. Cada um destes vínculos participará do cálculo do peso do documento dentro dos resultados.

Participam no cálculo do peso de um documento as relevâncias das classes seletoras que o selecionaram e o **TF-IDF** da mesma classe dentro do documento, segundo a formulação presente na Equação 10.

$$\mathrm{PesoD}(\mathrm{d_i}, \mathrm{S}) = \prod_{\mathrm{c}\in\mathrm{S}}(\mathrm{Relevancia_c}\ \mathrm{TF}-\mathrm{IDF_c}) \quad \textbf{(10)}$$

O peso de um documento $\mathbf{d_i}$ corresponde ao produtório da conjunção da relevância das classes seletoras ($\mathbf{Relevancia_c}$) e da relevância da classe presente no documento ($\mathbf{TF\text{-}IDF_c}$). As classes ***c*** estão presentes no conjunto **S**, que corresponde a todas as classes seletoras que estão no rol de classes que caracterizam o documento $\mathbf{d_i}$ do *corpus* **D**.

Por exemplo, se um dado documento $\mathbf{d_1}$ é caracterizado pela presença de uma classe com relevância **8** e outro documento $\mathbf{d_2}$ é caracterizado pela presença de outras duas classes com relevâncias **3** e **4**, assumindo-se para todos os casos o **TF-IDF = 1**, o cálculo do peso para o documento $\mathbf{d_1}$ corresponderia ao visto na Equação 11.

$$\mathrm{PesoD}(\mathrm{d_1}, \{\mathrm{c_1}\}) = 8\,.\,1 = 8 \quad \textbf{(11)}$$

Por outro lado, o cálculo do peso para o documento $\boldsymbol{d_2}$ equivaleria ao formulado na Equação 12.

$$\mathrm{PesoD}(\mathrm{d_2}, \{\mathrm{c_2}, \mathrm{c_3}\}) = (3\,.\,1).(4\,.\,1) = 12 \quad \textbf{(12)}$$

O propósito de relacionar todos os pesos de uma classe por meio de um produtório é de potencializar o **efeito combinatorial** das diversas classes identificadas, valorizando mais as ocorrências combinadas do que as isoladas.

A multiplicação pelo **TF-IDF** de cada classe seletora que ocorre no documento tem o propósito de levar em conta também o grau de discriminância que a mesma apresenta diante do documento recuperado em particular e do *corpus*.

Assim, o peso de uma classe seletora em um documento conjunga a relevância desta classe diante dos parâmetros fornecidos (certeza combinada com cobertura) e relevância da classe para caracterizar o documento.

O cálculo do peso é realizado para cada um dos documentos existentes no conjunto resultado, que então é ordenado decrescentemente por esta métrica. Como

resultado, os documentos com maiores pesos são mostrados antes, contribuindo para o propósito de ter no topo da lista os documentos avaliados como os mais desejados pelo usuário.

4.6 Avaliação da proposta por advogados consultores

Foi considerado importante avaliar a proposta e as capacidades do Motor de Busca Semântica em comparação com a tecnologia *full-text search* por meio da consulta a profissionais do Direito. A estratégia para tal avaliação foi de realizar demonstrações a dois advogados consultores para submeter o produto à sua operação.

Ambos os consultores são advogados profissionais atuantes. Um deles tem formação em Direito desde 2006, com três pós-graduações e também titulação de mestre. O outro tem formação em Direito desde 2005, com duas pós-graduações.

A estratégia foi dar liberdade para os dois profissionais e colher suas impressões resultantes da experimentação do produto. Os profissionais realizaram algumas consultas e relataram suas opiniões frente ao que puderam recuperar do *corpus* que estava à disposição. Em ambos os casos os entrevistados relataram:

a) a ocorrência do Problema da Recuperação da Informação é uma experiência diária em matéria de recuperação jurisprudencial;
b) consideraram a proposta de recuperação semântica interessante e potencialmente benéfica para a atividade de recuperação jurisprudencial.

Um dos entrevistados comentou que a vinculação de referências a leis no texto das decisões poderia contribuir para tornar a recuperação mais eficiente. Nesta oportunidade, o entrevistado comentou que apreciaria poder consultar simplesmente por números de parágrafos de leis, sem ter que referenciar quaisquer palavras diretamente. Neste caso, o Motor de Busca teria a responsabilidade de identificar as classes da ontologia que estivessem relacionadas com estes parágrafos, como uma alternativa ao fornecimento de palavras.

O outro entrevistado evidenciou a importância de recuperar as decisões levando-se em conta também o relator do processo, na forma de um filtro de pesquisa, visto que os advogados constroem suas estratégias também com base na forma de interpretação que certos magistrados tendem a empregar. Neste contexto, o relator

tanto poderia ser utilizado como critério de inclusão como de exclusão de decisões no resultado da busca.

Um dos entrevistados teve a oportunidade de perceber a diferença de volume de decisões retornado pelo protótipo quando comparado com o sítio do TJPR. A discrepância percebida causou estranheza daquele profissional, mas foi considerada dentro dos parâmetros pois:

a) um Motor de Busca Semântica deixa de trazer resultados que um motor *full-text* traria, visto que em alguns casos, palavras isoladas não formam conceitos e por isso os documentos onde ocorrem não são recuperados;
b) um Motor de Busca Semântica pode trazer resultados que um motor *full-text* não traria, visto que as classes podem ser vinculadas por meio de diversas representações léxicas diferentes, além de realizar o processo de expansão de consulta para abranger mais documentos.

Nesta oportunidade, este mesmo profissional percebeu que variações de singular e plural dos rótulos correspondem ao mesmo conceito, evidenciando uma das virtudes do Motor de Busca Semântica.

Por questões de prazo, as sugestões oferecidas pelos consultores não puderam ser implementadas e, portanto, foram registradas como oportunidades para trabalhos futuros.

4.7 Experimentos

Para medir o grau de eficiência do Motor de Busca Semântica, o que também contribui para avaliar o grau de resultado da ontologia **OntoLegis**, foram realizados experimentos comparativos com o serviço de recuperação jurisprudencial do TJPR, acessível por meio do endereço **http://portal.tjpr.jus.br/jurisprudencia**.

De forma a manter a comparação justa, todas a execuções sobre o sítio do TJPR foram executadas na modalidade “Pesquisa Detalhada”, na qual foram incluídas as seguintes opções de busca:

a) pesquisar “na Íntegra do Acórdão”, visto que o Motor de Busca Semântica indexou unicamente o texto dos Acórdãos e Decisões Monocráticas, não das ementas, que são resumos das decisões;

b) critério de ordenação “mais relevantes primeiro”, pois o Motor de Busca Semântica efetua a ordenação por relevância;
c) julgamentos com datas delimitadas pelo período de 05 de fevereiro de 2016 e 01 de novembro de 2016, que corresponde ao existente na base de dados semanticamente indexada.

O processo de comparação consistiu em submeter um conjunto de palavras com parâmetros a cada um dos sistemas de busca (Motor de Busca Semântica e sítio do TJPR) e avaliar indicadores dos três primeiros documentos retornados em cada caso. Os experimentos são detalhados nas seções a seguir.

4.7.1 Busca por “veículo dano material indenização”

As palavras submetidas foram “veículo”, “dano”, “material” e “indenização” que, no caso do Motor de Busca semântica, foram traduzidas para as seguintes classes já em ordem de relevância:

1) Indenização;
2) DanoMaterial;
3) VeículoAutomotor;
4) PréviaIndenização;
5) JustaIndenização;
6) IndenizaçãoTrabalhista;
7) AçãoDeIndenização;
8) DesapropriaçãoSemIndenização;
9) DanoMoralReflexo;
10) DanoÀVidaDeRelação.

A busca semântica com base nestas classes seletoras retornou **12.709** decisões, sendo que as três primeiras podem ser vistas no Quadro 14.

Quadro 14 – Resultados da busca por "veículo dano material indenização" submetida ao Motor de Busca Semântica

Número do Processo	**Classes identificadas**
0027421-29.2015.8.16.0030/0	7 x Indenização (TF-IDF = 6,00) 8 x DanoMaterial (TF-IDF = 13,23) 5 x VeículoAutomotor (TF-IDF = 7,34)
1562685-0	2 x Indenização (TF-IDF = 1,71) 3 x DanoMaterial (TF-IDF = 4,96) 12 x VeículoAutomotor (TF-IDF = 17,62) 3 x AçãoDeIndenização (TF-IDF = 2,34)
1304625-0	7 x Indenização (TF-IDF = 6,00) 6 x DanoMaterial (TF-IDF = 9,92) 1 x VeículoAutomotor (TF-IDF = 1,47) 5 x AçãoDeIndenização (TF-IDF = 3,90)

Fonte: o Autor (2017)

Na primeira coluna estão listados os números dos processos cujas decisões foram retornadas, enquanto que na segunda coluna estão os conceitos localizados, suas frequências naquele documento e o TF-IDF correspondente. Os resultados da busca semântica parecem estar compatíveis com a expectativa, onde a quantidade, diversidade e relevância de conceitos identificados influencia diretamente na posição relativa dos documentos recuperados.

Por outro lado, a busca no sítio do TJPR retornou **2.695** decisões, sendo que as três primeiras podem ser vistas no Quadro 15.

Quadro 15 – Resultados da busca por "veículo dano material indenização" submetida ao sítio do TJPR

Número do Processo	**Termos identificados**
0075975-43.2015.8.16.0014/0	1 x Veículo 3 x Dano 1 x Material 7 x Indenização
0004363-52.2015.8.16.0044/0	2 x Veículo 19 x Dano 5 x Material 9 x Indenização
0000545-04.2015.8.16.0041/0	2 x Veículo 12 x Dano 2 x Material 4 x Indenização

Fonte: o Autor (2017)

Apesar da busca do TJPR estar ajustada para ordenar seus resultados por relevância, a ordem apresentada das três primeiras decisões não parece corresponder à mera frequência e diversidade dos termos localizados. O sítio do TJPR não informa qual política é utilizada na ordenação por relevância.

Outro fato notável é a diferença numérica da quantidade de resultados recuperados (**2.695** na busca *full-text* contra **12.709** na busca semântica). Este fenômeno se explica por diversos motivos:

a) na busca semântica os parâmetros foram traduzidos para 10 classes, o que aumenta o espaço de resultados possíveis, aumentando a taxa de revocação;
b) a classe "**VeículoAutomotor**" é identificada por três rótulos diferenciados: "veículo", "veículo automotor" e "automóvel", permitindo que rótulos diferentes remetam a ocorrência do mesmo conceito;
c) a classe "**DanoMaterial**" é identificada por quatro rótulos diferenciados: "dano material", "danos materiais", "dano de natureza material" e "danos de natureza material", permitindo que rótulos diferentes remetam a ocorrência do mesmo conceito;
d) a classe "**DanoMoralReflexo**" foi selecionada a partir dos parâmetros e a busca passa a incluir também os documentos que apresentam ocorrência desta classe;
e) a classe "**IndenizaçãoTrabalhista**" foi selecionada a partir dos parâmetros e assim a busca passa a incluir documentos que apresentam ocorrência desta classe;
f) a busca por termos exatos não leva em conta a forma plural dos parâmetros colocados e nem expande a busca para outros conceitos relacionados como "**JustaIndenização**" ou "**PréviaIndenização**".

Percebeu-se, neste caso, que a busca semântica apresentou uma de suas virtudes, no sentido de ultrapassar as formas exatas dos parâmetros fornecidos, aceitando as múltiplas alternativas para identificar o conceito procurado.

Por outro lado, o processo de identificação de classes seletoras incluiu algumas classes que não parecem estar relacionadas com os interesses do consultante, o que aumentou erroneamente o espaço (reduzindo a taxa de precisão) de busca e remete à necessidade de aprimoramento.

4.7.2 Busca por “roubo próprio”

As palavras submetidas foram “roubo” e “próprio” que, no caso do Motor de Busca semântica, foram traduzidas para a classe “**RouboPróprio**”.

A busca semântica com base nesta classe seletora não encontrou nenhum documento.

A busca no sítio do TJPR retornou **4.201** decisões, sendo que as três primeiras podem ser vistas no Quadro 16.

Quadro 16 – Resultados da busca por "roubo próprio" submetida ao sítio do TJPR

Número do Processo	**Termos identificados**
1580677-6	2 Roubo 1 Próprio
1587819-2	1 Roubo 0 Próprio
1559356-9	1 Roubo 0 Próprio

Fonte: o Autor (2017)

Os resultados da busca no sítio do TJPR apontam aparente existência de apenas um documento que inclua as duas palavras, mas o fato de as mesmas apresentarem frequências diferentes sugerem que não são utilizadas combinadamente.

Uma avaliação do texto da decisão recuperada em primeiro lugar confirma a suposição. As três ocorrências dos parâmetros passados são:

a) “HABEAS CORPUS - ROUBO- AUSÊNCIA DE...”;
b) “...preso preventivamente pelo injusto de roubo.”;
c) “...cabível o recurso próprio, situação que...”.

Nenhuma das ocorrências apresenta a conjunção dos parâmetros, o que corresponderia à ocorrência da classe “**RouboPróprio**”. Este é um efeito frequente em motores *full-text search*. As demais decisões recuperadas incluem documentos que apresentam ao menos uma das palavras usadas como parâmetros.

Entende-se que aqui o Motor de Busca Semântica apresentou o grau de seletividade desejado para melhor corresponder aos anseios do consultante, pois este declarou por meio dos parâmetros que tinha interesse na conjunção das duas palavras.

4.7.3 Busca por “maus tratos menor”

As palavras submetidas foram “maus”, “tratos” e “menor” que, no caso do Motor de Busca semântica, foram traduzidas para as seguintes classes já em ordem de relevância:

1) MausTratos;
2) Menor.

A busca semântica com base nestas classes seletoras retornou **1.381** decisões, sendo que as três primeiras podem ser vistas no Quadro 17.

Quadro 17 – Resultados da busca por "maus tratos menor" submetida ao Motor de Busca Semântica

Número do Processo	**Classes identificadas**
1517683-1	1 x MausTratos (TF*IDF = 11,32) 4 x Menor (TF*IDF = 6,78)
1506075-2	33 x Menor (TF*IDF = 55,97)
1456511-6	1 x MausTratos (TF*IDF = 11,32) 2 x Menor (TF*IDF = 3,39)

Fonte: o Autor (2017)

Já a busca no sítio do TJPR retornou **177** decisões, sendo que as três primeiras podem ser vistas no Quadro 18. Atenção ao fato de que o terceiro documento na verdade é o quarto item recuperado pelo TJPR. O terceiro documento efetivo estava classificado como segredo de justiça e não pôde ser avaliado. Apesar do número do documento figurar entre os demais, o texto de sua decisão encontrava-se inacessível.

Quadro 18 – Resultados da busca por "maus tratos menor" submetida ao sítio do TJPR

Número do Processo	**Termos identificados**
1548002-9	1 x Maus 1 x Tratos 17 x Menor
1417397-8	1 x Maus 1 x Tratos 20 x Menor
0000516-84.2015.8.16.0030/0	Dados indisponíveis por tratar-se de segredo de justiça, cujo acórdão encontra-se inacessível

Fonte: o Autor (2017)

O Motor de Busca Semântica retornou mais resultados que o sítio do TJPR. Não é possível identificar o motivo, pois as classes "**MausTratos**" e "**Menor**" apresentam apenas os rótulos "maus tratos" e "menor", respectivamente. Neste sentido, não é possível explicar por qual motivo o sítio do TJPR não foi capaz de identificar aqueles documentos, visto que os critérios de busca especificam precisamente aquelas mesmas palavras.

Com relação ao Motor de Busca Semântica, percebeu-se que a maioria absoluta dos documentos (1.377 dos 1.381) continham apenas a classe "Menor" (representada pelo rótulo "menor"), não no conceito de "menor de idade", mas simplesmente com a função de advérbio de intensidade (por exemplo "menor complexidade").

O segundo documento trazido na busca semântica contém apenas a classe "Menor", que se explica pela frequência (33 ocorrências). Este volume de ocorrências fez com que o peso daquele documento superasse o terceiro colocado que apresenta frequência menor dos dois conceitos.

Os outros dois documentos que apresentam a conjunção dos dois conceitos encontram-se nas posições 11 (processo número 1488162-0, com 1 x MausTratos e 1 x Menor) e 12 (processo número 1283497-4, com 1 x MausTratos e 1 x Menor). Este efeito contraria a intenção de valorar documentos que apresentam conjunção de conceitos presentes, o que sugere que o método para determinação da relevância de documentos precisa ser revisto.

Este é também um exemplo que demostra que a simples identificação de ocorrência de um rótulo menos específico, como é o caso da palavra "menor", leva à incorreta classificação de um documento. Neste caso, para que o equívoco não tivesse lugar, o rótulo em questão teria que ter o sentido de "menor de idade", a ser reconhecido por meio de uma análise linguística de seu emprego em uma frase.

