Marcelo Schiessl

Lexicalização de Ontologias: o relacionamento entre conteúdo e significado no contexto da Recuperação da Informação

Brasília
27 de abril de 2015

Marcelo Schiessl

Lexicalização de Ontologias: o relacionamento entre conteúdo e significado no contexto da Recuperação da Informação

Tese apresentada ao Programa de Pós-graduação em Ciência da Informação da Universidade de Brasília como requisito parcial para a obtenção do título de Doutor em Ciência da Informação.

Orientadora: Prof.ª Dra. Marisa Bräscher

Universidade de Brasília — UnB
Faculdade de Ciência da Informação
Programa de Pós-Graduação
schiessl@unb.br
marcelo.schiessl@gmail.com

Brasília
27 de abril de 2015

Schiessl, Marcelo.

S332l Lexicalização de Ontologias: o relacionamento entre conteúdo e significado no contexto da Recuperação da Informação / Marcelo Schiessl; orientador Marisa Bräscher. — Brasília, 2015.

261 p.

Tese (Doutorado - Doutorado em Ciência da Informação) — Universidade de Brasília, 2015.

1. Ontologia. 2. Lexicalização de ontologias. 3. Web Semântica. 4. Processamento de linguagem natural. 5. Recuperação da Informação. I. Bräscher, Marisa, orient. II. Título.

FOLHA DE APROVAÇÃO

Título: "**LEXICALIZAÇÃO DE ONTOLOGIAS:** o relacionamento entre conteúdo e significado no contexto da Recuperação da Informação".

Autor (a): José Marcelo Schiessl

Área de concentração: Gestão da Informação

Linha de pesquisa: Organização da Informação

Tese submetida à Comissão Examinadora designada pelo Colegiado do Programa de Pós-graduação em Ciência da Informação da Faculdade em Ciência da Informação da Universidade de Brasília como requisito parcial para obtenção do título de **Doutor** em Ciência da Informação.

Tese aprovada em: 16 de Abril de 2015

Profª. Drª. Marisa Bräscher Basilio Medeiros
Presidente (UFSC)

Prof. Dr. Ailton Luiz Gonçalves Feitosa
Membro Externo (IESB)

Prof. Dr. Miguel Filho Ferreira de Oliveira
Membro Externo (CEF)

Prof.ª Drª. Fernanda Passini Moreno
Membro Interno (UnB/PPGCINF)

Prof. Dr. Rogério Henrique de Araújo Junior
Membro Interno (UnB/PPGCINF)

Profª. Drª. Lillian Maria Araújo de Rezende Alvares
Suplente (UnB/PPGCINF)

Dedico estte trabalho àqueles que buscam incessantemente o conhecimento para construir um mundo mais justo e melhor de se viver.

Agradecimentos

À minha esposa Solange e minhas filhas Ingrid e Karin por terem me proporcionado o melhor dos refúgios: cumplicidade, compreensão e amor incondicionais.

À professora Marisa Bräscher por ter sempre um conselho preciso e por ser a mão invisível que alinhavou e arrematou as arestas do trabalho.

Ao professor Robredo, *in memoriam*, que me doou tanta energia e vitalidade que eram suas marcas registradas. Apesar da partida inesperada, sua presença permeia toda esta investigação.

Aos professores Steffen Staab — orientador — e Matthias Tim — conselheiro — que me mostraram uma nova perspectiva para fazer Ciência durante o meu estágio de doutoramento no WeST Institute em Koblenz, Alemanha.

Aos professores Miguel, Rogério, Fernanda, Ailton e Lillian pela disponiblibidade para compor minha banca e pelas primorosas sugestões que tanto enriqueceram minha pesquisa.

Ao professor Mamede por sempre se mostrar acessível e pelas boas discussões durante a minha pesquisa.

À Rita Berardi, que conheci do outro lado do Atlântico, por compartilhar comigo momentos de angústias e de alegrias, enquanto nos adaptávamos à cultura alemã e por termos representado, bem, um pouquinho do Brasil em terras estrangeiras.

Ao amigo Mark Jordan, americano de alma brasileira, que, gentil e pacientemente, me socorreu prontamente nos meus embates com o idioma do tio Sam.

Ao apoio dado pelo suporte da FCI, especialmente à Marthinha, por tornar minha vida mais fácil no trato com as formalidades administrativas.

À família e aos amigos pelo apoio incondicional às minhas escolhas e por estarem presentes na minha jornada de vida.

Para ter certeza que não estou sendo injusto com alguém que, certamente, não mencionei pela inevitável falha na memória, agradeço especialmente a VOCÊ que, de alguma forma, possibilitou que meu sonho se tornasse realidade.

Die Grenzen meiner Sprache bedeuten die Grenzen meiner Welt. [1]

Ludwig Wittgenstein

[1] Os limites de minha linguagem significam os limites de meu mundo.

Resumo

Investiga as tecnologias da Web Semântica e as técnicas de Processamento de Linguagem Natural para a elaboração semiautomática de uma base léxico-ontológica, em Português, circunscrita ao domínio de risco financeiro que, incorporada ao modelo de recuperação da informação, visa melhorar a precisão. Identifica teorias, ferramentas e técnicas que propiciam a automatização de procedimentos que extraem elementos ontológicos e léxicos de bases estruturadas e não estruturadas. Esses elementos combinados na forma de base de dados apoiam a geração de índice léxico-semântico que fornece insumos para a proposição de modelo de recuperação da informação semântica. A metodologia adotada se apresenta em: levantamento de fundamentos teóricos e metodológicos, ferramentas e bases de dados ontológicas e textuais; construção de ontologia e base léxico-ontológica com foco no risco financeiro; elaboração de modelo de recuperação da informação semântica; avaliação do modelo realizada num corpus de informação relacionada ao risco financeiro que foi indexado nos moldes tradicionais e contemplando a informação semântica para, então, medir a precisão nas duas situações. Os resultados alcançados demonstram a utilização da metodologia, no domínio de risco financeiro em Português, para a elaboração da ontologia OntoRisco, da base léxico-semântica RiscoLex e da proposta de modelo de recuperação da informação semântica que mostrou resultados superiores aos modelos de recuperação da informação tradicionais, nos testes realizados. Conclui que os resultados satisfatórios mostram a aplicabilidade da proposta metodológica para o domínio em questão e aponta para a possibilidade de expandir a outros domínios com as devidas adaptações dos recursos utilizados. O estudo contribui para a área de representação e organização da informação e do conhecimento na medida em que fornece metodologia, aporte teórico e insumos para que profissionais continuem promovendo o desenvolvimento da Ciência da Informação.

Palavras-chave: Ontologia. Lexicalização de ontologias. Web semântica. Processamento de linguagem natural. Recuperação da informação.

Abstract

This research investigates Semantic Web technologies, and Natural Language Processing techniques in order to semi automatically develop a lexicon-ontological database in Portuguese. This base is intended to improve the precision measurement in the financial risk domain by coupling it into an information retrieval model. It identifies theories, tools and techniques that allow the automation of procedures that extract ontological and lexical elements from structured and non-structured databases. When these lexical and ontological elements are together in a database format, they give support to the generation of lexico-semantic index that can help the creation of semantic information retrieval models. The methodology used is as follows: surveys of methodological and theoretical fundamentals, tools, as well as ontological and textual databases; creation of an ontology and a lexicon-ontology base in the financial risk domain; elaboration of semantic information retrieval model; evaluation of the model using an information corpus related to financial risks – it was indexed by using traditional techniques, i.e. keywords, and also using semantic information; and then the precision of both situations was finally measured. The results achieved demonstrate the methodology used in the financial risk domain to elaborate the ontology, OntoRisco, using the lexico-semantic database, RiscoLex, and the semantic informational retrieval model proposed in this work. In the tests executed, the results were better than those presented by using traditional information retrieval models. It concludes that the results were quite satisfying since they show the applicability of the methodological proposal to the financial risk domain. In addition, it points the possibility of extending the proposal to other domains with only some adaptations.

Keywords: Ontology. Ontology lexicalization. Semantic web. Natural language processing. Information retrieval.

Lista de ilustrações

Figura 1 Representação gráfica de tripla 40

Figura 2 Qual o significado da mensagem? 48
Figura 3 O triângulo do significado e o sucesso da comunicação 50
Figura 4 Relação entre classes e entidades 52
Figura 5 Relação entre classes 52
Figura 6 Modelo Layer Cake 53
Figura 7 Representação da Catedral de Brasília 54
Figura 8 Exemplo em XML 56
Figura 9 Grafo de RDF com URIs 57
Figura 10 Grafo de RDF com URIs/Literal 58
Figura 11 RDF com rótulos em japonês 59
Figura 12 RDF e RDFS 60
Figura 13 Conhecimento Intensional x Extensional 61
Figura 14 Arquitetura de LD 64
Figura 15 Exemplo OWA 65
Figura 16 Relacionamentos de Subclasses entre OWL e RDF/RDFS 67
Figura 17 Extração de grafos 68
Figura 18 Espectro de Ontologias 79
Figura 19 Triângulo do Conceito 93
Figura 20 Triângulo Semiótico revisitado 94
Figura 21 Lematização de Termos 107
Figura 22 Representação Gráfica da Lei de Zipf 113
Figura 23 Significância de Termos 114
Figura 24 Espaço Vetorial em duas dimensões 115
Figura 25 Projeção de observações 117
Figura 26 Modelo Clássico de Busca 120
Figura 27 Processo de Recuperação da Informação 121

Figura 28 Visão macro do domínio de Risco Financeiro 142
Figura 29 Relação Ontologia e Linguagem Natural 149
Figura 30 Principais elementos 150
Figura 31 Ilustração da relação de equivalência entre Risco e Perigo 152
Figura 32 Ilustração da relação de hiponímia 152
Figura 33 Ilustração da Representação lexical em RDF 153

Figura 34 Fluxo de criação da RiscoLex 154
Figura 35 Conceitos e termos 159
Figura 36 Visão Geral do MoRIS 161
Figura 37 Anotação semântica 162
Figura 38 Bag of Words de Risco Financeiro 167
Figura 39 Validação do léxico do domínio 169
Figura 40 Campo semântico do termo “bem” 173
Figura 41 Hierarquia de Significados 176

Lista de tabelas

Tabela 1 Formato tradicional 107
Tabela 2 Representação do Corpus 108
Tabela 3 Representação Generalizada do Corpus 108
Tabela 4 Representação do *corpus* em Código Binário 109
Tabela 5 Representação do *corpus* em Frequência 109
Tabela 6 Representação do *corpus* ponderado 110
Tabela 7 Alta Dimensionalidade — Documento por Termo 111
Tabela 8 Exemplo da Lei de Zipf 113
Tabela 9 Matriz Termo-Documento 116
Tabela 10 Tabela de contingência 2X2 129

Tabela 11 Busca sintática X semântica 172
Tabela 12 Avaliação de resultados 178

Lista de quadros

Quadro 1 – Resultado SPARQL 69
Quadro 2 – Comparação de alguns termos e expressões da CI e da WS 71
Quadro 3 – Extração de unidades de informação por ordem de complexidade 100
Quadro 4 – Objetivos específicos X Revisão de Literatura 135
Quadro 5 – Etiquetas da Floresta Sintá(c)tica 147
Quadro 6 – Abordagem de Lesk 155
Quadro 7 – Etapas de pré-processamento 166
Quadro 8 – Significados de artigo 175
Quadro 9 – Consultas da avaliação 177
Quadro 10 – Objetivos específicos X dificuldades encontradas 183
Quadro 11 – Objetivos específicos X resultados 184

Lista de Abreviaturas

ABox Assertional Knwoledge

API Application Programming Interface

ARPANET Advanced Research Projects Agency Network

BC Banco Central do Brasil

BC Base de Conhecimento

BIM Binary Independency Model

BoW Bag of Words

bs busca semântica

bt busca tradicional

CC Ciência da Computação

CERN European Nuclear Research Center

CI Ciência da Informação

DAML DARPA Agent Markup Language

DCR Data Category Registry

DL Description Logic

DOI Digital Object Identifier

DVS Decomposição de Valores Singulares

ER Entity Relationship

HTML Hyper Text Markup Language

http Hyper Text Transfer Protocol

IA Inteligência Artificial

IDF inverse document frequency

IMDB Internet Movie Data Base

ISBN International Standard Book Number

ISKO International Society for Knowledge Organization

ISL Indexação por Semântica Latente

ISSN International Standard Serial Number

KIM Knowledge Information Management

LD Lógica Descritiva

lemon Lexicon Model for Ontologies

LOD Linked Open Data Cloud

LSI Latent Semantic Indexing

MBWS Motor de Busca para Web Semântica

NLTK Natural Language ToolKit

OC Organização do Conhecimento

OIL Ontology Inference Layer

OMW Open Multilingual Wordnet

OWA Open World Assumption

OWL Web Ontology Language

PLN Processamento de Linguagem Natural

POS Part of Speech

PWN Pricenton WordNet

RDF Resource Description Framework

RDFS Resource Description Framework Schema

RI Recuperação da Informação

RIH Recuperação da Informação Híbrida

SKOS Simple Knowledge Organization System

SPARQL Simple Protocol and RDF Query Language

SQL Structured Query Language

SRI Sistemas de Recuperação da Informação

SVD Singular Value Decomposition

TBox Terminological Knowledge

TF term frequency

TF/IDF term frequency/inverse document frequency

UML Unified Modeling Language

UNA Unique Name assumption

URI Uniform Resource Indentifier

URL Uniform Resource Locator

URN Uniform Resource Name

W3C World Wilde Web Consortium

WeST Institute for Web Science and Technologies

WS Web Semântica

WWW World Wide Web

Sumário

Introdução 27

I Preparação da pesquisa 31

1 Definição do problema 33
1.1 Panorama . . . 33

2 Objetivos 37
2.1 Objetivo Geral . . . 37
2.2 Objetivos Específicos . . . 37

3 Justificativa 39

II Revisão de Literatura 43

4 Revisão de Literatura 45
4.1 Web Semântica . . . 45
4.1.1 Breve panorama histórico . . . 46
4.1.2 O significado da informação na Web . . . 47
4.1.3 Da web de documentos à web de dados . . . 50
4.1.4 Tecnologias da Web Semântica . . . 53
4.1.4.1 Uniform Resource Identifier . . . 54
4.1.4.2 Resource Description Framework . . . 55
4.1.4.3 Resource Description Framework Schema . . . 59
4.1.4.4 Web Ontology Language . . . 62
4.1.4.5 SPARQL . . . 68
4.1.5 Web Semântica na Ciência da Informação . . . 70
4.1.6 Considerações . . . 72
4.2 Ontologia . . . 72
4.2.1 Definição . . . 73
4.2.1.1 Conceituação . . . 75
4.2.1.2 Especificação formal e explícita . . . 76
4.2.1.3 Compartilhamento . . . 76
4.2.2 O conjunto das partes . . . 77
4.2.2.1 Tipos e categorias de ontologia . . . 78

4.2.3 Considerações . . . 79
4.2.4 O despertar na Ciência da Informação . . . 80
4.2.4.1 Considerações . . . 82
4.3 Linguagem natural nos domínios das Ontologias . . . 83
4.3.1 A linguagem em zeros e uns . . . 85
4.3.1.1 Morfologia . . . 87
4.3.1.2 Sintaxe . . . 88
4.3.1.3 Semântica . . . 88
4.3.1.4 Pragmática . . . 89
4.3.2 A Linguística na Ciência da Informação . . . 90
4.4 Processamento de Linguagem Natural . . . 91
4.4.1 Corpus como Recurso . . . 92
4.4.1.1 A descoberta de padrões no texto . . . 95
4.4.2 Conceitos fundamentais . . . 97
4.4.2.1 Ambiguidade . . . 97
4.4.2.2 Noções elementares . . . 98
4.4.3 Características de um Documento . . . 99
4.4.4 Processamento do Texto . . . 100
4.4.4.1 Coleta de Dados . . . 100
4.4.4.2 Padronização de Documentos . . . 101
4.4.4.3 Tokenization . . . 102
4.4.4.4 Padronização de Conteúdo . . . 102
4.4.4.5 Criação de Bases de Apoio . . . 103
4.4.4.6 Part-Of-Speech Tagging . . . 105
4.4.4.7 Lematização . . . 106
4.4.5 A Representação Quantitativa do Texto . . . 107
4.4.6 Redução de dimensionalidade . . . 111
4.4.6.1 Lei de Zipf . . . 112
4.4.6.2 Significância das Palavras de Luhn . . . 113
4.4.6.3 Decomposição de Valores Singulares . . . 115
4.4.7 Considerações . . . 118
4.5 A Recuperação da Informação . . . 119
4.5.1 Conceitos fundamentais . . . 120
4.5.2 Modelos de RI . . . 121
4.5.2.1 Modelo Booleano . . . 122
4.5.2.2 Modelo Vetorial . . . 123
4.5.2.3 Modelo Probabilístico . . . 124
4.5.3 Recuperação da Informação na Web Semântica . . . 125
4.5.3.1 A Web como um grafo . . . 126

4.5.3.2 Modelos Semânticos para Recuperação da Informação . . . 126
4.5.4 Avaliação em RI . . . 128
4.5.5 Estado da arte . . . 130
4.5.6 Considerações . . . 133
4.6 Considerações Finais acerca da Revisão de Literatura . . . 134

III Metodologia e Resultados 137

5 Procedimentos metodológicos 139
5.1 Tipo de pesquisa . . . 139
5.2 Método de Abordagem . . . 140
5.2.1 Fontes de pesquisa . . . 140
5.2.1.1 Busca da informação . . . 140
5.2.2 Parte I – Fundamentação teórica . . . 141
5.2.2.1 Levantamento bibliográfico . . . 141
5.2.3 Parte II – Desenvolvimento . . . 142
5.2.3.1 O *corpus* e a ontologia . . . 142
5.2.3.2 Recursos computacionais . . . 146
5.3 Abordagem para responder a questão I . . . 148
5.3.1 Modelo de lexicalização de ontologia . . . 149
5.3.2 Abordagem de lexicalização . . . 154
5.4 Abordagem para responder a questão II . . . 157
5.4.1 A RiscoLex e a RI . . . 158
5.4.2 Modelo de Recuperação da Informação Semântica . . . 159
5.4.2.1 Consulta . . . 160
5.4.2.2 Busca . . . 161
5.4.2.3 Indexação . . . 161
5.4.2.4 Ranking . . . 163
5.4.3 Avaliação . . . 164
5.5 Resultados e discussão . . . 164
5.5.1 Questão I . . . 164
5.5.2 Questão II . . . 169
5.5.3 Avaliação . . . 177

6 Conclusões e recomendações 181

Referências 187

Anexos 201

ANEXO A Ontologia do Risco Financeiro (OntoRisco) 203

ANEXO B Base Lexical (RiscoLex) 223

Introdução

A capacidade da sociedade contemporânea de armazenar, organizar, selecionar e compreender a informação disponível está inexoravelmente vinculada às tecnologias. Das estantes de bibliotecas aos incontáveis repositórios digitais espalhados pelo mundo, nunca foi tão difícil encontrar a informação adequada. Paradoxalmente, jamais houve tanta facilidade em acessá-la. Virtualmente não há barreiras, quase tudo está a um clique no mouse de computadores pessoais que vasculham fontes disponíveis na *World Wide Web*, ou simplesmente Web.

A combinação de técnicas e teorias possibilitam a interação do usuário com o universo de informação nesse ambiente. A mais visível é a Recuperação da Informação (RI) que se apresenta na forma de motores de busca, como os populares Yahoo! e Google. Menos aparente, a Organização do Conhecimento (OC) mantém-se no ambiente dos profissionais da informação. Com o desenvolvimento acelerado, nas últimas décadas, das tecnologias da informação e de redes, ambas, RI e OC, têm ocupado cada vez mais destaque no âmbito da Ciência da Informação (CI).

Tais assuntos, assim como a própria CI, são interdisciplinares e requerem também o suporte teórico e prático de áreas como Linguística, Estatística e Ciência da Computação (CC). Essa reunião de áreas tem revelado profissionais que trafegam nas tênues linhas que estabelecem regiões fronteiriças e abordagens que caracterizam as disciplinas científicas. Os dois temas, RI e OC, têm em comum o fato de serem permeados pela história da classificação que dispõe de arcabouço teórico que pode sustentar novas teorias.

Essa conjunção de teorias e tecnologias propicia o ferramental necessário à evolução dos sistemas de organização do conhecimento, dedicados à organização, ao gerenciamento, à disseminação e à recuperação da informação; uma constante preocupação humana, pois, desde os tempos remotos, o homem tem formulado esquemas que visam elaborar sistemas de organização do conhecimento, tais como classificações, dicionários especializados, glossários, tesauros e ontologias.

A revolução da Web propiciou a massificação do acesso à informação. A outra revolução, ainda em curso, é a da Web Semântica (WS) que se baseia em princípios de que a informação eletrônica será não ambígua, os dados serão facilmente encontrados, reutilizáveis e interoperáveis, bem como os dispositivos serão onipresentes. A ideia é trazer a ubiquidade da Web para o dia a dia dos usuários, com documentos enriquecidos com informações semânticas de conteúdos de páginas da Web, criando, assim, um ambiente em que agentes, na forma de programas de computadores, naveguem na rede, coletem informações e executem tarefas complexas para usuários.

No entanto, extrapolar a habilidade humana — trivial — de interpretação semântica de conteúdos registrados para o universo binário e digital de computadores tem se constituído em um problema de difícil solução para a comunidade científica, que tenta transpor os limites nos quais as máquinas estão encarceradas, isto é, no universo daquilo que é calculável. Apesar dos avanços, contudo, elas ainda são prisioneiras do paradigma em que foram criadas que são os procedimentos sistemáticos de cálculo.

Dessa forma, por mais ambicioso que sistemas complexos, tais como ontologias, possam almejar o processamento semântico da informação, as atuais tecnologias restringem-nos à capacidade dos computadores de executar somente o processamento sintático, ou seja, a busca de padrões. Nesse caso, a proposta inicial e diferenciadora das ontologias de interagir, tanto com homens quanto com máquinas, fica comprometida e a participação humana na elaboração, organização e indexação do conteúdo não pode ser prescindida.

A linguagem da lógica é um subconjunto da linguagem natural. A primeira é utilizada principalmente na ciência e expressa o conteúdo sem ambiguidades, mas possui limitações relacionadas ao seu alcance na captura do significado. A outra, utilizada no dia a dia para comunicação das pessoas, é completa e consegue exprimir quaisquer significados, inclusive aqueles que ainda não existem. Nesse caso, criam-se expressões e termos, sempre que necessário, para denotar um conceito. Entretanto, ela é vaga, imprevisível e ambígua e essas características são determinantes para que a linguagem natural seja flexível e robusta.

Para que aplicações de Processamento de Linguagem Natural se beneficiem da precisão das ontologias e estabeleçam a ponte entre o conteúdo e o significado, a utilização de bases de dados léxicas e terminológicas são cruciais. Todavia, elas se encontram em "silos" de dados orientados a aplicações com formatos e interfaces específicas e, por causa disso, o acesso a elas é difícil.

Nesse contexto, particularmente no da iniciativa do *Linked Data* que é de conectar dados distribuidos na Web, são fornecidas soluções efetivas para a reutilização de dados vinculados por meio de *Uniform Resource Indentifier*[2] (URI). Contudo, bases de dados com essas características frequentemente carecem de informações sobre como os elementos ontológicos, tais como classes e propriedades, são percebidos morfológica e sintaticamente. Os rótulos (*labels*) utilizados para descrever essas URIs não possuem informações linguísticas como formas morfológicas, argumentos sintáticos ou variantes léxicas. Tais informações são decisivas para que se possa estabelecer ou restringir as interpretações de um léxico em face de um dado elemento ontológico.

Além disso, no espírito de representar o conhecimento de acordo com o formalismo exigido pela WS, utilizar esses *links* propicia a reutilização de léxicos relacionados, bem como a participação da comunidade na expansão desses léxicos de forma natural, i. e., em linguagem natural. Para tanto, pressupõe-se a adoção de padrões para a representação de

[2] é uma cadeia de caracteres compacta usada para identificar ou denominar um recurso na Internet. Fonte: http://pt.wikipedia.org/wiki/URI

ontologias e léxicos relacionados em formatos acessíveis às máquinas.

Assim, parece natural que a exploração dos recursos e tecnologias disponíveis seja o caminho para promover o equilíbrio entre elementos léxicos e ontológicos. Em particular, enquanto existem muitos recursos terminológicos, eles raramente são providos de informações semânticas para viabilizar a utilização de regras de inferências na WS. Da mesma forma, os recursos semânticos estão desprovidos de informações morfossintáticas.

Por conseguinte, considera-se que a representação léxica associada aos termos relacionados às ontologias seja vital para o desenvolvimento da WS. A idealização de um modelo léxico-ontológico contribuirá com a CI, haja vista que ele, entre outras características, incrementa tanto os atributos semânticos que adicionam expressividade às ontologias, quanto a capacidade de o usuário interagir na construção de definições de determinado domínio.

Esse é o núcleo desta pesquisa que busca representar a linguagem natural na forma adequada às ontologias e vice-versa. Ao incluir recursos linguísticos em um sistema de tratamento de linguagem natural pode-se proporcionar melhor interação com o usuário e melhoria na qualidade dos sistemas de recuperação da informação. Para tanto, pode-se valer dos recursos e tecnologias disponíveis para adaptá-los à língua portuguesa e propor um modelo que atuará de acordo com os princípios da Web Semântica.

A consecução deste trabalho, aliada a outras iniciativas de elaboração semiautomática de ontologias para o Português, culminará na proposição de novas metodologias e suporte teórico. Logo, cientistas da informação melhor aproveitariam a precisão e a velocidade dos métodos computacionais, na difícil tarefa de organizar o mundo sob a perspectiva da informação, de forma que ela esteja preparada em partes processáveis e, sem perder de vista o objetivo maior da CI, adequada ao usuário.

Parte I

Preparação da pesquisa

1 Definição do problema

1.1 Panorama

A explosão da informação e suas consequências são temas recorrentes na CI como, por exemplo, em Marcondes & Campos (2008) que afirmam que nunca foi tão difícil recuperar informação relevante, apesar da enorme disponibilidade. Em contrapartida, instituições gastam grandes somas para gerir o legado de informação da era pré-digital e para acompanhar o crescimento de acervos digitais.

Entretanto, após décadas, profissionais da área ainda não resolveram o problema da organização e da representação de informação para atender ao usuário satisfatoriamente. Várias iniciativas buscam soluções viáveis tanto econômica quanto tecnologicamente e, nesse sentido, uma área de pesquisa que tem atraído a atenção de pesquisadores está voltada para as ontologias.

As ontologias têm ocupado lugar de destaque em publicações da área, como se comprova em Gruber (1993a), Guarino (1997), Brewster et al. (2005), Lima-Marques (2006), Vital & Café (2011) e Schiessl & Bräscher (2012). Especificamente para a CI, Marcondes & Campos (2008) declaram que as ontologias têm recebido atenção especial por parte da comunidade, mas, em geral, as propostas de linguagens e ferramentas ainda não alcançaram um nível satisfatório para contribuir efetivamente na orientação do usuário em relação ao processo de construção de ontologias.

Por outro lado, a construção e a manutenção de ontologias mostram-se mais difíceis do que se poderia imaginar. Essas atividades requerem mão de obra especializada — rara e cara — que executa tarefas complexas tais como: definir e relacionar conceitos; delimitar domínios e criar regras de inferência. Conforme Maedche & Staab (2002b), Shamsfard & Barforoush (2003) e Shamsfard & Barforoush (2004), uma alternativa para superar essas dificuldades é a construção colaborativa e reuso da informação publicada, o que converge para os interesses da comunidade da WS.

Assim, desde o artigo de Berners-Lee, Hendler & Lassila (2001) que anunciava o mundo novo da WS, a ideia de utilizar a Web para vincular dados que até então não possuíam ligação, ou para diminuir as barreiras para vincular dados que estavam conectados por outros métodos, vem se consolidando como opção promissora para fornecer soluções às demandas por informação qualificada. Esse conjunto de dados vinculados na web é conhecido como *Linked Data*. Conforme a *World Wilde Web Consortium*[1] (W3C), o *Linked Data* está no

[1] http://www.w3.org/

fundamento da WS, pois trata da integração em larga escala e das inferências sobre dados disponíveis na Web.

De acordo com Heath & Bizer (2011), o *Linked Data* promove uma mudança de paradigma na qual, além de documentos ou páginas em *Hyper Text Markup Language* (HTML), podem ser publicados dados em formato padrão e aberto para estender a atual Web. Esses dados seriam relacionados entre si como forma de integração em escala global para possibilitar a criação da Web de dados. Assim, na última década, a comunidade científica tem construído a *Linked Open Data Cloud*[2] (LOD) que consiste de expressiva quantidade de dados em *Resource Description Framework* (RDF) — utilizado como um método geral para descrever ou modelar conceitualmente uma informação da web na forma de triplas (sujeito, propriedade, objeto) — que estão interligados por meio de um *Uniform Resource Identifier* (URI) — sequência de caracteres que permite a identificação única de recursos — e, de tal modo, possibilita o compartilhamento e a restrição de interpretações semânticas.

O reflexo desse esforço é o crescimento da LOD que já ultrapassa a quantidade de 30 bilhões de triplas[3] de dados em RDF publicados na Web. A ideia é que não se invente conteúdo, mas que se conecte esses conteúdos existentes independente de formato, espaço e tempo. Bases de conhecimento como *Internet Movie Data Base*[4] (IMDB) ou DBpedia[5], correspondente à Wikipedia[6] em RDF, são amplamente utilizadas em pesquisas para várias aplicações tais como, geração de linguagem natural, refinamento de buscas na web, interoperabilidade entre domínios etc. Gigantes como o Google[7] já começam a utilizar esses dados estruturados para proporcionar a busca semântica e melhorar os resultados.

Contudo, do ponto de vista prático, o usuário padrão desconhece o mundo estruturado da Web Semântica, que ainda permanece em grande parte inacessível, pois a maneira intuitiva de se acessar a informação relevante é por meio da linguagem natural, isto é, dados não estruturados segundo a ótica computacional. Nesse domínio, a CI busca soluções no campo da recuperação de informação que, de acordo com Bräscher (1999), envolve tanto operações para a adequação de documentos a serem recuperados, quanto para formulação de consultas de usuários a serem submetidas aos sistemas de recuperação da informação.

Rijsbergen (1979) declara que o armazenamento e a recuperação da informação, em princípio, são simples. Um usuário lê todos os documentos de um arquivo (armazenamento) e extrai (recupera) aqueles que satisfazem sua necessidade, isto é, que respondem à pergunta formulada por ele. Solução pouco prática para o usuário há trinta anos e, obviamente, agravada nos dias de hoje pela quantidade de documentos eletrônicos disponíveis, o que indica o caminho da automação como uma possível saída. Entretanto, replicar em máquinas o comportamento humano na leitura e na caracterização de documentos envolve aspectos

[2] http://linkeddata.org/
[3] http://www4.wiwiss.fu-berlin.de/lodcloud/state/
[4] http://www.imdb.com/
[5] http://dbpedia.org/About — representação em RDF de parte da Wikipedia.
[6] http://www.wikipedia.org/
[7] http://www.google.com/insidesearch/features/search/knowledge.html

subjetivos de inferência sobre a sintática e a semântica de textos para decidir sua relevância no processo.

A tarefa de organizar e representar a informação remete à subjetividade que abrange o processo de indexação, pois depende da experiência e da expertise dos profissionais envolvidos. Isso, aliado ao aumento exponencial de informações que um sistema deverá abranger e à rapidez necessária para tratar essas informações que se impõe aos profissionais, para dar vazão à demanda dos usuários, reflete na qualidade dos sistemas de recuperação de informação. Portanto, volume e velocidade são variáveis importantes no processo e, nesse sentido, a automação deve ser considerada.

Além disso, Bräscher (1999) afirma que a recuperação da informação vai muito além da verificação de uma sequência de caracteres, que representam os documentos armazenados, que coincida com a consulta do usuário, isto é, não é mero exercício de extração de palavras do texto, mas de relacionar o significado atribuído a elas. Por isso, os sistemas devem estar preparados para tratar fenômenos linguísticos que interferem na qualidade da recuperação. Portanto, deve se incorporar um componente de tratamento da linguagem que abranja os componentes morfológicos, lexicais, sintáticos, semânticos e pragmáticos.

Apesar da necessidade de inserção de componentes linguísticos nos sistemas de recuperação da informação, Baeza-Yates & Ribeiro-Neto (1999) ressaltam que os motores de busca ainda utilizavam técnicas estatísticas aplicadas ao conteúdo léxico dos textos, com critérios baseados no número e popularidade de indicações. Reymonet, Thomas & Aussenac-Gilles (2007) afirmam que essas ferramentas ainda são essencialmente as mesmas e que dessa forma não é possível descrever a informação no nível semântico. Atualmente, há algumas iniciativas que lidam com o problema, mas ainda fora do alcance do usuário comum.

A despeito da promessa da WS de estabelecer uma parceria entre pessoas e máquinas, Wilks & Brewster (2009) ressaltam que a representação do conhecimento — ontológica neste caso — deve estar vinculada a alguma LN para ser justificada. Acrescentam ainda que uma língua é um sistema de eventos raros, mas é um modelo completo. Por isso, atribuem a Spärck Jones[8] a afirmação de que "as palavras são autorrepresentativas e nada mais, nenhum outro símbolo pode substituí-las ou codificá-las com significados equivalentes". Charniak (1973) e Wilks (1977) corroboram com essa afirmação em formas diferentes, mas expressam que as palavras retêm informação essencial que não está presente em nenhuma outra representação.

Fica evidente que o mundo da WS e da LN precisam ser conciliados. Para utilizar o conhecimento nessas representações de conhecimento é necessária a existência de uma ponte entre os elementos de uma ontologia — classes, propriedades e indivíduos — e seus correspondentes em linguagem natural. Logo, para se capturar informações linguisticamente ricas sobre verbalizações de elementos simples ou complexos de uma ontologia, é necessário o conhecimento lexical, isto é, o conhecimento do conjunto de palavras acerca do domínio de

[8]Pesquisadora da Universidade de Cambridge com interesse em PLN e RI que trouxe grande contribuição com o conceito de *inverse document frequency* (IDF). http://en.wikipedia.org/wiki/Karen_Sp%C3%A4rck_Jones

interesse. Além disso, esse conhecimento deve estar acessível às máquinas e publicado de forma a facilitar seu reuso.

A configuração do cenário atual caracteriza-se, portanto, pela enorme quantidade de bases de dados estruturadas e textuais, mas separadas pelos formatos em que são publicadas para o consumo. Isso tem formado "silos" de dados sem comunicação uns com os outros, o que dificulta a utilização plena do conteúdo dessas bases. Os atuais sistemas de recuperação evoluíram de busca puramente sintática para outras formas que consideram aspectos semânticos na recuperação de documentos, mas com sérias restrições devido à falta de padrões que uniformizem a correspondência entre a LN e as estruturas ontológicas da WS.

O estabelecimento efetivo dessa ponte entre os dois mundos possibilitaria que consultas submetidas em LN pudessem buscar a semântica disponível na WS e fornecer alternativas para tratar um problema central de interesse da CI: a ambiguidade. Manning, Raghavan & Schütze (2009) afirmam que a ambiguidade de um termo facilmente introduz outros termos correlacionados que não são estatisticamente significantes e, por isso, interfere na qualidade dos sistemas de recuperação da informação. Logo, a elaboração de um modelo de sistema de recuperação da informação que inclua a lexicalização de ontologias, isto é, a vinculação dos elementos ontológicos com a LN, para tratamento da ambiguidade contribuirá tanto no campo teórico quanto no pragmático: no primeiro por fornecer uma metodologia e, no segundo, por facilitar o acesso às ontologias e, assim, propiciar a semântica necessária para organizar os crescentes repositórios de documentos digitais e facilitar a recuperação da informação.

De tal forma, as questões que motivam essa pesquisa são:

> **Questao I**: É possível construir automaticamente, por meio de processamento automático da linguagem natural, uma base de léxicos para o Português brasileiro, a partir de ontologia e de *corpus* vinculados, que possa ser lida por computadores e na qual se explicitem os componentes morfológicos, sintáticos e semânticos?
>
> **Questão II**: É possível melhorar a precisão[9] em sistemas de recuperação da informação em Português brasileiro com o uso dessa base de léxicos para desambiguação?

[9]Manning, Raghavan & Schütze (2009) definem precisão como a fração de documentos relevantes recuperados que satisfazem a necessidade de informação usuário. Para tanto, um documento é relevante se é aquele que o usuário percebe como portador da informação que atenda às suas necessidades pessoais.

2 Objetivos

2.1 Objetivo Geral

Criar automaticamente uma base de léxicos em Português brasileiro e no domínio de risco financeiro, contendo informações morfológicas, sintáticas e semânticas apropriadas para leitura por máquinas, permitindo a vinculação entre dados estruturados e não estruturados e integrá-la em um modelo de recuperação da informação com o objetivo de aumentar a precisão em relação aos modelos clássicos que utilizam somente palavras-chave.

2.2 Objetivos Específicos

1. Identificar tecnologias da Web Semântica apropriadas para a representação do conhecimento em determinado domínio;
2. Produzir ontologia para o domínio de risco financeiro;
3. Gerar uma base de léxicos em Português brasileiro que contenha os aspectos morfológicos, sintáticos e semânticos dos itens lexicais vinculados aos elementos ontológicos da indústria financeira;
4. Propor um modelo de recuperação de informação que utilize automaticamente a base de léxicos como forma de tratar a ambiguidade;
5. Testar o modelo e mensurar a precisão para comparar os resultados entre o modelo aqui proposto e um modelo clássico de recuperação da informação no domínio de risco financeiro.

3 Justificativa

O termo Recuperação da Informação pode ter significados muito amplos, porém, neste trabalho, se adota a seguinte definição:

> Recuperação da informação é encontrar material (usualmente documentos) de natureza não estruturada (usualmente textos) que satisfaça uma necessidade de informação dentro de grandes coleções (usualmente armazenadas em computadores).(MANNING; RAGHAVAN; SCHÜTZE, 2009, pag. 1. Tradução nossa.)[1]

Embora, atualmente, o termo recuperação da informação possa, por exemplo, remeter à busca de informações na Web por meio de motores de busca como Yahoo!, Google etc., o termo também abrange tarefas cotidianas como a busca de documentos em computadores pessoais ou nas caixas de entrada de e-mails. Portanto, a recuperação de documentos textuais pressupõe que um usuário interaja em linguagem natural com um sistema que seja capaz de identificar, no repositório, aqueles que sejam de interesse do usuário e disponibilizá-los em formato compreensível.

A adoção da linguagem natural como premissa básica para interface com qualquer sistema de tratamento da linguagem e de RI permite a interação direta e, portanto, a elaboração de perguntas que reflitam mais precisamente a necessidade do usuário. Contudo, isso influencia na complexidade e na característica de representação do conteúdo de documentos que está ligada à qualidade da recuperação (BRÄSCHER, 1999).

Atualmente, o texto é o recurso mais abundante em ambientes corporativos e na Web. Pessoas utilizam o texto como forma de comunicação e expressão de ideias de maneira intuitiva e corriqueira. Em consonância com essa visão, para Dahlberg (1978b), o ser humano emprega palavras como forma de tradução das idealizações de objetos do mundo, real ou abstrato. Dessa forma, utiliza a linguagem para relacionar os objetos aos conceitos. De fato, Cimiano, Völker & Buitelaar (2010) e Buitelaar et al. (2011) corroboram essa ideia ao afirmar que a linguagem humana é um modo primário de transferência de conhecimento e comunicação entre humanos.

Os sistemas que usualmente operam no modo sintático da linguagem são incapazes de detectar incoerências semânticas como nas frases “Eu como tijolo.” ou “O automóvel falou.”, mas que sintaticamente são perfeitas. Ainda, a LN está inexoravelmente ligada a ambiguidade como na frase “Ele está próximo ao banco”. Banco onde se senta ou onde se movimenta dinheiro? Palavras possuem múltiplos significados e a gramática também pode ser ambígua na

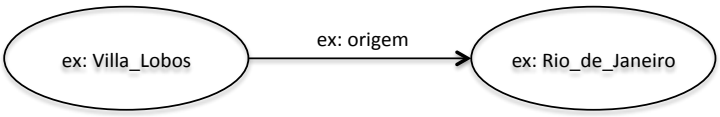
[1] *Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).*

estrutura e, por isso, na interpretação do significado. O exemplo "O homem viu a criança de binóculos" pode ser interpretado como o homem vendo a criança por meio de binóculos ou a criança portando um desses instrumentos. Esses exemplos, embora facilmente compreendidos por pessoas em quase todas as situações, não são tarefa trivial para sua decodificação em computadores. Isso pode acarretar dificuldades de definição de interpretações precisas em domínios técnicos.

De acordo com Guarino, Oberle & Staab (2009), a semântica, no contexto da WS, quer dizer significado. Isso possibilita o uso mais efetivo de dados subjacentes, pois ao leitor humano é deixada a tarefa de interpretar as lacunas e relacionamentos presentes nos textos. Frequentemente, as fontes disponíveis apresentam apenas palavras-chaves para serem visíveis aos motores de busca. O que pode ser visto como uma semântica limitada. Contudo, se as palavras-chaves estiverem relacionadas a outras com vínculos definidos, o contexto se forma e revela a semântica. Por exemplo, a palavra "banco" isoladamente é ambígua, mas vinculada a outras como "agência", "caixa eletrônico", "saque", "depósito" insere-se no contexto de "instituição financeira" e revela sua semântica.

Buitelaar et al. (2011) afirmam que um texto pode ser traduzido para formatos semanticamente estruturados como o RDF. Isso permite que esse dado estruturado possa ser representado em qualquer linguagem e manipulável por máquinas. Textos simples, isto é, descritivos com frases diretas, curtas e na voz ativa, podem ser traduzidos para uma ontologia com auxílio de algoritmos simples. Por exemplo: "Villa-Lobos nasceu no Rio de Janeiro" seria transformado para a tripla <sujeito, propriedade, objeto> [2] em RDF que está graficamente representado na Figura 1.

Figura 1: Representação gráfica de tripla



Fonte: Elaborado pelo Autor

Entretanto, o texto "O famoso carioca, reconhecido internacionalmente, é o compositor de obras como «12 estudos para violão» e «O trenzinho do caipira»" é uma forma mais complexa na qual se pode inferir a mesma coisa, dentre outras. Para tanto, o analista deve concluir que "carioca" é o atributo de quem nasce na cidade do Rio de Janeiro e que o autor destas

[2] Um exemplo real seria:

```
<?xml version="1.0"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
            xmlns:ex="http://www.example.org/" >
  <rdf:Description rdf:about="http://www.example.org/Heitor_Villa-Lobos">
    <ex:origem rdf:resource="http://www.example.org/Rio_de_Janeiro"/>
  </rdf:Description>
</rdf:RDF>
```

obras é Villa-Lobos. Assim, esse exemplo requer a representação de informações linguísticas para que sejam computadas as variantes sintáticas e morfológicas para viabilizar inferências semânticas e interpretação do significado.

De acordo com Guarino (1998), ontologias capturam o conhecimento, mas falham na captura da estrutura e uso de termos que são objetos da Terminologia e da Lexicologia. A estrutura e o uso de termos são fundamentais para expressar e se referir ao mesmo conhecimento em Linguagem Natural. Paradoxalmente, pesquisadores têm dado menos atenção ao tratamento adequado das questões relacionadas ao léxico e à linguística nas áreas de organização do conhecimento e de recuperação da informação.

Atualmente, não há como desprezar o volume de bases semanticamente estruturadas existentes e a possibilidade de reuso desses recursos. Paralelamente, a quantidade de recursos textuais que podem servir de insumo para enriquecer essas bases de conhecimento ou criar domínios específicos de acordo com a necessidade de usuários ou utilizá-las como resolução semântica parece ser um caminho promissor para o tratamento automático da ambiguidade.

Logo, a solução desse problema demanda um modelo formal de representação de conhecimento que abarque a semântica da ontologia, a terminologia[3] — que é usada para expressar esse conhecimento em LN — e informações linguísticas sobre termos e suas unidades lexicais. Tal modelo proporciona a participação de máquinas no processo de tradução e inferência. Portanto, além dos níveis semântico e terminológico, argui-se que a representação do nível léxical é necessária para a utilização adequada de ontologias no tratamento da linguagem e como forma de integração dos níveis terminológico e ontológico.

Mais especificamente, este trabalho busca ofertar alternativas para melhorar a representação e o acesso às informações financeiras dada a relevância econômica e social da indústria financeira. Os pilares desse setor são baseados em criar incentivos a investimentos que visem aumentar a produtividade; estimular conexões no comércio e negócios com vistas a facilitar a transferência de tecnologias e melhoria de uso dos recursos disponíveis; proporcionar acesso amplo aos ativos e mercados, visando construir uma base de ativos dos mais pobres e aumentar o retorno de tais ativos; reduzir o risco e a vulnerabilidade financeiras, possibilitando a inserção dos mais pobres — população não bancarizada — e o benefício em função disso, nesse setor.

Portanto, do ponto de vista da organização da informação, o desdobramento da crise mundial de 2008, que afetou as principais economias, incluindo o Brasil, demandou uma resposta no tratamento de informações do setor financeiro, que sofreu fortemente com a falta de padronização de formatos para a troca de informações entre entidades financeiras. O acordo de Basileia III[4], iniciado em 2010, estabeleceu regras para o tratamento científico do risco bancário, que inclui a criação de bases e a viabilização de transferência de conhecimento

[3] A acepção do termo aqui refere-se ao conjunto de termos utilizados em um domínio técnico ou científico.

[4] Os Acordos de Basileia III ou simplesmente Basileia III referem-se a um conjunto de propostas de reforma da regulamentação bancária. Fonte: http://pt.wikipedia.org/wiki/Basileia_III

entre as instituições. Tais regras devem ser adotadas para assegurar a estabilidade na estrutura de sistemas financeiros.

Assim, essa pesquisa se justifica i) pela importância do tema que versa sobre tratamento e recuperação da informação na CI; ii) pela relevância da WS na CI, uma vez que grande parte das pesquisas está concentrada na Ciência da Computação; iii) por possibilitar mais opções de recursos em Português brasileiro para futuras pesquisas que abordem o tema; iv) pela relevância do tema para sociedade, haja vista que se busca melhorar a qualidade da recuperação da informação; v) por iniciar uma base de conhecimento computável para Português brasileiro para o Risco Financeiro, que é um importante subconjunto do Sistema Financeiro e tem grande importância e impacto para a sociedade em geral.

Parte II

Revisão de Literatura

4 Revisão de Literatura

Este capítulo fornece o aporte teórico para fundamentar a pesquisa com base nos trabalhos de clássicos e nos recentes avanços da área. A cobertura da bibliografia consultada abrange cinco partes. A primeira investiga a Web Semântica compreendendo seu desenvolvimento histórico, definições e interpretação e relação com a Ciência da Informação. A segunda apresenta noções elementares de Linguística e a interação da linguagem natural com estruturas ontológicas. A terceira trata do tema ontologias com foco nos aspectos conceituais. A quarta aborda o Processamento da Linguagem Natural e suas principais técnicas. A última cobre a Recuperação da Informação abarcando um pouco de sua história, principais modelos e interações com outras áreas de interesse para a pesquisa.

4.1 Web Semântica

Imaginando a Web como um ambiente no qual se armazenam e se recuperam documentos, quando se digita um endereço, ou, mais comumente, *Uniform Resource Locator* (URL) em um *browser* , ou navegador, solicita-se um documento a um servidor. Este localiza o endereço e retorna o que foi solicitado em HTML. Esta linguagem define a sintaxe que diz ao computador como as informações contidas no documento são apresentadas ao usuário.

Assim, tem-se a *World Wide Web* (WWW) que possui servidores de documentos, protocolos de comunicação, rede de computadores, navegadores e buscadores que recuperam documentos para os usuários. Toda essa arquitetura funciona bem e tem atendido a maioria dos desejos do usuário comum. A deficiência é que a arquitetura descrita se mantém restrita ao plano da sintaxe e não é capaz de ir além disso, como, por exemplo, a semântica.

Portanto, o objetivo da Web Semântica, composta por várias camadas de metadados, lógica e segurança é habilitar a transferência de significados entre agentes computadorizados para que “entendam” o significado contido nas páginas da Web. Para tal, existe um esforço em transformar a Web de documentos — WWW — para a Web de dados — WS. Nessa, qualquer conceito pode ser um recurso identificado semanticamente e ter interação com máquinas. A passividade das solicitações feitas por usuários seria substituída por um comportamento ativo de agentes computadorizados que trafegam pela rede organizando e coletando informações que auxiliem a tarefa dos usuários.

Para tal, novas tecnologias têm sido implementadas para prover a semântica aos documentos. Isso possibilitaria que eventos corriqueiros, como uma festa de aniversário agendada com tais tecnologias, pudesse interagir ativamente com a agenda pessoal e dispositivo de GPS de um usuário para que promovesse a melhor opção, tais como lembretes para não esquecer,

melhores trajetos para o local e informação a respeito do evento como sugestão de presentes de acordo com a idade e o gosto do aniversariante.

4.1.1 Breve panorama histórico

A *Advanced Research Projects Agency Network* (ARPANET) é considerada a precursora da Internet. Ela se iniciou em 1969 pelo Departamento de Defesa dos EUA e, portanto, com fins militares. A ARPANET consistia em uma rede operacional de computadores para troca de pacotes. Nessa arquitetura, os militares disseminavam informações estratégicas e de pesquisas do governo em núcleos remotos com objetivo de eliminar a concentração de dados em um único local e assegurar que a informação não pudesse ser destruída em caso de caso de ataque nos tempos da Guerra Fria.

A parte da ARPANET dedicada à pesquisa se transformou na Internet que se conhece hoje. Ela é considerada como a primeira geração da Internet. Essa consistia na conexão a computadores remotos para baixar arquivos e somente então era possível lê-los no computador local. A característica dessa primeira geração é o processamento centrado em computador. Além disso, ela era apenas para especialistas, pois o acesso à informação era complexo e caro.

Somente em 1990, dois pesquisadores do *European Nuclear Research Center* (CERN), Tim Berners-Lee e Robert Cailliau, propuseram uma rede baseada em hipertextos para acessar diversos tipos de informações por meio de ligações que se assemelhariam a uma teia de nós, na qual o usuário poderia navegar de acordo com sua necessidade de informação. Essa teia ficou conhecida como *World Wide Web* ou simplesmente Web — teia em Inglês. A web se caracteriza pelo processamento centrado no documento. Por isso, a necessidade agora é um sistema para documentar tudo e gerenciar esse acervo. O acesso à informação é facilitado pela utilização de navegador para carregar o documento e pelo uso de *hyperlink* para vinculá-los.

Essa breve história circunscrita ao século XX e ao mundo anglo-saxão é a mais comum na literatura, principalmente na CC, mas a ideia original parece extrapolar esses limites e ser muito mais antiga. No século XVIII, os franceses Denis Diderot (1713 — 1784) e Jean-Baptiste le Rond d'Alembert (1717 — 1783) vislumbraram a construção de um conjunto de 28 volumes contendo 71.818 artigos e 3.129 ilustrações, cujo objetivo era reunir todo o conhecimento do mundo em uma Enciclopédia[1] . Além disso, os artigos possuíam *links* entre si e, assim, a ideia se constitui, em sua essência, muito similar à Web atual.

Santos (2008) relata que Paul Otlet, em 1934, publica a obra *Traité de Documentacion: le livre sur le livre: théorie et pratique* que antecede muitas das ideias reconhecidas sobre a organização de redes internacionais de cooperação para tratamento e troca de informações documentadas. As proposições como Princípio Monográfico, Classificação Decimal Universal

[1] http://en.wikipedia.org/wiki/Encyclop%C3%A9die

e a utilização de fichas padronizadas pressupõe um projeto de cunho universalista. A ideia era que se criasse um livro universal do conhecimento que nunca estaria completo, mas sempre em crescimento semelhante ao conhecimento humano. Rayward (2002), corroborado por Robredo (2012), ressalta a importância da obra de Otlet e alcunha esse livro universal de "Internet de papel" pela semelhança com a Internet[2] atual.

Bush (1945) propõe o primeiro sistema baseado em hipertextos que foi batizado de Memex. Ele imaginava esse sistema como uma forma de extensão da memória em que livros, registros e comunicações na forma de microfilmes seriam armazenados. Além disso, o processo de consulta a essas informações seria mecanizado e, portanto, aumentaria a velocidade e flexibilidade do processo. Ao autor é atribuída a previsão da explosão da informação em função da crescente produção científica após a II Guerra Mundial e os problemas que adviriam desse evento.

Finalmente, o sucesso da web atual deve muito ao trabalho desses pioneiros e sua popularização é atribuída ao desenvolvimento de tecnologias para auxiliar o usuário, como o uso do navegador que tornou o acesso à informação muito mais fácil, isto é, a um clique do mouse. Convencionou-se que a Web da década de 90 seria a Web 1.0. Essa se destinava aos profissionais da área. Em seguida, a Web 2.0, a partir de 2000, permitiu que pessoas comuns pudessem ser capazes de publicar e acessar informações na Web. Dessa popularização, tem-se, finalmente, um livro global e caótico de textos, imagens etc. Mas e a Web 3.0?

4.1.2 O significado da informação na Web

Para entender a Web 3.0 ou Web Semântica, é necessário responder a questão: o que é o significado? Allemang & Hendler (2008) alertam que semântica se relaciona com a compreensão da natureza do significado, mas o próprio termo tem diversos significados. De forma geral, ele está relacionado ao termo sintaxe. Intuitivamente, nas linguagens, sintaxe se refere ao "como" algo é dito, enquanto semântica é o significado daquilo que é expresso. Analisando a frase: "Eu amo CI!", tem-se a sintaxe regulando a função de todas as palavras e sinais, enquanto a semântica expressa o significado do que a frase realmente quer dizer. Podemos mudar a sintaxe reescrevendo a frase: "Eu ♡ CI", mas a semântica permanece a mesma.

A referência à sintaxe e à semântica está relacionada à comunicação da mensagem que se quer transferir. Pessoas fazem isso naturalmente utilizando a escrita, a fala, imagens ou equivalentes. Da mesma forma, na web, os conteúdos são adequados para o consumo humano, como na Figura 2:

Contudo, a menos que se saiba Vietnamita, é impossível inferir sobre o que se trata a mensagem. Essa é a mesma dificuldade para a máquina em capturar o significado em

[2]Apesar do uso frequente dos termos "Internet" e "web" indistintamente na literatura, convém ressaltar que o termo "Internet" refere-se à rede mundial de computadores ou uma rede de redes, no sentido de infraestrutura em rede. Enquanto a "web" refere-se ao modelo de compartilhamento de informações construído sobre a Internet.

Figura 2: Qual o significado da mensagem?

Fonte: Elaborado pelo Autor

relação a um texto, figura ou vídeo. Saber o que estes símbolos representam é a chave para decodificar o conteúdo, apreender o significado e conectar imagens e textos.

Pessoas possuem conhecimento contextual, conhecimento do mundo e experiência para resolver o problema. Por outro lado, entre máquinas, a comunicação é feita por formas artificiais e padronizadas que foram criadas para tal propósito. A web se baseia na linguagem de marcação *HyperText Markup Language* (HTML) que não consegue explicar o que a informação realmente significa. A consequência é que máquinas apenas manipulam sintaxe de forma que informações sejam trocadas entre elas, mas são incapazes de compreender o significado dessas mensagens.

Para responder o que é o significado, há que se perceber que ele não somente depende da sintaxe e da semântica, mas também do contexto, da pragmática e da experiência. Esses conceitos são descritos aqui apenas o suficiente para contextualizar a importância desses na Web Semântica. As definições são a reunião de anotações de discussões com o Grupo de Pesquisa do *Institute for Institute for Web Science and Technologies* (WeST)[3] e formalizadas de acordo com Fiorin (2002) e o dicionário online Priberam[4].

O termo sintaxe vem do Grego *súntaksis* e significa arranjo ou ordenação. Em Linguística, a sintaxe constitui o estudo de princípios e processos pelos quais as sentenças são construídas em determinada língua, isto é, como os itens lexicais se estruturam em uma sentença. Em linguagens formais, como a Lógica, ou nas linguagens de programação, refere-se ao conjunto de regras pelas quais expressões bem formadas, ou válidas, podem ser criadas a partir de um conjunto fundamental de símbolos. Em CI, a sintaxe está relacionada à estrutura normativa dos dados, pois ela define as regras de como expressões podem ser formadas, e quais são válidas. Isso é essencial para que se possa interpretar corretamente essas estruturas e, portanto, a correção das informações necessita, prioritariamente, da verificação da sintaxe para a determinar a validade das declarações ou sentenças.

O estudo do significado é função da semântica, palavra *sēmantiká* de origem grega. É a parte da linguística que se ocupa do sentido e do significado de palavras ou de símbolos. Na lógica, estuda as relações entre signos e respectivos referentes. Ela investiga a interpretação

[3]http://www.uni-koblenz-landau.de/campus-koblenz/fb4/west/staff — as anotações são, em grande parte, fruto de discussões com meu orientador Steffen Staab (supervisor) e conselheiro Matthias Thimm (advisor) no período de um ano referente ao estágio de doutoramento (sanduíche) na Universidade Koblenz-Landau, Alemanha.

[4]http://www.priberam.pt/Produtos/Dicionario.aspx

de signos ou símbolos utilizados por agentes ou comunidades dentro de circunstâncias e contextos particulares. Um importante aspecto da semântica é investigar como o significado de conceitos complexos podem ser derivados de conceitos simples a partir de regras da sintaxe. Em lógica, por exemplo, a semântica pode ser derivada de regras sintáticas pela demonstração da validade dessas regras para todas as possibilidades. Dessa forma, se relaciona diretamente ao contexto e à pragmática.

O termo contexto, que vem do Latim *contextus*, remete à noção de conjunto, combinação, mistura, tessitura ou entrelaçamento. Denota a maneira pela qual as ideias são concatenadas no discurso, ou seja, a vizinhança de um símbolo ou conceito em uma dada situação. Em Linguística, compreende o conjunto de elementos linguísticos à volta de som, palavra, locução, construção, frase, parte de discurso, etc. Ele é representado por todos os elementos de qualquer tipo de comunicação que define a interpretação do conteúdo comunicado.

Nesse sentido, ao contexto, é atribuída a função de preencher as lacunas do texto deixadas entre autor e leitor. Tudo que possa ser utilizado para entender o conteúdo faz parte do contexto. Ele ainda se divide em duas partes: i) geral, que se refere ao local ou tempo no qual a mensagem foi elaborada; ii) pessoal ou social, que se refere à relação entre emissor e receptor de uma mensagem. Logo, o significado pode ser determinado por essa relação, pois a mensagem pode ser compreendida de diferentes formas, conforme essa relação. Todos os aspectos do contexto devem ser utilizados para se entender a mensagem, mas ainda há intenção do emissor que deve ser considerada que objeto da pragmática.

Pragmática reflete a intenção pela qual a linguagem é utilizada para comunicar uma mensagem. Ela estuda as maneiras em que o contexto extralinguístico contribui para o significado. Ainda, ela denota o estudo da aplicação da linguagem em diferentes situações, isto é, dependendo da situação, tem-se diferentes tipos de mensagens sendo enviadas. Ela significa a intenção do emissor. O termo vem do Grego *pragmatikos*, que se refere à ação, prática ou alcance por meio da ação.

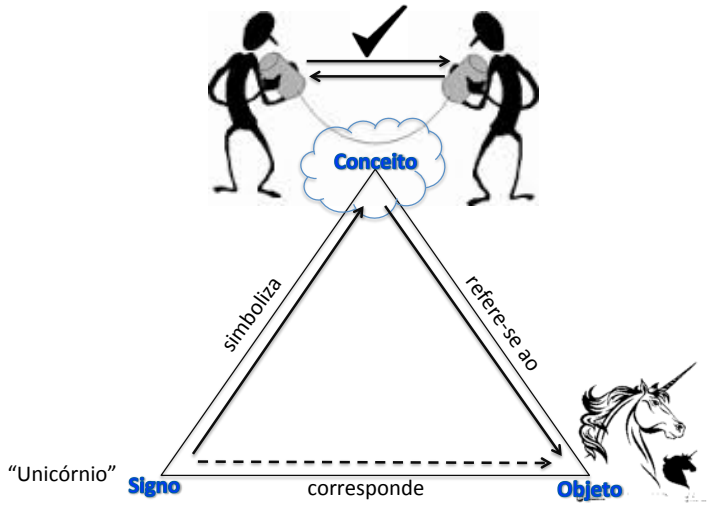
Por fim, para realmente se entender a mensagem, ainda é necessária a experiência em fazer isso. O termo experiência, do Latim *experientia*, sugere ensaio, prova ou tentativa. Dessa forma, experiência considera toda a informação a qual se aprende e se coloca no contexto do mundo no qual se vive. É o conhecimento obtido por meio da observação, da prática, da tentativa, etc., portanto, tem-se uma estreita relação com o senso comum e conhecimento do mundo.

Todos esses conceitos estão interligados para promover o sucesso na comunicação, visto que a mensagem deve estar correta (sintaxe) e o significado da mensagem transmitida deve ser interpretado (semântica) corretamente pelo receptor, isto é, compreendido com o mesmo sentido pretendido pelo emissor. Assim, no âmbito dessa pesquisa, a compreensão depende do contexto do emissor e do receptor, e da pragmática do emissor. Por sua vez, o contexto de ambos depende da experiência — ou conhecimento de mundo — do emissor e do receptor.

Na linguagem natural, a Figura 3 ilustra a abordagem de Ogden & Richards (1923), na

qual o significado é expresso por meio de um signo ou símbolo — o termo Unicórnio — que corresponde ao referente — objeto real, ou não; esse símbolo evoca um conceito — ou referência abstrata, uma representação mental, a imagem que traduz o senso comum de unicórnio — que se refere ao objeto. Portanto, a comunicação é realizada com sucesso quando o emissor e receptor percebem o mesmo referente, a compreensão se estabelece em ambos, pois compartilham o mesmo significado.