4.7.4 Busca por "busca apreensão inadimplência"

As palavras submetidas foram "busca", "apreensão" e "inadimplência" que, no caso do Motor de Busca semântica, foram traduzidas para as seguintes classes já em ordem de relevância:

1) Inadimplemento;
2) BuscaEApreensão;

3) Apreensão;
4) BuscaDomiciliar;
5) BuscaPessoal.

A busca semântica com base nestas classes seletoras retornou **3.183** decisões, sendo que as três primeiras podem ser vistas no Quadro 19.

Quadro 19 – Resultados da busca por “busca apreensão inadimplência” submetida ao Motor de Busca Semântica

Número do Processo	**Classes identificadas**
1535261-3	5 x Inadimplemento (TF*IDF = 15,24) 11 x BuscaEApreensão (TF*IDF = 24,89) 13 x Apreensão (TF*IDF = 24,05)
1540580-6	8 x Inadimplemento (TF*IDF = 24,38) 6 x BuscaEApreensão (TF*IDF = 13,58) 11 x Apreensão (TF*IDF = 20,35)
1477403-9	3 x Inadimplemento (TF*IDF = 9,14) 8 x BuscaEApreensão (TF*IDF = 18,10) 20 x Apreensão (TF*IDF = 37,00)

Fonte: o Autor (2017)

A busca no sítio do TJPR retornou **3.165** decisões, sendo que as três primeiras podem ser vistas no Quadro 20.

Quadro 20 – Resultados da busca por “busca apreensão inadimplência” submetida ao sítio do TJPR

Número do Processo	**Termos identificados**
1547115-7	8 Busca 8 Apreensão 2 Inadimplência
1549756-6	23 Busca 24 Apreensão 1 Inadimplência
1474053-7	1 Busca 1 Apreensão 2 Inadimplência

Fonte: o Autor (2017)

Neste caso a busca semântica conseguiu identificar maiores ocorrências da classe “Inadimplemento”, pois esta inclui os rótulos “inadimplemento” e “inadimplência”, o que aumentou a sensibilidade da busca a variações da palavra. Ainda assim, o volume de respostas foi menor, o que sugere maior precisão na recuperação. Por exemplo, o primeiro documento recuperado pela busca semântica

apresenta apenas ocorrências da palavra “inadimplemento”, que não foi recuperado pela busca *full-text* do sítio do TJPR.

Este experimento mostra que a busca *full-text* também pode incorrer na situação valorar mais um documento com menor variedade de parâmetros presentes. No caso do segundo documento, não existem ocorrências da palavra “inadimplência”, mas a alta frequência das outras duas fez com que este recebesse maior relevância que o terceiro colocado, que apresenta as três palavras em menor quantidade.

4.7.5 Busca por “habeas corpus homicídio”

As palavras submetidas foram “habeas”, “corpus” e “homicídio” que, no caso do Motor de Busca semântica, foram traduzidas para as seguintes classes já em ordem de relevância:

1) HabeasCorpus;
2) Homicídio;
3) PrimeiroHabeasCorpus;
4) HabeasCorpusPreventivo;
5) SegundoHabeasCorpus;
6) HabeasCorpusDeOfício;
7) HabeasCorpusLiberatório;
8) RecursoOrdinárioEmHabeasCorpus;
9) HomicídioDoloso;
10) HomicídioSimples;
11) HomicídioCulposo;
12) HomicídioQualificadoPrivilegiado;
13) HomicídioPrivilegiado;
14) HomicídioQualificado;
15) HabeasCorpusSubstitutivoDeRecursoOrdinário.

A busca semântica com base nestas classes seletoras retornou **2.300** decisões, sendo que as três primeiras podem ser vistas no Quadro 21.

Quadro 21 – Resultados da busca por “habeas corpus homicídio” submetida ao Motor de Busca Semântica

Número do Processo	**Classes identificadas**
1534379-6	8 x HabeasCorpus (TF*IDF = 13,99) 3 x Homicídio (TF*IDF = 9,60) 1 x PrimeiroHabeasCorpus (TF*IDF = 3,54) 1 x HabeasCorpusLiberatório (TF*IDF = 2,58) 2 x HomicídioQualificado (TF*IDF = 3,12)
1526583-5	7 x HabeasCorpus (TF*IDF = 12,24) 3 x Homicídio (TF*IDF = 9,60) 2 x HomicídioSimples (TF*IDF = 4,20) 2 x HomicídioQualificado (TF*IDF = 3,12)
1549849-6	7 x HabeasCorpus (TF*IDF = 12,24) 3 x Homicídio (TF*IDF = 9,60) 2 x HomicídioSimples (TF*IDF = 4,20) 2 x HomicídioQualificado (TF*IDF = 3,12)

Fonte: o Autor (2017)

Por outro lado, a busca no sítio do TJPR retornou **2.750** decisões, sendo que as três primeiras podem ser vistas no Quadro 22.

Quadro 22 – Resultados da busca por "habeas corpus homicídio" submetida ao sítio do TJPR

Número do Processo	**Termos identificados**
1506508-6/01	3 Habeas 3 Corpus 1 Homicídio
1548677-6	1 Habeas 1 Corpus 3 Homicídio
1546008-3	5 Habeas 5 Corpus 2 Homicídio

Fonte: o Autor (2017)

Apesar da busca do TJPR estar ajustada para ordenar os resultados por relevância, a ordem apresentada dos três primeiros resultados não parece corresponder à mera contagem de ocorrência dos termos localizados.

Por outro lado, a busca semântica valorou corretamente a relevância das repostas levando em conta a variedade da conjunção de conceitos presentes.

A busca semântica retornou uma quantidade menor de documentos (2.300 contra 2.750 da busca do sítio do TJPR), mas foi capaz de identificar formas

especializadas de homicídio, representadas pelas classes “**HomicídioQualificado**” e “**HomocídioSimples**” além de formas especializadas de “**HabeasCorpus**” (“**HabeasCorpusLiberatório**” e “**PrimeiroHabeasCorpus**”), o que sugere que se a busca tivesse sido por estes critérios, os resultados trazidos seriam mais específicos e mais interessantes para o usuário.

4.8 Avaliação do protótipo pelo Grupo de Estudos em Direito Eletrônico

Com o objetivo de enriquecer a avaliação do Motor de Busca Semântica implementado pelo sítio ConsultaJuris, um questionário para coleta de opinião foi disponibilizado. As respostas foram coletadas, tabuladas e estão apresentadas na próxima seção.

4.8.1 Análise de respostas

Ao todo foram registradas 10 (dez) participações no período de 02/01/2017 a 14/01/2017. O período, além de ser considerado curto, coincidiu também com o recesso dos tribunais e férias dos advogados, o que possivelmente contribuiu para uma baixa aderência à pesquisa.

As respostas ocorreram em dois períodos bem distintos, sendo o primeiro de 04/01/2017 a 06/01/2017 e o segundo de 07/01/2017 a 14/01/2017. Os períodos são considerados diferentes por que as versões do Motor de Busca Semântica foram também diferentes.

O primeiro período ofereceu respostas que influenciaram critérios de valoração das classes seletoras e permitiriam o aperfeiçoamento do processo de tradução de parâmetros em classes.

Por tratar-se de um volume pequeno, todas as respostas do primeiro período estão colocadas no Quadro 23.

Quadro 23 – Respostas coletadas no questionário de opinião existente no sítio ConsultaJuris no período de 4 a 6 de janeiro de 2017

#	Q1	Q2	Q3	Q4
01	Péssimo	Ruim	Sim	A descrição semântica deve reforçar ou negar termos. Por exemplo: Ao busca "roubo Próprio" ele deveria reforçar o conceito "rouboProprio", mas negar "rouboImproprio" (visto serem tipos diferentes). Buscas específicas tem resultado piores que a booleana.
02	Indiferente	Ruim	Sim	

03	Indiferente	Péssimo	Não	
04	Indiferente	Equivalente	Sim	O projeto é muito interessante. Faço algumas sugestões: - Permitir controlar os temas correlatos aplicados na busca, atribuindo pesos. - incluir operadores lógicos (E, OU, NÃO)

Fonte: o Autor (2017)

De forma geral, a percepção deixada pelas respostas é de que a experiência com o ConsultaJuris é pior do que o TJPR (duas respostas “Ruim”, uma “Péssimo” e uma “Equivalente”) e “Indiferente” em relação à recuperação jurisprudencial em geral.

A primeira resposta forneceu um exemplo que influenciou a forma como as classes seletoras são tratadas. Originalmente, o processo de tradução de parâmetros em classes não levava em conta a relação vertical entre classes, ou seja, a existência de uma superclasse entre as classes seletoras.

Tomando-se o exemplo fornecido pelo próprio respondente, o que se tinha é que as classes seletoras eram “**Roubo**” e “**RouboPróprio**”. A classe “**Roubo**” era mantida por ter um grau de relevância elevado (ainda que inferior a “**RouboPróprio**”) e em seguida era expandida, onde ocorria a inclusão de suas subclasses, “**RouboPróprio**” (já existente) e “**RouboImpróprio**”. Por causa da expansão de classes, os parâmetros eram, ao final, traduzidos para as classes “**Roubo**”, “**RouboPróprio**” e “**RouboImpróprio**”, incorrendo na situação relatada pelo respondente.

Com a análise de relação entre as seletoras **antes** da expansão, foi detectada que a classe “**Roubo**” era superclasse de “**RouboPróprio**” e por isso fora marcada para não ser expandida. Ela também foi removida por apresentar uma relevância menor que sua subclasse. Como resultado, apenas a classe “**RouboPróprio**” persistiu como seletora. Os efeitos desta medida estão demonstrados em um dos experimentos executados, descrito em detalhes na seção 4.7.2.

Um outro respondente sugeriu que o ConsultaJuris permitisse interferir nos pesos das classes seletoras para controlar indiretamente a ordenação dos resultados, além de permitir a utilização de operadores booleanos. Restrições de tempo e objetivos, no entanto, impediram sua adoção.

Após a atualização do sítio para a inclusão do novo critério de tradução de parâmetros, mas sem qualquer alteração na indexação dos documentos, um novo

conjunto de respostas foi coletado durante o segundo período. Estas respostas podem ser vistas no Quadro 24.

Quadro 24 – Respostas coletadas no questionário de opinião existente no sítio ConsultaJuris no período de 7 a 14 de janeiro de 2017

#	Q1	Q2	Q3	Q4
05	Indiferente	Equivalente	Sim	
06	Indiferente	Ruim	Sim	Indico a inclusão de mais tribunais da Justiça Federal e do Trabalho. Assim como dos tribunais superiores STJ e STF
07	Péssimo	Péssimo	Sim	É mais interessante ver apenas a ementa, e não o texto completo da decisão. A seleção através de ementas é bem mais rápida. Minha pesquisa retornou resultados fracos e é mais rápido pesquisar com as mesmas palavras chaves no portal do TJPR.
08	Bom	Bom	Sim	Muito fácil de ser utilizado.
09	Indiferente	Ruim	Sim	
10	Indiferente	Ruim	Sim	Não mostrar elementos de nível superior ou igual ao pesquisado. Exemplo: prisão domiciliar retorna prisão e prisão em flagrante, ou seja, nível superior e igual. Já prisão em flagrante, funciona bem, pois traz além do termo as modalidades de flagrante.

Fonte: o Autor (2017)

A percepção geral de eficiência do ConsultaJuris na recuperação de decisões não parece ter mudado substancialmente, visto constarem quatro ocorrências de “Indiferente” e uma de “Péssimo”, mas houve uma ocorrência de “Bom”. Quando comparado com o TJPR, a percepção de qualidade continua inferior à do TJPR, mas houve também uma ocorrência de “Bom”.

Um dos respondentes sugeriu o aumento da base de dados para inclusão de outros tribunais. Outro questionou a proposta de indexar o texto completo da decisão, sugerindo unicamente o uso da ementa. No entanto, este tema parece controverso, visto que um dos advogados consultores entrevistados (vide seção 4.6) argumentou precisamente o contrário.

Houve ainda um respondente que enalteceu a operação simples do ConsultaJuris.

Um último respondente fez um comentário considerado valioso. O mesmo comparou duas buscas, sendo uma “prisão domiciliar” e outra “prisão flagrante”. No primeiro caso o respondente afirma que além da classe seletora “PrisãoDomiciliar”,

também foi incluída na busca a classe “Prisão”. Neste contexto, o respondente julgou que a classe “Prisão” não deveria constar por seu uma classe de “nível superior”, ou seja, uma superclasse.

A análise da ontologia mostrou que não existe relação de hierarquia entre “**Prisão**” e “**PrisãoDomiciliar**”. Para esta última consta “**RegimeDeCumprimentoDePena**” como sua superclasse.

Neste caso, não foi possível para o tradutor de parâmetros remover a classe “**Prisão**”, pois a mesma apresenta relevância elevada e nenhuma relação de hierarquia com “**PrisãoDomiciliar**”.

O caso de “prisão flagrante” é diferente, pois os parâmetros são traduzidos para as classes “**Prisão**” e “**PrisãoEmFlagrante**” que, neste caso, apresentam relação de hierarquia. A primeira é superclasse da segunda. Neste caso, o tradutor de parâmetros foi capaz de identificar a relação e remover a classe “**Prisão**” do rol de seletoras. Este efeito é resultado direto da aplicação da sugestão da primeira resposta, envolvendo as classes “**Roubo**” e “**RouboPróprio**”.

A tabulação de todas as respostas permitiu concluir que de forma geral os resultados recuperados pelo ConsultaJuris foram indiferentes (7 avaliações), não demonstrando qualidade superior ou inferior a quaisquer outros sistemas de busca, mas houve duas avaliações como “péssimo” e uma como “bom”. Já quando comparado com o TJPR, o ConsultaJuris foi, de forma geral, considerado inferior, com duas ocorrências de “Péssimo”, cinco ocorrências de “Ruim”, duas ocorrências de “Equivalente” e uma ocorrência de “Bom”.

Ainda que constituindo uma amostra estatisticamente insignificante, as respostas foram suficientemente ricas para influenciar uma melhoria no motor de busca que encontrou resultado.

Diante de todos os processos levados a efeito no intuito de alcançar uma resposta para a questão de pesquisa enunciada e também os objetivos geral e específicos declarados, as análises realizadas mostraram que um Motor de Busca Semântica está em condições de responder ao anseio por uma recuperação jurisprudencial mais eficiente e precisa.

Outras explorações dos resultados atingidos e analisados estão pormenorizados na seção a seguir.

5 CONSIDERAÇÕES FINAIS

O referencial teórico demonstrou que o tema já vem recebendo atenção tanto fora do Brasil quando no exterior, mas permanece sem evoluções significativas. O mesmo referencial teórico, no entanto, apontou a Web Semântica como a área de estudos que parece prover, se não a solução, ao menos tecnologias promissoras no sentido de superar a limitação atual.

O Motor de Busca Semântica é apontado tanto pela Web Semântica como por diversos autores como a melhor tecnologia para a recuperação de informações em grandes volumes de dados.

No âmbito do Brasil, o relatório Justiça em Números 2016 demonstrou que o volume anual de produção de documentos jurisprudenciais apresenta tendência crescente e já atingiu o patamar de 27,2 milhões no ano de 2015, caracterizando este campo como um candidato a beneficiar-se da recuperação semântica.

Neste sentido, esta pesquisa objetivou construir uma ontologia própria para servir a um Motor de Busca Semântica a ser especialmente empregado na recuperação jurisprudencial. As próximas seções confrontam e discutem os objetivos determinados para a orientação da pesquisa e os resultados atingidos, mensurados e analisados.

5.1 Direcionamento da pesquisa

No contexto apresentado, a seguinte questão de pesquisa foi enunciada: **Quais conceitos precisam compor uma ontologia para que esta suporte um Motor de Busca Semântica que possa reduzir o Problema da Recuperação da Informação no âmbito da recuperação jurisprudencial no Brasil?**

Apesar de não haver propriamente uma padronização de formato para a elaboração dos documentos jurisprudenciais (Sentença, Decisão Monocrática e Acórdão), existe um vocabulário frequentemente utilizado pelos magistrados na redação daqueles documentos. A recorrência deste vocabulário motivou a **Secretaria de Jurisprudência** do **Supremo Tribunal de Justiça** (STJ) do Brasil, a conceber e manter um **Vocabulário Jurídico Controlado** (aqui referenciado como **Tesauro Jurídico do STJ**), que enumera, define e organiza um conjunto de mais de **13.000** termos e expressões técnicas frequentemente encontradas nas decisões.

Tal tesauro constituiu ponto de partida para a construção de uma ontologia peso leve chamada **OntoLegis**, que engloba e organiza conceitos e rótulos léxicos dedicados a suprir um Motor de Busca Semântica para a recuperação conceitual de documentos jurisprudenciais respondendo, assim, a questão de pesquisa.