Figura 3: O triângulo do significado e o sucesso da comunicação



Fonte: Adaptado de Ogden & Richards (1923)

Esse processo é abstrato e se dá no sistema cognitivo das pessoas, a semântica é implícita, mas é resolvida, na maioria das vezes, de forma simples na comunicação do dia a dia. O desafio, então, é responder o que isso significa para a informação publicada na web.

4.1.3 Da web de documentos à web de dados

A web tradicional pode ser vista como um livro gigante de dimensão global, pois temos páginas conectadas a outras que descrevem temas de interesses de usuários. Contudo, nessa web, não há semântica explícita, isto é, o significado permanece oculto na linguagem natural contida nos documentos da web. Isso se deve a fatores limitantes, tais como: i) ambiguidade na linguagem natural (abacaxi — fruta ou problema), ii) falta de contexto (uma imagem de uma pedra sem qualquer explicação), iii) conhecimento insuficiente ou muito específico (a palavra *Warscheinlichkeit*[5] para quem não entende Alemão ou "regressão" para um psicólogo ou para um estatístico), iv) diferentes culturas que levam a diferentes interpretações[6], v) conhecimento implícito que deduz a informação a partir da informação disponível (o termo "pátria de chuteiras") e vi) necessidade pessoal de informação em relação àquela apresentada.

[5]Probabilidade em Alemão

[6]Nos EUA, o sinal feito com a junção do polegar e o indicador, formando um círculo, e deixando os outros dedos esticados significa que está tudo bem ou *ok*, mas é ofensivo no Brasil

Desde o início, vislumbrando o potencial desse recurso global, o inventor da Web afirma:

> A Web de documentos legíveis por pessoas está sendo mesclada com uma web de dados legíveis por máquinas. O potencial dessa combinação de humanos e máquinas trabalhando juntos e se comunicando por meio da web pode ser imenso[7]. (BERNERS-LEE, 1998, pag. 1. Tradução nossa).

Diante disso, esse mesmo autor ainda ressalta:

> A Web foi concebida como um espaço de informacão que deveria ser útil não somente para a comunicação homem-homem, mas também para que máquinas fossem capazes de participar e de auxiliar[8]. (BERNERS-LEE et al., 1998, pag. 1. Tradução nossa).

O problema é que a participação de máquinas fica condicionada à compreensão da informação disponível nos documentos que estão na Web, pois pressupõe-se que elas leiam e interpretem corretamente o conteúdo para que entendam a mensagem. Logo, deve haver alguma forma de se explicitar a semântica contida nos textos em linguagem natural. Essa preocupação em representar o conhecimento por meio de conteúdos com metadados semânticos explicitamente anotados se constitui numa mudança de paradigma que é o embrião da Web Semântica.

Considerando o exemplo "Villa-Lobos brilha!". A expressão pode ser ambígua, pois existe mais de um Villa-Lobos. A dedução trivial para humanos sobre Villa-Lobos não está disponível na maneira que o computador entenda, isto é, não há semântica explicitamente anotada. Assim, o significado deverá estar de alguma forma explícito para que se tenha acesso à semântica. O primeiro passo é saber o que é Villa-Lobos? Um teatro em Brasília, uma pessoa ou qualquer outra coisa? Além disso, considerando que se descreve uma pessoa, seria o maestro Heitor, o guitarrista Dado ou o jogador de pôquer Nicolau? Na linguagem natural esse processo de decisão para escolher entre as várias possibilidades é a desambiguação.

Para se tomar uma decisão, utilizando o processamento por máquinas, o significado (semântica) de entidades e classes deve ser definido explicitamente. A Figura 4 mostra a relação entre os objetos. Ressalta-se que, para a máquina, só há o lado direito da figura, pois o lado esquerdo em LN não representa nada para ela. Isso quer dizer que a inferência se dá pela manipulação e interpretação dessas relações.

Portanto, a definição da relação entre entidades e classes é parte do significado que é expresso por meio desse tipo de estrutura. Logo, o que se pode inferir diretamente dessas relações é que Villa-Lobos é um artista e que, por consequência, é uma pessoa. Além disso, todo artista é uma pessoa, mas o contrário nem sempre é verdadeiro.

Nessa primeira representação, é possível deduzir que não se trata do jogador de pôquer Nicolau Villa-Lobos. A continuação do processo de descrição conduz à interpretação precisa

[7] *The web of human-readable document is being merged with a web of machine-understandable data. The potential of the mixture of humans and machines working together and communicating through the web could be immense.*

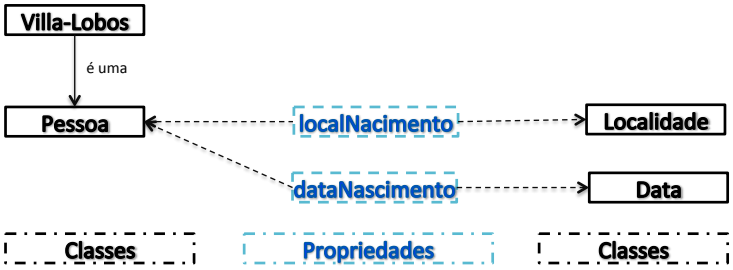
[8] *The Web was designed as an information space, with the goal that it should be useful not only for human-human communication, but also that machines would be able to participate and help.*

Figura 4: Relação entre classes e entidades

Fonte: Elaborado pelo Autor

do objeto analisado. Além disso, como se trata de uma pessoa, podem ser apresentadas relações com outras classes que individualizam a entidade e que já foram definidas em outras fontes de dados semânticos, como na Figura 5 por exemplo:

Figura 5: Relação entre classes



Fonte: Elaborado pelo Autor

Na seção 4.1.4.3, o assunto será tratado com mais detalhes. Contudo, o que se expõe aqui é que com esse tipo de representação, o significado da informação se torna explícito por meio de representações de conhecimento padronizadas e formais — estruturadas. Dessa forma, é possível processar o significado da informação automaticamente, relacionar e integrar dados heterogêneos e deduzir a informação implícita.

Portanto, se o conhecimento se torna explícito por meio de tecnologias da web, então, se cria algo que se chama de Web Semântica que:

> ... é uma extensão da web atual, na qual a informação é dada com significado bem definido para melhor possibilitar que pessoas e computadores trabalhem em cooperação[9] (BERNERS-LEE; HENDLER; LASSILA, 2001, p. 3. Tradução nossa.)

Para Allemang & Hendler (2008), a visão da WS se constitui em um sistema globalmente organizado, no qual a informação flui de um lugar para o outro de forma suave e, sobretudo, ordenada. Contudo, essa abordagem requer que a semântica seja, de alguma forma,

[9] *The Semantic Web is not a separate Web but an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation.*

explicitada para que se possa ter acesso a ela. Logo, a linguagem natural deve ser anotada explicitamente por meio de metadados semânticos que codificam o significado (semântica) de forma a serem lidos e interpretados corretamente por máquinas.

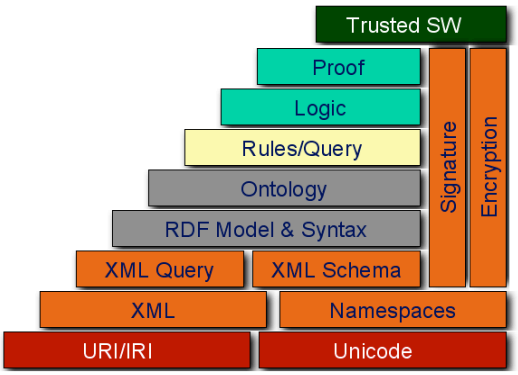
Finalmente, a ideia central da Web Semântica, então, é auxiliar as máquinas na leitura e na utilização da web. De acordo com Berners-Lee et al. (1998), isso transformará a atual Web, um livro gigante e global, em uma base de dados gigante e global. Tal tecnologia não introduz inteligência artificial, nem transformam as máquinas em seres conscientes, mas fornece ferramentas para que elas encontrem, troquem e "interpretem[10]" informações.

4.1.4 Tecnologias da Web Semântica

Como a informação e seu significado pode ser explicitamente expresso na Web? Isso pode ser feito utilizando tecnologias, desenvolvidas para Web, que são capazes de capturar a semântica de forma que ela possa ser lida e compreendida por máquinas. Compreender, neste sentido, quer dizer que a informação pode ser interpretada de forma correta.

As tecnologias, padronizadas pela *World Wide Web Consortium* (W3C)[11], têm sido desenvolvidas para viabilizar a WS e são o resultado do esforço da comunidade científica internacional. Segundo Prud'Hommeaux (2004), o conjunto dessas tecnologias foram agrupados em um esquema gráfico conhecido como *Layer Cake* que está representado na Figura 6:

Figura 6: Modelo Layer Cake



Fonte: http://www.w3c.it/talks/2005/openCulture/slide7-0.html

Para o propósito dessa pesquisa, não se abordarão todas as camadas, mas apenas aquelas relevantes para os objetivos já mencionados na seção 2.

[10]Interpretação — " Sentido em que se toma o que se ouve ou o que se lê, e que se julga ser o verdadeiro" em http://www.priberam.pt/. A interpretação está limitadas às regras de inferências adicionadas ao modelo de dados.

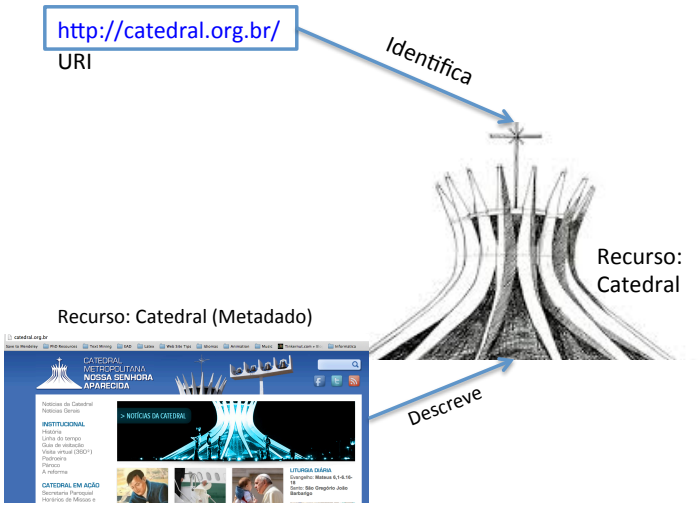
[11]http://www.w3.org/ — A W3C é uma comunidade internacional que desenvolve padrões abertos para garantir o desenvolvimento permanente da Web.

4.1.4.1 Uniform Resource Identifier

Nós, humanos, temos a habilidade inata de usar substitutos para falar sobre coisas como se eles fossem os próprios objetos. Por exemplo, se uma imagem de um carro nos fosse apresentada com a seguinte pergunta: "O que é isso?". Responderíamos, salvo raríssimas exceções, imediatamente que "isso é um carro" e não "Isso é uma imagem de um carro". Portanto, conforme o triângulo semiótico de Ogden & Richards (1923), usamos um substituto do verdadeiro ente que simboliza um conceito que está no nosso aparelho cognitivo, isto é, nomeamos as coisas. Esta habilidade se reflete, também, quando queremos representar um conhecimento na Web.

Quando se fala de coisas que se quer adicionar na Web, há que se pensar em nomear estas para que sejam identificáveis e recuperáveis. Para tal fim, utiliza-se um mecanismo chamado de *Uniform Resource Identifier* (URI). Para Heath & Bizer (2011), URI fornece um meio simples e extensível para identificar um recurso[12]. URI se refere aos diferentes tipos de identificadores de recursos que são construídos de acordo com um esquema padronizado. Além disso, se destina a identificar qualquer coisa que possa ser representada via URI e a distinguir um recurso — textos, imagens, vídeos, sons, conceitos concretos (carro, lua) ou abstratos (amor, divindade), etc. — de outros na Web.

Figura 7: Representação da Catedral de Brasília



Fonte: Elaborado pelo Autor

Considera-se, como exemplo, um conhecido ponto turístico da capital do Brasil. A Figura 7 apresenta uma imagem da Catedral de Brasília, isto é, um substituto. Portanto, o recurso do qual se fala é a Catedral como se fosse aquela que existe na Esplanada dos Ministérios em Brasília. Contudo, essa existe também na Web e possui uma URI que identifica o recurso que permanece no mundo real. Ao mesmo tempo, esta possui um tipo de representação na

[12] http://www.ietf.org/rfc/rfc3986 — ao leitor interessado, nesse documento se encontra a especificação completa de URI.

web que contém metadados que descrevem o objeto real e são apropriados para exibir a informação adequada ao consumo humano, nesse caso, uma página da Web.

Obviamente, a catedral do mundo digital é diferente da original. Entretanto, utilizamos a URI para nomeá-la na Web e, ao mesmo tempo, oferecer um lugar no qual ela possa ser descrita o suficiente para substituir a Catedral original e ser passível de representação em formatos adequados às máquinas que fornecem informação acessível aos humanos.

O conceito de URI já está amplamente difundido na CI, como na localização de páginas da Web via *Uniform Resource Locator* (URL), na identificação de livros via *International Standard Book Number* (ISBN) ou publicações seriadas via *International Standard Serial Number* (ISSN) ou, ainda, conteúdos digitais via *Digital Object Identifier* (DOI). Todos correspondem ao esquema padronizado que identifica e individualiza os objetos.

Finalmente, Allemang & Hendler (2008) ensinam que URI combina duas abordagens: i) localizador — *Uniform Resource Locator* (URL) que denota onde um recurso pode ser encontrado na web pela declaração de seu mecanismo primário de acesso (http, ftp etc.); ii) identificador — *Uniform Resource Name* (URN) que denota um identificador persistente para um recurso na Web sem definir sua localização ou forma de acesso. Por exemplo o ISBN que apenas nomeia e individualiza um recurso, mas não oferece acesso nenhum. Portanto, URL e URN são complementares e também tipos de URI.

4.1.4.2 Resource Description Framework

Tomando como exemplo a frase: "Machado de Assis é o pai da Academia Brasileira de Letras". É corriqueiro perceber que "Machado de Assis" é uma pessoa, "Academia Brasileira de Letras" é uma Instituição e que existe uma relação entre ambos. Fácil perceber ainda que "pai" é um tipo de parentesco, mas, nesse contexto, seria criador ou fundador de algo. Embora trivial aos humanos, há que se fornecer algum padrão que seja computável por máquinas. O que se pretende é tornar a frase legível ao computador.

Daconta, Smith & Obrst (2003) declaram que a *eXtensible Markup Language* (XML) faz parte do rol de ferramentas que se destinam a traduzir o conteúdo da web para as máquinas. Ela é uma linguagem de marcação como a HTML, mas com o propósito de adicionar *tags* (rótulos) que descrevem dados. Essas tags são invisíveis aos usuários, mas visíveis aos computadores. Elas fornecem informações legíveis aos robôs, ou *bots*, que coletam dados para os atuais motores de busca como Yahoo! ou Google.

Todavia, a liberdade da XML para criação e uso dessas tags é também uma restrição para a padronização, pois uma simples frase pode conter várias versões corretas. A expressão do exemplo poderia ser expressa, dentre outras possíveis, como na Figura 8.

A Figura 8 mostra que, para uma comunicação de sucesso, requer-se a conversão dos esquemas A e B para um esquema único. Atualmente, isso é resolvido com a utilização de *EXtensible Stylesheet Language Transformation* (XSLT), que estabelece uma camada esquemática que unifica diferentes esquemas XML. Porém, à medida que se aumenta a quantidade de

Figura 8: Exemplo em XML

Esquema A	Esquema B
<?xml version="1.0" encoding="UTF-8"?> <xml> <instituição> <fundador>Machado de Assis</fundador> <nome>Academia Brasileira de Letras</nome> </instituição> </xml>	<?xml version="1.0" encoding="UTF-8"?> <xml> <person name="Machado de Assis"> <instituiçãoFundada>Academia Brasileira de Letras</instituiçãoFundada> </person> </xml>

Fonte: Elaborado pelo Autor

informação para modelar o conhecimento, também se aumenta a complexidade da tradução e a dificuldade de manutenção.

Assim, Tim Berners-Lee[13] sugeriu que a comunidade, que colaborativamente construiu a web de documentos, direcionasse esforços para a construção de uma web de dados que obedecessem a um único padrão. Nela, qualquer conceito existente poderia ser descrito e relacionado a outros e, assim, todo conhecimento humano poderia ser representado. Para tal, ele aponta o uso de *Resource Description Framework* (RDF), que é reconhecido como o elemento essencial[14] da Web Semântica.

O termo RDF é assim decomposto: i) Resource significa que, em princípio, qualquer coisa pode ser expressa, mas deve ser identificada de forma única e ser referenciável via URI; Description refere-se à descrição de recursos pela representação de propriedades e relacionamentos entre esses recursos e representados na forma de grafos; e iii) Framework remete à ideia de combinação de protocolos baseados na web (http, URI, XML etc.), à base em modelos formais (semânticas) e à definição de todas as relações permitidas entre os recursos, de acordo com Daconta, Smith & Obrst (2003), Allemang & Hendler (2008).

RDF faz o que o nome sugere, isto é, fornece uma estrutura para descrever recursos utilizando um esquema simples para expressar fatos ou declarações. A ideia por trás do RDF é clara, todo conceito seria representado por uma tripla composta por sujeito, propriedade (ou predicado) e objeto. De fato, essa combinação é familiar a todo nativo de línguas ocidentais, pois é a forma intuitiva pela qual se constroem frases simples. O sujeito refere-se ao conceito que se quer descrever; a propriedade, aos atributos relacionados ao sujeito; o objeto é algo a que se refere com a propriedade. Utilizando essa simples ideia, pode-se descrever qualquer coisa.

Como ilustração, a representação do exemplo anterior em RDF seria segmentada em:

- sujeito: Machado de Assis;
- propriedade: é o pai da;

[13] http://www.ted.com/talks/tim_berners_lee_on_the_next_web
[14] building block — tradução nossa.

- objeto: Academia Brasileira de Letras.

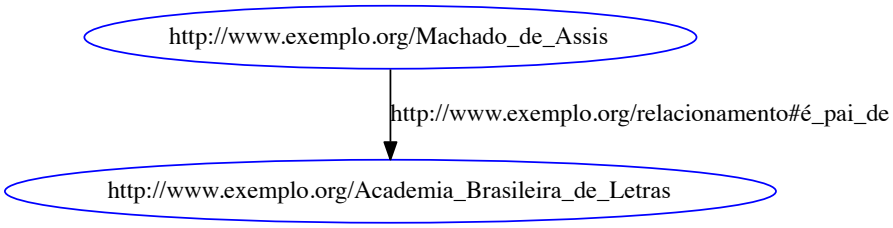
Até agora, a máquina já é capaz de identificar dois elementos e uma relação entre eles. Contudo, ainda não há especificações precisas sobre o que são os elementos, nem o tipo de relação entre eles. Para fornecer a forma específica de o quê ou quem são esses recursos, RDF utiliza as URIs para direcionar os computadores a documentos ou a objetos que representam esses recursos.

RDF utiliza URI para especificar sujeitos e propriedades. Estritamente falando, objetos também podem ser especificados por URIs, além de literais. O exemplo, então, poderia ser representado por:

- sujeito: http://www.exemplo.org/Machado_de_Assis;
- propriedade: http://www.exemplo.org/relacionamento#é_pai_da;
- objeto: http://www.exemplo.org/Academia_Brasileira_de_Letras.

Nessa representação, a URI fornece ao computador um ponto específico de referência para cada item da tripla. Visualmente, tem-se o grafo representado na Figura 9.

Figura 9: Grafo de RDF com URIs



Fonte: Elaborado pelo Autor

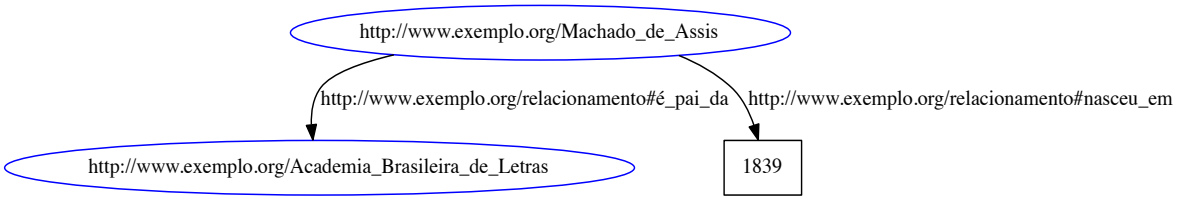
O objeto também pode ser expresso por literais, isto é, valor atribuído ao objeto que não existe separado dele. Os tipos de literais são padronizados pela W3C que apresenta uma extensa lista de tipos já padronizados com base nos XML *Schema datatypes*[15]. Esses valores abrangem caracteres, números, datas, horas e uma variedade de formas que eles podem ser apresentados. Por exemplo, acrescentando o ano de nascimento ao grafo da Figura 9, tem-se: "Machado de Assis nasceu em 1839 e é o pai da Academia Brasileira de Letras". A Figura 10 mostra a representação gráfica deste exemplo:

Destaca-se que, apesar de confortável e autoexplicativo, os grafos são muito utilizados para visualizar as entidades e relações entre elas, mas não são legíveis para máquinas. Logo, há várias formas de representar RDF[16] legível para o computador, mas, a seguir, se apresentam

[15] http://www.w3.org/2001/XMLSchema#

[16] Para o leitor interessado em detalhamento técnico, recomenda-se a leitura atenta de http://www.w3.org/RDF/.

Figura 10: Grafo de RDF com URIs/Literal



Fonte: Elaborado pelo Autor

somente as mais comuns com seguinte frase como exemplo: “Machado de Assis nasceu em 1839”.

N-triple notation é a forma de representação mais simples, pois se utiliza uma tripla por linha finalizando com um ponto. As URIs entre “<>” e literais entre aspas. Por exemplo:

Exemplo 4.1: Notação N-triple

```
<http://www.exemplo.org/Machado_de_Assis> <http://www.exemplo.org/relacionamento#e_pai_de> ‘‘1839’’.
```

RDF/XML notation é bastante comum na TI pela semelhança óbvia com a XML, porém nem um pouco intuitivo para profissionais de outras áreas. O mesmo exemplo a seguir:

Exemplo 4.2: Notação RDF/XML

```
<?xml version="1.0" encoding="utf-8" ?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
            xmlns:ex="http://www.exemplo.org/relacionamento#">
<rdf:Description rdf:about="http://www.exemplo.org/Machado_de_Assis">
  <ex:nasceu_em>1839</ex:nasceu_em>
</rdf:Description>
```

A linguagem mais intuitiva é *Terse RDF Triple Language (Turtle) notation* que é baseada em *N-triple*, mas com recursos que facilitam a leitura e escrita para humanos. Segue o exemplo:

Exemplo 4.3: Notação Turtle

```
@prefix ex: <http://www.examplo.org/> .
ex:Machado_de_Assis <http://www.exemplo.org/relacionamento#nasceu_em> "1839" .
```

Como se nota no exemplo, além da tripla ser representada em uma linha finalizada com o ponto, tem-se o recurso de utilizar prefixos para evitar a escrita da URI completa. Ao invés de <http://www.exemplo.org/Machado_de_Assis>, basta utilizar o prefixo definido “ex” na primeira linha e eliminar os símbolos “<>”. A substituição para ex:Machado_de_Assis facilita a leitura e torna a expressão mais clara para humanos.

4.1.4.3 Resource Description Framework Schema

A expressão codificada em RDF — sujeito, predicado, objeto — é capaz de expressar declarações simples como visto nos exemplos nas principais notações, mas de onde vem o significado no RDF? O que se sabe até então é que há um sujeito e um objeto relacionados de alguma forma por meio de uma propriedade, mas há uma grande contribuição da linguagem natural, nos rótulos das URIs, no processo humano de compreensão do relacionamento entre sujeito e objeto. Como seria se esses rótulos estivessem em uma língua desconhecida? Ainda há algum significado a ser apreendido?

Figura 11: RDF com rótulos em japonês

Fonte: Elaborado pelo Autor

No exemplo da Figura 11, sabe-se apenas que a tripla em RDF relaciona, por meio de uma propriedade, um sujeito a um objeto, mas, a menos que se saiba Japonês, não está conectado com nenhum significado pré-definido que dê alguma pista sobre o quê se relaciona com o quê. Logo, é necessário algo que represente o conhecimento em um nível mais abstrato, menos conectado à linguagem natural, como o *Resource Description Framework Schema* (RDFS) que é, muitas vezes, referido como *RDF Vocabulary Description Language.*

Allemang & Hendler (2008) declaram que o RDFS é uma linguagem que define o vocabulário utilizado no RDF. Ela permite a definição de classes que reúnem um grupo de entidades que possuem algo em comum. Ainda, possibilita a definição de propriedades e suas restrições. Além disso, define hierarquia entre classes (subclasses e superclasses) e propriedades (subpropriedades e superpropriedades).

De acordo com W3C[17], o RDFS é uma extensão semântica do RDF e fornece mecanismos para descrever grupos de recursos relacionados e as relações entre eles. Daconta, Smith & Obrst (2003) apresentam os componentes[18] principais desse vocabulário que são divididos em:

Classes

- rdfs:Resources — esta é a classe de todas as coisas, pois todas as outras são subclasses dela;
- rdfs:Class — define um grupo de entidades relacionadas que compartilham um conjunto de propriedades;
- rdfs:Literal — representa valores constantes como textos e números;

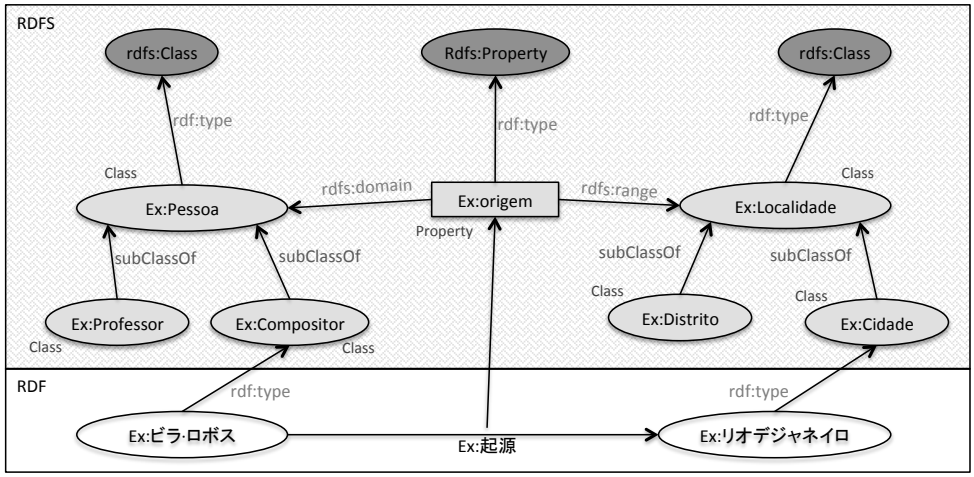
[17] http://www.w3.org/TR/rdf-schema/
[18] Há uma lista completa em http://www.w3.org/TR/rdf-schema/

- rdf:Property — define uma propriedade de uma classe e a gama de valores que se pode representar.

Propriedades

- rdfs:domain — define a quais classes uma propriedade pertence;
- rdfs:range — define o conjunto de valores possíveis para uma propriedade;
- rdf:type — uma propriedade padrão para definir que um sujeito RDF é do tipo definido em um RDF schema;
- rdfs:subClassOf — especifica que uma classe é uma especialização de outra;
- rdfs:subPropertyOf — declara que todos os recursos relacionados por uma propriedade são também relacionados por outras;
- rdfs:label — atributo que define um rótulo de classes legível por humanos.

Figura 12: RDF e RDFS



Fonte: Elaborado pelo Autor

Como exemplo, a Figura 12 adiciona mais elementos para representar a tripla em Japonês. O que se infere é que a entidade "Ex:ビラ·ロボス" é um compositor que é uma classe. Esta é uma subclasse de pessoa que também é outra classe. A propriedade "Ex:起源", por sua vez, é traduzida pela utilização de *namespace*[19]. Ela fixa um domínio (pessoa) e um alcance (localidade) que estabelece a relação de pessoa que tem origem em um lugar. Completando a tripla, o objeto "Ex:リオデジャネイロ" é uma classe cidade. Esta também é subclasse da classe localidade.

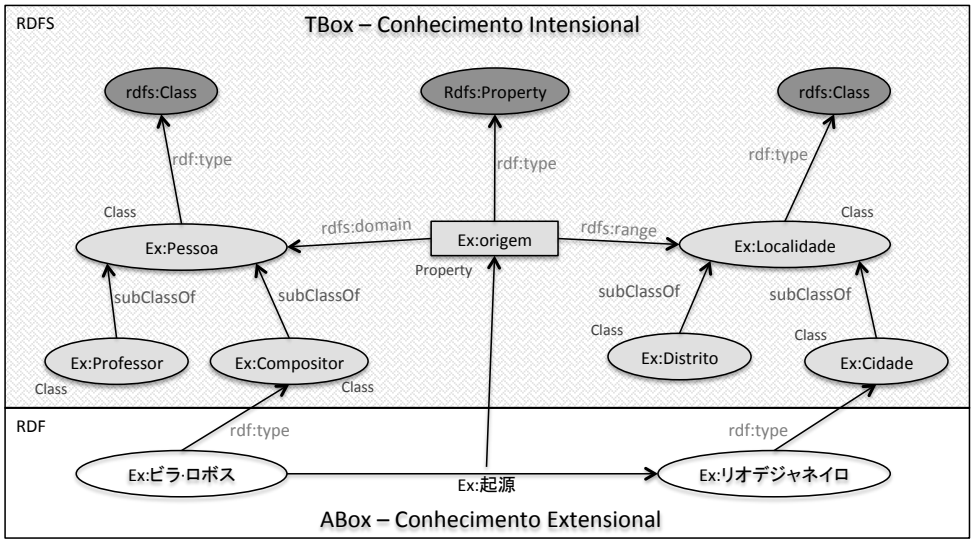
[19]Normalmente, o termo *namespace* não é traduzido na literatura da área. Ele fornece um método simples para qualificar nomes de atributos e elementos utilizados em documentos RDF/XML pela associação aos nomes identificados com URI.

Assim, mesmo desconhecendo os termos da tripla, é possível ter noção do significado dela. Pela adição, associação e comparação entre as classes e propriedades é possível colocar alguma semântica no modelo de dados. Portanto, quanto mais elementos na camada esquemática do modelo, RDFS, mais possibilidade de extrair inferências a respeito das triplas em RDF.

Conforme Nardi & Brachman (2003), RDFS e RDF combinam dois tipos de conhecimento: i) o conhecimento intensional (generalizado) que permanece no nível abstrato — conceitual — e trata sobre o próprio modelo de dados, isto é, a relação entre entidades gerais como classes e propriedades; e ii) o conhecimento extensional (especializado) que trata de especificação das entidades ou da instanciação de classes. Como resultado, os relacionamentos entre entidades na camada de especialização são refletidos na camada de generalização formando uma base de conhecimento RDF(S).

A Figura 13 mostra a representação entre camada de especialização e generalização. A camada de especialização é comumente referida como ABox, que vem de *Assertional knowledge*. Por exemplo, Heitor Villa-Lobos seria a particularização, instanciação ou especificação da classe pessoa. A camada de generalização é chamada de TBox, de *Terminological knowledge*, em que se localizam as abstrações do domínio que viabilizam inferências sobre o modelo de dados. Assim, nessa camada, o relacionamento entre classes e propriedades é o que realmente introduz a semântica no modelo de dados, do qual se evolui uma ontologia.

Figura 13: Conhecimento Intensional x Extensional



Fonte: Elaborado pelo Autor

Resumindo, a semântica dos elementos de uma base de conhecimento RDF(S) é fornecida em termos de suas propriedades e seus valores. Isso quer dizer que se pode fazer deduções com relações de hierarquias entre classes, propriedade e pelo estabelecimento de restrições conectadas às propriedades, como domínio e alcance. Desta forma, RDF e RDFS fornecem semântica suficiente para representar conhecimento, embora apenas em um nível superficial.

Mas, considerando a afirmação: "um pouco de semântica percorre um longo caminho"[20], julga-se que, com essas ferramentas da WS, os sistemas de informação estão habilitados a percorrerem esse caminho.

4.1.4.4 Web Ontology Language

Como qualquer idioma, a fluência depende do domínio de vocabulário e da capacidade de inferir diferentes significados atribuídos a um termo de acordo com o contexto. Não basta prover dicionários às máquinas, mas também fornecer documentos que descrevem todos os termos e a lógica para estabelecer relações entre eles.

Na Web Semântica, esses documentos vêm na forma de "esquemas" e ontologias. Ambas são relacionadas e possuem o objetivo de prover o vocabulário humano de forma aceitável ao computador. Genericamente, os esquemas são métodos para organizar a informação. Ontologias fornecem o vocabulário que descrevem objetos e como eles se relacionam.

O RDFS adiciona classes, subclasses e propriedades aos recursos criando uma estrutura básica de linguagem. Além desse, outro esquema é o *Simple Knowledge Organization System* (SKOS), que classifica recurso em termos de abrangência ou restrição, permite a utilização de rótulos alternativos ou preferidos que viabilizam a publicação de tesauros e glossários na Web.

Contudo, ainda tem-se a dificuldade de prever relações conflituosas ou de negação, como classes que são disjuntas por natureza, como "homem" e "mulher", pois indivíduos da classe "mulher" não serão permitidos na classe "homem" e vice-versa. Isso quer dizer que, em RDFS, é impossível afirmar se há inconsistências. Por outro lado, a noção de *Open World Assumption* (OWA) pressupõe que se não está explícito, então é possível. Da mesma forma, não há pressuposição de nomes únicos, isto é, pessoa A deve estar explicitamente expresso que não é pessoa B. Enfim, a especificação de entidades e relacionamentos deve ser exaustiva, a não ser que se consiga adicionar regras de inferência numa camada mais abstrata que possa estabelecer limites e restrições generalizadas à base de dados.

Então, para estruturas de classes e propriedades mais complexas, foi desenvolvida a *Web Ontology Language* (OWL). Ela utiliza o RDF e RDFS como base e adiciona mais vocabulário para descrever grupos de coisas, tais como classes, fatos sobre essas classes, relacionamentos entre classes e instâncias, e características desses relacionamentos. É voltada para o processamento de conteúdo da Web e foi projetada para ser legível por aplicações de computadores. Além disso, permite a criação de regras de inferências e axiomas para permitir deduções por meio de ferramental da lógica (W3C, 2014).

Historicamente, a OWL é o resultado do esforço conjunto das comunidades de defesa da Europa e EUA que procuravam por formatos de dados que seriam autodescritivos e dinâmicos

[20] A frase: "A little semantics goes a long way" é frequentemente atribuída ao pioneiro da WS James A. Hendler, mas ele mesmo não tem certeza da autoria, conforme http://www.cs.rpi.edu/~hendler/LittleSemanticsWeb.html

o suficiente para que agentes autônomos computadorizados pudessem atuar sobre eles. Nessa busca, simultaneamente, perceberam que era preciso desenvolver uma nova linguagem. Assim, os europeus criaram a *Ontology Inference Layer* (OIL) e os americanos, a *DARPA Agent Markup Language* (DAML). Mais tarde, a combinação de OIL e DAML se tornou uma especificação da W3C que ficou conhecida como OWL (WILKS; BREWSTER, 2009; W3SCHOOL, 2013).

A sintaxe de OWL 1.0 é baseada em RDF/XML, contudo, por causa do vocabulário adicional e de maneiras particulares para formatar dados, ela constitui um modelo OWL semanticamente mais elaborado, com uma linguagem nativa e reconhecida como uma das aplicações mais importante da lógica descritiva atualmente. De acordo com Antoniou & van Harmelen (2009), a OWL 1.0 se classifica em três sublinguagens:

- OWL Full é totalmente compatível com RDF, isto é, qualquer documento RDF também é um documento OWL. Ela permite máxima expressividade e abrange todo vocabulário disponível para a OWL. Entretanto, não é decidível e, portanto, não oferece garantia de "reasoning" efetivo e completo.

- OWL DL é um subconjunto da OWL full que restringe a maneira como os construtores de OWL e RDF podem ser usados. Permite o uso completo do núcleo da linguagem, mas com algumas limitações sobre restrições de classes. É a mais utilizada, pois possibilita o melhor custo-benefício, isto é, a melhor expressividade e desempenho, enquanto garante-se que todas as declarações sejam computáveis e conclusivas. DL significa Description Logic — Lógica Descritiva (LD) em Português — devido a sua correspondência com as lógicas de descrição. A OWL DL é a mais importante aplicação da LD.

- OWL Lite é um subconjunto da OWL DL e, assim, exclui alguns construtores que são apropriados para representações mais complexas. Dessa forma, existem limitações de como uma classe pode ser declarada e de como se atribuir restrições a ela. A OWL Lite, basicamente, se aplica à classificação hierárquica e a restrições simples. A principal vantagem é a facilidade de aprendizagem e de implementação.

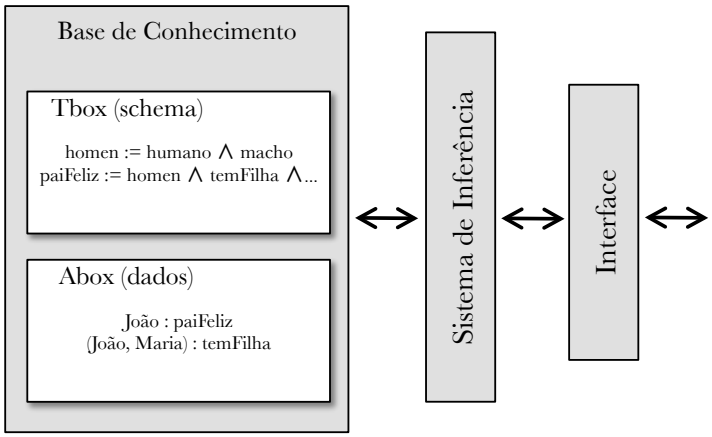
OWL Full pode ser vista como uma extensão do RDF, enquanto OWL Lite e DL como extensões de visão restrita do RDF. Cada uma dessas sublinguagens é uma extensão de seu predecessor. Do ponto de vista prático, usuários devem escolher a versão que melhor responde as suas necessidades. Todavia, como OWL lite e DL oferecem suporte garantias computacionais, são mais utilizadas, em especial, a DL pela sua correspondência com a LD que permite descrever conceitos e semânticas baseadas em lógica para determinado domínio de maneira formal e bem estruturada.

Segundo Baader, Horrocks & Sattler (2009), LD é baseada na lógica de primeira ordem, que é um sistema de inferência dedutiva com fundamento na matemática. O que significa que, com OWL, podem-se expressar fatos e se ancorar em um sistema de provas fundamentadas

na matemática para descobrir as implicações daqueles fatos. Tal ferramental possibilita representar o conhecimento em consonância com a visão humana e, ao mesmo tempo, formal no nível semântico. Logo, se expressa um conhecimento rigoroso, sem ambiguidade, semanticamente rico e interpretável por máquinas.

Horrocks (2002) afirma que Lógicas Descritivas são uma família de linguagens de representação do conhecimento frequentemente utilizadas em modelagem de ontologias. Ainda, elas fornecem meios para modelar os relacionamentos entre entidades em um domínio de interesse. Este é descrito em termos de classes, propriedades e indivíduos. Adicionalmente, a arquitetura da Lógica Descritiva constitui uma Base de Conhecimento (BC) — que possui o Conhecimento Intensional (*Terminological Knowledge* — TBox) e o conhecimento extensional (*Assertional Knwoledge* — ABox), um sistema de inferência e uma interface, conforme Figura 14:

Figura 14: Arquitetura de LD



Fonte: Adaptado de Horrocks (2002)

Krötzsch, Simancik & Horrocks (2012) ensinam que, nesse tipo de arquitetura, há uma clara separação dos conhecimentos intencional e extensional. O TBox é construído por declarações que descrevem conceitos gerais e propriedades. Por exemplo, representar a equivalência entre pessoa e humano — "pessoa ≡ humano" — ou definição de homem — "homem ≡ pessoa ∩ ¬fêmea". Enquanto o ABox captura o conhecimento sobre indivíduos, ou seja, os conceitos aos quais pertencem e como se relacionam entre si. Por exemplo: a declaração "Maria é mãe de Pedro" é expressa em LD como mãe(Maria, Pedro). O sistema de inferência possibilita a dedução de novos fatos a partir da BC, as provas de que as declarações são corretas e se todo conteúdo da BC é consistente e, finalmente, a interface permite que o usuário acesse o conhecimento contido nessa BC.

Um exemplo simples de OWL/RDF:

Exemplo 4.4: Modelo OWL simples

```
<owl:thing rdf:ID=''fci-unb''/>
```

Apesar da tentação humana de inferir além do que está expresso, a máquina não tem nenhum outro tipo de interpretação além daquele que esse simples modelo representa: que existe um indivíduo rotulado de "fci-unb" e que ele é uma *Thing* (coisa).

O vocabulário OWL DL permite que se descreva classes e propriedades que pertencem ao domínio de forma que se estenda monotonicamente a representação do conhecimento de interesse. De acordo com Antoniou & van Harmelen (2009), ressalta-se que a OWL herda muitas propriedades da LD. Dois pontos são fundamentais para garantir a solidez e consistência dos cálculos de inferência, tais como i) *Open-world assumption* (OWA), que estipula que não se pode deduzir a falsidade de uma declaração pela falta dela, isto é, se não se apresenta uma declaração como sendo verdadeira, simplesmente não é possível afirmar que ela seja falsa; ii) *Unique Name assumption* (UNA), que estabelece que dois indivíduos diferentes possuem nomes diferentes. Isso não é o caso na OWL, pois ela fornece vocabulário para fazer dois indivíduos equivalentes ou distintos.

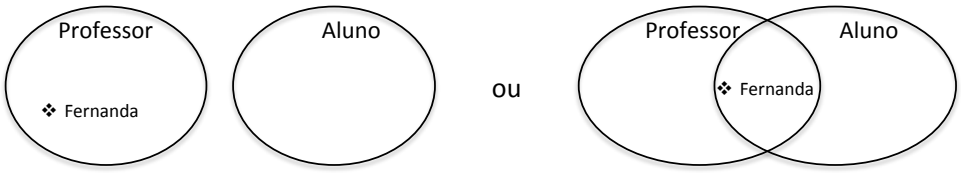
Assim, considerando as implicações de herança da LD, apresenta-se o exemplo 5 para representar duas classes, professor e aluno, e instâncias associadas às classes, tem-se:

Exemplo 4.5: Modelo OWL simples — OWA

```
<owl:Class rdf:ID=‘‘professor’’/>
<owl:Class rdf:ID=‘‘aluno’’/>
<owl:professor rdf:ID=‘‘Fernanda’’/>
```

O exemplo 4.5 leva a maioria das pessoas a concluir que Fernanda pertence somente a classe professor, como ilustra o lado esquerdo da Figura 15:

Figura 15: Exemplo OWA



Fonte: Elaborado pelo Autor

Porém, considerando a OWA, a dedução correta é que "não se sabe" se as classes Professor e aluno possuem intersecção no modelo apresentado. Por isso, a interpretação pode ser qualquer uma dessas da Figura 15.

Portanto, um sistema baseado em lógica sempre fará inferências ou interpretará as implicações das declarações da base de conhecimento. No caso do exemplo, se se quer estabelecer que Professor não pode ser Aluno, então, é necessário que se faça essa declaração explícita, isto é, estabelecer que as classes são disjuntas. Isto requer que se acrescente uma declaração tal como:

Exemplo 4.6: Modelo OWL com restrição explícita

```
<owl:Class rdf:ID=‘‘Professor’’/>
<owl:Class rdf:ID=‘‘Aluno’’/>
     <owl:disjointWith rdf:resource=‘‘#Professor’’/>
<owl:professor rdf:ID=‘‘Fernanda’’/>
```

A princípio, parece complicado ter que especificar todo tipo de restrição, mas, na verdade, um sistema dotado deste tipo de motor de inferência permite que se responda o que deve e o que não deve ser verdade sobre quais indivíduos podem ou não podem ser membros de algumas classes. Qualquer instância adicionada ao sistema estará sujeita ao processo lógico de verificação da restrição expressa no modelo. Isso garante que futuras adições que porventura tenham qualquer contradição sejam rejeitadas. Logo, para produção e manutenção de grandes bases de conhecimento, esse tipo de ferramenta é fundamental para a garantir a qualidade.

A W3C (2014) declara que a OWL 1 definiu dois principais dialetos, OWL DL e OWL Full, e um subconjunto sintático, OWL Lite. Contudo, verificou-se que isso não era suficiente para lidar com as necessidades identificadas posteriormente nas ontologias OWL publicadas.

Assim, o W3C OWL *Working Group*[21] produziu a OWL 2 que refina e estende a OWL 1 e herda as características da linguagem, decisões de projeto e caso de uso da OWL 1. Ela é uma recomendação W3C desde 2009. A nova versão adiciona várias novas características que incluem o *syntactic sugar*, que facilita a escrita de padrões comuns, o aumento do poder de expressividade para propriedades, suporte estendido para *datatypes*, capacidades de metamodelagem simples e anotações estendidas. Além disso, define vários perfis que são subconjuntos da OWL 2 que melhoram requisitos de desempenho e facilitam a implementação.

Por isso, definiram-se três subconjuntos sintáticos ou perfis, como se segue:

- OWL 2 DL, assim como na OWL DL, é considerada a versão mais significante da especificação da OWL 2. Ela é uma versão com restrições sintáticas da OWL Full. Essas restrições produzem uma linguagem totalmente determinística e mais prática para implementação;

- OWL 2 Full é a principal versão, pois é considerada a extensão direta do RDFS, porém introduz a possibilidade de resultados não determinísticos;

- OWL 2 EL é um perfil que foi definido para fornecer um comportamento otimizado para ontologias grandes e complexas que dependem de definições de classes complexas;

- OWL 2 QL captura o poder de expressividade utilizado em ontologias simples, e a das *Entity Relationship* (ER) e *Unified Modeling Language* (UML), ou simplesmente, ER/UML. Isso significa que foi formulada para capturar a semântica de bases de dados e da UML e, portanto, pretende ajudar na integração de dados com a utilização de OWL;

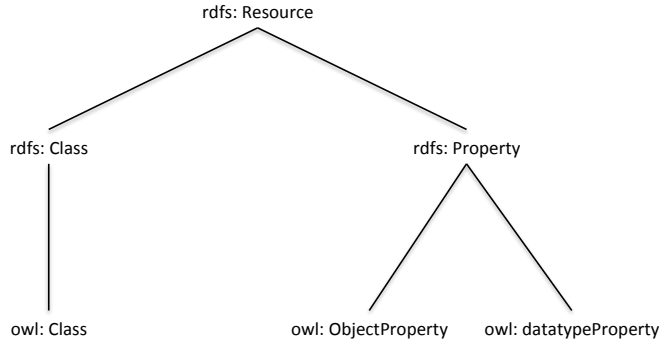
[21] http://www.w3.org/2007/OWL/wiki/OWL_Working_Group

- OWL 2 RL foi projetado para acomodar aplicativos que podem trocar a expressividade plena pela eficiência e pelas aplicações RDF(S) que necessitam de expressividade da OWL 2. Foi construído para otimizar a intersecção de programas baseados em regras com lógicas descritivas. A principal meta desse perfil é maximizar a expressividade da linguagem para representar o conhecimento.

Comparativamente, tem-se que a OWL é, então, um fragmento semântico da Lógica de Primeira Ordem. A OWL 1 se divide em OWL 1 Lite $\subseteq$ OWL 1 DL $\subseteq$ OWL 1 Full, enquanto a OWL 2 em OWL 2 EL, OWL 2 RL, OWL 2 QL $\subseteq$ OWL 2 DL $\subseteq$ OWL 2 *Full*.

Conforme Antoniou & van Harmelen (2009), todas as variedades OWL utilizam RDF nas respectivas sintaxes, declaram instâncias da mesma forma que em RDF, e possuem construtores OWL que são especializações de suas contrapartidas RDF, como mostra a Figura 16.

Figura 16: Relacionamentos entre Subclasses OWL e RDF/RDFS



Fonte: (ANTONIOU; VAN HARMELEN, 2009, p. 95.)

Assim, a OWL 2 possui uma estrutura interna que permite descrever os objetos que se quer representar, ora colocando-os em categorias, ora dizendo algo sobre suas relações. Todos os constituintes — indivíduos, classes e propriedades — são entidades. As propriedades se subdividem em *object properties* que relaciona objetos a outro objeto, e *datatype properties*, que atribui valores ao objeto, por exemplo: idade.

Finalmente, procurou-se apresentar, aqui, uma noção geral dos pontos de maior relevância para compreensão da abrangência e da utilidade do tema para o desenvolvimento da pesquisa. Além disso, há várias indicações de consulta[22] para o leitor mais interessado no detalhamento do assunto. O essencial é que, para formular o conhecimento de forma explícita, é útil assumir que ele é constituído de peças elementares na forma de declarações ou enunciados. Uma ontologia OWL é, fundamentalmente, uma coleção dessas peças. A partir disto, a OWL captura este tipo de conhecimento por meio de formas que ela pode representar.

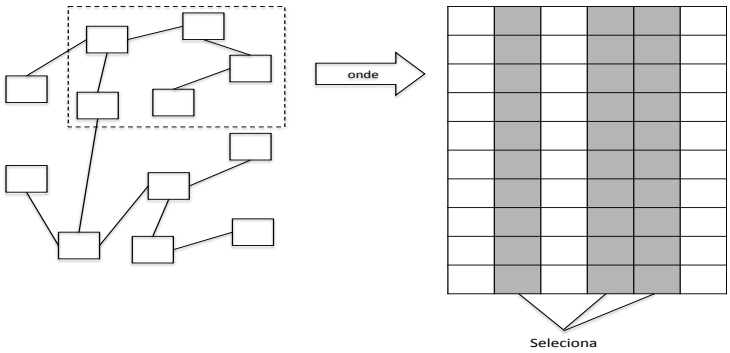
[22] http://www.w3.org/TR/2012/REC-owl2-primer-20121211/

4.1.4.5 SPARQL

Além de armazenar o conhecimento em formatos específicos para a WS, é necessária alguma forma de recuperar essas informações. Para tal, foi desenvolvida a *Simple Protocol and RDF Query Language* (SPARQL), uma especificação W3C. Ela foi inspirada na *Structured Query Language* (SQL), que é a linguagem padrão para sistemas de gerenciamento de bancos de dados relacionais. A versão mais recente foi liberada em março de 2013, SPARQL1.1.

SPARQL é, ao mesmo tempo, uma linguagem padrão de consulta e um protocolo de acesso aos dados. Isso significa que ela viabiliza não somente o acesso às triplas — sujeito, predicado, objeto — RDF, ou frequentemente chamada de grafos RDF, mas qualquer fonte de dados que possa ser mapeada em RDF. Ela permite extrair dados (semi)estruturados, explorar dados utilizando consultas por relações desconhecidas, executa combinações complexas de bases de dados heterogêneas com consultas simples, transforma dados RDF de um vocabulário para outro e constrói novos grafos RDF a partir de consultas em outros vocabulários.

Figura 17: Extração de grafos



Fonte: Adaptado de DuCharme (2011)

A Figura 17 ilustra a seleção de um padrão de triplas em uma base de dados e sua contrapartida no resultado que se quer visualizar. Dessa forma, SPARQL permite que o usuário formule as consultas por meio de vocabulário e sintaxe específicas. Em outras palavras, dado um padrão de tripla qualquer, um processador SPARQL considera conjuntos de triplas no modelo RDF — alvo que corresponda ao padrão formulado na consulta.

SPARQL compreende uma estrutura inspirada em SQL como mostra o exemplo 4.7:

Exemplo 4.7: Formato de uma consulta genérica

```
PREFIX onto: http://dbpedia.org/ontology/ % especifica o namespace para abreviar
SELECT ?artista ?album         % var para selecionar
  FROM http://dbpedia.org/       % grafo a ser consultado
WHERE               % o grafo correspondente
 { ?album onto:producer :Liminha .
   ?album onto:musicalArtist ?artista . }
```

A sintaxe utiliza comandos similares aos utilizados em SQL combinados com a parte específica para WS. O exemplo 7 pode ser traduzido da seguinte forma:

- A primeira parte contém um artifício — PREFIX — muito conveniente para fornecer abreviações para *namespaces*. O que facilita na hora de escrever e ler as consultas sem necessariamente ter que lidar com os extensos URIs. Aqui, a sequência "http://dbpedia.org/ontology/" pode ser substituída por "onto";
- As variáveis — iniciadas com "?" — são vinculadas aos termos RDF e serão escolhidas via comando SELECT para apresentação do resultado;
- Além disso, pode-se indicar com o FROM de qual grafo se quer selecionar. Nesse caso, está indicado a "http://dbpedia.org/";
- Triplas RDF em WHERE estabelecem as condições para que as variáveis sejam selecionadas. No exemplo, a primeira tripla estipula que se quer álbuns que tem como produtor (*producer*) um indivíduo definido na ontologia da DBpedia que se chama "Liminha". Na tripla seguinte, declara-se que, dos álbuns que atendem à primeira condição, também se quer os artistas que gravaram tais álbuns;
- A consulta SPARQL é transferida para um SPARQL *endpoint*[23] que executa a consulta e retorna uma tabela com os valores relacionados às variáveis selecionadas, como mostra o Quadro 1.

Quadro 1: Resultado SPARQL

Artista	**Album**
:Titãs	:Enquanto_Houver_Sol
:Titãs	:Provas_de_Amor
:Titãs	:Vossa_Excelência
:Guilherme_Arantes	:Planeta_Água
:Os_Paralamas_do_Sucesso	:Alagados
:Titãs	:O_Inferno_São_Os_Outros

Fonte: Elaborado pelo Autor

Existem muitas opções que podem otimizar as consultas e uma extensa sintaxe a ser aprendida na utilização da SPARQL. Ela possibilita formular consultas complexas que auxiliam na exploração de bases de dados em RDF. A ideia aqui é apenas mostrar o conceito básico e geral para quem se inicia no mundo da WS. Para aqueles leitores que querem se aprofundar no assunto, o site da W3C — http://www.w3.org/TR/rdf-sparql-query/ — é sempre uma excelente fonte de consulta, pois possui documentação completa das tecnologias da WS, além de tutoriais e exemplos que demonstram as possibilidades da SPARQL.

[23]Um serviço Web que aceita consultas SPARQL pela Web e retorna os resultados em formatos diversos de acordo com a necessidade do usuário. A seguir, a URL que foi submetida ao endpoint SNORQL: http://dbpedia.org/snorql/?query=PREFIX+onto%3A+%3Chttp%3A%2F%2Fdbpedia.org%2Fontology%2F%3E%0D%0ASELECT+%3Fartista+%3Falbum+%0D%0A+++FROM+%3Chttp%3A%2F%2Fdbpedia.org%2F%3E%0D%0AWHERE%0D%0A{+%3Falbum+onto%3Aproducer+%3ALiminha+.%0D%0A%3Falbum+onto%3AmusicalArtist+%3Fartista+.+}%0D%0A.

4.1.5 Web Semântica na Ciência da Informação

A Web mudou a forma como o homem interage com a informação. Em praticamente duas décadas, houve uma revolução cultural que mudou o paradigma das fronteiras que separam tempo e espaço na perspectiva da informação. A possibilidade de criar infinitos repositórios de documentos, trouxeram, também, o caos para o profissional da informação e parece não haver saída sem ajuda da Tecnologia da Informação (TI). Nesse contexto, surge a visão conciliadora da Web Semântica que propõe um ambiente auto-organizável com a parceria entre homens e máquinas.

O autores Souza & Alvarenga (2004) afirmam que a proposta da WS é provocada pela dificuldade de se identificar de forma precisa a pertinência de documentos com tecnologias essencialmente voltadas para a apresentação de dados e exibição de conteúdos na Web. Em função disso, a descrição do conteúdo informacional é deficiente e pouco útil ao consumo por humanos e máquinas. Os autores mencionam a falta de estratégia de indexação na web e, em decorrência disso, a ineficiência da estratégia de recuperação da informação baseada, primariamente, em palavras-chaves. Assim, ainda acrescentam que a visão da WS seria atingida com a padronização de tecnologias, que abarquem infraestrutura, apresentação de dados, linguagens e descrição de metadados que viabilizem a criação de uma "língua" comum e o compartilhamento da informação para qualquer tipo de usuário — agentes computadorizados ou pessoas.

A perspectiva da representação e da recuperação de informação na Web é o foco de Feitosa (2005). Especialmente, em relação à informação legislativa, este autor busca trazer, para o contexto da CI, a aplicação de técnicas e conceitos de WS, de Terminologia e de Indexação. Como resultado do trabalho, foi proposta uma metodologia para a organização, a representação e a recuperação das normas legislativas com a utilização de controle terminológico e conceitual na organização e recuperação de informações legislativas, e com a utilização de tecnologias da Web Semântica.

Pickler (2007) considera a organização caótica do conteúdo na Web e de que forma a utilização de ontologias, para representar o tema das páginas da Web ou para controlar o vocabulário de forma semelhante aos tesauros, poderiam ser solução para adequação dos conteúdos neste ambiente. Ela conclui que, apesar de as ontologias fornecerem vocabulário comum a um domínio, elas são mais complexas e flexíveis que os tesauros, e, portanto, ferramentas com características distintas. O espaço da pesquisa em CI que compreende as ontologias e a WS é a investigação de Marcondes & Campos (2008). A CI tem longa tradição teórica, metodológica e práticas que possuem intersecção com a proposta da WS e para a construção de ontologias. Há vários caminhos para o desenvolvimento de ontologias: a padronização de linguagens para representação, elaboração de ontologias específicas referenciadas por outra mais genérica e consensuais. Há, também, vários antigos questionamentos relacionados ao conhecimento que precisam ser respondidos para que a CI contribua com todo seu potencial na visão da WS.

Robredo (2010) estuda qual é a situação da WS, no contexto da CI, em aplicações concretas. Ele propõe o foco na informação dividida em "predominantemente não documentarias" (as que mais têm se desenvolvido com base na Web semântica) e "predominantemente documentárias" (nas quais caberia depositar as maiores esperanças).

A primeira refere-se à parceria entre comunidade da WS da indústria, na qual busca acelerar a transferência de conhecimento da academia para aplicação direta na atividade industrial. Essa colaboração permite também a oportunidade aos pesquisadores da WS de refinar as pesquisas orientadas pela demanda na atividade industrial. O autor lamenta a incipiência desse tipo de parceria ainda existente no Brasil. A segunda caracteriza-se pela relação com atividades clássicas dentro da CI, como biblioteconomia, arquivologia, documentação, gestão da informação e do conhecimento, recuperação da informação etc. O autor frisa um importante ponto que é o reaparecimento de conceitos e princípios, muito utilizados na CI, com novas roupagens. O Quadro 2 mostra essas relações:

Quadro 2: Comparação de alguns termos e expressões da CI e da WS

Web / Web semântica	Ciência da Informação
Agentes de recomendação	Politicas de aquisição; Lei de Bradford
Ambiguidade / desambiguação	Sinonímia; polissemia. (nos tesauros: use; veja também) [para tal termo use tal outro]
Compartilhamento	Catálogos coletivos
Formato de comunicação e intercâmbio de dados	Norma ISO 2709, velha de mais de 40 anos, é atualizada periodicamente
Identificadores / Localizadores (URLs, URIs...)	Número de chamada, etc.
Inferência	Curvas estocásticas sobre séries históricas.
Interoperabilidade	OCLC, etc.
Linguagens de marcas; hyperlinks	Remissivas
Metadados	Dublin Core (metadados e qualificadores)
Ontologias (pobres filósofos se remexendo nos seus túmulos seculares!)	Clusters temáticos; métricas da informação
Open systems/ Sistemas abertos	OPACs
Parsers	Indexação automática
Relações entre conceitos	Descritores compostos; adjacência; proximidade
Reúso	Pesquisa bibliográfica; este trabalho
Reverse file	Arquivo invertido
Tags	Etiquetas; tags; campos de dados; Unesco/Unisist; CDS/ISIS; MARC
Taxonomias	Sistemas de classificação; tesauros
Template	Planilha / Folha de entrada

Fonte: (ROBREDO, 2010, p. 33)

Fica evidente a pluralidade das áreas de conhecimento como Linguística, Filosofia, Computação, Indexação, Classificação etc. na WS. Isso caracteriza que a CI está, inexoravelmente, entrelaçada com a WS, tanto para fornecer aporte teórico quanto técnicas que foram testadas por décadas. Observa-se, no Quadro 2, que as denominações, uma vez traduzidas para a nomenclatura da área, fazem parte da rotina de trabalho de profissionais de informação. Além disso, evidencia o envolvimento do cientista da informação na aquisição de mais e melhores tecnologias da informação para fortalecer a sinergia com outras áreas conhecimento que também tem a informação como sua principal matéria-prima.

Em suma, o nascimento da WS no berço da CC tende a afastar os pesquisadores da CI pela sua proposta excessivamente computacional. Entretanto, o que se verifica, na realidade,

é que o assunto é tão antigo quanto qualquer outra tentativa de organizar e representar o conhecimento humano. Como tal, a afinidade da CI com o tema é total e ela tem muito a contribuir e preencher os espaços com as teorias e técnicas já testadas. No fundo, temos um sistema de informação, de abrangência global, que precisa de profissionais com formação consistente no gerenciamento da informação e que, em tempos de tecnologia moderna, podem contar com a parceria das máquinas.

4.1.6 Considerações

Esse é um breve resumo de como tecnologias semânticas tornam o conteúdo da web mais "inteligente", declarando, explicitamente, a semântica e o significado das informações e, assim, possibilitando que elas sejam legíveis por máquinas.