A respeito da ontologia OntoLegis é possível afirmar:

a) foi concebida a partir da combinação da ontologia **JurisTJPR** e do **Tesauro Jurídico** do STJ;
b) é composta por **10.210** classes caracterizadas por **12.595** rótulos;
c) é uma ontologia polierárquica, ou seja, algumas classes apresentam mais de uma superclasse;
d) apresenta **116** classes (0,0011%) com ambiguidade léxica, ou seja, que compartilham as mesmas expressões características;
e) está armazenada na forma de um documento OWL, denominado **OntoLegis-2016-06.owl**, compatível com a recomendação **Semantic Web** do **World Wide Web Consortium**;

A ontologia OntoLegis foi construída a partir de dois recursos informacionais previamente validados por seus idealizadores, ou seja, a ontologia JurisTJPR proposta por Molinari em sua dissertação de mestrado e o Tesauro Jurídico do STJ que é mantido por profissionais especialistas na área, a partir do estudo das próprias decisões proferidas e publicadas pelo Supremo Tribunal de Justiça.

Entende-se que a combinação destes dois artefatos fornece um conjunto de conceitos organizados e anotados por rótulos léxicos que respondem à questão de pesquisa, visto que, a partir de experimentos, apresentam suficiente poder de cobertura para, ao menos como ponto de partida, classificar semanticamente o *corpus* de decisões existentes nos tribunais do país.

5.2 Resposta aos objetivos

De forma a promover esforços para responder à questão de pesquisa, o seguinte objetivo geral foi estabelecido para nortear esta pesquisa científica: **elaborar uma ontologia que suporte um Motor de Busca Semântica que possa reduzir o Problema da Recuperação da Informação no domínio da consulta jurisprudencial no Brasil**.

No intuito de reduzir a complexidade do objetivo geral e estabelecer etapas para o processo de pesquisa, os seguintes objetivos específicos foram enunciados:

a) elaborar uma ontologia preliminar a partir do Tesauro Jurídico do STJ;
b) elaborar uma ontologia para a recuperação jurisprudencial pela combinação da ontologia preliminar com a ontologia JurisTJPR, proposta por Molinari (2011);
c) construir e avaliar um Motor de Busca Semântica experimental que utilize a ontologia final elaborada.

As seções seguintes descrevem os processos realizados para alcançar tais objetivos.

5.2.1 Elaboração da ontologia preliminar

O processo de elaboração da ontologia peso leve preliminar correspondeu à conversão dos descritores presentes no tesauro jurídico fornecido pelo STJ para construções equivalentes em uma ontologia peso leve.

Um descritor pode estar relacionado com outros descritores, alguns mais genéricos e outros mais específicos, constituindo assim uma taxonomia de descritores. Podem também apresentar sinônimos, tanto em terminologia técnica quanto em formas de sentido equivalente, mas com uso não recomendado em documentos jurisprudenciais. Finalmente, descritores podem ter uma lista de outros descritores relacionados.

O objetivo específico de produzir uma ontologia peso leve a partir de um tesauro foi atingido pela aplicação do **Cenário 5** da metodologia **NeOn**. Este cenário instrui a aplicação de um *design pattern* para a engenharia de ontologias[17].

O *design pattern* em questão apresenta instruções para o mapeamento das construções existentes em um tesauro para suas correspondentes em uma ontologia. O resultado é uma ontologia peso leve, adequada ao uso de um Motor de Busca Semântica,

Com a facilitação trazida pela construção de um aplicativo específico, foi possível executar o Cenário 5 da metodologia NeOn, o que possibilitou a conversão

[17] http://ontologydesignpatterns.org, acesso em 17 jan. 2017

do Tesauro do STJ em uma ontologia. A ontologia peso leve final apresenta **10.210** classes. Estas mantiveram a organização taxonômica previamente estabelecida pelo tesauro, que já estava validada pela **Secretaria de Jurisprudência** do STJ.

Cada uma das classes da ontologia apresenta ao menos um rótulo, mas em diversos casos uma mesma classe é caracterizada por mais de um rótulo. As classes têm vários rótulos porque existem formas alternativas para a caracterização léxica de uma classe, como a existência de sinônimos, formas plurais, siglas, abreviaturas e formas leigas.

A ontologia peso leve preliminar foi armazenada em um documento OWL chamado **Vocabulario-2016-06.owl** e constituiu o resultado do atingimento do primeiro objetivo específico enunciado.

5.2.2 Combinação das ontologias Vocabulario-2016-06 e JurisTJPR

A ontologia peso leve preliminar armazenada no documento OWL chamado **Vocabulario-2016-06.owl** é compatível com as recomendações do W3C para Web Semântica, o que o habilita a ser utilizada em quaisquer aplicações compatíveis com este padrão, notadamente os editores de ontologia como por exemplo o Protégé 5.

A ontologia proposta por Molinari (2011) já se encontrava no formato OWL, mas precisou ser complementada para dispor de rótulos para suas classes. Além disso, algumas classes precisaram ter seus nomes adaptados para se enquadrarem à convenção utilizada pela ontologia Vacabulario-2016-06, que apresenta preposições em seus nomes.

Com as duas ontologias em compatibilidade de convenções, o software Protégé foi utilizado para proceder com a combinação de ambas, o que produziu uma nova ontologia peso leve denominada **OntoLegis**, que foi armazenada no documento OWL chamado **OntoLegis-2016-06.owl**.

Um *corpus* de **59.386** decisões foi recuperado a partir do próprio sítio de Internet do **Tribunal de Justiça do Paraná** (período de **05/02/2016** a **01/11/2016**) para o cálculo de estatísticas descritivas que avaliam o grau de cobertura da ontologia. As análises mostraram que todos os documentos deste *corpus* apresentam ao menos uma classe da ontologia, estabelecendo **100%** de cobertura.

Diante das evidências objetivas alcanças, entende-se que esta nova ontologia está em plenas condições de alimentar um sistema de indexação e recuperação

semânticas e constitui o artefato que realiza o atingimento do segundo objetivo específico da pesquisa.

5.2.3 Construção e avaliação do Motor de Busca Semântica

Com base na ontologia OntoLegis construída, um Motor de Busca Semântica foi construído e submetido a três formas distintas de avaliação:

a) a apresentação a dois advogados consultores permitiu perceber o grau de aceitação à proposta e também coletar recomendações e exemplos de buscas, que serviram de insumo para aprimoramentos, tanto nesta pesquisa como em trabalhos futuros;
b) a comparação de cinco diferentes buscas com o recurso de recuperação jurisprudencial oferecido pelo sítio do TJPR, utilizando um *corpus* comum, permitiu a identificação de virtudes e fragilidades no Motor de Busca Semântica;
c) a experimentação pública com preenchimento de um questionário de quatro questões levou à elaboração de um sistema de tradução de parâmetros que aproveita a estrutura taxonômica dos conceitos presentes na ontologia, o que resultou diretamente em melhoria percebida por outro respondente; ademais, as respostas serviram para medir o grau de satisfação geral do Motor de Busca Semântica baseado na ontologia OntoLegis, o que possibilitou a percepção de suas fraquezas atuais e de seu potencial como possível solução para o Problema da Recuperação da Informação.

A análise dos experimentos mostrou que as relações semânticas estabelecidas pela ontologia efetivamente permitem a seleção mais criteriosa de documentos (aumento da precisão). Por outro lado, a indexação semântica mostrou fragilidade quando existe ambiguidade léxica, fazendo com que mais de uma classe seja selecionada pela ocorrência do mesmo rótulo (redução da precisão).

Também foi possível identificar casos onde a estrutura taxonômica da ontologia levou o motor de busca a tomar decisões consideradas menos interessantes pelos experimentadores, produzindo resultados que não correspondiam aos seus anseios.

A conjunção das três formas de avaliação a que foi submetido o motor de busca permitiu ter a noção de que existe espaço para melhorias tanto em seus processos inferenciais quando na própria ontologia que os guia. Ao mesmo tempo, estas mesmas avaliações ofereceram evidências para reconhecer o potencial de ganho de precisão e eficiência que a tecnologia de recuperação semântica oferece.

Diante destas conclusões, considera-se que o terceiro objetivo específico foi atingido.

5.3 Trabalhos e pesquisas futuros

Os experimentos realizados, assim como a submissão ao uso por advogados consultores e experimentação pública permitiram coletar um conjunto de situações não incluídas no âmbito desta pesquisa. Também foram identificadas situações de fragilidade nos processos realizados que contribuíram para tornar os efeitos menos eficazes do que o desejado.

Todas estas situações foram encaradas como insumos de valor que abre espaço e identificam oportunidades de melhorias a serem conquistadas em outras pesquisas enunciadas nas seções a seguir.

5.3.1 Solução de ambiguidades

A situação que ofereceu maior impacto no desempenho do Motor de Busca Semântica foi a incorreta vinculação de classe durante o processo de indexação semântica. Esta incorreção se deu por ocasião da identificação errônea de classes a partir da ocorrência de seus rótulos.

Um caso emblemático envolve a classe “**EstadoDoSergipe**”, que foi identificada na quase totalidade dos documentos indexado por causa da ocorrência do pronome “**se**”, confundido com a sigla de unidade federativa “**SE**”, presente na lista dos rótulos daquela classe.

A solução para este impasse passa pela identificação da função sintática de cada palavra antes de sua comparação com os rótulos das classes, no processo de indexação semântica. Neste sentido a presença da palavra “se” sob a função de pronome não deveria motivar a identificação da classe “**EstadoDoSergipe**”.

Caso similar se deu com a classe “**Desembargador**”, que foi identificada por causa de seu rótulo “des”, que é a abreviatura da palavra “desembargador”. O mesmo rótulo também identifica a classe “**DireitoEspecialDeSaque**”, pois está incluído como

sua sigla. Neste caso, as duas classes são vinculadas ao documento, sendo que uma delas necessariamente não o deveria.

Técnicas de rotulação sintática (do inglês, *POS tagging*) podem ser indicadas para suporte a este tipo de desambiguação. Tais técnicas são utilizadas tanto em pesquisas envolvendo Processamento de Linguagem Natural quando Mineração de Texto (WEISS et al., 2010).

5.3.2 Revisão das relações entre classes

Ao menos um caso de relação faltante entre classes foi identificado, no caso da busca por “prisão domiciliar”. Neste caso, a tradução de parâmetros incluiu a classe “Prisão”, tomada como incorreta pelo consultante.

Assumindo que a relação efetivamente deveria existir, a mesma poderia ser incluída sem maiores comprometimentos para a ontologia, pois o processo de indexação semântica já está preparado para a incidência de mais de uma superclasse. Neste caso, bastaria que a classe “**Prisão**” passasse a ser uma superclasse de “**PrisãoDomiciliar**”.

Apesar de facilidade técnica em solucionar o caso, a questão envolve a compreensão semântica dos conceitos envolvidos. É necessário que um ou mais profissionais do Direito avaliem o caso no sentido de efetivamente corroborarem que a relação existe e pode ser colocada na ontologia.

Este é um exemplo de relação hierárquica suspeita que foi identificada, mas é necessário avaliar se existem outros casos, pois os experimentos foram suficientes para evidenciar que a eficácia da busca semântica é fortemente influenciada pela estrutura taxonômica dos conceitos.

Neste sentido, esta fragilidade poderia ser superada pela avaliação de um ou mais profissionais do Direito comprometidos com a proposta de aprimorar a ontologia e assim aprimorar também a recuperação semântica por esta viabilizada.

5.3.3 Aprimoramento dos relacionamentos entre classes

O Tesauro Jurídico do STJ fornece anotações de relacionamento genérico entre seus descritores. Estes relacionamentos foram transferidos para a ontologia, mas por sua natureza não específica, não foram aproveitados nem no processo de indexação semântica e nem no processo de recuperação.

Neste sentido, tais relacionamentos poderiam ser aprimorados para conter semântica mais específica que auxiliasse a identificação, indexação e recuperação semânticas, o que permitiria tirar ainda mais proveito do conhecimento armazenado na ontologia.

5.3.4 Vinculação com código de leis

Um dos advogados consultores sugeriu que o uso de leis fosse considerado tanto no processo de indexação quando de recuperação jurisprudencial. Na oportunidade, o profissional em questão relatou que em muitos casos, a identificação do assunto sendo tratado por uma decisão é diretamente identificado quando existe a citação de uma lei e seu respectivo parágrafo.

Neste sentido, a ontologia poderia ser expandida para contemplar leis e parágrafos de leis e vinculá-los aos conceitos por meio de seus rótulos ou de classes especialmente relacionadas. É importante relembrar que o Tesauro do STJ traz em seu prefixo “CAT” a lista de compêndios legais vinculados ao descritor. Estes poderiam servir como ponto de partida para a implementação desta melhoria.

5.3.5 Expansão da amplitude do *corpus*

Um dos respondentes sugeriu que mais tribunais fossem incluídos no *corpus* indexado pelo ConsultaJuris. Neste sentido, o componente de importação precisaria ser expandido para identificar outros formatos de sítios dos outros tribunais.

Outra forma de procurar aumentar o conteúdo do *corpus* seria de importar documentos armazenados sob o padrão LexML (http://www.lexml.gov.br), que facilitaria sobremaneira sua recuperação e posterior indexação.

5.3.6 Aprimoramento da tradução de parâmetros

Ao menos um experimento mostrou que a tradução de parâmetros ainda permite a seleção de classes que não guardam relação com os interesses principais da busca. O cálculo de relevância das seletoras existe precisamente para procurar mitigar este tipo de desvio, mas houve casos em que não foi suficiente para desfavorecer classes sem relação com os interesses do consultante.

Um caminho possível para a minoração deste problema será a avaliação de não relação entre as seletoras e a desconsideração de classes menos relevantes que

não estão relacionadas com as mais relevantes. Esta estratégia tiraria proveito da estrutura taxonômica que em outras ocasiões provou-se valiosa.

5.3.7 Aprimoramento da sintaxe de busca

Um dos respondentes sugeriu o uso de operadores booleanos, enquanto que outro ofereceu a ideia de dar ao usuário a oportunidade de decidir quais classes devem ser utilizadas como seletoras. Outra sugestão foi a de indicar quais classes não podem aparecer no documento, ou seja, uma seletora inversa.

Estas são sugestões que não implicam na estrutura da ontologia, mas em aperfeiçoamento da experiência de uso do ConsultaJuris e são possíveis graças à indexação semântica. Neste sentido, entende-se que as mesmas podem de fato tornar a ferramenta ConsultaJuris mais atraente e elevar seu grau de satisfação. Em alguns casos, são também um subproduto das possibilidades trazidas pela indexação semântica.

REFERÊNCIAS

ADELMANN, A. et al. **Semantic Search over the Web**. Graz University of Technology. Austria, 2013

AFONSO, A. R. **B2:** um sistema para indexação e agrupamento de artigos científicos em português brasileiro utilizando computação evolucionária. 2013. 158 (Doutorado em Ciência da Informação) - Programa de Pós-graduação em Ciência da Informação, Universidade de Brasília, Brasília.

BEPPLER, M. D.; FERNANDES, A. M. D. R. Aplicação de Text mining para a Extração de Conhecimento Jurisprudencial. In: **Anais SULCOMP,** v. 1, n. 1, 2012.

BOCHEREAU, L.; BOURCIER, D.; BOURGINE, P. Extracting legal knowledge by means of a multilayer neural network application to municipal jurisprudence. In: **Proceedings of the 3rd international conference on Artificial intelligence and law**, 1991, Charleston, USA. ACM. p.pp. 288-296.

BRASIL. **Justiça em Números 2016**. Brasília: Conselho Nacional de Justiça, 2016. Relatório

BUENO, T. C. D. A. et al. JurisConsulto: retrieval in jurisprudencial text bases using juridical terminology. In: **Proceedings of the 7th international conference on Artificial intelligence and law**, 1999, ACM. p. 147-155.

CHEN, Y.-L.; LIU, Y.-H.; HO, W.-L. A text mining approach to assist the general public in the retrieval of legal documents. In: **Journal of the American Society for Information Science & Technology,** v. 64, n. 2, p. 280-290, 2013.

CLARKE, S. G. D.; ZENG, M. L. From ISO 2788 to ISO 25964: The evolution of thesaurus standards towards interoperability and data modelling. **Information Standards Quarterly** (ISQ), v. 24, n. 1, 2012.

DINIZ, M. H. **Compêndio de introdução à ciência do direito**: introdução à teoria geral do direito, à filosofia do direito, à sociologia jurídica e à lógica jurídica: norma jurídica e aplicação do direito. 25. São Paulo: Saraiva, 2012.

DRAGONI, M.; DA COSTA PEREIRA, C.; TETTAMANZI, A. G. A conceptual representation of documents and queries for information retrieval systems by using light ontologies. **Expert Systems with applications,** v. 39, n. 12, p. 10376-10388, 2012.