A Web Semântica, conforme Berners-Lee, Hendler & Lassila (2001), não substitui a Web atual, mas estende as possibilidades de utilização. Por mais paradoxal que possa parecer, esta é construída se utilizando computadores, entretanto, direcionada para pessoas. Todo conteúdo, tais como imagens e textos em linguagem natural, ou mesmo o leiaute para apresentação da informação, é destinado ao consumo humano. Pois, por mais útil que o computador possa ser, ele não é capaz de ler, de compreender e de inferir relações da maneira como pessoas o fazem.

O desenvolvimento, principalmente na última década, das tecnologias que suportam a visão da WS encoraja a pesquisa nesse campo e estimula a interdisciplinaridade por causa do impacto em todos os níveis da sociedade moderna. A materialização da WS é, sem dúvida, uma quebra de paradigma na forma como lidamos com a informação. A máxima, creditada à Berners-Lee, é que em um mundo de dados interligados e codificados, mesmo com razoável semântica, não seria necessária a criação de bases de dados, mas apenas de conexão delas, isto é: não crie, conecte!

Apesar do forte viés computacional, a WS é um sonho percorrido há tempos e, portanto, de forte correlação com as pesquisas da CI. Contudo, é inegável a contribuição da ubiquidade da Web e da parceria perfeita para atingir os propósitos da área. Muitos questionamentos permeiam o cabedal teórico da área que abrange as ciências exatas, humanas e sociais, e, ainda, permanecem sem repostas. Mas, assim é a ciência que constrói novas camadas de saber sobre outras ou, muitas vezes, as substituem com a refutação de velhos argumentos.

4.2 Ontologia

O termo Ontologia nasceu na Filosofia e é tema de debate filosófico há alguns séculos. As discussões acerca da definição do termo e do significado vêm desde a época em que Aristóteles iniciava sua empreitada em classificar as coisas do mundo. Nessa disciplina, Ontologia representa a existência da essência, ou melhor, do Ser no mundo. Além dos limites da

Filosofia, outras áreas se apropriaram do termo e alteraram lhe a abrangência e o significado, como a CI, que trabalha ancorada nos avanços computacionais e na fundamentação filosófica, adotando, entretanto, o paradigma da Ciência da Computação que modela o mundo em frações calculáveis. É razoável que pesquisadores dessas áreas adotem o termo no sentido de descrever o mundo naquilo que pode ser representado no universo computacional.

Desde o início dos anos 90, a ontologia tem se tornado um assunto de pesquisa mais frequente, primeiramente nas comunidades de Inteligência Artificial (IA) com interesse na engenharia do conhecimento e Processamento de Linguagem Natural, seguida pela CI, que se interessa pela natureza, formas de organização e representação do conhecimento e da informação. A razão desse "sucesso" é, em grande parte, devido à promessa de compartilhamento de conhecimento consensual de um domínio e à interação entre homens e máquinas.

Assim, o termo "ontologia" tem sido utilizado por comunidades científicas de áreas como a CC e CI. Essa popularidade vem com os avanços tecnológicos que proporcionam a massificação e o compartilhamento de informações digitais em âmbito global e a consequente necessidade de organizá-las para, então, recuperá-las. Cada comunidade, entretanto, interpreta o termo da maneira que lhe convém e que atende suas necessidades.

4.2.1 Definição

Guarino & Giaretta (1995) e Lima-Marques (2006) propõem o uso de "Ontologia" — com a letra "O" inicial maiúscula, para denotar uma disciplina filosófica; enquanto todas as outras "ontologias", que se relacionam às bases de conhecimento projetadas para representar conhecimento compartilhado, são escritas com a letra "o" inicial minúscula.

Na Filosofia, Ontologia origina se na Metafísica[24] que, para Aristóteles, é a Filosofia Primeira que trata do estudo do "Ser" enquanto ser. Smith (2003) afirma que, apesar da ideia milenar, o termo foi cunhado somente em 1613, independentemente, pelos filósofos Rudolf Göckel, na obra *Lexicon philosophicum* e Jacob Lorhard em seu *Theatrum philosophicum.*

Como conceito filosófico, ela pode ser descrita ainda como a Ciência da Existência. Apoiando-se na definição de Chauí (2003), o termo Ontologia é formado por dois radicais gregos: *onto* que significa "o Ser na forma mais pura e real da existência" e *logia*, "estudo ou conhecimento". Assim, conforme Smith (2004), Ontologia significa estudo ou conhecimento do Ser, dos entes ou das coisas tais como são em si mesmas, real e verdadeiramente. Neste sentido, Ontologia tenta responder ao questionamento: Qual é a essência do Ser?

Ao contrário das ciências experimentais, que objetivam a descoberta e modelagem da realidade sob certa perspectiva, Ontologia se concentra na natureza e estrutura das coisas em si, independente de quaisquer outras considerações, até mesmo se elas realmente existem. Por exemplo, uma ontologia de unicórnios ou qualquer coisa fictícia: embora não existam

[24] Literalmente significa o que vem após a Física (SMITH, 2003)

de fato, a natureza e estrutura delas podem ser descritas em termos de relações e categorias gerais (GUARINO; OBERLE; STAAB, 2009).

Por outro lado, o uso dominante do termo ontologia na Ciência da Computação, refere-se a um tipo especial de objeto de informação ou artefato computacional. Nesse sentido, ela idealiza uma representação formal de recurso. De acordo com Gruber (1993b), a explicação de existência neste caso é pragmática: para sistemas computacionais, aquilo que existe é o que pode ser representado. Uma ontologia, então, é um modelo, uma representação do conhecimento.

Campos (2010) adiciona que os mecanismos de representação do conhecimento — responsáveis por facilitar os processos de formalização de objetos e respectivas relações em contextos pré-definidos — permitem a sistematização de conceitos e a elaboração de definições consistentes que se destinam a viabilizar inferências sobre o domínio.

Dentro de um determinado domínio, uma ontologia não se relaciona apenas à representação computacional. Ela também deve refletir o consenso sobre o conhecimento desse domínio. Assim sendo, os termos específicos de uma realidade específica, e a relação entre eles, podem ser fornecidos por uma ontologia (CORAZZON, 2012). Nesse sentido, ela é entendida como a padronização de conceitos e suas relações.

Dessa forma, outras áreas do conhecimento apropriaram-se do termo Ontologia e o sentido filosófico inicial migrou para algo menos abstrato que o entendimento do "Ser". Segundo Castel (2002), o campo da IA foi elaborado com o sentido de processo cognitivo artificial, de tal forma que uma representação da realidade está relacionada à percepção humana. Logo, a ontologia busca a divisão da realidade em pequenas partes para que seja factível entendê-la e processá-la.

Muitas definições foram apresentadas nas últimas décadas, mas a que melhor caracteriza a essência da ontologia, ou pelo menos é a mais citada no contexto da CC e da CI, é baseada na proposta de Gruber que se desdobra nas seguintes definições:

Definição 1 — *Gruber (1993b) propôs que ontologia é uma especificação de uma conceituação*[25].

Definição 2 — *Borst (1997) complementou afirmando que ontologia é uma especificação de uma conceituação compartilhada.*

Definição 3 — *Studer, Benjamins & Fensel (1998) combinaram essas duas definições estabelecendo que ontologia é uma especificação explícita e formal de uma conceituação compartilhada.*

Nesse mesmo trabalho em que se apresenta a definição 3 , os autores mostraram uma noção concisa, destacando a compreensão dos termos que compõem a definição com:

[25] Apesar de verificar em outros trabalhos a tradução para "conceitualização", optou-se por "conceituação" que remete ao mesmo significado: ato ou efeito de conceituar. Parece ser uma tradução do termo "*conceptualisation*" mais imediata e, portanto, facilita a leitura.

> Uma "conceituação" refere-se a um modelo abstrato de algum fenômeno do mundo por ter identificado conceitos relevantes de tal fenômeno. "Explícito" significa que tipos de conceitos utilizados e limitados no seu uso são explicitamente definidos. Por exemplo, no domínio médico, os conceitos são doenças e sintomas, as relações entre eles são causais e a limitação ou restrição é que a doença não pode ser a própria causa. "Formal" refere-se ao fato de que ontologias devem ser passíveis de leitura por máquinas, o que exclui a linguagem natural. "Compartilhado" reflete a noção de que uma ontologia captura o conhecimento consensual, isto é, ele não está privado ao indivíduo, mas à aceitação por um grupo[26]. (STUDER; BENJAMINS; FENSEL, 1998, p.25. Tradução nossa).

A noção de ontologia como forma de especificação é fundamental na elaboração conceitual e construção da Web semântica, pois o estabelecimento de limites a conceitos específicos e a definição das relações entre eles são essenciais para que a máquina possa "inferir" o significado da informação. No entanto, a definição 3 necessita de noção mais precisa dos termos que a compõe. O termo "conceituação", por exemplo, carrega ambiguidade e a transfere para o termo ontologia. Em função disso, procederemos à análise mais detalhada sobre os conceitos que ele abrange.

4.2.1.1 Conceituação

Guarino & Welty (2009) alertam para o fato de que uma conceituação trata de conceitos. Uma Ontologia, em sua essência, deve tratar de um conceito independente de seu estado no mundo, isto é, da intensão de um conceito que Dahlberg (1978b) sugere como a soma total de suas características. E também a soma total dos respectivos conceitos genéricos e das diferenças específicas ou características especificadoras.

Adicionalmente, Sowa (2000) acrescenta que intensão é o significado intrínseco ou conjunto de todos os atributos e propriedades de um conceito. Isto é, aquilo que generaliza; a extensão é o conjunto de conceitos mais específicos aos quais a intensão se aplica, ou seja, aquilo que diferencia e singulariza.

Isso significa que o foco deve ser no conceito fundamental das coisas que independem das características acessórias delas. Por exemplo, uma cadeira, em seu sentido mais elementar, é um conceito intensional. Já uma cadeira de aço é uma extensão do conceito fundamental. Entende-se então que a superclasse cadeira é imutável. As subclasses cadeira são alteradas conforme as mudanças em suas características.

[26] A 'conceptualisation' refers to an abstract model of some phenomenon in the world by having identified the relevant concepts of that phenomenon. 'Explicit' means that the type of concepts used, and the constraints on their use are explicitly defined. For example, in medical domains, the concepts are diseases and symptoms, the relations between them are causal and a constraint is that a disease cannot cause itself. 'Formal' refers to the fact that the ontology should be machine readable, which excludes natural language. 'Shared' reflects the notion that an ontology captures consensual knowledge, that is, it is not private to some individual, but accepted by a group.

4.2.1.2 Especificação formal e explícita

Nas aplicações práticas, como nas comunicações humanas, utilizam-se linguagens para referir aos elementos de uma conceituação. A linguagem utilizada compromete-se com a conceituação na medida em que permite o acesso ao intangível, isto é, ao conhecimento. De fato, conceituação refere-se à idealização de algo que se mantém inacessível na mente das pessoas. Para tanto, há que se descrever, detalhadamente, as características de determinadas entidades, ou o relacionamento entre elas, de forma a eliminar a ambiguidade e a viabilizar a estrutura, ou forma, adequada para acesso de pessoas ou máquinas. Em outras palavras, tornar o conceito implícito em explícito.

Dessa forma, possibilita-se que algo abstrato e restrito à mente de um indivíduo, ou de um grupo, possa ser expresso em linguagem adequada que restringirá as interpretações àquela relacionada ao domínio de interesse. Isto é, direciona-se o conceito ao modelo de mundo pretendido e excluem-se todos os outros.

Conclui-se, com base em Guarino & Welty (2009), que o grau de especificidade de uma contextualização depende da riqueza do domínio do discurso, do vocabulário escolhido e da expressividade da linguagem adotada para estabelecer um conjunto bem definido de sentenças que definirão os axiomas.

Desse comprometimento entre a linguagem utilizada e a conceituação pretendida, emerge o conceito de compromisso ontológico. Guarino & Welty (2009) afirmam que a noção de compromisso ontológico é uma extensão da noção padrão de modelo. Essa é uma descrição extensional de significado, aquela, uma descrição intensional. Portanto, o compromisso ontológico compreende a substituição da noção de modelo para conceituação.

A instituição de compromisso ontológico manifesta, então, um posicionamento referente aos objetos que se reconhecem como fundamentais numa visão de mundo e que viabilizam o diálogo sobre esse mundo, e, ainda, em função de como esses objetos são caracterizados. Essa visão de mundo, além de estar explícita nas relações apresentadas numa ontologia, também estará nas definições desse domínio.

4.2.1.3 Compartilhamento

Há que se destacar que a escolha de uma ontologia é uma decisão que tem por base o uso pretendido e a aceitação – ou compartilhamento – de determinada visão da realidade, uma vez que ela não abrange todos os domínios do conhecimento. Dessa maneira, o processo de escolha de determinada ontologia é orientado para a adequação da necessidade de indivíduos ou grupos.

Ressalta-se, também, que, em virtude do caráter formal da notação usada para a representação, a normalização do domínio pode eliminar incoerências envolvendo as ambiguidades inerentes à linguagem. Nesse sentido, as ontologias estabelecem um vocabulário comum e representam o conhecimento específico de forma explícita e em elevado nível de generalização

que lhes garante um desejado potencial de reutilização.

Guarino & Welty (2009) argumentam que alguém pode questionar a impossibilidade de se compartilharem conceituações inteiras, já que são privativas da mente das pessoas. De fato, o que se compartilha são aproximações de conceituações que são limitadas ao conjunto de conceitos e de relações explicitadas. Assim, reconhece-se que tais conceituações são parcialmente compartilhadas.

4.2.2 O conjunto das partes

Goméz-Peréz & Benjamins (1999) afirmam, referenciando Gruber (1993a), que o conhecimento nas ontologias pode ser formalizado utilizando cinco tipos de componentes: classes, relações, funções, axiomas e instâncias apresentados a seguir:

a) Conceitos são utilizados no sentido de Dahlberg (1978a)„ isto é, sentido amplo que pode descrever qualquer coisa sobre a qual se fala. De forma sucinta, eles representam as ideias básicas sobre o que se busca formalizar. Sowa (2000) ensina que conceitos podem ser referenciados como "categorias" no sentido filosófico, "domínio" na teoria de banco de dados, "tipos" na inteligência artificial e na lógica e "classes" nos sistemas orientados a objeto. Qualquer que seja a denominação, a seleção dos conceitos determina qualquer coisa que pode ser representada no universo computacional. A capacidade de generalização dessas estruturas está diretamente relacionada a distorções, restrições ou incompletude na seleção dos conceitos que constituem o núcleo da maioria das ontologias. Um conceito representa um grupo de indivíduos que compartilham características comuns que podem ser mais ou menos específicas, e isso remete à noção de extensão e de intensão de conceitos de Dahlberg (1978a). Esses conceitos ontológicos são fornecidos pela observação das coisas do mundo ou pelo raciocínio que dá sentido às abstrações. Eles podem, portanto, descrever objetos reais ou abstratos, tarefas, teorias, funções, estratégias, processo etc. Adicionalmente, um conceito pode ter um sub-conceito, frequentemente referido como subclasse ou um tipo de referência que estabelece um processo de herança das características do conceito mais amplo, mas que possui outras que o diferenciam. Isso pode ser ilustrado na inclusão de homem no conceito de mamífero.

b) Relações representam os tipos de interações e enlaces entre conceitos de um domínio. Elas descrevem a forma como um conceito se relaciona com outro. De forma generalizada, qualquer subconjunto de um produto cartesiano é relação. Tais relações podem ser estabelecidas hierarquicamente ou não. Por exemplo, hepatite é "subconceito de" doença e está "conectado" a tratamento num domínio médico. Nas relações, há sempre relações de aridade, ou lugares, variadas. No exemplo, "Kant é filósofo" existe apenas um lugar a ser preenchido, isto é, alguém ou "x é filósofo". Nesse caso, a sentença expressa uma característica ou atributo do elemento ou sujeito e, portanto, a expressão

denota a aridade 1 ou um lugar, mas não é uma relação, pois essa ocorre entre dois conceitos ou mais. Assim, quando se trata de relações binárias, "Marcos é pai de Sara", considera-se uma relação de aridade 2. Essa relação é um subconjunto de um conjunto maior que compreende todos os pares ordenados de pais e filhos, por exemplo, {<Adão, Abel>, <Adão, Caim>, <D. João VI, D. Pedro I>, <Marcos, Sara>...}. Generalizando esse raciocínio obtém-se qualquer subconjunto formado por pares, triplas, quádruplas e n-uplas. Quando se considera a relação "Santa Catarina situa-se entre o Paraná e o Rio Grande do Sul", tem-se uma relação de aridade 3 e assim por diante.

c) Funções são casos especiais de relações nas quais se determina um elemento decorrente do cálculo de uma função que considera outros elementos constantes da ontologia. Nesse âmbito, existem no mínimo dois conceitos, pois existe um domínio e um tipo especial de relação – função – que atua no contradomínio. Por exemplo, uma função "encontra-Primogênito" para determinar se um elemento é "Primogênito", deve-se levar em conta outros tipos de relações existentes tal como a ordem de nascimento de filhos de um casal – mais velho ou não – e a quantidade de irmãos. Sales, Campos & Gomes (2008) ensinam que o nome "função", usual na Ciência da Computação, é referenciado a relações complexas na CI, as quais correspondem a todas as relações associativas.

d) Axiomas são usados para modelar sentenças que estabelecem as relações que os elementos ontológicos devem cumprir. Pode-se estabelecer que um elemento estabeleça uma relação simétrica com outro, por exemplo, x é irmão de y se, e somente se, y é irmão de x. Esse tipo de declaração sempre será verdadeira numa dada interpretação.

e) Instâncias conduzem à individualização de um elemento. Por exemplo, Heitor Villa-Lobos é uma instância do conceito homem.

O estabelecimento metódico desses componentes constitui a essência da expressividade e da abrangência de uma ontologia. Todos esses componentes devem ser igualmente planejados e executados com parcimônia para que ela cumpra o papel de representar o conhecimento de um domínio.

4.2.2.1 Tipos e categorias de ontologia

De acordo com Guarino (1998), as ontologias podem ser organizadas conforme o nível de generalização que compreende suas definições:

Upper (top-level) ontology — representa conceitos muito gerais, independentes de domínio específico. Por exemplo, a árvore do conhecimento de Aristóteles para representar o mundo em categorias. Esse é o tipo mais abrangente, mais geral.

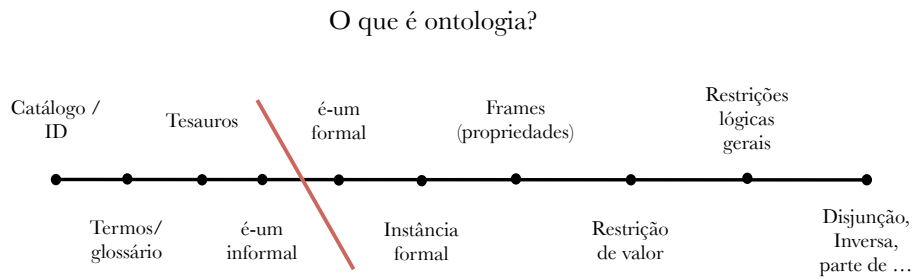
Domain Ontology — representa conceitos fundamentais de acordo com um domínio genérico como uma Ontologia de bebidas.

Task Ontology — representa conceitos fundamentais de acordo com uma atividade ou tarefa genérica como a fabricação de automóveis. Assim, como a Domain Ontology, o nível de generalização é intermediário.

Application Ontology — representa coneitos de um domínio especializado/específico. Ex. Ontologia de risco operacional. Esse é o nível mais específico.

Lassila & McGuinness (2001) categorizaram as ontologias por nível de expressividade em relação à semântica como mostra a Figura 18:

Figura 18: Espectro de Ontologias



Fonte: Adaptado de Lassila & McGuinness (2001)

Nessa classificação, os autores consideram, da esquerda para direita, que as nuances de sistemas de categorização que variam de objetos linguísticos com definições de significados informais, como vocabulários controlados — lista de termos, glossários, tesauros — até objetos lógicos com alto grau de formalização.

4.2.3 Considerações

Compartilhar conhecimento é se expressar no âmbito de uma língua comum. Expressar-se em uma língua comum é utilizar símbolos e conceitos comuns — sintaxe —, ter um acordo com os significados expressos por tais símbolos e conceitos — semântica —, desenvolver uma classificação de conceitos — taxonomia —, associações e relações desses conceitos — tesauros —, regras e conhecimento sobre quais relações são permitidas e quais fazem sentido — ontologia.

Para representar o conhecimento, é necessário representá-lo formalmente de modo que seja possível computá-lo, isto é, processar algo com esse conhecimento. Essa é precisamente a proposta das ontologias. Viu-se, nesta seção, detalhes sobre a definição de ontologias e suas consequências para representar o conhecimento. Para a CI, expressar conhecimento é formular um modelo abstrato de realidade por meio de uma linguagem legível por máquinas e que possa ser compartilhada. Por ora, considera-se a praticidade das ferramentas, desenvolvidas para esse ambiente, que preenchem os requisitos de representação do conhecimento.

Finalmente, o tema ontologias é controverso tanto no aspecto teórico, quanto no pragmático, pois existem diversas versões sobre o mesmo tema. Uma unanimidade no campo é que a elaboração das ontologias é complexa, cara e demorada. A compreensão das ontologias é vital para desenvolvimento da Web Semântica, por fornecer vocabulário e dispositivos de inferência para serem utilizados pelas máquinas, imitando, assim, o raciocínio humano para criação de novas classes a partir de deduções lógicas.

4.2.4 O despertar na Ciência da Informação

Uma relação intuitiva da CI com ontologias remete-nos aos achados de Dahlberg (1978b) com a Teoria do Conceito. Nada mais natural, pois essa teoria traz para o seio da CI o principal componente que as ontologias se propõem a representar: o conceito. Além disso, a autora impulsionou o assunto para área, presidindo a fundação, em 1989, da *International Society for Knowledge Organization* (ISKO) que protagonizou a consolidação de um importante campo de pesquisa conhecido como Organização do Conhecimento.

De fato, Sales, Campos & Gomes (2008) mencionam que representar o conhecimento é elemento essencial para viabilizar ontologias, pois compreende a análise semântica, de fundamental importância para a organização de sistemas de conhecimento, que se destinam a aperfeiçoar a recuperação da informação, lugar de destaque na pesquisa ligada à organização do conhecimento.

No âmbito da CI, dois trabalhos podem ser considerados como pioneiros em destacar a inserção do termo ontologia. Ambos apontam para o que parece ser uma "recente" preocupação de áreas como a Ciência da Computação e a Linguística, mas que já vem sendo enfrentada pelos pesquisadores da área.

No primeiro, em um trabalho que descreve com muita clareza as oportunidades e barreiras que a engenharia do conhecimento enfrenta e, por sua vez, possui forte relação com os problemas arrostados por cientistas da informação, Vickery (1997) destaca a utilização frequente do termo ontologia na literatura da CI e a influência na profissão.

Ela conclui o trabalho considerando que:

> A analogia com classificações bibliográficas e tesauros é obvia, embora existam igualmente diferenças porque os usos pretendidos para ontologias não são os mesmos para classificações e tesauros [...] este novo desenvolvimento de ferramentas no "nível do conhecimento" mostra mais uma vez a compreensão crescente da importância da análise semântica no processamento da informação. Os problemas que os cientistas da informação têm enfrentado por tanto tempo são agora enfrentados por uma comunidade mais abrangente de engenheiros do conhecimento[27] (VICKERY, 1997, p.284–285. Tradução nossa).

[27] The analogy with bibliographic classifications and thesauri is obvious, although there are equally obvious differences because the uses intended for ontologies are not the same as for classifications and thesauri. [...] this new development of tools at the 'knowledge level' shows once again the growing understanding of the importance of semantic analysis in information processing. The problems with which information scientists have for so long been struggling are now being faced by a wider community of knowledge engineers.

No segundo, Soergel (1999) aborda um assunto semelhante e demonstra uma preocupação com a falta de integração entre as várias comunidades científicas que pesquisam o mesmo objeto, porém com viés natural de suas formações e necessidades de pesquisa:

> A classificação tem sido usada há tempos nos sistemas de informação e bibliotecas para fornecer uma orientação ao usuário no sentido de contemplar sua necessidade de informação e estruturar os resultados de busca, funções largamente ignoradas pelas comunidades de recuperação de textos, mas que agora recebem crescente atenção no contexto de ajudar usuários a lidar com a grande soma de informação na Web. Recentemente, outras áreas, tais como a Inteligência Artificial, Processamento de Linguagem Natural e engenharia de softwares descobriram a necessidade de classificar, o que leva ao surgimento do que essas áreas chamam de ontologias[28] (SOERGEL, 1999, p.1120. Tradução nossa.).

Uma década depois, corroborando com Soergel e Vickery, sobre a integração entre comunidades científicas, Robredo (2010) em um dos seus últimos trabalhos, menciona que para que a CI possa lidar com a demanda ocasionada pela Web Semântica e, implicitamente, com ontologias, a área deve seguir aproveitando e refinando suas metodologias e exercendo sua vocação interdisciplinar, interagindo com outras disciplinas:

> Parece, pois, indispensável abordar, nas matérias que se relacionam com as descrições, o ensino de novas técnicas de leitura, de raciocínio. Sem entrar em detalhes, parece conveniente reaprender a pensar, a ler, a entender. Uma pitada de filosofia, de linguística, de semântica, de teoria das classificações (taxonomias, hierarquias, relações dos conceitos) e, naturalmente, lógica, merece ser considerada. [...] E, finalmente, um melhor conhecimento do básico das tecnologias da informação e da comunicação para poder trocar ideias e buscar caminhos para trabalhar e pesquisar em conjunto com outros profissionais, que também se interessam pela informação e seu processamento e uso (ROBREDO, 2010, p.41).

Assim, desde o final dos anos 90, o estudo das ontologias tem ocupado a agenda de profissionais da informação, haja vista que Marcondes & Campos (2008) relatam que enquanto o foco da ciência é explicar fenômenos naturais, a ontologia se ocupa da descrição, organização e sistematização do conhecimento alcançado pelas descobertas científicas. O que parece ser um lugar natural da CI dado que ela se ocupa, desde o nascedouro, da organização de domínios do conhecimento.

Os autores ainda destacam que a CI, ciência dos conteúdos registrados, já se ocupa tradicionalmente em responder questões teóricas, metodológicas e práticas que foram atualmente propostas pela Web Semântica e mais especificamente para construção de ontologias. Campos (2004) corrobora essa afirmação ao mencionar que áreas de pesquisa da CI, como a

[28] Classification has long been used in library and information systems to provide guidance to the user in clarifying her information need and to structure search results for browsing, functions largely ignored by the text retrieval community but now receiving increasing attention in the context of helping users to cope with the vast amount of information on the Web. Fairly recently, other fields, such as AI, natural language processing, and software engineering, have discovered the need for classification, leading to the rise of what these fields call ontologies

organização e modelização de domínios de conhecimento, fornecem aportes metodológicos que possibilitam o acesso e a organização de repositórios de documentos e o processamento semântico de informações em computadores.

Por fim, reafirmando a interdisciplinaridade da CI e seu domínio de atuação, que é o conhecimento e a informação registrados, e a relação com tecnologias e serviços relacionados que facilitam gestão e uso, Saracevic (2009) declara:

> A CI tem várias características gerais que são o motivo de sua evolução e existência. Elas são compartilhadas com muitas áreas modernas; primeira, a CI é interdisciplinar por natureza. Contudo, com vários avanços, relações com várias disciplinas, mudam com o passar do tempo. A interdisciplinaridade está longe de acabar; segunda, a CI está inexoravelmente conectada à tecnologia da informação. Um imperativo tecnológico compele e encoraja sua evolução, assim como de outras áreas e, além disso, da sociedade da informação como um todo; terceira, a CI é, com muitos outros campos, participante ativa na evolução da sociedade da informação. A CI tem uma forte dimensão humana e social, acima e além da tecnologia; quarta, enquanto a CI tem um forte componente de pesquisa que conduz avanços na área, ela possui um igualmente forte, senão mais forte, componente profissional orientado na direção de serviços de informação em diversos de ambientes. Muitas inovações vêm de profissionais na área; quinta, a CI também está conectada à indústria da informação, um ramo vital, altamente diversificado e um ramo global da economia[29] (SARACEVIC, 2009, p.15. Tradução nossa).

Pelo exposto, é desejado e natural que as ontologias, como instrumentos de representação e organização do conhecimento, ocupem lugar de destaque na pesquisa e incrementem a interação da CI com outras áreas do conhecimento.

4.2.4.1 Considerações

No contexto da CI, o termo ontologia não se refere somente à explicação sistemática da existência, mas da formalização e explicitação de conceitos de determinado domínio do conhecimento, podendo ser portável para atender às necessidades de indivíduos ou grupos.

De forma resumida, entende-se que uma conceituação é uma abstração, uma visão simplificada do mundo que se quer representar. Nesse sentido, as ontologias são acordos sociais que, entre outras coisas, permitem o intercâmbio de dados entre programas por conta da padronização com consequência da formalização, simplificam a unificação de

[29] Information science has several general characteristics that are the leitmotif of its evolution and existence. These are shared with many modern fields. First, information science is interdisciplinary in nature. However, with various advances, relations with various disciplines are changing over time. The interdisciplinary evolution is far from over. Second, information science is inexorably connected to information technology. A technological imperative is compelling and encouraging the evolution of information science, as is the evolution of a number of other fields, and moreover, of the information society as a whole. Third, information science is, with many other fields, an active participant in the evolution of the information society. Information science has a strong social and human dimension, above and beyond technology. Fourth, while information science has a strong research component that drives advances in the field, it also has an equally strong, if not an even stronger, professional component oriented toward information services in a number of environments. Many innovations come from professionals in the field. Fifth, information science is also connected with information industry, a vital, highly diversified and global branch of the economy.

representações diferentes, por trazer consenso aos conceitos representados e facilitam a comunicação entre pessoas pela normalização de jargões e idiossincrasias da área.

A tecnologia foi a grande catalizadora de pesquisas nessa área. O legado herdado pela CI provocou grandes avanços na organização e representação da informação. A aplicação das ontologias no universo da Web contribui para que a "interpretação" automática do conteúdo semântico de páginas dos sítios na Internet seja possível e amplie a colaboração entre máquinas e homens.

Como forma de sintetizar tudo o que é abordado com as nuances de cada área, descreve-se uma ontologia como modelo conceitual que, de alguma forma, envolve o uso de categorias semânticas, de conceitos e suas relações significativas em um domínio específico e, ainda, que utilize linguagem formal e lógica que viabilize a inferência automática.

Portanto, entende-se a ontologia como o meio de representação de determinada realidade pela conceituação sistemática, compartilhada e formal, dentro de um domínio do conhecimento, mas sem restrições disciplinares, buscando a padronização dos termos e conceitos. Assim, propõe-se uma visão ampla da ontologia, não propriamente como um artefato, mas como uma base conceitual.

4.3 Linguagem natural nos domínios das Ontologias

A importância da linguagem natural no contexto das ontologias é um tema que provoca alguns debates, mas não se descarta sua importância nessa conjuntura. A primeira é muito flexível, dinâmica, instrumento de comunicação entre os humanos e carregada de ambiguidades. A segunda, por outro lado, é formal e estática, na medida do possível, e não admite múltiplas interpretações para conceitos definidos em um domínio. Sob a perspectiva da Linguística, a conceituação, que é nuclear para a definição de ontologia, é um resultado da legitimação coletiva de um nicho sociocultural que nos cerca e que, portanto, reforça a importância da linguagem natural como elemento estruturante na representação do conhecimento. Melo & Bräscher (2011, p.93) adicionam:

> [...] a conceptualização encontra-se condicionada pela experiência de nosso corpo, do mundo externo e de nossa relação com o mundo. [...] poderíamos generalizar que alterações na ordem do real e nas representações do real exigem o compartilhamento pela maioria dos membros de uma organização, sociedade, cultura etc.

Essa discussão remete à excelência humana na utilização da linguagem natural e inclui uma parte da história da inteligência artificial que, há décadas, previu um agente inteligente o suficiente para compreender as armadilhas da subjetividade da língua. No atual estágio de evolução da área, este "ser" artificial, dotado das capacidades cognitivas do ser humano, ainda está por nascer, ou ser criado. O fato é que muito se avançou e o campo da Linguística, em parceria com a Computação, tem fornecido aplicações e teorias que conduzem pesquisas para a independência das máquinas.

Sob a perspectiva de ontologias, essa capacidade de processar a linguagem natural é especialmente desejada, dado que elas são criadas para manipulação, tanto por pessoas quanto por agentes computadorizados. É importante, portanto, que se faça a distinção entre os objetos linguísticos e ontológicos, bem como o universo ao qual cada um está restrito.

Segundo Cimiano, Völker & Buitelaar (2010), a principal diferença entre uma ontologia e um léxico, uma base de dados lexicais ou tesauros, é que esses são objetos linguísticos, enquanto ontologias são teorias lógicas, não linguísticas.

As relações estabelecidas nos objetos linguísticos estão no âmbito das relações lexicais tais como hiperonímia, hiponímia, meronímia, antonímia, sinonímia etc., ou seja, são definidas linguisticamente. Por outro lado, as relações percebidas nas ontologias são definidas no âmbito lógico e, como ensina Dahlberg (1978a), conceitos e propriedades são definidos em bases lógicas expressas por meio de condições suficientes e necessárias representadas por axiomas.

Hirst (2009) enfatiza que, afinal, uma ontologia é um conjunto de categorias de objetos ou ideias do mundo, com certas relações entre elas e, logo, não é um objeto linguístico. Um léxico, por outro lado, depende, por definição, da linguagem natural e do sentido da palavra.

Portanto, as relações dos objetos linguísticos agrupados em categorias não são definidas no âmbito lógico, mas linguisticamente e que, por conseguinte, não correspondem necessariamente aos conceitos que são definidos extensionalmente e intensionalmente sob o paradigma das bases de conhecimento. Contudo, destaca-se que a semântica de um conceito em uma ontologia, e mais particularmente em lógicas descritivas, é extensional. Todavia, conceitos também possuem dimensão intensional, axiomatizada em termos de condições suficientes e necessárias, e, muitas vezes, é acompanhada pela definição em linguagem natural. Essa, por sua vez, serve basicamente como interface ao usuário — humano — que necessita de uma caracterização precisa da intensão dos conceitos de maneira que se assegure de que serão utilizados da forma pretendida.

Outro ponto importante que diferencia ontologias da linguagem natural trata da ambiguidade inerente à última, mas que não pode existir na primeira. Cimiano, Völker & Buitelaar (2010) afirmam que as ontologias, como sistemas simbólicos, dependem de símbolos únicos para identificar significados por meio de suas conexões lógicas com outros símbolos. De tal forma, sob essa visão, conceitos são simplesmente símbolos não interpretados que podem ser manipulados conforme um conjunto de regras bem definidas.

Suscita-se, portanto, a afirmação de que pessoas interagem com as ontologias, explorando-as, povoando-as, consultando-as etc. De tal forma, conceitos em uma ontologia são designados usualmente por definições e rótulos — *labels* — em linguagem natural para serem utilizados por humanos. Völker, Hitzler & Cimiano (2007) declaram que uma ontologia sem rótulos em linguagem natural atribuídos às classes ou propriedades seria quase inútil, visto que sem esse tipo de fundamento, seria difícil ou mesmo impossível para humanos associar uma ontologia a suas conceituações, isto é, a ontologia não seria interpretável por humanos.

Naturalmente, essas definições e rótulos trazem consigo a ambiguidade inerente à linguagem natural e, portanto, os conceitos seriam ambíguos. Entretanto, essa ambiguidade relaciona-se somente aos rótulos e às definições de conceitos expressos em linguagem natural, não aos conceitos propriamente ditos, uma vez que são estruturas simbólicas, abstratas e sem ambiguidades, em seu sentido intensional.

Assim, conceitos podem ter definições mais ou menos precisas, o que pode oferecer margem para interpretações distintas do sentido intensional do conceito. Nesse caso, o que se questiona é se o conjunto de definições lógicas e axiomáticas é suficiente, isto é, se está em nível de granularidade satisfatório para definir e restringir a interpretação de forma inequívoca.

Uma importante conclusão que se obtém a partir dessa discussão é que as ontologias refletem as estruturas do mundo tal com elas são, independente da linguagem. Hirst (2009) sugere que a linguagem faz distinções que não são relevantes do ponto de vista ontológico. Por outro lado, há categorias que não são necessariamente lexicalizadas numa dada linguagem, mas certamente existem — ex. a classe de objetos que podem ser "montados". Assim, ontologias nunca deveriam ser específicas para uma certa linguagem, pois como elas estão relacionadas à existência, e não às lexicalizações, são, portanto, independentes da linguagem.

4.3.1 A linguagem em zeros e uns

Algumas correntes quantitativas e qualitativas debatem a eficiência ou, até mesmo, a necessidade dos métodos de um ou de outro para tratar a linguagem natural. Ambas concordam que os volumes de documentos eletrônicos são demasiados e demandam a utilização de procedimentos computacionais.

Uma ilustração desse debate pode ser vista no artigo com título provocador "*Do we need Linguistics When We Have Statistics?* [30]", mas que, na realidade, discute a abordagem computacional, comparada ao julgamento humano, para encontrar locuções no texto de forma a identificar grupos de adjetivos conceitualmente relacionados. Isto é, a aplicação de métodos linguísticos em sistemas estatísticos com o objetivo de melhorar o desempenho.

Apesar do título, o autor declara:

> [...] muitas formas de conhecimento linguístico tem uma contribuição positiva significante para o desempenho do sistema. Nós atribuímos (essa melhora) ao efeito combinado dos módulos de conhecimento linguístico a habilidade de nosso sistema executar uma classificação bem ajustada de adjetivos nas classes semânticas[31] (HATZIVASSILOGLOU, 1996, p.67. Tradução nossa).

[30] Precisamos de Linguística se temos Estatística? . Tradução nossa.

[31] ... many forms of linguistic knowledge have a significant positive contribution to the performance of the system. We attribute to the combined effect of the linguistic knowledge modules the ability of our system to perform fine-tuned classification of adjectives into semantic classes.

Esse trabalho foi "respondido" seis anos mais tarde com título também provocador "*Do we Need Statistics when we have Linguistics?*"[32] no qual o autor contrapõe as diferenças entre métodos quantitativos e qualitativos:

> Entre a comunidade de linguistas, métodos estatísticos ou, de modo geral, técnicas quantitativas são majoritariamente ignoradas ou evitadas pela falta de treinamento, medo ou hostilidade. As razões: i) tais técnicas simplesmente não se relacionam com Linguística, Filologia ou Humanidades; Estatística pertence ao domínio das ciências, Matemática e assemelhados; e/ou ii) existe um sentimento de que estes métodos destroem a "mágica" do texto literário[33] (GÓMEZ, 2002, p.234. Tradução nossa).

Diferenças a parte, o consenso é que a combinação de métodos quantitativos e qualitativos tem acrescentado muito aos processos de gestão da informação. Não há como negar a contribuição inestimável para a CI de autores, de formações variadas e com objetivos diversos em suas áreas de atuação, como Shannon (1948), Zipf (1949), Chomsky (1956), Baeza-Yates & Ribeiro-Neto (1999) e Saussure (2000). A história do Processamento de Linguagem Natural é permeada pela cooperação interdisciplinar que tem fornecido subsídios para o estabelecimento da área.

Para a CI, o Processamento de Linguagem Natural é muito relevante na pesquisa de Recuperação da Informação. À medida que os acervos eletrônicos aumentam, também cresce a procura por métodos capazes de tratar textos escritos em linguagem natural e indexá-los automaticamente. As técnicas já estão bem estabilizadas e prestam um grande serviço.

Lancaster (1968), já em 1968, traduz com bastante objetividade, declarando que a Recuperação da Informação é o termo, embora impreciso, convencionalmente aplicado ao tipo de atividade de representação, armazenamento e organização de informações com o propósito de recuperá-la posteriormente. Assim, um sistema de recuperação da informação não informa ao — isto é, altera o conhecimento do — usuário sobre o assunto de sua busca. Ele informa simplesmente a existência, ou não, e a localização de documentos relacionados a sua solicitação.

Baeza-Yates & Ribeiro-Neto (1999) atualizam a definição afirmando que a Recuperação da Informação lida como representação, armazenamento, organização e acesso de itens de informação. Estas informações podem ser referências de documentos: todo documento físico ou um simples parágrafo, páginas da Web, áudio, imagens, fotografias, vídeos etc.

Essa é a parte visível que a maioria dos usuários, e alguns profissionais, enxergam na recuperação da informação. Entretanto, a organização da informação até o ponto de ser recuperada por sistemas envolve atividades de processamento do texto para torná-lo compreensível aos aparatos computacionais. Portanto, de forma geral, o Processamento de

[32] Precisamos de Estatística se temos Linguística?. Tradução nossa.

[33] Among the linguistic community, statistical methods or more generally quantitative techniques are mostly ignored or avoided because of the lack of training, fear and dislike too. The reasons: (1) these techniques are just not related to linguistics, philology or humanities; statistics falls into the province of sciences, mathematics and the like; and/or (2) there is a feeling that these methods may destroy the "magic" in literary text.

Linguagem Natural desenvolve e pesquisa formas para o tratamento da linguagem utilizando procedimentos computacionais e linguísticos.

A natureza humana busca padrões em tudo. Esses podem ocorrer desde a observação do ciclo lunar até a observação de estruturas geométricas complexas. Essa avidez humana nos impõe a busca pela lógica existente em qualquer ambiente aparentemente desordenado e caótico. Nesse âmbito, um texto – com sua multiplicidade de elementos – apresenta uma ordem de construção que gera o padrão de um gênero textual (MELO; BRÄSCHER, 2011).

Savoy & Gaussier (2010) e Manning & Schütze (1999) afirmam que o componente básico da linguagem natural são palavras ou termos. Os quais constituem frases, parágrafos, capítulos e textos completos. Esses possuem padrões e estruturas — implícitas e explícitas — passíveis de serem reconhecidas e são considerados, na perspectiva da Linguística, a unidade da linguagem natural. Os autores acrescentam que, para execução do Processamento de Textos em Linguagem Natural, é importante examinar as camadas que o constituem e observar o papel que cada uma delas desempenha. São elas: morfologia, sintaxe e semântica. Silva et al. (2007) acrescentam mais uma como sendo a pragmática.

4.3.1.1 Morfologia

O objetivo do PNL no passo morfológico é a unificação de variantes morfológicas em uma mesma forma. Por exemplo, o verbo "ser" substituindo todas as conjugações de tempos verbais como "é", "sou", "fui", "serei" etc. Ou "ator", "atores", "atriz", "atrizes" para "ator".

Para tal fim são aplicados alguns algoritmos — chamados *stemmers* — que identificam as diversas formas e as transformam em uma única. Uma característica desse passo é a forte relação com o idioma a ser aplicado, pois não há um algoritmo universal, mas específico para determinada língua.

Existem vários algoritmos disponíveis, inclusive para o Português, que são essencialmente baseados em dicionários ou em análise de padrão morfológico ou *stemming procedures*[34] — baseados em detecção de padrões dos termos ou em probabilidades.

Savoy & Gaussier (2010) chamam a atenção para a dificuldade que pode haver em decorrência de variações e de erros ortográficos. Nesse caso, um deslize na escrita como "casa" e "caza" ou uma variação entre o Português brasileiro e o europeu como "fato" ou "facto" podem produzir resultados surpreendentes.

Neste sentido, é interessante observar a recente mudança na ortografia do Português, em função do acordo ortográfico da língua portuguesa de 1990[35] — que está em vigor desde 2009 no Brasil paralelamente ao antigo, mas que será obrigatório a partir de 2016 — impacta esse processo e deve demandar esforço para comunidade científica para "traduzir" as mudanças ocorridas para a forma de escrita atual como, por exemplo, "vôo" e "voo" ou "infra-estrutura" e "infraestrutura".

[34] Procedimentos para "encontrar" forma não flexionada. (tradução nossa)

[35] Informações sobre o acordo em http://www.portaldalinguaportuguesa.org/acordo.php

4.3.1.2 Sintaxe

A análise sintática investiga as relações entre os agrupamentos de termos e as funções gramaticais que cada um desempenha em frases ou sentenças de enunciado completo. Esse tipo de ordenação não é aleatório, mas obedece a regras, ou a uma gramática, que norteiam a língua em questão. O objetivo é capturar a essência dos elementos terminológicos combinados, tais como: sujeito, adjetivo, advérbios etc.

Manning & Schütze (1999) adicionam que linguistas agrupam termos de uma língua em classes que possuem comportamento sintático similar. Essas são chamadas categorias sintáticas ou gramaticais ou mais comumente como *Part of Speech*[36] (POS) no meio da comunidade de Processamento de Linguagem Natural.

Analisadores sintáticos, geralmente, identificam dependências entre palavras e suas funções para rotulá-las. É comum a utilização bases de dados contendo termos com as devidas anotações ou rótulos para o processo de automação desse passo. A função é traduzir estruturas como: "O pássaro é amarelo" para O (artigo definido masculino) pássaro (substantivo masculino) é (verbo de ligação) amarelo (predicativo do sujeito).

Silva et al. (2007) sintetizam informando que o processamento sintático não utiliza somente informações sintáticas que postula. Declaram que a autonomia em relação ao componente morfológico e semântica é evidente e isso determina a relevância que desempenha a análise sintática.

4.3.1.3 Semântica

Uma intuitiva metáfora, alusão a uma pintura, introduzindo a noção de semântica é apresentada a seguir:

> [...] percebe-se que fios de cores diferentes são entrelaçados em determinada disposição que a aparência final é uma amálgama de tonalidades que gera uma bela representação. Assim é um texto, ou parte dele, que surge igualmente da tessitura de signos diversos postos em relação uns com os outros dando forma a períodos. [...] Cada parágrafo constitui-se numa malha de associações de significados. Ele é uma unidade, um "todo" e não mera somatória de sentenças, [...] Assim, o significado literal dos vários lexemas que entram na composição das frases são "transcendidos" em favor da emergência de sentidos (significados contextuais) (MELO; BRÄSCHER, 2011, p.75–78).

Nesse sentido, Manning & Schütze (1999) definem a semântica como o estudo do significado das palavras, construções e discursos. Ainda, dividem-na em semântica lexical — que investiga o significado individual das palavras — e em como esses significados individuais combinados formam o significado de construções frasais ou sentenças.

Savoy & Gaussier (2010) destacam que, na falta de sistemas robustos o suficiente para fornecer uma análise semântica de construções frasais, a maioria dos trabalhos tem focado nas semânticas lexicais.

[36] Partes do discurso. (tradução nossa)

Complementarmente, Auger & Barrière (2008) esclarecem que as relações semânticas são o núcleo de qualquer sistema representacional e são elementos-chave para possibilitar a próxima geração de sistemas de processamento de informação com capacidade semântica e lógica. Aquisição, descrição e formalização de relações semânticas são necessidades essenciais para muitas aplicações de Processamento de Linguagem Natural.

4.3.1.4 Pragmática

Bräscher (1999) ensina que o componente pragmático relaciona-se com o contexto externo ou contexto extralinguístico que se relaciona à situação de enunciação, a fatores socioculturais, ao conhecimento enciclopédico. Pode-se interpretar como a investigação daquilo que realmente se pretende declarar quando se diz alguma coisa, isto é, o foco está no uso, não no significado.

Esse tema é debatido frequentemente na área por permanecer, segundo alguns autores, no limiar da semântica. Entretanto, parece que a menção do mundo extralinguístico extrapola as formas e estruturas — mundo linguístico — enquanto se observa que "a língua recupera de uma situação comunicativa diversos fatores que implicam a determinação de certa compreensão das palavras e sentenças" (SILVA et al., 2007, p.21).

Os autores ainda adicionam que o Processamento de Linguagem Natural privilegia o texto e não o discurso. Entretanto, os marcadores discursivos que trazem coesão e coerência ao texto podem ser identificados na forma de, por exemplo, conectivos que permitem a unidade dos sentidos. Essa noção é essencial para a eliminação de ruídos ou contradições.

De forma a exemplificar e estabelecer o domínio da pragmática, McCleary & Viotti (2009) mostram o seguinte exemplo: "A porta está aberta".

Situação 1 : Sem contexto – forma-se o conceito de um objeto físico que serve para marcar a entrada e saída de uma sala.

Situação 2 : Sala de aula com alunos conversando sem parar e o professor diz: " A porta está aberta" — nesse caso o conceito é de expulsar alguém da sala.

Situação 3 : Sala de aula e alguém bate na porta pelo lado de fora e o professor diz "A porta está aberta" — trata-se evidentemente de um convite para entrar.

Situação 4 : Sala de aula e há muito barulho do lado de fora e o professor se dirige para um aluno próximo à porta e diz: "A porta está aberta" — o significado é agora um pedido para que se feche a porta.

As autoras concluem o raciocínio, declarando que a análise da conceituação formada pelo uso dessa sentença é domínio da Pragmática.

Por fim, nesse ponto de vista, também se apoiam no pensamento de Wittgenstein e suas observações que sustentam a premissa de que a determinação do significado das palavras

se dá por meio da descrição de seu uso. Ou seja, o significado é moldável à situação de comunicação. Isso significa dizer que ele se insere nas questões essenciais das ontologias que são: conceituação, compartilhamento e consenso.

4.3.2 A Linguística na Ciência da Informação

Para Saussure (2000), o signo linguístico é o objeto da Linguística. Este é constituído da ligação de significado e de significante que se associam a um conceito e uma imagem acústica, respectivamente. Ambas são entidades abstratas, ou melhor, intelectuais no aparelho cognitivo dos falantes de uma língua.

Portanto, usa-se o signo para falar sobre as coisas que nos rodeiam, por exemplo, o signo ou palavra "cadeira" remete à representação mental — significado — privativa da mente tal qual a sequência fonológica ou o som produzido que identifica a "cadeira" — significante — em um idioma. Nesse momento está estabelecida uma relação simbólica entre significado e significante. Da mesma forma, essa relação se estende para textos.

Essa noção é compatível com ideia de conceituação abordada na seção 4.2.1.1 Conceituação que trata da definição de ontologia e ocupa lugar de destaque na Teoria do Conceito, formulada por Dahlberg (1978a) que, aliada a ideia[37] de Saussure (2000), via na linguagem um princípio de classificação, traz a discussão para um tema importante na CI que é a Teoria da Classificação.

Perceber conceitos na perspectiva de categorias é essencial para a organização de quaisquer estruturas conceituais e consequentemente para representar o conhecimento. A associação de conceitos aos signos organiza a experiência humana, por meio da língua, na tarefa de perceber o mundo agrupando as entidades, as situações, os eventos e relações em várias categorias diferentes (MCCLEARY; VIOTTI, 2009).

As autoras mencionam que Aristóteles é considerado o pai da noção clássica de categoria, na qual, as categorias são estabelecidas em função de traços ou atributos necessários e suficientes que entidades devem preencher para participar de uma dada categoria. Entretanto, essa abordagem, apesar da utilização nos dias de hoje, possui limitações, pois não há bastantes atributos necessários e suficientes para definir todas as categorias.

Nesse sentido, Wittgenstein (1999), o segundo, critica a categorização aristotélica por ela exigir limites bem definidos, o que na prática não se verifica. Ele observa que a formação do conceito acontece com a observação ou a experiência de fatos diferentes nas quais se percebem as semelhanças que ele chama de semelhanças de família.

> Pode-se dizer que o conceito "jogo" é um conceito com contornos imprecisos. – "Contudo, um conceito impreciso é realmente um conceito?" – Uma fotografia pouco nítida é realmente a imagem de uma pessoa? Sim, pode-se substituir com vantagem uma imagem pouco nítida por uma nítida? Não é a imagem pouco nítida justamente aquela de que, com frequência, precisamos?

[37] Para Saussure (2000), a língua é uma ferramenta que auxilia o ser humano a classificar as entidades físicas e abstratas do mundo e as relações entre elas.

> [...] Ver algo comum. Suponha que eu mostre a alguém diferentes quadros coloridos e diga: "A cor que você vê em todos se chama ocre". Essa é uma elucidação que é compreendida, enquanto a outra a procura e vê o que é comum àqueles quadros. Pode então olhar para o algo comum, apontar para ele (WITTGENSTEIN, 1999, p.54).
>
> [...]
>
> Reconhecemos que aquilo que chamamos de "frase", "linguagem", não é unidade formal que me representa, mas a família de estruturas mais ou menos aparentadas entre si (WITTGENSTEIN, 1999, p.64).

McCleary & Viotti (2009), apoiadas nas afirmações de Wittgenstein, declaram que, por causa dos limites difusos entre elementos de categorias, essas são representadas pelo conjunto de atributos que caracterizam o membro prototípico, isto é, aquele que reúne o consenso. Por exemplo, numa categoria "mamífero", um "cão" poderia representar um membro prototípico, por reunir características que se reconhecem nos "mamíferos". Por outro lado, "ornitorrinco" ou "morcego" certamente não seriam membros prototípicos da categoria.

Dessa forma, os fenômenos linguísticos permeiam a existência humana. Eles são essenciais para que o ser humano possa estabelecer vínculos entre conceitos das coisas que o rodeiam com ideia de categorias que são a base da organização da informação. Por sua vez, organizar compreende a representação e a classificação dos entes do mundo que viabilizam a estruturação dos espaços informacionais e a recuperação da informação relevante. Por tudo isso, a linguística ocupa lugar de destaque nos estudos relacionados à informação.

4.4 Processamento de Linguagem Natural

O Processamento de Linguagem Natural (PLN) é a tecnologia para lidar com o produto mais presente no cotidiano: a linguagem humana. Ela se apresenta em quase todos os tipos de mídia, tais como páginas na web, e-mails, tweets, documentação, jornais, artigos científicos etc. A evolução do PLN nas últimas décadas faz parte do cotidiano de todos, como os populares corretores ortográficos e gramaticais em editores de textos, filtros de spam. Assim, a PLN desempenha um papel fundamental na evolução da CI, em especial na RI.

Na tentativa de descobrir padrões implícitos no texto, Hearst (1999) constata que o texto exprime uma fonte de informação rica, extensa e compilada apropriadamente para consumo humano, mas complexa e inadequada para decodificação automática pelas máquinas. Nesse sentido, buscam-se alternativas para que o processo de cognição humana na leitura e interpretação de textos possa ser transposto, na forma de algoritmos, para que a máquina seja capaz de imitar o comportamento humano.

De acordo com Manning & Schütze (1999), o estudo da Linguística tem uma contribuição fundamental para solucionar esse problema, pois busca caracterizar e explicar a diversidade de observações linguísticas, às quais pessoas estão expostas, seja em diálogos, seja na escrita, seja em qualquer outro meio. Assim, ela se ocupa com a vertente cognitiva sobre como o homem adquire, produz e entende a linguagem; tenta entender a relação entre o discurso

linguístico e o mundo; e com a compreensão das estruturas linguísticas com as quais o homem se comunica.

Paralelamente, o desenvolvimento da informática tem possibilitado grandes avanços no estudo das linguagens naturais. Nesse contexto, apresenta-se a Linguística Computacional que reúne as competências necessárias das áreas da Linguística e da Computação, com foco na construção de sistemas especialistas em reconhecimento e produção de informação em linguagem natural. Mais especificamente, os estudos de PLN que têm como meta a interpretação e geração de informação nos diferentes aspectos da língua: sons, palavras, sentenças e discurso nos níveis estruturais, de significado e de uso Vieira & Lima (2001).

O PLN já existe há décadas e, nesse ínterim, desenvolveram-se várias técnicas tipicamente linguísticas, isto é, as sentenças do texto são separadas em partes gramaticais (sujeito, verbo, etc.) utilizando uma gramática formal ou um léxico, então a informação resultante é interpretada semanticamente e usada para extrair informação sobre o que foi escrito Kao & Poteet (2005). Corroborando a visão e a complementando, Dale (2010) afirma que o trabalho em PLN tende a ver o processo de análise da linguagem decomposto em fases, nas quais se espelham as distinções linguísticas teóricas entre sintaxe, semântica e pragmática.

4.4.1 Corpus como Recurso

Em um mundo ideal, um sistema automatizado deve ser capaz de construir representações precisas do uso, significados etc. de palavras a partir de fontes suficientes de dados. Isso é útil tanto para pessoas quanto para máquinas. Nesse contexto, a capacidade de encontrar informação precisa dentro do universo de informações muitas vezes irrelevantes é objetivo primordial da CI que tem produzido cada vez mais pesquisas nesse campo.

Cimiano, Völker & Buitelaar (2010) consideram a linguagem humana como atividade primária de transferência de conhecimento que ocorre, em grande parte, por meio da escrita. Além disso, a evolução tecnológica proporciona cada vez mais o acesso a grandes quantidades de textos nas mais variadas formas e fontes. Pragmaticamente, é possível obter coleções de textos substanciais relacionados ao domínio ao qual se está interessado. Entidades acadêmicas, públicas ou privadas possuem vastos repositórios de relatórios, manuais, e-mails e cartas e a utilização dessas coleções de texto como fonte parece um caminho natural.

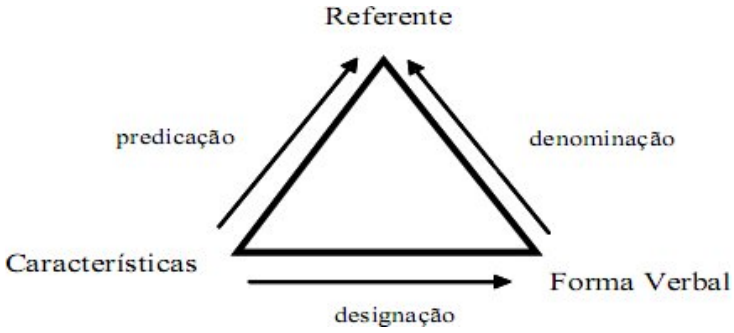
Uma pergunta relevante que pode surgir é quanto conhecimento está no texto? A resposta não é trivial, pois, como qualquer representação, um texto é um recorte da realidade que contém o conhecimento de mundo e os conceitos do autor que deve interagir com o leitor. Dessa forma, tem-se que:

> [...] todos os nossos conceitos são abstrações da realidade no sentido de que eles são produtos e instrumentos da habilidade humana de pensar e falar sobre a realidade limitados ao conhecimento dessa realidade. Contudo,

> eles diferem quanto ao grau de abstração, variando do mais específico e individual ao mais geral [38] Dahlberg (1978a, p.145. Tradução nossa.).

Assim, a produção textual é vista como uma forma de manter e evidenciar o conhecimento, mas somente a parte que está efetivamente explícita. Nesse sentido, Dahlberg (1978a) e Dahlberg (1979) mostra que o homem utiliza palavras para traduzir os pensamentos sobre objetos que estão a sua volta. Adicionalmente, emprega a linguagem para vincular os objetos aos respectivos conceitos. Assim, declaram-se os atributos necessários ou possíveis dos objetos para se alcançarem as características fundamentais dos conceitos.

Figura 19: Triângulo do Conceito



Fonte: Dahlberg (1978a, p.144)

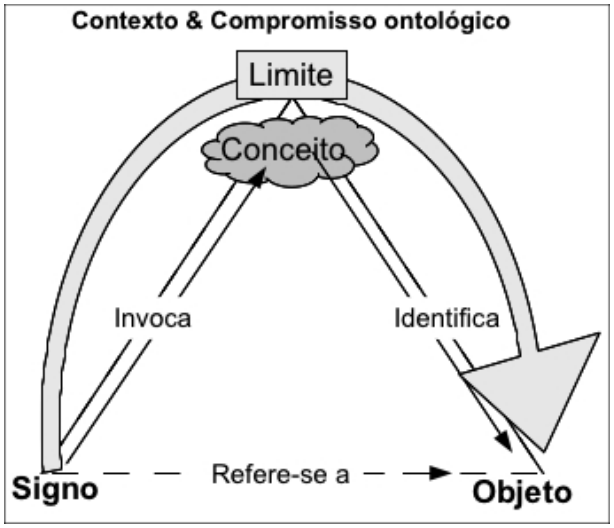
A Figura 19 mostra uma forma de esquematizar as condições necessárias para se identificar um conceito. Nela, são expressas as relações entre os elementos que compõem o triângulo: as características ou o conjunto de propriedades ou predicações atribuídas ao referente, que representa o objeto real — concreto ou abstrato — e a forma verbal utilizada por membros de um domínio que denota o referente.

A autora explica que na parte superior do triângulo, o referente indica onde estão as fontes de criação do conceito e no canto esquerdo inferior, o "significado" — aqui representando as características — que simboliza sua excelência na conceituação. Por fim, no canto direito inferior, o termo como a última parte a ser determinada que expressa forma verbal que simboliza o conceito. Além disso, nesse triângulo, observa-se que o conceito é, então, a conjunção desses elementos como um todo e não a junção de partes distintas, isto é, não há conceito sem a presença dos três elementos que o compõem.

Uma visão similar comum no contexto das ontologias, mas que pode dar a noção de quanto o significado pode estar expresso no texto, é apresentada por Guarino, Oberle & Staab (2009). Os autores revisitaram o trabalho de Ogden e Richards, *The Meaning of the meaning*, de 1923, no qual esses autores apresentavam um triângulo semiótico como uma forma de sistematizar o conhecimento, bem como as estratégias de estudo do significado.

[38] [...] all our concepts-.are abstractions of reality in the sense that they are abstractions of reality in the sense that they are products and instruments of man's ability to think and speak about reality to the extent permitted by his knowledge of reality. They differ, however, in the degree of abstraction, ranging from the most specific and individual ones to specific ones to the most general ones.