FERAUCHE, T.; DE ALMEIDA, M. A. Aprendizado de classificadores das ementas da Jurisprudência do Tribunal Regional do Trabalho da 2ª. Região-SP. In:VI WorkShop de Pesquisa do Centro Estadual de Eucação Tecnológica Paula Souza, **Anais...** São Paulo, 2011.

GARCIA, G. F. B. **Introdução ao Estudo do Direito**. 3 ed. São Paulo: Editora Método, 2015.

GIL, A. C. **Como elaborar projetos de pesquisa**. 4. ed. São Paulo: Atlas, 2002.

GRIFFO, C.; ALMEIDA, J. P. A.; GUIZZARDI, G. Towards a Legal Core Ontology based on Alexy's Theory of Fundamental Rights. In:MWAIL, ICAIL 2015, **Anais...** 2015.

GRUBER, T. R. Toward principles for the design of ontologies used for knowledge sharing. In: **International journal of human-computer studies,** v. 43, n. 5, p. 907-928, 1995.

GUERRA, G. N. **Modelo de reputação e ontologia aplicados à rede social científica do observeunb**. Dissertação (Mestrado em Informática) — Departamento de Ciênca da Computação, Universidade de Brasília, 2012.

GÓMEZ-PÉREZ, A.; DE FIGUEROA BAONZA, M. C. S.; VILLAZÓN, B. **Neon methodology for building ontology networks**: Ontology specification. Tese (Doutorado) - Universidad Politécnica de Madrid, 2008.

GÓMEZ-PÉREZ, A.; SUÁREZ-FIGUEROA, M. C. NeOn Methodology for Building Ontology Networks: a Scenario-based Methodology. In:Proceedings of the International Conference on Software, Services & Semantic Technology. Sofia, **Anais...** 2009.

HEFLIN, J. **An Introduction to the OWL Web Ontology Language**. Lehigh University. National Science Foundation (NSF), 2007. E-book: Disponível em <http://www.cse.lehigh.edu/~heflin/IntroToOWL.pdf.>. Acesso em: 17 jan. 2017.

KASSIM, J. M.; RAHMANY, M. Introduction to semantic search engine. In:2009 International Conference on Electrical Engineering and Informatics, **Anais...** 2009, p.380-386.

KAVITHA, V.; HANUMANTHAPPA, M.; PRAKASH, B. Ontology Based Search Engine. **International Journal of Emerging Trends & Technology in Computer Science,** v. 4, n. 5, 2015.

KISHORE, V. et al. Maximizing the Search Engine Efficiency. **International Journal of Advanced Research in Computer Science and Electronics Engineering (IJARCSEE),** v. 1, n. 6, p. pp: 58-64, 2012.

LANCASTER, F. W. **Indexação e resumos:** teoria e prática. 2. ed. Brasília: Briquet de Lemos, 2004.

MARTÍNEZ-GONZÁLEZ, M. M.; ALVITE-DÍEZ, M.-L. **On the evaluation of thesaurus tools compatible with the Semantic Web**. Journal of Information Science, 2014.

MAXIMILIANO, C. **Hermenêutica e Aplicação do Direito**. 20 ed. Rio de Janeiro: Grupo Editorial Nacional, 2011.

MOENS, M.-F. Innovative techniques for legal text retrieval. **Artificial Intelligence and Law,** v. 9, n. 1, p. 29-57, 2001.

MOENS, M.-F.; DE BUSSER, R. I. K. First steps in building a model for the retrieval of court decisions. **International Journal of Human-Computer Studies,** v. 57, n. 5, p. 429, 2002

MOLINARI, A. H. **Indexação de acórdãos por meio de uma ontologia jurisprudencial populada a partir de um corpus jurídico real**. Dissertação (Mestrado) - CPGEI, UTFPR, 2011.

NEON. **NeOn Methodology in a nutshell**. Neon Project 2010. E-book: Disponível em < http://www.neon-project.org/nw/NeOn_Book>. Acesso em: 17 jan. 2017

PARREIRAS, F. S. **Semantic Web and Model-Driven Engineering**. Piscataway, New Jersey: John Wiley & Sons, 2012.

PEREIRA, E. C.; BUFREM, L. S. Princípios de organização e representação de conceitos em linguagens documentárias. **Encontros Bibli**, Florianópolis: v. 10, v. 20, 2005.

PRESSUTI, V.; GANGEMI, A. Content ontology design patterns as practical building blocks for web ontologies. In:International Conference on Conceptual Modeling, **Anais...** , Barcelona, Spain: 2008. p.128-141.

RAMOS JÚNIOR, H. S. **Uma ontologia para representação do conhecimento jurídico-penal no contexto dos delitos informáticos**. Dissertação (Mestrado em

Gestão do Conhecimento) - Centro Tecnológico, Universidade Federal de Santa Catarina, 2008.

RESNIK, P. Using information content to evaluate semantic similarity in a taxonomy.In: **In Proceedings of IJCAI-95**. Montreal, Canada, 1995.

ROBERTSON, S. Understanding inverse document frequency: on theoretical arguments for IDF. **Journal of documentation,** v. 60, n. 5, p. 503-520, 2004

SCHWEIGHOFER, E.; WINIWARTER, W. Intelligent information retrieval: Konterm-automatic representation of context related terms within a knowledge base for a legal expert system. In:25th Anniversary Conference of the Istituto per la documentazione giuridica of the CNR: Towards a Global Expert System in Law. **Anais…** Padua, Italy 1994.

SERBENA, C. A. Interfaces atuais entre a E-Justiça e a Q-Justiça no Brasil. **Revista de Sociologia e Política**, 2003, v. 21, n. 45, p. 47-56.

SILVA, A. D. **Análise das relações semânticas em tesauros jurídicos brasileiros**: orientações das normas e aplicação prática. Trabalho de Graduação - Curso de Biblioteconomia, Centro de Ciências da Educação, Universidade Federal de Santa Catarina, Florianópolis, 2013.

SIMPERL, E. P. B.; MOCHOL, M.; BÜRGER, T. Achieving Maturity: the State of Practice in Ontology Engineering in 2009. **IJCSA,** v. 7, n. 1, p. 45-65, 2010.

SLYPE, G. V.; HÍPOLA, P.; MOYA ANEGÓN, F. **Los lenguajes de indización:** concepción, construcción y utilización en los sistemas documentales. Madrid: Pirámide, 1991.

SPARCK JONES, K. A statistical interpretation of term specificity and its application in retrieval. **Journal of documentation,** v. 28, n. 1, p. 11-21, 1972.

THEODORO JÚNIOR, H; **Curso de Direito Processual Civil**, 25 ed, Editora Forense Ltda, Rio de Janeiro, 2015

TRIBUNAL DE JUSTIÇA DO PARANÁ, Disponível em <https://portal.tjpr.jus.br/jurisprudencia/>. Acesso em: 17 jan. 2017.

WEBER, R. Intelligent jurisprudence research: a new concept. In:Proceedings of the 7th international conference on Artificial intelligence and law, **Anais…** 1999. p.164-172.

WEISS, S. M.; INDURKHYA, N.; ZHANG, T. **Fundamentals of predictive text mining**. First Edition. Sidney, Australia: Springer Science & Business Media, 2010.

APÊNDICE A – Descritores com complemento que receberam tratamentos

Este Apêndice apresenta uma lista de todos os descritores com complementos entre parênteses que receberam algum tipo de tratamento em nível de tesauro. Em todos os casos, o complemento entre parênteses foi removido, sendo que estratégias específicas de tratamento foram adotadas.

As estratégias de tratamento foras as seguintes:

a) **Remoção direta**: o complemento é simplesmente removido;
b) **Conversão de sigla**: o complemento foi removido e adicionado na forma de um sinônimo; em caso de duplicação de termos consequente da remoção do complemento, os descritores duplicados são combinados, passando a ser um só.

#	Descritor	Operação
01	60 (SESSENTA SALÁRIOS MÍNIMOS)	Remoção direta
02	ACIDENTE VASCULAR CEREBRAL (AVC)	Conversão de sigla
03	ACORDO GERAL SOBRE TARIFAS ADUANEIRAS (GATT)	Conversão de sigla
04	ACORDO GERAL SOBRE TARIFAS ADUANEIRAS E COMÉRCIO (GATT)	Conversão de sigla
05	ACRE (AC)	Conversão de sigla
06	ADICIONAL DE QUALIFICAÇÃO (AQ)	Conversão de sigla
07	ADVERTÊNCIA (ESTATUTO DA CRIANÇA E DO ADOLESCENTE)	
08	AGÊNCIA DE FISCALIZAÇÃO DO DF (AGEFIS-DF)	Conversão de sigla
09	AGÊNCIA NACIONAL DE ENERGIA ELÉTRICA (ANELL)	Remoção direta
10	AGÊNCIA NACIONAL DE SAÚDE (ANS)	Conversão de sigla
11	AGÊNCIA NACIONAL DE SAÚDE SUPLEMENTAR (ANS)	Conversão de sigla
12	AGÊNCIA NACIONAL DE TRANSPORTES TERRESTRES (ANTT)	Conversão de sigla
13	AGÊNCIA NACIONAL DE VIGILÂNCIA SANITÁRIA (ANVISA)	Conversão de sigla
14	ARMAS (SÍMBOLO NACIONAL)	Conversão de sigla
15	ASSOCIAÇÃO (LEI DE TÓXICOS)	Conversão de sigla
16	ATO DAS DISPOSIÇÕES CONSTITUCIONAIS TRANSITÓRIAS (ADCT)	Conversão de sigla
17	ATO OBSCENO (CRIME)	Remoção direta
18	AUTO DE PRISÃO EM FLAGRANTE (APF)	Conversão de sigla
19	AUTO DE PRISÃO EM FLAGRANTE DELITO (APFD)	Conversão de sigla
20	BANCO CENTRAL (BACEN)	Conversão de sigla
21	BANCO CENTRAL DO BRASIL (BACEN)	Conversão de sigla
22	BÔNUS DO TESOURO NACIONAL (BTN)	Conversão de sigla

23	BÔNUS DO TESOURO NACIONAL FISCAL (BTNF)	Conversão de sigla
24	CADASTRO INFORMATIVO DE CRÉDITOS NÃO QUITADOS (CADIN)	Conversão de sigla
25	CADASTRO INFORMATIVO DE CRÉDITOS NÃO QUITADOS DO SETOR PÚBLICO FEDERAL (CADIN)	Conversão de sigla
26	CADASTRO INFORMATIVO DOS CRÉDITOS NÃO QUITADOS DE ÓRGÃOS E ENTIDADES FEDERAIS (CADIN)	Conversão de sigla
27	CADASTRO NACIONAL DE INFORMAÇÕES SOCIAIS (CNIS)	Conversão de sigla
28	CARTEIRA DE TRABALHO E PREVIDÊNCIA SOCIAL (CTPS)	Conversão de sigla
29	CASA DE PROSTITUIÇÃO (CRIME)	Remoção direta
30	CENTRALIZAÇÃO DE SERVIÇOS DOS BANCOS S/A (SERASA)	Conversão de sigla
31	CERTIDÃO DE DÍVIDA ATIVA (CDA)	Conversão de sigla
32	CERTIDÃO POSITIVA DE DÉBITO COM EFEITOS DE NEGATIVA (CPD-EN)	Conversão de sigla
33	CÓDIGO DE TRÂNSITO BRASILEIRO (CTB)	Conversão de sigla
34	CÓDIGO NACIONAL DE TRÂNSITO (CNT)	Conversão de sigla
35	CONSELHO DE ARQUITETURA E URBANISMO DO BRASIL (CAU/BR)	Conversão de sigla
36	CONSELHO DE ARQUITETURA E URBANISMO DOS ESTADOS E DO DISTRITO FEDERAL (CAUS)	Conversão de sigla
37	CONSELHO DE ARQUITETURA E URBANISMO DOS ESTADOS E DO DISTRITO FEDERAL (CAU/UF)	Conversão de sigla
38	CONSELHO FEDERAL DE ENGENHARIA E AGRONOMIA (CONFEA)	Conversão de sigla
39	CONSELHO NACIONAL DE SEGUROS PRIVADOS (CNSP)	Conversão de sigla
40	CONSELHO NACIONAL DE TRÂNSITO (CONTRAN)	Conversão de sigla
41	CONSELHO NACIONAL DO SEGURO PRIVADO (CNSP)	Conversão de sigla
42	CONSELHO REGIONAL DE ECONOMIA (CORECON)	Conversão de sigla
43	CONSELHO REGIONAL DE ENGENHARIA E AGRONOMIA (CREA)	Conversão de sigla
44	CONTRIBUIÇÃO DE INTERVENÇÃO NO DOMÍNIO ECONÔMICO (CIDE)	Conversão de sigla
45	CONTRIBUIÇÃO PARA FINANCIAMENTO DA SEGURIDADE SOCIAL (COFINS)	Conversão de sigla
46	CONTRIBUIÇÃO PARA O FINANCIAMENTO DA SEGURIDADE SOCIAL (COFINS)	Conversão de sigla
47	CONTRIBUIÇÃO PROVISÓRIA SOBRE MOVIMENTAÇÃO FINANCEIRA (CPMF)	Conversão de sigla
48	CONTRIBUIÇÃO PROVISÓRIA SOBRE MOVIMENTAÇÃO OU TRANSMISSÃO DE VALORES E DE CRÉDITOS E DIREITOS DE NATUREZA FINANCEIRA (CPMF)	Conversão de sigla
49	CONTRIBUIÇÃO SOCIAL SOBRE LUCRO LÍQUIDO (CSLL)	Conversão de sigla
50	CONTRIBUIÇÃO SOCIAL SOBRE O LUCRO LÍQUIDO (CSLL)	Conversão de sigla
51	DECLARAÇÃO DE CONTRIBUIÇÕES E TRIBUTOS FEDERAIS (DCTF)	Conversão de sigla
52	DEMONSTRATIVO DE CRÉDITO PRESUMIDO (DCP)	Conversão de sigla
53	DEMONSTRATIVO DO CRÉDITO PRESUMIDO (DCP)	Conversão de sigla
54	DEPARTAMENTO DE ESTRADAS DE RODAGEM (DER)	Conversão de sigla
55	DEPARTAMENTO DE TRÂNSITO (DETRAN)	Conversão de sigla
56	DEPARTAMENTO ESTADUAL DE TRÂNSITO (DETRAN)	Conversão de sigla
57	DEPARTAMENTO NACIONAL DE ESTRADAS DE RODAGEM (DNER)	Conversão de sigla

58	DEPARTAMENTO NACIONAL DE INFRAESTRUTURA DE TRANSPORTES (DNIT)	Conversão de sigla
59	DEPARTAMENTO NACIONAL DE TRÂNSITO (DENATRAN)	Conversão de sigla
60	DISTRIBUIÇÃO DE AVISO (CONTRAVENÇÃO)	Remoção direta
61	DISTRIBUIÇÃO DE LISTA (CONTRAVENÇÃO)	Remoção direta
62	DOCUMENTO DE ARRECADAÇÃO DE RECEITAS FEDERAIS (DARF)	Conversão de sigla
63	DOCUMENTO DE ARRECADAÇÃO DE RECEITAS JUDICIÁRIAS (DARJ)	Conversão de sigla
64	EMPRESA BRASILEIRA DE CORREIOS E TELÉGRAFOS (ECT)	Conversão de sigla
65	ESCRITO OBSCENO (CRIME)	Remoção direta
66	ESCRITÓRIO CENTRAL DE ARRECADAÇÃO E DISTRIBUIÇÃO (ECAD)	Conversão de sigla
67	ESCRITURAÇÃO DE INDÚSTRIA (CONTRAVENÇÃO)	Remoção direta
68	ESCRITURAÇÃO DE PROFISSÃO (CONTRAVENÇÃO)	Remoção direta
69	EXIBIÇÃO DE LISTA DE SORTEIO (CONTRAVENÇÃO)	Remoção direta
70	FATOR ACIDENTÁRIO DE PREVENÇÃO (FAP)	Conversão de sigla
71	FUNDAÇÃO DE SEGURIDADE SOCIAL (GEAP)	Conversão de sigla
72	FUNDAÇÃO GETÚLIO VARGAS (FGV)	Conversão de sigla
73	FUNDO DE ASSISTÊNCIA AO TRABALHADOR RURAL (FUNRURAL)	Conversão de sigla
74	FUNDO DE COMPENSAÇÃO DE VARIAÇÕES SALARIAIS (FCVS)	Conversão de sigla
75	FUNDO DE EQUALIZAÇÃO DA SINISTRALIDADE DA APÓLICE DO SEGURO HABITACIONAL DO SISTEMA FINANCEIRO DA HABITAÇÃO (FESA)	Conversão de sigla
76	FUNDO DE EQUALIZAÇÃO DE SINISTRALIDADE DA APÓLICE (FESA)	Conversão de sigla
77	FUNDO DE FINANCIAMENTO AO ESTUDANTE DO ENSINO SUPERIOR (FIES)	Conversão de sigla
78	FUNDO DE FINANCIAMENTO ESTUDANTIL (FIES)	Conversão de sigla
79	FUNDO DE GARANTIA DO TEMPO DE SERVIÇO (FGTS)	Conversão de sigla
80	FUNDO DE GARANTIA POR TEMPO DE SERVIÇO (FGTS)	Conversão de sigla
81	FUNDO NACIONAL DE SEGURANÇA E EDUCAÇÃO DE TRÂNSITO (FUNSET)	Conversão de sigla
82	GRATIFICAÇÃO POR ENCARGO DE CURSO OU CONCURSO (GECC)	Conversão de sigla
83	GUARDA DE LISTA DE SORTEIO (CONTRAVENÇÃO)	Remoção direta
84	GUIA DA PREVIDÊNCIA SOCIAL (GPS)	Conversão de sigla
85	GUIA DE RECOLHIMENTO DA PREVIDÊNCIA SOCIAL (GRPS)	Conversão de sigla
86	GUIA DE RECOLHIMENTO DA UNIÃO (GRU)	Conversão de sigla
87	GUIA DE RECOLHIMENTO DO FGTS E INFORMAÇÕES À PREVIDÊNCIA SOCIAL (GFIP)	Conversão de sigla
88	GUIA DE RECOLHIMENTO DO FUNDO DE GARANTIA DO TEMPO DE SERVIÇO E INFORMAÇÕES À PREVIDÊNCIA SOCIAL (GFIP)	Conversão de sigla
89	IMPOSTO DE RENDA PESSOA FÍSICA (IRPF)	Conversão de sigla
90	IMPOSTO DE RENDA PESSOA JURÍDICA (IRPJ)	Conversão de sigla
91	IMPOSTO DE TRANSMISSÃO CAUSA MORTIS E DOAÇÃO DE QUAISQUER BENS OU DIREITOS (ITCD)	Conversão de sigla
92	IMPOSTO SOBRE A PROPRIEDADE DE VEÍCULOS AUTOMOTORES (IPVA)	Conversão de sigla

93	IMPOSTO SOBRE A PROPRIEDADE PREDIAL E TERRITORIAL URBANA (IPTU)	Conversão de sigla
94	IMPOSTO SOBRE A PROPRIEDADE TERRITORIAL RURAL (ITR)	Conversão de sigla
95	IMPOSTO SOBRE A TRANSMISSÃO DE BENS IMÓVEIS (ITBI)	Conversão de sigla
96	IMPOSTO SOBRE A TRANSMISSÃO DE BENS IMÓVEIS E DE DIREITOS A ELES RELATIVOS (ITBI)	Conversão de sigla
97	IMPOSTO SOBRE OPERAÇÕES DE CRÉDITO, CÂMBIO E SEGURO, OU RELATIVAS A TÍTULOS OU VALORES MOBILIÁRIOS (IOF)	Conversão de sigla
98	IMPOSTO SOBRE OPERAÇÕES RELATIVAS À CIRCULAÇÃO DE MERCADORIAS E SOBRE PRESTAÇÕES DE SERVIÇOS DE TRANSPORTE INTERESTADUAL E INTERMUNICIPAL E DE COMUNICAÇÃO (ICMS)	Conversão de sigla
99	IMPOSTO SOBRE OPERAÇÕES RELATIVAS À CIRCULAÇÃO DE MERCADORIAS E SOBRE PRESTAÇÕES DE SERVIÇOS DE TRANSPORTE INTERESTADUAL E INTERMUNICIPAL E DE USECOMUNICAÇÃO (ICMS)	Conversão de sigla
100	IMPOSTO SOBRE PRODUTOS INDUSTRIALIZADOS (IPI)	Conversão de sigla
101	IMPOSTO SOBRE SERVIÇOS DE QUALQUER NATUREZA (ISS)	Conversão de sigla
102	IMPOSTO TERRITORIAL RURAL (ITR)	Conversão de sigla
103	IMPRESSÃO DE ANÚNCIO (CONTRAVENÇÃO)	Remoção direta
104	IMPRESSÃO DE BILHETE (CONTRAVENÇÃO)	Remoção direta
105	IMPRESSÃO DE LISTA (CONTRAVENÇÃO)	Remoção direta
106	ÍNDICE DE PREÇOS AO CONSUMIDOR (IPC)	Conversão de sigla
107	ÍNDICE DE PREÇOS AO CONSUMIDOR AMPLO (IPCA)	Conversão de sigla
108	ÍNDICE DE PREÇOS AO CONSUMIDOR AMPLO- ESPECIAL (IPCA-E)	Conversão de sigla
109	ÍNDICE DE PREÇOS AO CONSUMIDOR SÉRIE R (IPC-R)	Conversão de sigla
110	ÍNDICE GERAL DE PREÇOS DO MERCADO (IGPM)	Conversão de sigla
111	ÍNDICE NACIONAL DE CUSTO DA CONSTRUÇÃO (INCC)	Conversão de sigla
112	ÍNDICE NACIONAL DE PREÇOS AO CONSUMIDOR (INPC)	Conversão de sigla
113	ÍNDICE NACIONAL DE PREÇOS AO CONSUMIDOR AMPLO (IPCA)	Conversão de sigla
114	ÍNDICE NACIONAL DE PREÇOS AO CONSUMIDOR AMPLO-ESPECIAL (IPCA-E)	Conversão de sigla
115	INSTITUTO BRASILEIRO DO MEIO AMBIENTE (IBAMA)	Conversão de sigla
116	INSTITUTO BRASILEIRO DO MEIO AMBIENTE E DOS RECURSOS NATURAIS RENOVÁVEIS (IBAMA)	Conversão de sigla
117	INSTITUTO DO AÇÚCAR E DO ÁLCOOL (IAA)	Conversão de sigla
118	INSTITUTO NACIONAL DE COLONIZAÇÃO E REFORMA AGRÁRIA (INCRA)	Conversão de sigla
119	INSTITUTO NACIONAL DE PROPRIEDADE INDUSTRIAL (INPI)	Conversão de sigla
120	INSTITUTO NACIONAL DO SEGURO SOCIAL (INSS)	Conversão de sigla
121	JOGO DE AZAR (CONTRAVENÇÃO)	Remoção direta
122	LEI DE INTRODUÇÃO ÀS NORMAS DO DIREITO BRASILEIRO (LINDB)	Conversão de sigla
123	LOTERIA ESTADUAL (CONTRAVENÇÃO)	Remoção direta
124	LOTERIA ESTRANGEIRA (CONTRAVENÇÃO)	Remoção direta
125	MATRÍCULA DE INDÚSTRIA (CONTRAVENÇÃO)	Remoção direta

126	MATRÍCULA DE PROFISSÃO (CONTRAVENÇÃO)	Remoção direta
127	MONOETILENOGLICOL (MEG)	Conversão de sigla
128	OBJETO OBSCENO (CRIME)	Remoção direta
129	OUTORGA ONEROSA DE ALTERAÇÃO DE USO (ONALT)	Conversão de sigla
130	PROGRAMA DE ALIMENTAÇÃO DO TRABALHADOR (PAT)	Conversão de sigla
131	PROGRAMA DE FORMAÇÃO DO PATRIMÔNIO DO SERVIDOR PÚBLICO (PASEP)	Conversão de sigla
132	PROGRAMA DE INTEGRAÇÃO SOCIAL (PIS)	Conversão de sigla
133	PROGRAMA DE INTEGRAÇÃO SOCIAL E PROGRAMA DE FORMAÇÃO DO PATRIMÔNIO DO SERVIDOR PÚBLICO (PIS/PASEP)	Conversão de sigla
134	PROGRAMA DE RECUPERAÇÃO FISCAL (REFIS)	Conversão de sigla
135	PUBLICIDADE DE SORTEIO (CONTRAVENÇÃO)	Remoção direta
136	REDE FERROVIÁRIA FEDERAL (RFFSA)	Conversão de sigla
137	REDE FERROVIÁRIA FEDERAL S.A. (RFFSA)	Conversão de sigla
138	REDE FERROVIÁRIA FEDERAL S/A (RFFSA)	Conversão de sigla
139	REDE FERROVIÁRIA FEDERAL SA (RFFSA)	Conversão de sigla
140	REDE FERROVIÁRIA FEDERAL SOCIEDADE ANÔNIMA (RFFSA)	Conversão de sigla
141	REGIME ESPECIAL DE REINTEGRAÇÃO DE VALORES TRIBUTÁRIOS PARA AS EMPRESAS EXPORTADORAS (REINTEGRA)	Conversão de sigla
142	REQUISIÇÃO DE PEQUENO VALOR (RPV)	Conversão de sigla
143	RISCO AMBIENTAL DE TRABALHO (RAT)	Conversão de sigla
144	SEGURO CONTRA ACIDENTES DO TRABALHO (SAT)	Conversão de sigla
145	SEGURO DE ACIDENTE DO TRABALHO (SAT)	Conversão de sigla
146	SEGURO DE ACIDENTES DO TRABALHO (SAT)	Conversão de sigla
147	SEGURO OBRIGATÓRIO DE DANOS PESSOAIS CAUSADOS POR VEÍCULOS (DPVAT)	Conversão de sigla
148	SEGURO OBRIGATÓRIO DE DANOS PESSOAIS CAUSADOS POR VEÍCULOS AUTOMOTORES DE VIA TERRESTRE (DPVAT)	Conversão de sigla
149	SERVIÇO DE PROTEÇÃO AO CRÉDITO (SPC)	Conversão de sigla
150	SINAIS DE PERIGO (CONTRAVENÇÃO)	Remoção direta
151	SISTEMA FINANCEIRO DA HABITAÇÃO (SFH)	Conversão de sigla
152	SISTEMA FINANCEIRO HABITACIONAL (SFH)	Conversão de sigla
153	SISTEMA INTEGRADO DE ADMINISTRAÇÃO FINANCEIRA DO GOVERNO FEDERAL (SIAFI)	Conversão de sigla
154	SISTEMA INTEGRADO DE PAGAMENTO DE IMPOSTOS E CONTRIBUIÇÕES DAS MICROEMPRESAS E EMPRESAS DE PEQUENO PORTE (SIMPLES)	Conversão de sigla
155	SISTEMA NACIONAL DE ATENDIMENTO SOCIOEDUCATIVO (SINASE)	Conversão de sigla
156	SISTEMA NACIONAL DE POLÍTICAS PÚBLICAS SOBRE DROGAS (SISNAD)	Conversão de sigla
157	SISTEMA ÚNICO DE SAÚDE (SUS)	Conversão de sigla
158	SUBSTÂNCIA AVARIADA (CRIME)	Remoção direta
159	SUBSTÂNCIA NOCIVA À SAÚDE (CRIME)	Remoção direta
160	SUBSTITUIÇÃO DESTINADA À FALSIFICAÇÃO (CRIME)	Remoção direta

161	SUPERINTENDÊNCIA NACIONAL DE PREVIDÊNCIA COMPLEMENTAR (PREVIC)	Conversão de sigla
162	TABELA DE INCIDÊNCIA DO IMPOSTO SOBRE PRODUTOS INDUSTRIALIZADOS (TIPI)	Conversão de sigla
163	TARIFA DE ABERTURA DE CRÉDITO (TAC)	Conversão de sigla
164	TARIFA DE EMISSÃO DE CARNÊ (TEC)	Conversão de sigla
165	TARIFA DE RENOVAÇÃO DE CADASTRO (TRC)	Conversão de sigla
166	TARIFA DE USO DO SISTEMA DE DISTRIBUIÇÃO DE ENERGIA ELÉTRICA (TUSD)	Conversão de sigla
167	TARIFA DE USO DO SISTEMA DE TRANSMISSÃO DE ENERGIA ELÉTRICA (TUST)	Conversão de sigla
168	TAXA DE CONTROLE E FISCALIZAÇÃO AMBIENTAL (TCFA)	Conversão de sigla
169	TAXA DE ILUMINAÇÃO PÚBLICA (TIP)	Conversão de sigla
170	TAXA DE SERVIÇOS ADMINISTRATIVOS (TSA)	Conversão de sigla
171	TAXA REFERENCIAL (TR)	Conversão de sigla
172	TERMO DE AJUSTAMENTO DE CONDUTA (TAC)	Conversão de sigla
173	TÍTULOS DA DÍVIDA AGRÁRIA (TDA)	Conversão de sigla
174	TR (ANELL)	Conversão de sigla
175	TR (CADIN)	Conversão de sigla
176	TR (TCFA)	Conversão de sigla
177	TRANSPORTE DE AVISO OU LISTA (CONTRAVENÇÃO)	Remoção direta
178	UNIDADE DE REFERÊNCIA DE PREÇOS (URP)	Conversão de sigla
179	UNIDADE REAL DE VALOR (URV)	Conversão de sigla
180	VANTAGEM PESSOAL NOMINALMENTE IDENTIFICADA (VPNI)	Conversão de sigla

APÊNDICE B – Lista de classes que necessitaram desambiguação por renomeação

Este Apêndice apresenta uma lista de todas as classes que foram derivadas do tesauro jurídico do Supremo Tribunal de Justiça e que necessitaram de desambiguação por meio de renomeação, visto que a simples retirada e combinação do complemento ao descritor do tesauro geraria a mudança indesejada do próprio descritor.

É uma operação de natureza semântica e não léxica, na qual classes adicionais foram criadas para abrigar em seu nome o próprio complemento ou uma derivação deste.

A tabela a seguir lista todas as ocorrências deste tipo de operação. Atenção à classe "**Tragonomia**" que manteve seu nome, mas foi associada à classe "**Crea**".

#	Classe original	Classe renomeada
01	AçãoDeResponsabilidade (direito comercial)	AçãoDeResponsabilidadeComercial
02	Afiançado (direito processual penal)	AfiançadoPenal
03	Assistência (direito civil)	AssistênciaCivil
04	Ausência (direito civil)	AusênciaCivil
05	Ausente (direito civil)	AusenteCivil
06	Autor (direito autoral)	AutorDeObra
07	Autoria (direito autoral)	AutoriaDeObra
08	BisInIdem (direito tributário)	BisInIdemTributário
09	Brasão (símbolo nacional)	BrasãoNacional
10	Cédula (júri)	CédulaDeJúri
11	Comissário (concordata)	ComissárioDeConcordata
12	Compensação (direito civil)	CompensaçãoCivil
13	Concedente (concessão comercial)	ConcedenteComercial
14	Concessionário (concessão comercial)	ConcessionárioComercial
15	Concurso (licitação)	ConcursoLicitatório
16	Confederação (direito civil e comercial)	ConfederaçãoComercial
17	Confusão (direito civil)	ConfusãoCivil
18	Conveniência (direito administrativo)	ConveniênciaAdministrativa
19	Depósito (contrato)	DepósitoContratual
20	Desabamento (crime)	DesabamentoCriminoso
21	DesastreFerroviário (crime)	DesastreFerroviárioCriminoso

22	Deserção (crime)	DeserçãoMilitar
23	Desmoronamento (crime)	DesmoronamentoCriminoso
24	Detenção (direito civil)	DetençãoCivil
25	Direito (ciência do direito)	CiênciaDoDireito
26	Disponibilidade (direito administrativo)	DisponibilidadeAdministrativa
27	DissoluçãoDeSociedadeDeFato (comercial)	DissoluçãoDeSociedadeComercialDeFato
28	EmissãoDeFumaça (contravenção)	EmissãoDeFumaçaCriminosa
29	EmissãoDeGás (contravenção)	EmissãoDeGásCriminosa
30	EmissãoDeVapor (contravenção)	EmissãoDeVaporCriminosa
31	Encampação (direito administrativo)	EncampaçãoAdministrativa
32	Encampação (direito comercial)	EncampaçãoComercial
33	Epidemia (crime)	EpidemiaCriminosa
34	Erro (vício de consentimento)	ErroDeConsentimento
35	EstabelecimentoSimilar (direito penal)	EstabelecimentoPenalSimilar
36	Estadia (termo de marinha)	EstadiaMarítima
37	Exoneração (direito administrativo)	ExoneraçãoAdministrativa
38	Explosão (crime)	ExplosãoCriminosa
39	Federação (direito civil)	FederaçãoCivil
40	Federação (direito constitucional)	FederaçãoConstitucional
41	Fiança (direito processual penal)	FiançaPenal
42	Filiação (direito civil)	FiliaçãoCivil
43	Foro (enfiteuse)	ForoEnfitêutico
44	Garantia (direito civil)	GarantiaCivil
45	GuiaDeRecolhimento (preso)	GuiaDeRecolhimentoPenal
46	Habilitação (direito civil)	HabilitaçãoCivil
47	Habitação (direito real)	HabitaçãoReal
48	Impedimento (direito processual)	ImpedimentoProcessual
49	Improbidade (direito do trabalho)	ImprobidadeDeTrabalho
50	Incêndio (crime)	IncêndioCriminoso
51	Inundação (crime)	InundaçãoCriminosa
52	Invenção (coisa achada)	InvençãoDeCoisa
53	Leilão (licitação)	LeilãoLicitatório
54	MoedaFalsa (crime)	MoedaFalsaCriminosa
55	Paciente (medicina)	PacienteDeMedicina
56	PerigoDeDesabamento (contravenção)	PerigoDeDesabamentoCriminoso
57	PerigoDeDesastreFerroviário (crime)	PerigoDeDesastreFerroviárioCriminoso
58	PerigoDeInundação (crime)	PerigoDeInundaçãoCriminosa