Figura 20: Triângulo Semiótico revisitado



Fonte: Guarino, Oberle & Staab (2009, p.16)

O triângulo semiótico revisitado, da Figura 20, propõe que agentes se comprometem a uma ontologia. Ao mesmo tempo, limitam as conclusões porventura associadas à comunicação de signos determinados, pois nem as relações permanecem nem as consequências lógicas do uso de signos estão implícitas na teoria lógica que especifica a ontologia. Isso só é possível por causa do comprometimento ontológico entre linguagem e conceituações. Dessa forma, as possíveis interpretações entre signos, conceitos e entidades do mundo real são convenientemente limitadas.

Com base nessas visões, pode-se pensar na escrita como signos da linguagem que expressam a visão de mundo de autores de textos. Esses signos necessitam de interpretação que remete aos conceitos e a seus correspondentes no mundo: os objetos. Essa interpretação, por sua vez, opera dentro dos limites permitidos pelo comprometimento ontológico entre a linguagem e os conceitos. Assim, a abordagem de métodos computacionais para extrair significados e descobrir padrões em *corpus* faz sentido. Diante do volume atual de informação disponível, extrair padrões dos membros prototípicos já pode ser de grande ajuda.

Freitas (2007) declara que a perspectiva wittgensteiniana influenciará as estatísticas do significado, pois com o *moto* "o significado está no uso", a aproximação ocorre pela substituição de "uso" pelo "corpus"; especificamente pelas adjacências de uma palavra. Ou seja, a probabilidade de ocorrência de palavras com significados semelhante em textos semelhantes é maior, enquanto as polissêmicas ocorrem em textos diferentes.

A autora ainda acrescenta:

> Assumindo, com Wittgenstein, que as palavras só têm sentido no uso, o lexicógrafo deve recorrer ao *corpus* como se fosse ele, o lexicógrafo, um "instrumento" cuja função é organizar o que está no *corpus* e "traduzir" esta organização para a linguagem de definição de dicionário. (FREITAS, 2007, p.28)

A perspectiva de Wittgenstein sobre a linguagem tem tomado corpo entre linguistas que buscam legitimar a abordagem com base em *corpus* para fundamentar teorias e metodologias com o auxílio do processamento da linguagem natural. Garrão (2006, p.139), fazendo referência à afirmação de Wittgenstein sobre a incompletude e parcialidade da linguagem na qual ele faz um paralelo entre a linguagem e uma cidade, declara:

> A questão frequentemente intocada é a de que a língua em si não é completa; sempre é possível acrescentar mais uma casa ou mesmo uma rua, o que torna um *corpus* um fragmento de algo já fragmentado. Portanto, não há como fugir desse paradoxo uma vez que a completude da língua também é algo inatingível.

Neste sentido, os recursos quantitativos propiciam bons recursos na detecção de estruturas convencionais da língua. Adiciona-se, ainda, que os avanços tecnológicos fornecem subsídios para se constatar que uma visão probabilística da língua é factível e vantajosa. Por isso, o *corpus* tanto representa uma base de dados para identificar tais convenções, como desempenha uma função preditiva ao fornecer os ambientes linguísticos tipicamente relacionados a elas.

Logo, parece razoável afirmar que o texto é de fato uma fonte confiável para alimentar os Sistemas de Informação. Contudo, o texto não está prontamente acessível aos computadores, pois é por meio da leitura e dos processos cognitivos que pessoas apreendem as informações contidas nesse formato. Por outro lado, métodos computacionais lidam com dados em formatos numéricos ou simbólicos preparados para as máquinas. Assim, o PLN oferece recursos para o reconhecimento e a produção da informação apresentada em linguagem natural e a manipulação da informação textual via máquinas como se verá adiante.

4.4.1.1 A descoberta de padrões no texto

A questão que pode surgir é: se ontologias são objetos lógicos, como construí-las a partir de *corpus* linguístico? Conforme se viu, existe a relação da língua com categorias. Essas expressam agrupamentos de coisas do mundo, isto é, da realidade. Nesse contexto, as ontologias descrevem a realidade e, portanto, são independentes da língua. Entretanto, os rótulos que descrevem os conceitos são construídos por objetos linguísticos — palavras ou termos — e esses são dependentes da língua.

Um texto é, frequentemente, chamado de dado não estruturado por não se apresentar na forma de uma base de dados convencional, estilo planilhas com colunas e linhas, nos domínios da informática. O que dizer da estrutura fornecida pela gramática de uma língua? Certamente, o termo não estruturado não faz justiça aos limites impostos ao texto pela regra que rege uma língua, a gramática.

E é com base nas regras gramaticais, que estruturam a língua, que a linguística computacional tem se esforçado em traduzir os padrões léxico-sintáticos para o mundo binário — de zeros e uns — das máquinas, com o objetivo de extrair automaticamente relações semânticas

de textos. Em um trabalho inovador e resultado surpreendente, Hearst (1992) tem inspirado muitas investigações sobre o tema por variadas comunidades científicas, inclusive da CI.

> Existem muitas formas que a estrutura de uma língua pode indicar o significado de itens lexicais, mas a dificuldade permanece em encontrar construções que frequente e confiavelmente indiquem a relação de interesse[39] (HEARST, 1992, p.540. Tradução nossa).

Ela propôs a extração automática da relação lexical de hiponímia na língua inglesa, a partir de um corpus, reconhecendo certos padrões léxico-sintáticos que satisfazem as seguintes condições:

a) Eles ocorrem frequentemente e em muitos gêneros literários;

b) Eles (quase) sempre indicam a relação de interesse;

c) Eles podem ser reconhecidos com pouca ou nenhuma pré-codificação do conhecimento.

Apesar do sucesso desse trabalho, a descoberta e proposição dos padrões é realizada manualmente e, portanto, limitada quanto às possibilidades de emprego de outros padrões. Baseando-se nessa restrição, treze anos mais tarde, Snow, Jurafsky & Ng (2005), propuseram a aquisição de relações de hiponímia ou hiperonímia utilizando a aprendizagem de máquina para substituir o trabalho manual e ampliar as possibilidades de padrões, induzidos por "raciocínio" automático, que facilitam a aquisição deste tipo de relação.

Em língua portuguesa, Freitas (2007) segue o trabalho proposto por Hearst, que fornece seis pistas textuais, e o adapta para nosso idioma. A ideia de Hearst é comparar sintagmas nominais que apresentem uma estrutura sintática que possa se incluir na relação de hiponímia. Por exemplo, na frase: "Havia muitos tipos de cães tais como poodle, labrador, pastor-alemão etc." Hearst identificou que a expressão "tais como" — em língua inglesa — se apresentava como um padrão para identificar a relação de hiponímia. Isto é, "cão" seria o hiperônimo e "poodle, labrador e pastor-alemão" seriam hipônimos. Assim, o primeiro padrão sugerido foi:

$$NP_0\ such\ as\ NP_1, NP_2..., (and|or)NP_i,$$

no qual NP_0 corresponde ao hiperônimo e os demais, $NP_s(NP_1, NP_2...NP_i)$, aos hipônimos.

O trabalho de Freitas (2007) descartou três regras de Hearst por não terem aplicação no Português. Assim, utilizou três padrões originais que foram traduzidos e adaptados para o idioma. Além disso, propôs mais três padrões específicos de nossa língua que, ao final, totalizaram seis padrões.

A autora destaca o resultado positivo da aplicação de padrões léxico-sintáticos na automação da aquisição de relações semânticas. Ressalta a importância de se trabalhar com domínios específicos como forma de aumentar o desempenho das regras e aponta o espaço para pesquisas que investiguem outros tipos de relação semântica aplicadas ao Português.

[39] *There are many ways that the structure of a language can indicate the meanings of lexical items, but the difficulty lies in finding constructions that frequently and reliably indicate the relation of interest.*

4.4.2 Conceitos fundamentais

Um fato surpreendente sobre o conhecimento linguístico é que a maioria das tarefas de processamento da linguagem pode ser vista como a resolução da ambiguidade nos variados tipos. Jurafsky & Martin (2009) afirmam que para entender o comportamento complexo da linguagem, são necessários os seguintes conhecimentos linguísticos:

- Fonética e fonologia — conhecimento sobre sons linguísticos;
- Morfologia — conhecimento das unidades mínimas de significado das palavras;
- Sintaxe — conhecimento da relação de estrutura entre as palavras;
- Semântica — conhecimento do significado;
- Pragmática — conhecimento da relação dos significados com os objetivos e intenções do falante;
- Discurso — conhecimento sobre as unidades linguísticas maiores que um enunciado único.

4.4.2.1 Ambiguidade

Dizer que algo é ambíguo significa que há múltiplas estruturas linguísticas alternativas para uma determinada entrada — termo, expressão ou frase. Cimiano, Unger & McCrae (2014) afirmam que a ambiguidade compreende todos os casos nos quais expressões de linguagem natural possuem mais de um significado. Isso pode ser atribuído a significados lexicais alternativos, a propriedades estruturais ou à combinação de ambos. A primeira apresenta-se quando o mundo lexical e do significado permite múltiplas interpretações do mesmo termo, como "banco". A segunda, decorrente de ambiguidades sintáticas, exemplo "Eu lhe contei" — "contei a você" ou "contei a ele".

Bräscher (2002) declara que a ambiguidade ocorre por vários fatores, tais como polissemia, homografia, policategorização, relação contextual e estrutura sintática da frase. A primeira decorre da atribuição de mais de um significado a uma palavra — letra: de música ou signo do alfabeto; a segunda, de dois ou mais vocábulos de significados distintos que convergem para a mesma representação ortográfica — manga: fruta ou parte da camisa; a terceira, a palavra pertence a mais de uma categoria gramatical — verão: substantivo ou verbo; a quarta, pela interpretação em função do contexto — vetor: em Matemática ou em Biologia; a última, pela estrutura da frase e as relações entre seus constituintes — Eu acertei o homem de pijama: eu vestia o pijama ou o homem o vestia.

Considerando a sentença "Eu fiz sua pata"[40], um falante fluente pode inferir vários sentidos que exemplificam a ambiguidade em algum nível:

[40] O exemplo extraído de Jurafsky & Martin (2009) em inglês: *I made her duck* adaptado para o português perde um dos significados originais, mas adiciona outros.

a) Eu cozinhei uma ave aquática para você.

b) Eu cozinhei uma ave aquática que lhe pertence (ou a ele/ela).

c) Eu criei o pato (de gesso, plástico, madeira etc.) que você possui.

d) Eu cuidei das unhas das mãos da pessoa com quem falo (informal, gíria).

e) Eu usei minha vara mágica e criei uma pata do nada que lhe pertence.

Esses diferentes significados ocorrem por causa da ambiguidade. Primeiro, as palavras *pata* e *sua* são morfologicamente ambíguos. "Pata" pode ser o feminino de pato ou o "pé" de um animal. Segundo, o pronome possessivo "sua" pode significar que o diálogo se dá na segunda pessoa (tu/você) ou na terceira pessoa (ele/ela). Finalmente, o verbo "fazer" é semanticamente ambíguo, pois, segundo o dicionário Priberam[41], possui quarenta significados, entre eles, criar, preparar, cozinhar e cuidar que remetem ao exemplo aqui dado.

Esses vários significados são decididos de forma razoavelmente simples num diálogo, mas o processo de decisão em um mecanismo automático de inferência deve ser precedido de modelos que alimentam esses motores de inferência. Decidir se "banco" é um substantivo ou um verbo — primeira conjugação do verbo "bancar" — pode ser resolvido com anotações da classe lexical, ou *part-of-speech tagging*. Da mesma forma, decidir as várias possibilidades de interpretação do verbo "fazer" pode ser resolvido por desambiguação do sentido da palavra. Enfim, há uma variedade de tarefas que são apropriadas para o tratamento da ambiguidade que ocorre no âmbito lexical, sintático e semântico e que serão tratadas no decorrer dessa pesquisa.

4.4.2.2 Noções elementares

Todo documento na acepção de informação registrada está sujeito a diferentes abordagens. De acordo com Miranda (2003), seria razoável apontar duas direções complementares e interdependentes: uma direcionada para o conteúdo enquanto tal e a outra para a estrutura do próprio documento. Nesse sentido, a PLN pode ter uma abordagem voltada tanto para a compreensão do conteúdo, quanto para a análise da estrutura de documentos, isto é, análise estatística descritiva, e ambas visam identificar padrões implícitos em uma grande coleção de documentos.

Segundo Bird, Klein & Loper (2009) e Woodfield (2004), é importante explorar essas duas vertentes de análise e, para tanto, alguns conceitos básicos são necessários:

a) *Token* é uma sucessão contígua de caracteres que não contém um separador. Um separador é um caractere especial tal como um espaço em branco ou sinal de pontuação.

[41] http://www.priberam.pt/dlpo/fazer

b) Termo é composto por um ou mais *tokens* com significado específico numa dada linguagem.

c) Documento consiste em um grupo de termos.

d) *Corpus* é uma coleção de documentos.

Uma abordagem baseada em estatísticas descritivas pode viabilizar a descrição de um *corpus* pelo reconhecimento das características físicas dos documentos. Tais características podem incluir o tamanho do documento em relação ao número de palavras, sentenças, caracteres, distribuição do comprimento de palavras, a contagem de frequência absoluta e relativa de palavras-chave e assim por diante. Esse tipo de análise favorece a identificação de características que separam ou distinguem um documento de outros em uma coleção.

4.4.3 Características de um Documento

Um documento consiste essencialmente de elementos como letras, palavras, sentenças, parágrafos, pontuação e possíveis itens estruturais tais como capítulos e secções. Esses elementos podem ser contados como, por exemplo, o número de caracteres, palavras e sentenças ou, ainda, resumidos na forma de medidas estatísticas como média, mediana ou variância. Esse conjunto de medidas são chamados de resumos estatísticos descritivos.

Resumos estatísticos convertem textos em vetores numéricos. Essas representações de características quantitativas do texto são poderosas ferramentas para exploração dos dados e ainda podem fornecer informações importantes para adequação do *corpus* e da escolha de técnicas computacionais mais apropriadas. Entretanto, limitar a análise somente a esses valores significa ignorar as características linguísticas que são dominantes na identificação de documentos. Portanto, há que se considerar a quantidade de informação perdida quando se ignoram os termos e seus significados em determinados contextos.

Nesse processo de conversão, é necessário que o texto seja, inicialmente, transformado em unidades de informação que possam, ao mesmo tempo, conter a informação original do texto e estar apropriada ao manuseio por técnicas estatísticas ou computacionais. Essa extração de unidades de informação foi apresentada por Wakefield (2004) e complementada por Bird, Klein & Loper (2009) e Woodfield (2004) segundo a ordem de complexidade, como se observa no Quadro 3.

Neste contexto, a extração de unidades de informação representa um ou mais passos no processo de tratamento linguagem natural e estão destacadamente relacionados a uma linguagem específica. Importante observar também que a PLN empresta alguns algoritmos para solucionar problemas relacionados à ambiguidade presente na grande maioria dos textos.

Quadro 3: Extração de unidades de informação por ordem de complexidade

Atividade	Descrição	Exemplo
Extração de Token	O texto é separado em palavras sem considerar seu significado.	Ciência da Informação ⇒ ciência, da, informação
Extração de Termos	O texto é separado em palavras considerando seu significado dentro de uma linguagem específica.	Token + linguagem específica ⇒ termo (mineração de texto)
Extração de Conceitos	Conceito indica assuntos ou tópicos contidos em um texto e a extração de conceitos revela somente se tais conceitos estão presentes em um texto sem que haja a preocupação com seus detalhes.	Um artigo pode tratar de astronomia, carros ou câncer
Extração de Entidades	As entidades representam pessoas, localidades ou coisas em um texto, frequentemente, na forma de substantivos.	Pessoa ⇒ Senhor Coelho Animal ⇒ coelho
Extração de fato pontual	Um fato pontual relaciona uma entidade a uma ação. Esses pares podem estar separados por outros termos e, portanto, devem extraídos por sistemas capazes de reconhecer os fenômenos linguísticos.	sujeito ⇒ ação terrorista ⇒ explodiu
Extração de fato complexo	A compreensão envolve associações e inferências características de humanos	compreensão de linguagem natural

Fonte: Adaptado de Wakefield (2004)

4.4.4 Processamento do Texto

Uma simples mensagem enviada por correio eletrônico pode apresentar alguns desafios sob a perspectiva da máquina. Por exemplo, um texto mal formulado — com ideias confusas, erros ortográficos, abreviaturas fora do padrão, jargão técnico e peculiaridades do autor — contribui para aumentar a complexidade na tarefa de decodificação automática. Conforme Jurafsky & Martin (2009), o sucesso na remoção de ambiguidades na fase de preparação dos dados está diretamente relacionado ao êxito no tratamento dos algoritmos que lidam com a linguagem natural.

4.4.4.1 Coleta de Dados

Por motivos óbvios, uma das primeiras tarefas é a coleta de dados. A questão geral aqui é encontrar ou criar um *corpus* que seja uma amostra representativa da população de interesse. Dependendo da especificidade do trabalho, encontrar o dado linguístico apropriado pode se tornar um grande obstáculo. Por exemplo, fazer um estudo sobre uma língua com pouco recurso apropriado, como o Português. Nesse caso, o que seria uma simples tarefa de escolher uma fonte de dados pode se tornar numa penosa tarefa de garimpagem de material relevante

ao objetivo do estudo.

Os acervos de documentos relevantes ao estudo podem ser previamente conhecidos se, por exemplo, se analisa um *corpus* de um domínio bem definido como é o caso relacionado à Medicina ou a uma documentação de uma empresa específica. Por outro lado, encontrar os dados pode ser visto como parte do problema, como no caso de estudos direcionados a documentos na web. Há que se levar em conta características como o formato, a língua, a disponibilidade — Deep Web[42] — e a tecnologia necessária para fazer esse tipo de coleta.

A leitura automatizada de fontes textuais implica dificuldades práticas em relação aos textos que podem se apresentar em vários formatos e idiomas. Quando estendemos essa tarefa aparentemente simples, para a web, a complexidade pode ser multiplicada várias vezes. Nesse sentido, (WEISS et al., 2005) afirmam que a principal tarefa nessa fase é eliminar ruídos e assegurar que a amostra seja de boa qualidade. Assim como nos dados não textuais, a parcimônia deve ser levada a serio, pois a intervenção humana pode comprometer a integridade dos dados no processo de coleta.

4.4.4.2 Padronização de Documentos

Considerando que o PLN é realizado com auxilio de computadores, a tarefa de padronização pode ser vista como a transformação de arquivos de textos brutos, uma sequência de bits digitais, em arquivos homogêneos com conteúdo composto de sequências bem definidas de unidades significativas do ponto de vista linguístico e apropriado ao processamento automático. Essa etapa é caracterizada pela triagem de documentos e pode ser realizada tanto na apresentação estética do documento, quanto no detalhamento de propriedades mais técnicas como o sistema de codificação.

Segundo Palmer (2010), a definição da codificação de caracteres é fundamental para se ter um arquivo legível por máquinas. Numa língua como o Português que possui caracteres especiais como “ç, ã, á, à” entre outros, essa codificação deve estar definida em um arquivo que relaciona um ou mais bytes para os caracteres conhecidos. Estendendo o processo, esse tipo de codificação está relacionado à identificação do idioma, pois idiomas como o Chinês e o Português, por exemplo, implicam estratégias e aplicações diferentes para o processamento linguístico.

Outro estágio é a definição do formato que será utilizado para todos os documentos. Essa tarefa é importante, visto que alguns aplicativos necessitam de formatos pré-definidos, pois a leitura de um arquivo PDF pode não ser compatível com um arquivo DOC. Ainda, a preocupação com o conteúdo pode melhorar a experiência no decorrer do processo, pois pode ser necessário o descarte de elementos não relevantes ou não desejáveis para o propósito da análise, tais como figuras, cabeçalhos, rodapés, marcações de linguagens da web como HTML, vídeos, tabelas, links, entre outros.

[42] http://pt.wikipedia.org/wiki/Deep_Web - também chamada de Web invisível é a parte da web que está oculta aos mecanismos de indexação dos mecanismos de busca padrão como Google e Yahoo!.

A complexidade envolvida nessa fase é razoavelmente baixa, mas de extrema importância para o produto final. Atualmente, a maioria dos pacotes de programas para computadores voltados para o PNL oferecem soluções que não demandam conhecimentos especializados por parte do usuário. Vale a pena lembrar que essa tarefa pode representar um tempo precioso no processo, mas o resultado dessa triagem de documentos é um *corpus* textual bem definido, organizado por línguas e apropriado para manipulações por máquinas e análises mais sofisticadas.

4.4.4.3 Tokenization

Um texto possui um fluxo ordenado de palavras que seguem as normas linguísticas de um idioma para que ele faça sentido para o leitor. No entanto, para o propósito de manipulação automática do texto, é necessária a divisão do texto em unidades conhecidas como tokens. Tais unidades podem ser palavras, números ou sinais de pontuação. A esse processo dá-se o nome de tokenization.

Em Português, o que diferencia um token do outro são os espaços entre eles e frequentemente os algoritmos que executam a divisão do texto em tokens utilizam o espaço como delimitador. Jurafsky & Martin (2009) alertam que aqui, também, se requer cuidado, pois há termos compostos que quando separados possuem significados diferentes. Por exemplo, "Ponto Frio" representa uma conhecida rede de varejo, contudo, os termos "Ponto" ou "Frio" não remetem à ideia de uma rede de varejo.

Segundo Manning & Schütze (1999) e Palmer (2010) os sinais de pontuação são simplesmente retirados, mas é importante observar que os sinais de pontuação podem trazer informação sobre a macro estrutura do texto e, por isso, não devem ser negligenciados, como em textos financeiros que possuem diversos padrões de datas e números apresentados com pontos e vírgulas. Dessa forma, para obtenção de melhores resultados, deve-se adequar o programa que executa o trabalho de separação dos termos em função do texto que será tratado, caso contrário, muito trabalho deverá ser executado nos tokens adquiridos.

Em algum tipo de trabalho, o profissional pode estar interessado não em termos isolados, mas numa sequência que possua um significado como uma frase. Manning & Schütze (1999), Jurafsky & Martin (2009) mencionam que o processo de tokenization pode ser restrito a palavras, segmentação de termos, ou de frases. Nesta, para reconhecer automaticamente os limites de uma frase geralmente se utilizam os sinais de ponto, interrogação, exclamação e outros, contudo a tarefa é mais complexa do que parece e esse tema tem sido muito explorado.

4.4.4.4 Padronização de Conteúdo

Além de formatos e idiomas, os textos devem passar por uma padronização do conteúdo. Uma série de variações podem ocorrer no padrão de escrita, sem que se altere o significado, como Sr. e Senhor. Depois da popularização das redes sociais virtuais e mensagens de texto

via celulares, esse tipo de variação tem aumentado diariamente. Termos que, na norma culta, não seriam aceitáveis são muito frequentes no cotidiano de qualquer pessoa. Expressões como “D+” significando “demais” ou “vc’ para “você” são perfeitamente compreensíveis e aceitáveis na linguagem escrita informal.

Para o computador, Sr. e Senhor são dois termos diferentes até que se padronize o conteúdo, isto é, indique a forma que será empregada em todo *corpus* para utilização de algoritmos computacionais. Além disso, outros procedimentos para padronizar a coleção de documentos podem contribuir para facilitar o tratamento automático da linguagem como:

- Conversão de abreviações não padronizadas em termos válidos para evitar a ambiguidade. Ex. Gal. $\Rightarrow$ Galicismo, Gal. $\Rightarrow$ Galego;
- Inferência de potenciais problemas em relação à mesma palavra escrita, ora em letras maiúsculas, ora em minúsculas. Isso é necessário, pois a representação interna dos computadores, isto é, a representação em bytes é diferente, como por exemplo, Documento $\neq$ documento;
- Identificação de seções para documentos bem estruturados. Ex. Introdução, referência, sumário;
- Identificação de palavras-chave em documentos;
- Identificação e anotação de expressões sinônimas: "carro", "caminhão", "ônibus", "veículo”;
- Tratamento de sinais de pontuação que podem indicar a macro estrutura textual e, portanto, não devem ser simplesmente descartados. Ex. “,”, ”.”, ”;”, “:”, “<”, “>”, “!”, “?”;
- Compilação de estatísticas dos termos para fins de exploração. Ex. comprimento de termos, frequência de termos, frequência de combinação de termos.

4.4.4.5 Criação de Bases de Apoio

Na leitura de qualquer texto, o leitor se depara com uma série de símbolos, termos e expressões que ele reconhece e, espera-se, outra pequena parte, que lhe exige a consulta de uma fonte externa, como um dicionário, uma enciclopédia ou qualquer outro meio, para entender o significado daquilo que não fazia parte de seu conhecimento. A situação é análoga para o PLN. Muito trabalho para produzir conhecimento linguístico já foi realizado e, atualmente, há uma quantidade considerável de recursos linguísticos, tais como dicionários e tesauros que podem ser utilizados para auxiliar no trabalho em PLN.

As listas são muito comuns e auxiliam no processamento, pois o objetivo delas é diminuir a quantidade de termos na análise. São elas:

- Lista de Stopwords — são listas de apoio que contém as stopwords que são as palavras que ocorrem com muita frequência em uma dada língua. Um exemplo dessas palavras são os artigos ou as preposições que, na maioria dos textos, apresentam maior ocorrência. Uma vez que esses termos ocorrem em todos os documentos, seu poder de discriminação é muito pequeno e, por isto, eles são descartados da análise sem prejuízo de perda de informação;
- Lista de Startwords — ao contrário das listas de stopwords, essa lista contém os termos que caracterizam o domínio do assunto a ser pesquisado. Somente os termos nela contidos serão considerados. Esse tipo de lista é utilizado quando se examinam coleções altamente especializadas, dominadas pelo jargão técnico e são construídas por especialistas de domínio, com o objetivo de otimizar o tratamento da linguagem natural.

Uma preocupação natural é unificar todas as palavras que possuem o mesmo significado. Uma solução é utilizar dicionários de sinônimos ou tesauros para auxiliar na identificação e transformação de termos sinônimos em uma única representação. Por exemplo, os termos:

aldeia, cidade, localidade, lugarejo, povoação, povoado, vila e vilarejo $\Rightarrow$ terra.

Esse tipo de tratamento é importante para redução da quantidade de termos nos documentos. Para um número qualquer de termos sinônimos, a máquina trata-os como termos não correlacionados e, para fins estatísticos, cada termo é computado individualmente. Esse tipo de comportamento não é interessante para a captura do conceito do documento. Convertidos os termos sinônimos para uma única representação, a análise será feita em apenas um termo que pode ampliar sua relevância no documento.

Para tratar a ambiguidade, Jurafsky & Martin (2009) argumentam que uma ideia razoável é usar a definição de um dicionário confiável que é provavelmente um bom indicador do sentido, ou sentidos, que determinado termo possui. Como por exemplo “massa”:

a) Quantidade de matéria (física);

b) Qualidade do que é bom (gíria brasileira);

c) Grande reunião de pessoas.

Neste caso, em um trabalho no domínio de física teórica, a primeira opção seria a escolha natural para definir o significado do termo massa.

Outra utilização de dicionários de apoio seria na correção de erros ortográficos. O procedimento é análogo ao dicionário de sinônimos que, nesse caso, contempla os erros de digitação ou prováveis erros ortográficos mais comuns. Exemplos:

largatixa $\Rightarrow$ lagartixa, indentidade $\Rightarrow$ identidade ou poblema $\Rightarrow$ problema.

A correção ortográfica é útil para uniformizar o conteúdo e, consequentemente, otimiza a análise do conceito geral de documentos. Todavia, dependendo do objetivo da análise, esse passo deve ser executado com parcimônia, pois palavras com erros ortográficos podem indicar autoria.

Em caso de textos históricos, a datação do texto deve ser verificada para que não seja modificado o seu conteúdo sem as devidas considerações. A escrita de determinada época não deveria sofrer correções automáticas, pois seriam descaracterizadas. Exemplo:

assucar ⇒ açúcar, pharmácia ⇒ farmácia ou chlorophylla ⇒ clorofila.

De maneira similar, podem-se corrigir abreviaturas que estejam fora do padrão, como em:

17hrs ⇒ 17h, 10:30 h ⇒ 10h30min. ou Ms. ⇒ M.e[43]

Outro caso especial são os termos compostos por duas ou mais palavras, como "Casa Civil". Se há alguma regra no sistema a indicar que o termo "casa" deve ser substituído por "morada", não deverá ser aplicada no exemplo "Casa Civil", pois alteraria completamente o significado do termo. Uma solução para reconhecer esses termos automaticamente seria o cadastramento das formas compostas para sua correta identificação.

4.4.4.6 Part-Of-Speech Tagging

A expressão *part-of-speech tagging* poderia ser traduzida para anotação de partes do discurso, entretanto a literatura especializada no Brasil a utiliza frequentemente em inglês, aliás abreviando para POS *tag. Part of Speech* é a anglicização do termo latino *pars orationis* Nivre (2005), que em português seria a classe gramatical das palavras. Já o termo *tagging* quer dizer atribuir uma etiqueta, uma anotação ou um rótulo. Portanto, POS *tagging* poderia ser entendida como a atribuição de uma etiqueta (*tag*) correspondente à classe gramatical de um termo em um texto.

A língua portuguesa possui uma norma bem regulamentada e específica, as palavras são organizadas em classes gramaticais ou, mais comumente chamadas, na linguística computacional, como partes do discurso. Nesse sentido, Manning & Schütze (1999) e Jurafsky & Martin (2009) ensinam que POS tagging é a tarefa de identificar e anotar no *corpus* se cada termo é um substantivo, verbo, advérbio, adjetivo e assim por diante.

Essa anotação é geralmente utilizada na remoção de ambiguidades de termos homógrafos usados em contextos diferentes.

As sentenças:

Ela **nada** 200m. (verbo nadar)
Ela **nada** ganhou. (advérbio de negação)

[43] Apesar de muito comum, a abreviatura correta de "Mestre" é M.e, ou para o contexto acadêmico, Sc.M e S.M para Mestre em Ciência, pois Ms é a abreviatura de "manuscrito", segundo Academia Brasileira de Letras. http://www.academia.org.br/abl/cgi/cgilua.exe/sys/start.htm?sid=22

mostram que a palavra "nada", apesar da escrita e pronúncia idênticas, possui significados e funções gramaticais diferentes.

Dessa maneira, o mesmo termo utilizado como classes gramaticais diferentes, deve ser tratado como termo diferente tantas quantas forem suas classificações. Na escolha do termo preferido, no dicionário de sinônimos, a informação da classe gramatical do termo é determinante para sua conversão, pois um termo só será convertido para seu sinônimo, caso pertença à mesma classe gramatical. Caso contrário, serão considerados como termos distintos.

4.4.4.7 Lematização[44]

Num processo de padronização dos termos, é razoável que se queira que o termo livro não seja diferente de livros, já que se referem ao mesmo conceito. Analogamente, formas que se diferenciam segundo o gênero, como pesquisador e pesquisadora, também podem ser reduzidas a uma única forma. Ainda, as formas verbais variam de acordo com suas flexões e, portanto, poderiam ser padronizadas com o mesmo critério, por exemplo as formas conjugadas é, sou, era e fui para a forma infinitiva Ser.