59	Praça (militar)	PraçaMilitar
60	Pregão (licitação)	PregãoLicitatório
61	PrincípioDaAutonomia (direito comercial)	PrincípioDaAutonomiaComercial
62	PrincípioDaEspecialidade (direito internacional)	PrincípioDaEspecialidadeInternacional
63	PrincípioDaLegalidade (direito penal)	PrincípioDaLegalidadePernal
64	Prova (concurso público)	ProvaEmConcursoPúblico
65	Prova (ensino)	ProvaDeEscola
66	ProvaEscrita (concurso público)	ProvaEscritaEmConcursoPúblico
67	ProvaObjetiva (concurso público)	ProvaObjetivaEmConcursoPúblico
68	ProvaOral (concurso público)	ProvaOralEmConcursoPúblico
69	Referência (direito administrativo)	ReferênciaAdministrativa
70	Remissão (direito civil)	RemissãoCivil
71	Remissão (direito tributário)	RemissãoTributária
72	Remissão (estatuto da criança e do adolescente)	RemissãoDeMenor
73	Requerimento (jurisdição voluntária)	RequerimentoDeJurisdiçãoVoluntária
74	Resistência (crime)	ResistênciaCriminosa
75	Simples (direito tributário)	ImpostoSimples
76	SociedadeDeFato (comercial)	SociedadeComercialDeFato
77	Substituição (disposição testamentária)	SubstituiçãoTestamentária
78	Título (concurso público)	TítuloEmConcursoPúblico
79	Tragronomia (crea)	Tragometria
80	Transcrição (direito civil)	TranscriçãoCivil
81	Uso (direito civil)	UsoCivil

APÊNDICE C – Lista de classes com complementação de rótulos

Este Apêndice apresenta uma lista de todas as classes que receberam formas variantes dos rótulos oriundos do processo de conversão a partir do tesauro do Supremo Tribunal de Justiça.

Os rótulos foram adicionados para aprimorar a capacidade do Motor de Busca Semântica na identificação das classes, pois com mais variações textuais vinculadas a cada classe, estas terão maior chance de identificação.

É uma operação de natureza léxica e não semântica, mas foi realizada como uma complementação da ontologia peso leve resultante do tesauro jurídico, denominada de **Vocabulario-2016-06**.

A tabela a seguir lista apenas os rótulos que foram adicionados, excluindo as formas já existentes no próprio tesauro do STJ.

#	Classe	Rótulos
01	AçãoDeCobrança	ações de cobrança
02	AçãoDeNulidade	ações de nulidade ações declaratórias de nulidade
03	AçãoPenal	ações penais
04	Acórdão	acórdãos arestos
05	Agravado	Agravados Agravada agravadas
06	Agravante	agravantes
07	Agravo	agravos
08	Animal	animais
09	AnimalBravio	animais bravios animais bravos animal bravo
10	AnimalDeCria	animais de cria animal de criação animais de criação
11	AnimalDePequenoPorte	animais de pequeno porte
12	AnimalDeTração	animais de tração
13	AnimalDoméstico	animais domésticos
14	AnimalPerigoso	animais perigosos

15	AnimalSilvestre	animais silvestres
16	Apelado	apelados apelada apeladas
17	Apelante	apelantes
18	Comarca	comarcas
19	Competência	competências
20	ConflitoDeCompetência	conflitos de competência
21	Contrabando	crime de contrabando crime do contrabando
22	CriadorDeAnimais	criadores de animais criadores dos animais criador de animal criador do animal criadores do animal
23	Dano	danos
24	DanoÀImagem	danos à imagem
25	DanoAmbiental	danos ambientais
26	DanoAoErário	danos ao erário
27	DanoÀVidaDeRelação	danos à vida de relação
28	DanoColetivo	danos coletivos
29	DanoMaterial	danos materiais dano de ordem material danos de ordem material dano de natureza material danos de natureza material
30	DanoMoral	danos morais dano de ordem moral danos de ordem moral dano de natureza moral danos de natureza moral
31	DecisãoMonocrática	decisões monocráticas
32	Descaminho	crime de descaminho crime do descaminho
33	Desembargador	desembargadora des desª
34	Embargado	embargados embargada embargadas

35	Embargante	embargantes
36	Embargos	embargo
37	Ementa	ementas
38	Fato	fatos
39	Impetrado	impetrados impetrada impetradas
40	Impetrante	impetrantes
41	Juiz	juízes juíza juizas
42	Jurisprudência	jurisprudências
43	Interessado	interessados interessada interessadas
44	Legislação	legislações
45	Local	locais lugares
46	MandadoDeSegurança	mandados de segurança
47	Ministro	ministro ministra ministras
48	Presidente	presidentes
49	Prestação	prestações
50	Recorrente	recorrentes
51	Recorrido	recorridos recorrida recorridas
52	Reclamado	reclamados reclamada reclamadas
53	Reclamante	reclamantes
54	RecursoJudicial	Recursos judiciais
55	Relator	relatores relatora relatoras
56	Requerente	requerentes
57	Requerido	requeridos requerida requeridas

58	Réu	réus
59	RevisãoCriminal	revisões criminais
60	Revisor	revisores revisora revisoras
61	Serviço	serviços
62	Suscitado	suscitados suscitada suscitadas
63	Suscitante	suscitantes
64	PrestaçãoDeServiço	prestação de serviços prestação do serviço
65	TrânsitoEmJulgado	transitado em julgado transitada em julgado
66	Vítima	vítimas
67	Voto	votos

APÊNDICE D – Lista de classes comuns entre ontologias JurisTJPR e Vocabulario-2016-06

Este Apêndice apresenta uma lista de todas as classes que existiam simultaneamente nas ontologias **JurisTJPR** e **Vocabulario-2016-06**.

Em alguns casos as classes apresentavam formas ligeiramente diferentes na grafia de seus nomes, visto que cada ontologia empregou critérios diferentes a respeito deste atributo. Nestes casos, o nome da classe na ontologia JurisTJPR foi alterado para que no processo de combinação (do inglês, *merge*) o Protégé percebesse a correspondência entre as duas classes.

A tabela listando todas as ocorrências de classes em comum e suas eventuais renomeações encontra-se a seguir. A segunda coluna mostra o nome da classe como existente na ontologia **JurisTJPR** e a terceira coluna sua respectiva na ontologia **Vocabulario-2016-06**.

#	Nome original	Nome convertido
01	AcaoBusca	AçãoDeBusca
02	AcaoCobranca	AçãoDeCobrança
03	AçaoNulidade	AçãoDeNulidade
04	AcaoPenal	AçãoPenal
05	Acordao	Acórdão
06	Agravado	Agravado
07	Agravante	Agravante
08	Agravo	Agravo
09	Apelado	Apelado
10	Apelante	Apelante
11	Comarca	Comarca
12	Competencia	Competência
13	ConflitoCompetencia	ConflitoDeCompetência
14	ConflitoJurisicao	ConflitoDeJurisdição
15	ConflitoNegativo	ConflitoNegativoDeCompetência
16	DecisaoMonocratica	DecisãoMonocrática
17	Desembargador	Desembargador
18	Embargado	Embargado
19	Embargante	Embargante
20	Embargo	Embargos
21	Ementa	Ementa

22	Fato	Fato
23	HabeasData	HabeasData
24	HabeasCorpus	HabeasCorpus
25	Impetrado	Impetrado
26	Impetrante	Impetrante
27	Interessado	Interessado
28	Juiz	Juiz
29	Jurisprudencia	Jurisprudência
30	Legislacao	Legislação
31	Local	Local
32	MandadoSeguranca	MandadoDeSegurança
33	Ministro	Ministro
34	Presidente	Presidente
35	Reclamado	Reclamado
36	Reclamante	Reclamante
37	Recorrente	Recorrente
38	Recorrido	Recorrido
39	Recurso	RecursoJudicial
40	Relator	Relator
41	Requerente	Requerente
42	Requerido	Requerido
43	Reu	Réu
44	RevisaoCriminal	RevisãoCriminal
45	RevisaoTaxa	TaxaDeRevisão
46	Revisor	Revisor
47	Suscitado	Suscitado
48	Suscitante	Suscitante
49	Vitima	Vítima
50	Voto	Voto

APÊNDICE E – Lista dos 100 documentos com maior quantidade de classes identificadas

Este Apêndice apresenta uma lista dos 100 documentos com maior vinculação de classes.

A tabela a seguir lista cada um dos documentos, identificados pelo número do processo, apresentando quantas classes ao todo, apresentando também as dez com maior frequência foram identificadas. Neste caso os textos das decisões foram suprimidos por questão de espaço.

#	Processo	Qtd.	Classes
01	1451495-7	387	JurisprudênciaDominante AçãoDePrestaçãoDeContas TerceiraTurma Legislação Junho Segurado ProduçãoDeProva Citação ForoEnfitêutico Caput
02	1456640-2	365	PrestaçãoDeInformações JurisprudênciaDominante Legislação SaldoDevedor ConstituiçãoFederal Fiança DesequilíbrioContratual ApelaçãoCível EstadoDoRioGrandeDoSul Pagamento
03	1519496-6	364	Legislação SaldoDevedor Norma

			ValorResidualGarantido TerceiraTurma DecisãoExtraPetita ConstituiçãoFederal PoderAquisitivo ApelaçãoCível Cálculo
04	1447302-8	364	DecisãoExtraPetita SaldoDevedor JurisprudênciaDominante TerceiraTurma Segurado PrestaçãoDeInformações AçãoDePrestaçãoDeContas OrdemLegal ConstituiçãoFederal Norma
05	1459598-5	363	JurisprudênciaDominante TerceiraTurma Segurado ProduçãoDeProva PrestaçãoDeInformações AçãoDePrestaçãoDeContas Norma PrimeiraTurma ApelaçãoCível Citação
06	1486843-2	358	Câmbio AçãoDePrestaçãoDeContas JurisprudênciaDominante SessãoDeJulgamento PrimeiraTurma Norma TerceiraTurma

			Aumento Legislação Pagamento
07	1469661-6	355	Legislação Promulgação ConstituiçãoFederal PrimeiraTurma RepercussãoGeral PrincípioDaEspecialidade DireitoFundamental EstadoDeDireito Dúvida Norma
08	1448026-7	351	JurisprudênciaDominante ConstituiçãoFederal PrimeiraTurma Litisconsorte DecisãoExtraPetita PolíticaSalarial SaldoDevedor DemonstrativoDoDébito Dúvida TerceiraTurma
09	1412321-4	348	JurisprudênciaDominante DecisãoExtraPetita PolíticaSalarial SaldoDevedor ConstituiçãoFederal Norma ProduçãoDeProva Legislação PrimeiraTurma ApelaçãoCível

10	1460612-7	342	JurisprudênciaDominante ConstituiçãoFederal Norma AçãoDePrestaçãoDeContas TerceiraTurma Legislação DecisãoExtraPetita ApelaçãoCível Citação Rio
11	1476512-9	338	DecisãoExtraPetita SaldoDevedor JurisprudênciaDominante TerceiraTurma AçãoDePrestaçãoDeContas OrdemLegal Pagamento ApelaçãoCível Citação Caput
12	1583179-7	336	DesequilíbrioContratual Legislação ErroInescusável SaldoDevedor ValorResidualGarantido PoderAquisitivo TerceiraTurma ProduçãoDeProva Dúvida ConstituiçãoFederal
13	1486144-4	334	MovimentaçãoFinanceira RecursoManifestamenteInadmissível JurisprudênciaDominante SessãoDeJulgamento

			PrimeiraTurma Norma AçãoDePrestaçãoDeContas TerceiraTurma Aumento SupressãoDeInstância
14	1558037-5	333	Aumento SupressãoDeInstância Legislação Dúvida SaldoDevedor TerceiraTurma Condenado Pagamento Conta ProRata
15	1475365-6	331	SaldoDevedor DecisãoExtraPetita Liberdade Dúvida Legislação TerceiraTurma Norma ContratoAcessório ConsignaçãoEmPagamento ApelaçãoCível
16	1469568-0	330	Junho Legislação ObrigaçãoTributária TerceiraTurma PrimeiraTurma Norma Liberdade Dúvida

			ContratoAcessório SaldoDevedor
17	1511937-0	327	AçãoDePrestaçãoDeContas RecursoManifestamenteInadmissível JurisprudênciaDominante Legislação Autenticação MovimentaçãoFinanceira TerceiraTurma SaldoDevedor Liberdade Aumento
18	1459180-3	327	SaldoDevedor Liberdade JurisprudênciaDominante DecisãoExtraPetita PolíticaSalarial TerceiraTurma PequenaPropriedadeRural ProduçãoDeProva Legislação ApelaçãoCível
19	1510755-4	324	JurisprudênciaDominante NaturezaJurídica ObrigaçãoTributária Junho Aumento ConstituiçãoFederal Norma PoloPassivo EstadoDoAcre ForoEnfitêutico
20	1510500-9	321	AçãoDePrestaçãoDeContas SaldoDevedor

			DecisãoExtraPetita RecursoManifestamenteInadmissível JurisprudênciaDominante Legislação MovimentaçãoFinanceira Dúvida Aumento TerceiraTurma
21	1472257-7	319	Aumento Legislação Dúvida ConstituiçãoFederal TerceiraTurma ApelaçãoCível EstadoDoRioGrandeDoSul EstadoDoAcre Pagamento BuscaEApreensão
22	1484602-3	316	Aumento Liberdade Dúvida Legislação TerceiraTurma Norma ContratoAcessório ObrigaçãoTributária SaldoDevedor ApelaçãoCível
23	1518696-2	310	Aumento Norma DecisãoExtraPetita RepercussãoGeral TerceiraTurma ConstituiçãoFederal

			Legislação AçãoCondenatória ApelaçãoCível Cálculo
24	1473333-6	310	Legislação ConsignaçãoEmPagamento Liberdade Dúvida TerceiraTurma Norma ContratoAcessório ApelaçãoCível EstadoDoRioGrandeDoSul Citação
25	1494774-7	309	ConstituiçãoFederal JurisprudênciaDominante Legislação EnsinoFundamental UniversidadeDeSãoPaulo Liberdade EstadoDeDireito PrimeiraTurma DireitoFundamental ApelaçãoCível
26	1477374-3	308	MovimentaçãoFinanceira RecursoManifestamenteInadmissível JurisprudênciaDominante SessãoDeJulgamento PrimeiraTurma ProduçãoDeProva TerceiraTurma Norma SaldoDevedor Legislação

27	1494490-6	307	Curador Aumento Norma PrimeiraTurma ConstituiçãoFederal RepercussãoGeral NaturezaJurídica Tentativa Citação Pagamento
28	1479411-9	307	Aumento SaldoDevedor JurisprudênciaDominante ConstituiçãoFederal Norma AçãoDePrestaçãoDeContas TerceiraTurma Legislação ApelaçãoCível ForoEnfitêutico
29	1469485-6	307	SaldoDevedor Legislação ValorResidualGarantido Dúvida ConstituiçãoFederal TerceiraTurma Condenado Pagamento EstadoDoRioGrandeDoSul ProRata
30	1507706-6	306	Aumento RecursoManifestamenteInadmissível JurisprudênciaDominante RepercussãoGeral

			Norma ConstituiçãoFederal Legislação DecisãoExtraPetita ProduçãoDeProva AçãoDePrestaçãoDeContas
31	1481968-4	306	ConstituiçãoFederal JurisprudênciaDominante Legislação EnsinoFundamental UniversidadeDeSãoPaulo Liberdade EstadoDeDireito PrimeiraTurma DireitoFundamental ApelaçãoCível
32	1496865-1	305	Legislação ObrigaçãoTributária Liberdade Dúvida Norma TerceiraTurma ConsignaçãoEmPagamento SaldoDevedor ApelaçãoCível EstadoDoRioGrandeDoSul
33	1479175-8	303	SupressãoDeInstância TerceiraTurma SaldoDevedor RepercussãoGeral ConstituiçãoFederal Legislação TaxaDeJurosDeLongoPrazo Pagamento

			BuscaEApreensão ApelaçãoCível
34	1470711-8	302	RepercussãoGeral PoderAquisitivo ConstituiçãoFederal RecursoManifestamenteInadmissível JurisprudênciaDominante Dúvida Legislação Cidade Junho SupressãoDeInstância
35	1519872-6	302	ConsignaçãoEmPagamento DecisãoExtraPetita Norma Legislação TerceiraTurma RepercussãoGeral Dúvida ApelaçãoCível EstadoDoRioGrandeDoSul Cálculo
36	1574861-1	301	Norma Legislação Liberdade Dúvida RepercussãoGeral TerceiraTurma ContratoAcessório ApelaçãoCível EstadoDoRioGrandeDoSul Representado
37	1521639-2	299	Aumento Liberdade

			Dúvida Legislação Norma TerceiraTurma ContratoAcessório ApelaçãoCível EstadoDoRioGrandeDoSul Citação
38	1509573-5	298	Liberdade Dúvida Legislação TerceiraTurma Norma ConsignaçãoEmPagamento ApelaçãoCível EstadoDoRioGrandeDoSul ForoEnfitêutico BuscaEApreensão
39	1129651-2	297	JurisprudênciaDominante TerceiraTurma DanoAmbiental Norma AutoriaDeObra DecisãoExtraPetita PolíticaSalarial SaldoDevedor ProduçãoDeProva ApelaçãoCível
40	1553208-4	297	ProduçãoDeProva Legislação Norma SaldoDevedor TerceiraTurma ApelaçãoCível

			Pagamento Condenado BuscaEApreensão Conta
41	1447902-8	296	SaldoDevedor ProduçãoDeProva JurisprudênciaDominante PrimeiraTurma Dúvida ConstituiçãoFederal ApelaçãoCível Pagamento Caput ParteDispositiva
42	1507081-4	296	Legislação ProduçãoDeProva AçãoDePrestaçãoDeContas Dúvida ConstituiçãoFederal TerceiraTurma Pagamento ApelaçãoCível EstadoDoAcre ProRata
43	1483557-9	296	Legislação Promulgação ConstituiçãoFederal PrimeiraTurma RepercussãoGeral PrincípioDaEspecialidade DireitoFundamental EstadoDeDireito ArguiçãoDeInconstitucionalidade Condenado

44	1527920-2	296	ConsignaçãoEmPagamento ValorResidualGarantido SaldoDevedor Dúvida TerceiraTurma Citação Pagamento ApelaçãoCível EstadoDoAcre Contraprestação
45	1476304-7	295	JurisprudênciaDominante Legislação AçãoDePrestaçãoDeContas Dúvida TerceiraTurma ApelaçãoCível Cálculo Citação Pagamento Caput
46	1507719-3	295	DecisãoExtraPetita Legislação Dúvida ConstituiçãoFederal TerceiraTurma Citação Embargos ApelaçãoCível EstadoDoAcre EstadoDoRioGrandeDoSul
47	1474152-5	294	Aumento Legislação Junho ObrigaçãoTributária

			SaldoDevedor Liberdade Dúvida TerceiraTurma Norma ApelaçãoCível
48	1509250-7	294	Norma Legislação Liberdade Dúvida TerceiraTurma ConsignaçãoEmPagamento ApelaçãoCível EstadoDoRioGrandeDoSul ForoEnfitêutico Conta
49	1467824-5	293	Fiança JurisprudênciaDominante MovimentaçãoFinanceira TerceiraTurma ProduçãoDeProva ConstituiçãoFederal Norma AçãoDePrestaçãoDeContas Locação ContratoAcessório
50	1173582-3	292	RepercussãoGeral PoderAquisitivo ConstituiçãoFederal RecursoManifestamenteInadmissível JurisprudênciaDominante Dúvida Legislação Junho

			SupressãoDeInstância ForoEnfitêutico
51	1471169-8	291	SaldoDevedor DecisãoExtraPetita TaxaBásicaFinanceira Legislação Aluno Aumento ProduçãoDeProva Dúvida ConstituiçãoFederal Norma
52	1481399-9	291	AtoJurídico PoloPassivo ProduçãoDeProva Norma Legislação Dúvida ParteDispositiva ApelaçãoCível EstadoDoAcre Conta
53	1356508-7	290	Legislação Liberdade Dúvida Norma TerceiraTurma ContratoAcessório ApelaçãoCível EstadoDoRioGrandeDoSul BuscaEApreensão Pagamento
54	1468488-3	290	DemonstrativoDoDébito Legislação

			RecursoManifestamenteInadmissível JurisprudênciaDominante SaldoDevedor PoloPassivo ConstituiçãoFederal Fiança ForoEnfitêutico ProRata
55	1486935-5	290	PoloPassivo Norma PrimeiraTurma FraudeFiscal NaturezaJurídica TerceiraTurma Pagamento InteressePrivado EstadoDoRioGrandeDoSul Citação
56	1473174-7	288	MovimentaçãoFinanceira JurisprudênciaDominante SessãoDeJulgamento Norma AçãoDePrestaçãoDeContas SupressãoDeInstância Legislação Aumento ApelaçãoCível PedidoGenérico
57	1442069-8	288	DecisãoExtraPetita Legislação SaldoDevedor ValorResidualGarantido Segurado Junho

			Dúvida JurisprudênciaDominante Norma EstadoDoRioGrandeDoSul
58	1570990-1	288	ProduçãoDeProva Legislação Dúvida SaldoDevedor ValorResidualGarantido Pagamento ApelaçãoCível EstadoDoAcre EstadoDoRioGrandeDoSul BuscaEApreensão
59	1490807-5	286	Aumento JurisprudênciaDominante ConstituiçãoFederal Aposentadoria PrimeiraTurma Legislação Norma SessãoDeJulgamento RepercussãoGeral Dúvida
60	1452946-3	285	ConstituiçãoFederal Aumento NaturezaJurídica Norma FundoDeDireito AtoAdministrativo Junho Cálculo Pagamento Citação

61	1556600-0	285	Legislação Dúvida ConstituiçãoFederal TerceiraTurma Norma SaldoDevedor ProRata Pagamento Certidão EstadoDoRioGrandeDoSul
62	1488022-1	284	PoloPassivo ConstituiçãoFederal Legislação Liberdade RepercussãoGeral NaturezaJurídica Dever Certidão FlagranteDelito Pagamento
63	1512190-1	284	Legislação Liberdade Dúvida Norma TerceiraTurma ContratoAcessório ApelaçãoCível EstadoDoRioGrandeDoSul ForoEnfitêutico Representado
64	1496190-9	284	Legislação Dúvida ConstituiçãoFederal SaldoDevedor

			EstadoDoRioGrandeDoSul BuscaEApreensão Cálculo ApelaçãoCível Caput EstadoDoAcre
65	1502172-0	283	SaldoDevedor ErroInescusável RecursoManifestamenteInadmissível JurisprudênciaDominante ProduçãoDeProva Legislação PrimeiraTurma Norma TerceiraTurma EspecificaçãoDeProvas
66	1508825-0	281	DecisãoExtraPetita Norma Legislação Liberdade Dúvida TerceiraTurma ApelaçãoCível EstadoDoRioGrandeDoSul BuscaEApreensão Pagamento
67	1479946-7	281	TerceiraTurma Legislação Norma SaldoDevedor DecisãoExtraPetita RepercussãoGeral DesequilíbrioContratual ApelaçãoCível

			Conta EstadoDoRioGrandeDoSul
68	1472103-4	280	AçãoDePrestaçãoDeContas Fiança RecursoManifestamenteInadmissível JurisprudênciaDominante PrimeiraTurma Curador ProduçãoDeProva SaldoDevedor Norma TerceiraTurma
69	1460960-8	278	AutoridadeCoatora ConstituiçãoFederal Legislação Junho Quimioterapia Norma DireitoFundamental PoloPassivo NaturezaJurídica JurisprudênciaDominante
70	1372260-2	278	RecursoManifestamenteInadmissível JurisprudênciaDominante ProduçãoDeProva Norma AçãoCondenatória PrimeiraTurma Caput Pagamento EstadoDoRioGrandeDoSul DepósitoBancário
71	1520838-1	278	Dúvida TerceiraTurma

			Legislação ConstituiçãoFederal SupressãoDeInstância PoloPassivo ApelaçãoCível Pagamento EstadoDoRioGrandeDoSul Rio
72	1508037-0	278	JurisprudênciaDominante Norma SaldoDevedor CódigoPenal RecursoManifestamenteInadmissível Legislação ConstituiçãoFederal ApelaçãoCível Cálculo Citação
73	1490582-3	278	Aumento JurisprudênciaDominante ConstituiçãoFederal Legislação Aposentadoria PrimeiraTurma Norma SessãoDeJulgamento RepercussãoGeral Pagamento
74	1509941-3	277	JurisprudênciaDominante NaturezaJurídica ObrigaçãoTributária Junho PoloPassivo PrimeiraTurma

			ConstituiçãoFederal EstadoDoAcre ForoEnfitêutico Neto
75	1496905-0	277	ConstituiçãoFederal PrimeiraTurma RepercussãoGeral DireitoFundamental EstadoDeDireito ArguiçãoDeInconstitucionalidade Pagamento Conta ParteDispositiva LeiEmTese
76	1577438-4	276	ProduçãoDeProva Legislação Dúvida TerceiraTurma ValorResidualGarantido SaldoDevedor Condenado EstadoDoRioGrandeDoSul EstadoDoAcre ApelaçãoCível
77	1505987-3	275	SaldoDevedor DecisãoExtraPetita Liberdade Dúvida Legislação TerceiraTurma Norma ApelaçãoCível EstadoDoRioGrandeDoSul ForoEnfitêutico

78	1475880-8	275	PoloPassivo ProduçãoDeProva Norma SaldoDevedor Locação ImprensaOficial Dúvida TerceiraTurma ApelaçãoCível Caput
79	1459276-4	274	SaldoDevedor Norma JurisprudênciaDominante ProduçãoDeProva TerceiraTurma PoloPassivo ApelaçãoCível Cálculo Pagamento Caput
80	1507760-0	273	Legislação SaldoDevedor PoderAquisitivo TerceiraTurma JurisprudênciaDominante Pagamento ApelaçãoCível ProRata Caput EstadoDoAcre
81	1457478-0	273	JurisprudênciaDominante PrimeiraTurma Legislação PoloPassivo

			Dúvida TerceiraTurma Pagamento Certidão Caput ParteDispositiva
82	1502140-8	273	SaldoDevedor Norma RecursoManifestamenteInadmissível JurisprudênciaDominante TerceiraTurma Legislação CréditoPrivilegiado ApelaçãoCível Citação Pagamento
83	1477481-3	273	ValorResidualGarantido DecisãoExtraPetita SaldoDevedor Legislação TerceiraTurma ApelaçãoCível Pagamento Caput Embargos MinistérioDaFazenda
84	1514799-2	273	Legislação ObrigaçãoTributária AtoJurídico Liberdade Dúvida Norma TerceiraTurma ContratoAcessório

			ApelaçãoCível EstadoDoRioGrandeDoSul
85	1500144-8	272	Liberdade Dúvida Legislação TerceiraTurma Norma SaldoDevedor ApelaçãoCível EstadoDoRioGrandeDoSul BuscaEApreensão Cálculo
86	1494103-8	272	RecursoManifestamenteInadmissível JurisprudênciaDominante Legislação GuiaDeRecolhimentoDoFgtsEInformaçõesDaPrevidênciaSocial PrimeiraTurma ConstituiçãoFederal Norma PoloPassivo SupressãoDeInstância ObrigaçãoTributária
87	1475185-8	271	ProduçãoDeProva Norma TerceiraTurma ImprensaOficial Dúvida ErroInescusável Legislação EstadoDoRioGrandeDoSul Citação ApelaçãoCível

88	1453465-7	270	PoloPassivo ProduçãoDeProva Legislação ConstituiçãoFederal Segurado Junho TeoriaDaAsserção LegitimidadeProcessual TerceiraTurma Dúvida
89	1461586-6	268	JurisprudênciaDominante TerceiraTurma DecisãoExtraPetita ApelaçãoCível DanoInReIpsa Pagamento Caput ParteDispositiva DepósitoBancário RecursoOrdinárioEmMandadoDeSegurança
90	1458563-8	267	PoloPassivo Litisconsorte AtoAdministrativo ConstituiçãoFederal Norma DireitoFundamental PrimeiraTurma JurisprudênciaDominante ApelaçãoCível ForoEnfitêutico
91	1470321-4	267	Norma DeverFuncional Liberdade NaturezaJurídica

			ConstituiçãoFederal ApelaçãoCível Citação Dever ForoEnfitêutico Pagamento
92	1480441-4	267	PrimeiraTurma Aumento ConstituiçãoFederal PoloPassivo ObrigaçãoTributária AtoAdministrativo Matriz Liberdade RepercussãoGeral JurisprudênciaDominante
93	1495254-4	266	ValorResidualGarantido Legislação Dúvida ConstituiçãoFederal ApelaçãoCível Pagamento Contraprestação Negociação EstadoDoRioGrandeDoSul Rio
94	1506313-7	266	AçãoDePrestaçãoDeContas ProduçãoDeProva DecisãoExtraPetita TerceiraTurma SupressãoDeInstância SaldoDevedor Junho RecursoManifestamenteInadmissível

			JurisprudênciaDominante ApelaçãoCível
95	1426678-7	266	JurisprudênciaDominante SaldoDevedor Aumento TerceiraTurma DecisãoExtraPetita AçãoDePrestaçãoDeContas ProduçãoDeProva Pagamento ForoEnfitêutico Caput
96	1436596-3	266	ConstituiçãoFederal Legislação Aumento NaturezaJurídica PrimeiraTurma PoderAquisitivo RepercussãoGeral DecisãoExtraPetita Rio Pagamento
97	1421035-2	266	JurisprudênciaDominante SaldoDevedor Aumento TerceiraTurma DecisãoExtraPetita AçãoDePrestaçãoDeContas ProduçãoDeProva Pagamento ForoEnfitêutico Caput
98	1472425-5	265	Legislação ObrigaçãoTributária

			Liberdade Dúvida TerceiraTurma Norma SaldoDevedor ApelaçãoCível EstadoDoRioGrandeDoSul Rio
99	1475563-2	265	Dúvida Norma SaldoDevedor Aumento JurisprudênciaDominante Legislação ApelaçãoCível Pagamento BuscaEApreensão Caput
100	1467029-0	265	CódigoPenal EnsinoFundamental DireitoFundamental ConstituiçãoFederal Dúvida Legislação Imprudência Aumento ApelaçãoCível PenaDeMulta

APÊNDICE F – Lista das 100 classes mais recorrentes no *corpus* de testes

Este Apêndice apresenta uma lista das 100 classes mais frequentemente identificadas dentro do *corpus* de avaliação.

A tabela a seguir lista cada uma das classes (coluna "**Classe**"), com suas frequências absoluta (coluna "**Freq.**") e relativa (coluna "**Freq. r**") em relação tamanho do *corpus* e o valor da métrica ***Inverse Document Frequency*** (coluna "**IDF**") correspondente.

#	Classe	Freq.	Freq. R	IDF
01	EstadoDeSergipe	58859	0,9911	0,00
02	Desembargador	38382	0,6463	0,19
03	Artigo	37108	0,6249	0,20
04	CódigoDeProcessoCivil	36906	0,6215	0,21
05	OrganizaçãoSocial	36726	0,6184	0,21
06	Autos	35970	0,6057	0,22
07	RecursoJudicial	35636	0,6001	0,22
08	Relator	34752	0,5852	0,23
09	Prosseguimento	32681	0,5503	0,26
10	DireitoEspecialDeSaque	32317	0,5442	0,26
11	EstadoDoParaná	30395	0,5118	0,29
12	RecursoEspecial	30298	0,5102	0,29
13	DireçãoEAssessoramentoSuperior	29131	0,4905	0,31
14	Decisão	25675	0,4323	0,36
15	Lei	24499	0,4125	0,38
16	SuperiorTribunalDeJustiça	24146	0,4066	0,39
17	DecisãoMonocrática	23888	0,4022	0,40
18	Pedido	22260	0,3748	0,43
19	Pagamento	21692	0,3653	0,44
20	Presidente	20854	0,3512	0,45
21	Acórdão	20776	0,3498	0,46
22	ImpostoSobreImportação	20244	0,3409	0,47
23	Relatório	20137	0,3391	0,47
24	Inciso	20068	0,3379	0,47
25	Recorrente	19254	0,3242	0,49
26	PoderJudiciário	19175	0,3229	0,49
27	Valor	19102	0,3217	0,49
28	Julgamento	18868	0,3177	0,50
29	Análise	18864	0,3177	0,50

30	AusênciaCivil	18812	0,3168	0,50
31	Bem	18500	0,3115	0,51
32	Custas	18025	0,3035	0,52
33	Admissibilidade	16683	0,2809	0,55
34	Data	16590	0,2794	0,55
35	Súmula	16220	0,2731	0,56
36	Condenação	16092	0,2710	0,57
37	Mérito	15997	0,2694	0,57
38	Juiz	15891	0,2676	0,57
39	TribunalDeJustiça	15786	0,2658	0,58
40	Juízo	15021	0,2529	0,60
41	HonoráriosAdvocatícios	14537	0,2448	0,61
42	Recorrido	14372	0,2420	0,62
43	Jurisprudência	14346	0,2416	0,62
44	Civel	14219	0,2394	0,62
45	RecursoInominado	13757	0,2317	0,64
46	Rua	13614	0,2292	0,64
47	AgravoDeInstrumento	13537	0,2279	0,64
48	Feito	13386	0,2254	0,65
49	Ministro	13263	0,2233	0,65
50	Fato	13242	0,2230	0,65
51	Assunto	13178	0,2219	0,65
52	Direito	12991	0,2188	0,66
53	CiênciaDoDireito	12991	0,2188	0,66
54	Fundamentação	12477	0,2101	0,68
55	Classe	12334	0,2077	0,68
56	Razões	12135	0,2043	0,69
57	Serviço	12088	0,2035	0,69
58	Prazo	11844	0,1994	0,70
59	Cobrança	11674	0,1966	0,71
60	Comarca	11566	0,1948	0,71
61	Resolução	11539	0,1943	0,71
62	SegundaTurma	11496	0,1936	0,71
63	Banco	11184	0,1883	0,73
64	Objeto	11031	0,1858	0,73
65	Autor	11000	0,1852	0,73
66	AutorDeObra	11000	0,1852	0,73
67	Acordo	10966	0,1847	0,73
68	Agravante	10834	0,1824	0,74

69	Apelacao	10724	0,1806	0,74
70	Caput	10592	0,1784	0,75
71	Contrato	10584	0,1782	0,75
72	ApelaçãoCível	10545	0,1776	0,75
73	Consumidor	10520	0,1771	0,75
74	Magistrado	10303	0,1735	0,76
75	Tribunal	10281	0,1731	0,76
76	TurmaRecursal	10276	0,1730	0,76
77	Alegação	10087	0,1699	0,77
78	ParteDispositiva	9838	0,1657	0,78
79	ProvaEmConcursoPúblico	9798	0,1650	0,78
80	ProvaEscolar	9797	0,1650	0,78
81	Prova	9797	0,1650	0,78
82	ProvaDeEscola	9797	0,1650	0,78
83	ExameEscolar	9797	0,1650	0,78
84	EstadoDoRioGrandeDoSul	9767	0,1645	0,78
85	Março	9735	0,1639	0,79
86	EstadoDeSãoPaulo	9722	0,1637	0,79
87	AgravoRegimental	9716	0,1636	0,79
88	DanoMoral	9703	0,1634	0,79
89	Agravo	9683	0,1631	0,79
90	Indenização	9613	0,1619	0,79
91	Civil	9551	0,1608	0,79
92	TítuloEmConcursoPúblico	9464	0,1594	0,80
93	Crédito	9326	0,1570	0,80
94	SupremoTribunalFederal	9229	0,1554	0,81
95	CódigoDeDefesaDoConsumidor	9130	0,1537	0,81
96	Adimplemento	9056	0,1525	0,82
97	Pena	9012	0,1518	0,82
98	Requisito	8902	0,1499	0,82
99	Reforma	8454	0,1424	0,85
100	Objetivo	8229	0,1386	0,86

APÊNDICE G – Lista das 100 classes com menor recorrência no *corpus* de teste

Este Apêndice apresenta uma lista das 100 classes menos frequentemente identificadas dentro do *corpus* de avaliação, excluindo-se as classes com frequeência nula.

A tabela a seguir lista cada uma das classes (coluna "**Classe**") com seus rótulos (coluna "**Rótulos**"), frequência (coluna "**Freq.**") e o valor da métrica ***Inverse Document Frequency*** (coluna "**IDF**") correspondente.

#	Classe	Rótulos	Freq.	IDF
01	ProgramaDeGarantiaDaAtividadeAgropecuária	proagro programa de garantia da atividade agropecuária	1	4,77
02	ContribuiçãoDeIntervençãoNoDomínioEconômico	cide contribuição de intervenção no domínio econômico	1	4,77
03	SegundoVínculo	segundo vínculo	1	4,77
04	TermoDeRenegociação	termo de renegociação	1	4,77
05	ConflitoDeAtribuições	conflito de atribuições	1	4,77
06	TaxaDeServiçosAdministrativos	taxa de serviços administrativos tsa	1	4,77
07	SegundoHabeasCorpus	segundo habeas corpus	1	4,77
08	Entrevistado	Entrevistado	1	4,77
09	BotijãoDeGás	botijão de gás	1	4,77
10	PrisãoPerpétua	prisão perpétua	1	4,77
11	PrimeiroAno	primeiro ano	1	4,77
12	DireitoDoTrabalhador	direito do trabalhador	1	4,77
13	AçãoConstitutiva	ação constitutiva	1	4,77
14	AposentadoriaVoluntária	aposentadoria espontânea	1	4,77

		aposentadoria voluntária		
15	CrimeContraAHonra	crime contra a honra	1	4,77
16	Cantor	cantor	1	4,77
17	GásLiquefeitoDePetróleo	gás de cozinha gás liquefeito de petróleo glp	1	4,77
18	AcumulaçãoRemunerada	acumulação de cargos acumulação remunerada	1	4,77
19	CrimePróprio	crime próprio	1	4,77
20	MáquinaIndustrial	máquina industrial	1	4,77
21	RegimeDeEstimativa	regime de estimativa	1	4,77
22	LocaçãoParaTemporada	locação para temporada	1	4,77
23	ViolênciaArbitrária	violência arbitrária	1	4,77
24	Silo	silo	1	4,77
25	IncidenteDeSanidadeMental	incidente de sanidade mental	1	4,77
26	CâmaraDeCompensação	câmara de compensação carteira de compensação	1	4,77
27	Provisionamento	provisionamento	1	4,77
28	CrimeConexo	crime conexo	1	4,77
29	PerturbaçãoDaOrdem	perturbação da ordem	1	4,77
30	MarcaEspecífica	marca específica	1	4,77
31	PrimeiroSemestre	primeiro semestre	1	4,77
32	SonorizaçãoAmbiente	sonorização ambiental sonorização ambiente sonorização mecânica	1	4,77

33	PrazoJudicial	prazo judicial	1	4,77
34	AbonoDePermanênciaEmServiço	abono de permanência em serviço adicional de permanência	1	4,77
35	ObraLiterária	obra literária	1	4,77
36	DomínioDireto	domínio direto	1	4,77
37	TítuloLíquidoECerto	título líquido e certo	1	4,77
38	PrincípioDaExcepcionalidade	princípio da excepcionalidade	1	4,77
39	Preponente	preponente	1	4,77
40	ConselhoNacionalDeAssistênciaSocial	cnas conselho nacional de assistência social	1	4,77
41	AtoObsceno	ato obsceno	1	4,77
42	TeoriaDosMotivosDeterminantes	teoria dos motivos determinantes	1	4,77
43	Enfiteuta	enfiteuta foreiro	1	4,77
44	Esterilidade	esterilidade	1	4,77
45	AfirmaçãoFalsa	afirmação falsa	1	4,77
46	NotaDeEmpenho	nota de empenho	1	4,77
47	HerdeiroNecessário	herdeiro necessário	1	4,77
48	LucroBruto	lucro bruto	1	4,77
49	AtoIneficaz	ato ineficaz	1	4,77
50	DeficienteFísico	deficiente físico	1	4,77
51	AçãoQuantiMinoris	ação de redução de preço ação estimatória ação quanti minoris	1	4,77
52	AutorizaçãoEspecialDeTrânsito	autorização especial de trânsito	1	4,77
53	IntegraçãoSocial	integração social	1	4,77

54	DocumentoDigitalizado	documento digitalizado	1	4,77
55	RiscoAdministrativo	risco administrativo	1	4,77
56	IneficáciaAbsolutaDoMeio	ineficácia absoluta do meio	1	4,77
57	PrincípioDaLiberdadeContratual	princípio da liberdade contratual	1	4,77
58	PromoçãoPorAntiguidade	promoção por antiguidade	1	4,77
59	LançamentoPorDeclaração	lançamento por declaração	1	4,77
60	AdoçãoÀBrasileira	adoção à brasileira	1	4,77
61	PromissárioCessionário	compromissário cessionário promissário cessionário	1	4,77
62	AlongamentoDaDívidaRural	alongamento da dívida rural	1	4,77
63	ReceitaBrutaAnual	receita bruta anual	1	4,77
64	SaídaDeMercadoria	saída de mercadoria	1	4,77
65	AbonoAnual	abono anual	1	4,77
66	Órtese	órtese	1	4,77
67	Retransmissão	retransmissão	1	4,77
68	ExploraçãoFlorestal	exploração florestal	1	4,77
69	SuperintendenteRegional	superintendente regional	1	4,77
70	CorporaçãoMilitar	corporação militar	1	4,77
71	ClubeSocial	clube social	1	4,77
72	SentençaRescindida	sentença rescindida	1	4,77
73	InsanidadeMental	incapacidade mental insanidade mental	1	4,77
74	TerrenoAlheio	terreno alheio	1	4,77

75	RéuIncerto	réu incerto	1	4,77
76	RegimeDisciplinarDiferenciado	rdd regime disciplinar diferenciado	1	4,77
77	ContaDeDepósitoPopular	conta de depósito popular depósito popular	1	4,77
78	Empate	empate	1	4,77
79	Triplicata	triplicata	1	4,77
80	RiscoProfissional	risco profissional	1	4,77
81	RelativamenteIncapaz	relativamente incapaz	1	4,77
82	ProvimentoDaCorregedoria	provimento da corregedoria	1	4,77
83	InscriçãoDefinitiva	inscrição definitiva	1	4,77
84	Evicto	evicto	1	4,77
85	Dublagem	dublagem	1	4,77
86	AcidenteFerroviário	acidente de trem acidente ferroviário	1	4,77
87	ListaSêxtupla	lista sêxtupla	1	4,77
88	ViolaçãoDeCorrespondência	violação de correspondência	1	4,77
89	MedicinaVeterinária	medicina veterinária	1	4,77
90	DomínioPleno	domínio pleno	1	4,77
91	Áustria	áustria república da áustria	1	4,77
92	ReceptaçãoCulposa	receptação culposa	1	4,77
93	Barragem	barragem	1	4,77
94	Nubente	nubente	1	4,77
95	Etanol	etanol	1	4,77
96	ConselhoFederalDeEngenhariaEAgronomia	confea conselho federal de engenharia arquitetura e agronomia conselho federal	1	4,77

		de engenharia e agronomia		
97	EmbargosÀAdjudicação	embargos à adjudicação	1	4,77
98	PatrimônioCultural	patrimônio cultural	1	4,77
99	AgênciaDeViagem	agência de viagem	1	4,77
100	RepúblicaDoSenegal	república do senegal senegal	1	4,77

APÊNDICE H – Pseudocódigo dos principais algoritmos implementados

Este Anexo apresenta os pseudocódigos de diversos algoritmos implementados ao longo das várias etapas da construção de elaboração da ontologia **OntoLegis** e do **Motor de Busca Semântica** que a emprega.

Combinação de linhas múltiplas no arquivo do Tesauro Jurídico do STJ

Como descrito na seção 4.1.1, propósito deste algoritmo é detectar a presença de seções com definições longas de texto que estejam ocupando várias linhas e reorganizá-las todas em uma linha apenas. Este tipo de situação ocorreu **6.746** vezes no documento em texto plano gerado a partir do Tesauro Jurídico do STJ, principalmente com o prefixo “NOTA”.

Visando a simplificação do algoritmo que iria converter o tesauro em ontologia, tais ocorrências receberam este tratamento.

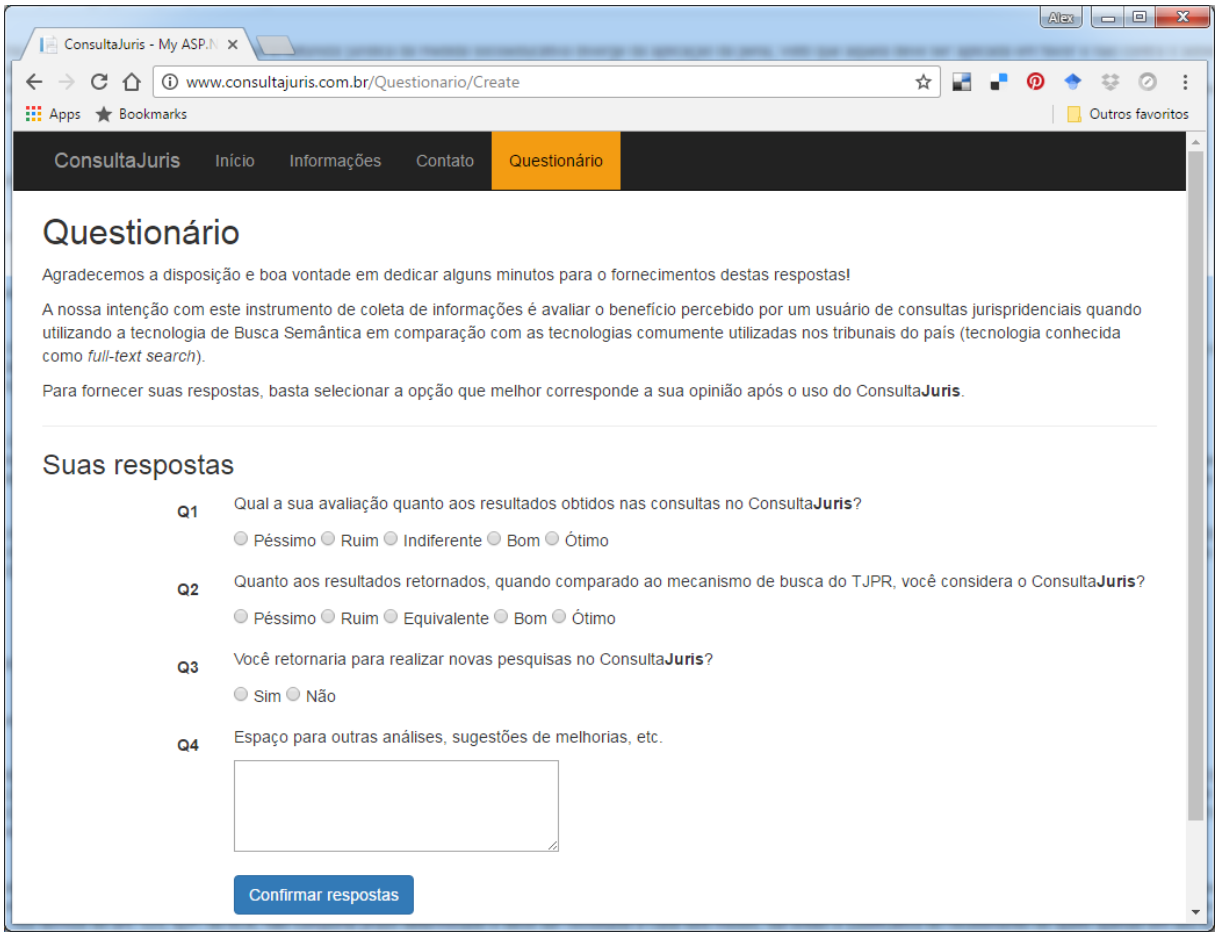
```
Início

EmDescritor ← falso
EmRotulo ← falso
Para cada linha do documento Faça
  Se primeiro caractere da linha não é tabulação Então
    Se EmDescritor Então
      Concatenar linha atual ao fim da linha anterior
      Remover linha atual
    Senão
      EmDescritor ← verdadeiro
      EmRotulo ← falso
    Fim-Se
  Senão
    Se EmRotulo Então
      Concatenar linha atual ao fim da linha anterior
      Remover linha atual
    Senão
      EmDescritor ← falso
      EmRotulo ← verdadeiro
    Fim-Se
  Fim-Se
Fim-Para

Fim
```

Fonte: o Autor (2017)

APÊNDICE I – Página do questionário existente no sítio ConsultaJuris

Este Anexo apresenta o questionário utilizado como instrumento de coleta de opiniões para medir o grau de satisfação que o Motor de Busca Semântica implementado no sítio ConsultaJuris atingiu ao ser experimentado por um público de advogados, que são usuários primários da recuperação jurisprudencial.

O questionário, tal qual é mostrado no sítio ConsultaJuris pode ser visto na figura a seguir.



Fonte: o Autor (2017)