Assim, a ideia é encontrar os lexemas ou lemas das formas flexionadas. Isto é, a lematização converte todas essas variações para uma forma padrão que pode ser o radical da palavra. Segundo Baeza-Yates & Ribeiro-Neto (1999), a lematização retira do termo o afixo (prefixo e sufixo), além de reduzir suas variantes, ou seja, desinências[45] e vogal temática, ao mesmo radical que expressa um conceito comum.

Segundo Jurafsky & Martin (2009), a maior vantagem de tal procedimento possibilita um termo de uma consulta, submetida a um motor de busca, encontrar os documentos que contêm todas as variações morfológicas do termo. Por outro lado, tal abordagem pode retirar distinções que são úteis. Por exemplo, os termos informação, informações, informar, informado e informando em uma única forma padrão que seria "inform". Tal procedimento pode ser muito útil em algumas aplicações que levam em conta a quantidade de vezes que o termo se repete.

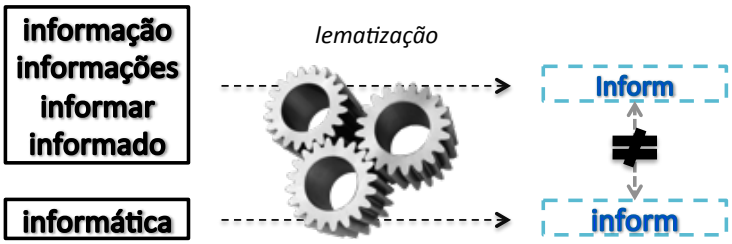
Conforme a Figura 21, o procedimento pode cometer alguns erros como a conversão do termo informática — que se relaciona com computação, tecnologia — para "inform" que se originou do termo informação. Outra consideração são os termos que não possuem formas no singular, por exemplo, "costas" — ou dorso — convertido para "costa" — equivalente a litoral — muda o sentido completamente.

Do ponto de vista da intuição linguística, esse procedimento parece melhorar o processo de RI, mas, surpreendentemente, Salton (1989) mostra em um estudo empírico na comunidade de RI que a lematização não auxilia no desempenho de sistemas de RI clássicos, nos quais o

[44] Frequentemente escrito em Inglês: *Stemming*

[45] São elementos mórficos que se apõem ao radical para assinalar as flexões da palavra (gênero, número, modo, tempo, pessoa) (TERRA, 2002)

Figura 21: Lematização de Termos



Fonte: Elaborado pelo Autor

desempenho é medido como uma média ao longo das buscas realizadas. Ele alerta que o processo de lematização frequentemente elimina mais informação do que deveria, contudo a sobriedade na escolha dos algoritmos que fazem a lematização ainda podem ajudar a melhorar o processo.

4.4.5 A Representação Quantitativa do Texto

Os procedimentos descritos até agora visam minimizar a quantidade de termos extraídos dos textos e preparados para o processamento. O objetivo é facilitar a análise e reduzir o custo computacional. Para tal, uma parte fundamental para representar a linguagem natural está na adequação do texto ao formato reconhecido por algoritmos computacionais, isto é, na transformação do texto em números, sem que se perca a informação nele codificada.

O formato de planilhas geralmente é o mais utilizado. Dessa forma, cada registro, ou linha, refere-se a um caso e cada coluna, ou campo, refere-se a um atributo específico.

Tabela 1: Formato tradicional

PESSOA	IDADE	PESO	ALTURA
Marcos	26	68	178
Ivone	22	57	162
...	...	...	...
Sueli	40	65	153

Fonte: Elaborado pelo Autor

A Tabela 1 mostra os indivíduos nas linhas e as características desses indivíduos nas colunas. Todos os campos identificam o registro especifico completamente, ou seja, um indivíduo é composto pela totalidade de seus atributos. Essa é a maneira convencional de se apresentarem os dados e utilizá-los em procedimentos computacionais.

Esse raciocínio pode ser estendido aos dados textuais. Dessa forma, cada documento é um indivíduo e os termos contidos nesse documento são as características dispostas nas colunas da tabela. Como por exemplo:

- Doc1 — “Informação é segurança”;

- Doc2 — “O caminho é a informação”;
- Doc3 — “A segurança da informação é a política”.

Uma representação possível desse *corpus* na forma de planilha está ilustrada na Tabela 2:

Tabela 2: Representação do Corpus

Documento	Termos							
	a	**informação**	**é**	**o**	**caminho**	**da**	**segurança**	**política**
D_1	a	p	p	a	a	a	p	a
D_2	p	p	p	p	p	a	a	a
D_3	p	p	p	a	a	p	p	p

Fonte: Elaborado pelo Autor

Nesse exemplo, tem-se uma representação binária informando a presença “p” ou ausência “a” do termo na qual é possível observar todos os documentos e todos os termos relacionados existentes na coleção.

Generalizando, sejam n documentos, a quantidade de documentos na coleção representados por $D = \{D_1, D_2, ..., D_n\}$, e m termos, ou atributos, presentes no *corpus* representados por $T = \{T_1, T_2, ..., T_m\}$. Cada documento D é representado por m termos existentes no documento. Cada termo pode ser uma palavra simples ou composta. Então, a_{nm} representa a influência do atributo m no documento n que pode estar representada pela indicação da presença do termo, pela frequência do termo em relação ao documento ou pela frequência[46] do termo em relação à coleção de documentos. Dessa forma, para qualquer *corpus* pode ser representado como mostra a Tabela 3:

Tabela 3: Representação Generalizada do Corpus

Documento	Termos			
	T_1	T_2	...	T_n
D_1	a_{11}	a_{12}	...	a_{1m}
D_2	a_{21}	a_{22}	...	a_{2m}
...	...	...	...	...
D_n	a_{n1}	a_{n2}	...	a_{nm}

Fonte: Elaborado pelo Autor

As estratégias para quantificar os termos na tabela são variadas e, em alguns casos, podem levar em conta simplesmente a existência do termo, em outros, a frequência do termo em relação ao documento ou, ainda, a frequência do termo em relação à coleção. Cada representação privilegia uma característica em detrimento de outra, o tipo de representação é dependente da aplicação.

[46] Quantidade de vezes que o termo ocorre.

Representação Binária — considera-se a existência do termo. Se o valor a_{ij} é igual a 1, então o termo t_j ocorre no documento d_i. Caso contrário, a_{ij} igual a 0, para todo $j \in \{1, \ldots, M\}$ e $i \in \{1, \ldots, N\}$. Conforme apresentado na equação 4.1.

$$a_{ij} = \begin{cases} 1, & \text{se } t_j \text{ ocorre em } D_i, \\ 0, & \text{se } t_j \text{ não ocorre em } D_i. \end{cases} \tag{4.1}$$

Reescrevendo a Tabela 2, tem-se a representação na Tabela 4.

Tabela 4: Representação do *corpus* em Código Binário

Documento	Termos							
	a	**informação**	**é**	**o**	**caminho**	**da**	**segurança**	**política**
D_1	0	1	1	0	0	0	1	0
D_2	1	1	1	1	1	0	0	0
D_3	1	1	1	0	0	1	1	1

Fonte: Elaborado pelo Autor

Esse tipo de representação da Tabela 4 apenas informa a existência do termo, não levando em consideração a quantidade de vezes em que ele aparece. Em algumas aplicações, essa simples representação é suficiente. Contudo, pode ser interessante saber quantas vezes cada termo ocorre ao invés de simplesmente saber se ele ocorre ou não no documento. Nesse caso, ao invés de 1 e 0, coloca-se a frequência observada do termo referente ao documento em que foi encontrado.

Representação por Frequência — considera-se a quantidade de vezes que o termo aparece. Essa medida é repetidamente apresentada com tf, do inglês "*term frequency*". Esse tipo de representação dá a ideia da importância de um termo proporcional à quantidade de vezes que ele aparece.

O termo a_{ij} é atribuído do valor de $tf(t_j, d_i)$ que é a frequência do termo t_j no documento d_i.

$$a_{ij} = tf(t_f, d_i) \tag{4.2}$$

Sua representação é representada na Tabela 5.

Tabela 5: Representação do *corpus* em Frequência

Documento	Termos							
	a	**informação**	**é**	**o**	**caminho**	**da**	**segurança**	**política**
D_1	0	1	1	0	0	0	1	0
D_2	1	1	1	1	1	0	0	0
D_3	2	1	1	0	0	1	1	1

Fonte: Elaborado pelo Autor

A contagem simples, conforme equação 4.2, leva em consideração apenas a presença do termo no documento, e não reflete como os documentos são comparados entre si dentro do corpus.

Documentos maiores tendem a ter frequências mais altas que documentos menores e, portanto, pode ser necessário criar uma medida que leve em conta a presença do termo em relação aos outros documentos da coleção.

Nesse sentido, Spärck Jones (1972) estudou os efeitos da especificidade de termos, ao longo de coleções de textos, em processos de recuperação da informação. O estudo considera a utilização de pesos que quantifiquem a distribuição de termos em todos os documentos do corpus, isto é, uma medida que vai além da simples frequência de um termo. Assim, a medida ***inverse document frequency*** (idf) é um fator de escala para a importância do termo em relação aos outros documentos da coleção.

$$idf(j) = log\left(\frac{N}{df(j)}\right) \tag{4.3}$$

Na equação 4.3, $df(j)$ é o número de documentos que contém o termo j e N, a quantidade de documentos do corpus. Onde $j \in \{1, \ldots, M\}$, e $df(j) \in \{1, ..., N\}$.

Logo, se $df(j) = N$, Isto é, o termo j ocorre em todo documento da coleção, dessa forma tem-se $log\left(\frac{N}{N}\right) = log(1) = 0$, portanto, essa medida favorece termos que aparecem em poucos documentos. Então, combinando a equação 2 e 3 temos:

$$a_{ij} = tf(t_f, d_i) * idf(j) \tag{4.4}$$

Da equação 4.4 infere-se que quando um termo aparece em vários documentos, sua importância é reduzida, pois $idf(j)$ se aproxima de 0. Caso contrário, ela é aumentada.

Geralmente, deseja-se que os documentos da coleção sejam tratados com a mesma importância, independente de seu tamanho. Essa medida possibilita que os atributos dos termos tanto para documentos maiores quanto para menores possam ser comparados na mesma escala.

Considerando que coleção na Tabela 6 tenha apenas 3 documentos, aplicando-se a equação 4.4 tem-se a representação conhecida como ***tf-idf.***

Tabela 6: Representação do *corpus* ponderado

Documento	Termos							
	a	**informação**	**é**	**o**	**caminho**	**da**	**segurança**	**política**
D_1	0,00	0,00	0,00	0,00	0,00	0,00	0,48	0,00
D_2	0,18	0,00	0,00	0,48	0,48	0,00	0,00	0,00
D_3	0,35	0,00	0,00	0,00	0,00	0,48	0,48	0,48

Fonte: Elaborado pelo Autor

A tf-idf é uma das ponderações mais utilizadas por causa das propriedades mencionadas. Ainda existem, contudo, diversas variações de ponderações que podem ser aplicadas, dependendo do objetivo do trabalho.

Resumidamente, na escrita, alguns termos destacam os aspectos semânticos de um documento mais que outros, isto é, existem termos, em determinados contextos, que são mais significativos do que outros em relação ao conceito intrínseco do texto. Em linhas gerais, os pesos utilizados ajudam distinguir termos considerados mais importantes em relação à captura de características de documentos (BAEZA-YATES; RIBEIRO-NETO, 1999).

O leitor interessado em mais detalhes e outras técnicas de ponderação pode verificar em Yang & Pedersen (1997) que analisam distintas ponderações existentes e como elas influenciam no resultado final. Manning & Schütze (1999) e Weiss et al. (2005) apresentam os aspectos teóricos de alguns pesos e sobre como utilizá-los.

4.4.6 Redução de dimensionalidade

Percebe-se que ao transformar um texto em uma tabela com cada coluna representando uma palavra faz com que ela alcance facilmente mais de mil colunas. À medida que se trabalha com textos mais complexos essa quantidade aumenta exponencialmente e isso tende a prejudicar o desempenho do processamento. A saída então é tentar diminuir a quantidade de palavras sem perder informação.

De acordo com Wives (2001), um texto é uma sequência de termos logicamente encadeados formando locuções, frases, orações, parágrafos, capítulos etc. Cada tipo de documento possui características expressas pelos termos mais ou menos apropriadas para descrevê-lo. A escolha dessas características mais relevantes é determinante para a representação individualizada e sem perda de informação de documentos da coleção.

Independente da medida escolhida, a tabela de representação terá um número para indicar a presença do termo e o número zero para indicar a ausência dele, conforme a Tabela 7.

Tabela 7: Alta Dimensionalidade — Documento por Termo

Documento	Termos					
	T_1	T_2	T_3	T_4	...	T_{1200}
D_1	1	1	0	0	...	0
D_2	0	0	1	0	...	1
...	...	...	...	...	...	...
D_n	1	0	0	0	...	0

Fonte: Elaborado pelo Autor

Constatam-se dois problemas neste tipo de representação que são:

- Um número muito grande de termos — ao transformar o documento em palavras dispostas nas colunas da tabela, o número de termos será certamente elevado. Por exemplo, a frase "A Ciência da Informação se correlaciona com a Ciência da Computação" possui nove termos distintos. Expandindo a contagem para um documento é de se esperar que essa tabela possua centenas ou milhares de colunas representando os termos do texto.

- Uma grande quantidade de zeros — espera-se que os termos identificados em um documento não sejam os mesmos identificados em um segundo. Por exemplo, escolhendo-se um texto contendo 1.000 termos e um outro, com 1.200 termos é provável que haja a interseção de uma certa quantidade destes termos que serão preenchidos com o número 1, conforme ilustrado a Tabela 7, e o restante seja preenchido com zeros. Por exemplo, supondo que 400 termos sejam identificados nos 2 textos, logo serão assinalados com o número 1 tanto no texto A quanto no texto B. Dessa forma, 600 termos serão identificados somente no texto A e 800, somente no texto B. Estendendo esse raciocínio para uma grande quantidade de documentos teremos uma tabela com uma enorme quantidade de colunas contendo zero, isto é, ausência do termo.

4.4.6.1 Lei de Zipf

Os termos encontram-se distribuídos nos textos de um corpus, obedecendo a um padrão em relação à frequência de alguns deles e suas posições em uma lista ordenada. Zipf (1949), professor de linguística em Harvard (1902—1950), observou que essa relação aplica-se em vários fenômenos humanos, aos quais chamou de O Princípio do Menor Esforço[47].

A aplicação dessa lei consiste na contagem f e na ordenação r dos termos de uma dada coleção de textos . Logo, o produto da frequência de cada termo f e sua ordem r na lista de termos é aproximadamente uma constante k tal que:

$$f * r = k \tag{4.5}$$

Isso significa que se o termo mais frequente se repete 1000 vezes, então o $2^{\underline{o}}$ termo mais frequente se repetiria $\frac{k}{2}$ vezes que é 500, e, por conseguinte, o $3^{\underline{o}}$ termo seria contado $\frac{k}{3}$, que totaliza 333 e assim por diante.

Tabela 8 ilustra a aplicação da lei de Zipf.

Assim, trata-se da frequência de palavras em textos que obedecem uma distribuição específica. A Figura 22 mostra uma representação da lei de Zipf na qual se observa que ela é uma constatação empírica e apresenta uma descrição da distribuição de frequências de palavras na linguagem humana: há poucos termos que são muito comuns, uma quantidade média de termos de frequência intermediária e muitos termos que ocorrem poucas vezes.

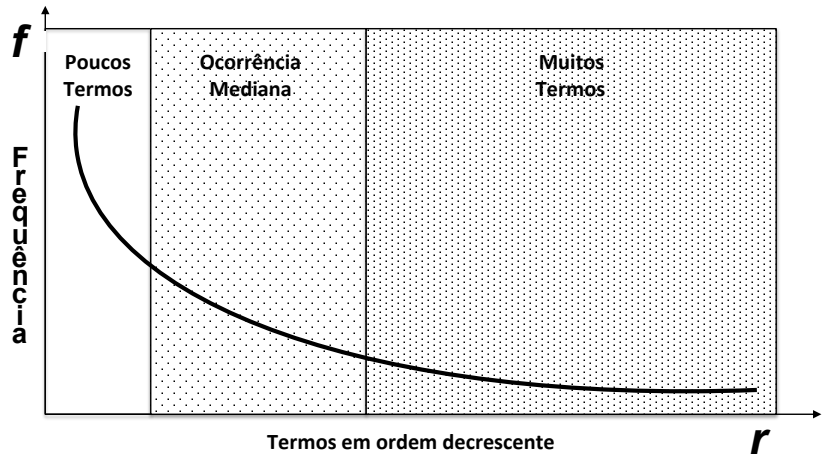
[47] *The Principle of the Least Effort*, tradução nossa

Tabela 8: Exemplo da Lei de Zipf

Palavras	Frequência (f)	Ordenação (r)	Constante $(k = f * r)$
A	1000	1	1000
Informação	500	2	1000
Busca	333	3	1000
Ciência	250	4	1000
...	...	...	...
Texto	1	1000	1000

Fonte: Elaborado pelo Autor

Figura 22: Representação Gráfica da Lei de Zipf



Fonte: Adaptado de Zipf (1949)

No domínio do PNL, essa lei tem grande aplicação, pois auxilia na seleção de pontos de cortes para remoção de palavras com baixo poder de discriminação de documentos e contribui para a redução da dimensionalidade. Ressalte-se que a escolha do ponto de corte é feita de maneira arbitrária. Deve-se levar em conta, também, a experiência do analista que busca encontrar o ponto ótimo entre quantidade mínima de termos e menor perda de informação.

Por fim, Manning, Raghavan & Schütze (2009) afirmam que essa intuição que fundamenta a lei de Zipf pode ser estendida para além do documento textual tradicional e que, com pequenas transformações algébricas, essa lei caracteriza a distribuição de links nas páginas da web. Contudo, o ajuste de dados à lei de Zipf não é particularmente bom, mas o suficiente para servir como um modelo para distribuição de termos.

4.4.6.2 Significância das Palavras de Luhn

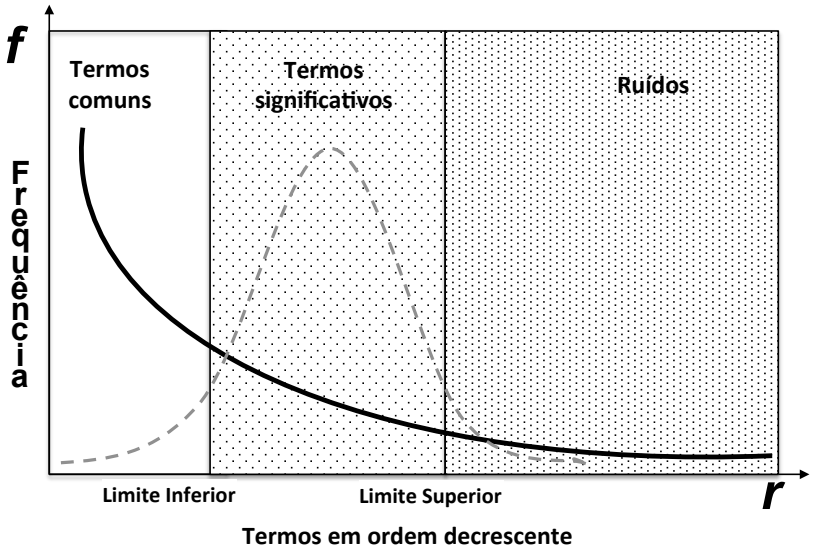
Luhn (1958) em seu artigo *The Automatic Creation of Literature Abstract* propõe a identificação automática de tópicos em artigos com o propósito de criar resumos automáticos. Ele afirma que a divisão de textos em capítulos, parágrafos, orações, frases, etc. são manifesta-

ções físicas da associação de ideias do escritor. Assim, na linguagem escrita, as ideias mais associadas intelectualmente são implementadas por palavras mais associadas fisicamente.

Um escritor normalmente repete palavras à medida que constrói a argumentação em torno de determinado assunto. Embora não considere as relações lógicas ou semânticas, O autor defende que a frequência de um termo em um documento fornece uma medida útil para determinar a significância de uma palavra.

Segundo Moens (2000), Luhn descobriu que padrões de distribuição de termos poderiam fornecer informação significativa sobre o conteúdo de um documento. Altas frequências de termos tendem a ser comuns, mas irrelevantes para enfatizar o conteúdo. Por outro lado, uma ou duas ocorrências de um termo em textos relativamente longos também podem indicar baixa relevância em relação ao conteúdo do texto.

Figura 23: Significância de Termos



Fonte: Adaptado de Luhn (1958, pág. 11)

Pragmaticamente, utiliza-se a abordagem de Zipf para criar uma lista de termos ordenada decrescentemente e, então, o critério de Luhn. A ideia é que existem pontos de corte que podem ser calculados por meio de métodos estatísticos ou arbitrados de acordo com a experiência de analistas de domínio. Esses pontos delimitam as ocorrências dos termos que são potencialmente significativos para a identificação do tema.

Conforme Figura 23, os termos à esquerda do limite superior são comuns e aqueles à direita do limite inferior são raros e, portanto, não contribuem significativamente para o conteúdo do texto. Os termos dentro dos limites superior e inferior são os mais significativos. Existe, ainda, uma curva que Luhn chamou de poder de decisão de termos significantes, que expressam a capacidade de discriminar o conteúdo, ilustrando que os termos, em uma ordem de significância imaginária que se inicia próxima de zero, vão crescendo em habilidade de discriminação até atingirem o pico na metade entre os limites superior e inferior e então

começam a diminuir simetricamente até o último termo.

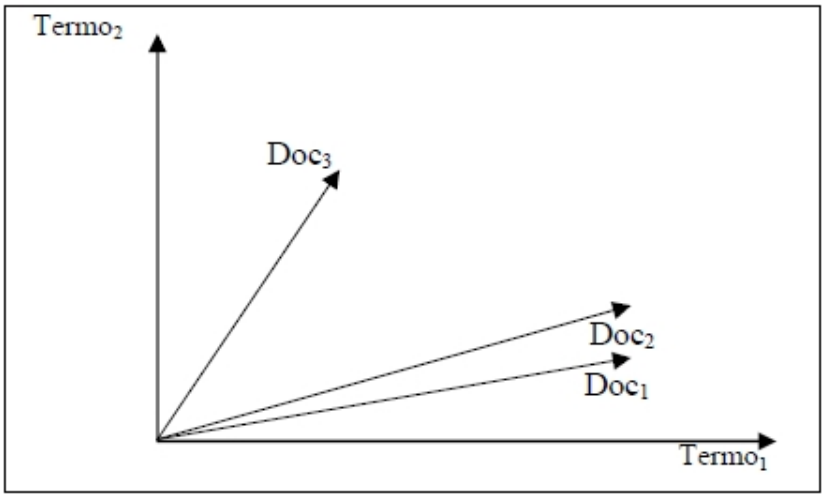
Certa arbitrariedade está envolvida na escolha de tais limites. Não há oráculo que forneça esses valores e eles tendem a ser estabelecidos por tentativa e erro. Destaca-se que essa análise não se aplica somente aos termos, mas também às frases e termos lematizados (RIJSBERGEN, 1979).

4.4.6.3 Decomposição de Valores Singulares - DVS[48]

Um dos problemas intrínsecos aos descartes de termos para reduzir a dimensionalidade é a perda da informação. A conversão de vários termos sinônimos para um termo preferido é um bom exemplo de redução de dimensionalidade sem perda de informação. O ideal é que esse mesmo raciocínio possa ser estendido aos termos semanticamente relacionados, de forma que possam ser combinados em um só, mantendo a informação original com uma dimensão menor.

Segundo Manning & Schütze (1999), as técnicas de redução de dimensionalidade extraem um grupo de objetos que existem no espaço com muitas dimensões e os representa no espaço com poucas, em geral, duas ou três dimensões com a finalidade de visualização. O modelo de espaço vetorial é uma dessas técnicas que, por sua simplicidade conceitual e utilização de proximidade espacial para denotar similaridade semântica entre documentos, é frequentemente utilizada na RI.

Figura 24: Espaço Vetorial em duas dimensões



Fonte: Adaptado de Manning, Raghavan & Schütze (2009, pág. 121)

A Figura 24 mostra a representação em duas dimensões correspondentes aos termos $Termo_1$ e $Termo_2$ e três documentos — Doc_1, Doc_2 e Doc_3 — no espaço. A proximidade de vetores é calculada pelo ângulo, isto é, quanto menor o ângulo entre dois vetores, mais próximos semanticamente eles são. Nesse exemplo, os documentos 1 e 2 possuem uma

[48] *Singular Value Decomposition* (SVD), tradução nossa.

proximidade espacial muito maior que com o documento 3, isto indicaria que os documentos 1 e 2 são mais similares semanticamente.

O espaço vetorial original é constituído de termos únicos que ocorrem nos documentos e, mesmo em uma coleção de textos de tamanho moderado, eles podem chegar a dezenas ou centenas de milhares. Entretanto, para grande parte de algoritmos computacionais, isto é um fator proibitivo. Por causa disso, a redução de dimensionalidade sem perda de informação é uma necessidade (YANG; PEDERSEN, 1997).

Frequentemente, por questões práticas, utiliza-se a matriz de representação dos dados da forma termo-documento como na Tabela 9.

Tabela 9: Matriz Termo-Documento

Termos	Documentos				
	D_1	D_2	D_3	...	D_m
T_1	1	1	0	...	0
T_2	0	0	0	...	1
T_3	1	0	0	...	1
...	...	...	...	...	...
T_n	1	0	0	...	0

Fonte: Elaborado pelo Autor

A Tabela 9 indica n dimensões de termos distribuídos por meio de m documentos de tal forma que a meta será extrair dimensões que contenham relações semânticas entre os documentos, isto é, a combinação de termos que componham conceitos relacionados aos documentos

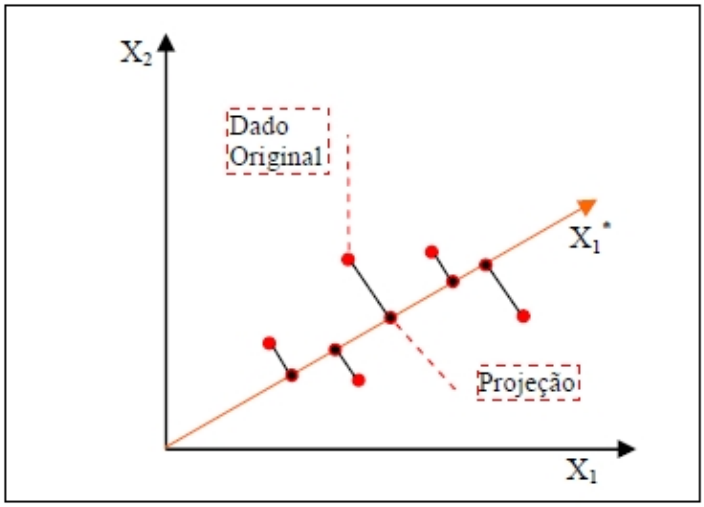
Desse modo, a DVS, é uma técnica matemática de redução de dimensionalidade para formar novas variáveis que são combinações lineares das variáveis originais. A finalidade é utilizar o menor número possível de novas variáveis que contenham as informações das variáveis originais. Como resultado, tem-se poucas variáveis com a menor perda de informação possível.

Formalmente, tem-se um espaço n-dimensional (n termos) que é projetado sobre um espaço k-dimensional, onde $k < n$. Segundo Sharma (1996) o objetivo é formar novas variáveis que são combinações lineares das variáveis originais. Na prática, deve se observar que k deve ser muito menor que n para que valha a pena o emprego da técnica.

Portanto, a projeção transforma um vetor de documentos no espaço n-dimensional de termos para um vetor em um espaço k-dimensional reduzido. Como ilustração, Figura 25, é equivalente a representar geometricamente duas variáveis originais em apenas uma, ou seja, DVS é muito semelhante a ajustar uma reta, um objeto unidimensional, a um conjunto de observações que existe no plano — bidimensional (MANNING; SCHÜTZE, 1999).

Khattree & Naik (2000) afirmam que a seleção de projeções em dimensão menor é feita geralmente pela otimização de características interessantes nos dados originais através de

Figura 25: Projeção de observações



Fonte: Adaptado de Sharma (1996, pág. 65)

todas as direções de projeções, isto é, a captura da máxima variação dos dados em uma quantidade menor de variáveis.

A DVS é computada pela decomposição de uma matriz A que contenha os vetores de documentos com cada coluna correspondente a um documento, isto é, o elemento a_{ij} da matriz registra a frequência do termo i no documento j. Quaisquer dos pesos discutidos anteriormente podem ser utilizados para popular a matriz.

Segundo Sharma (1996), Manning & Schütze (1999), Khattree & Naik (2000) e Woodfield (2004), a matriz A é decomposta no produto entre das matrizes T, S e D.

$$A_{txd} = T_{txn} * S_{nxn} * (D_{dxn})^t \tag{4.6}$$

em que

$t =$ quantidade de termos,

$d =$ quantidade de documentos,

$n = min(t, d)$,

$D^t =$ é a transposta de D.

Cada matriz T e D é ortogonal, que significa que a matriz $T^t * T = I$ e $D^t * D = I$. S é uma diagonal.

A DVS pode ser vista como um método para percorrer os eixos do espaço n-dimensional tal que o primeiro eixo percorre a maior variação entre os documentos, a segunda dimensão percorre a dimensão com a segunda maior variação e assim por diante. O número máximo de dimensões é a quantidade de termos da matriz A.

As matrizes T e D representam os termos e documentos, respectivamente, nesse novo espaço. A diagonal de S contém os valores singulares de A em ordem decrescente de maneira que o i-ésimo valor singular corresponde à quantidade de variação ao longo do i-ésimo eixo.

A projeção da DVS é a combinação linear das linhas e termos da matriz termo-documento original. A ideia de combinação linear pode ser entendida como uma extensão de média ponderada dos termos. Essa média ponderada produz o conceito contido nos documentos e, daqui, considera-se que a DVS converte termos em conceitos.

Matematicamente, essa projeção forma o subespaço k-dimensional que representa o melhor ajuste para descrever os dados originais. A projeção de colunas da matriz termo-documento é um método para representar cada documento por k conceitos distintos. Em outras palavras, a coleção de documentos é mapeada no espaço k-dimensional no qual cada dimensão é reservada para cada conceito. Da mesma forma, cada linha, ou termo, pode ser projetada sobre as k primeiras colunas de S. Enfim, a técnica da DVS encontra a projeção ótima para um espaço reduzido, de maneira que representa os termos e documentos da melhor forma possível em um espaço dimensional menor.

Finalmente, a aplicação da DVS na área de Recuperação da Informação é chamada de Indexação por Semântica Latente (ISL), ou usualmente em inglês, *Latent Semantic Indexing* (LSI). Essas novas dimensões são uma melhor representação de documentos e de consultas. O nome "latente" é uma metáfora devido ao fato de que essas novas dimensões são a representação verdadeira, pois a ISL recupera a estrutura semântica original do espaço e suas dimensões originais.

4.4.7 Considerações

O PLN talvez seja, atualmente, um dos temas que mais atrai pesquisadores de áreas diversas. Aí se elencam cientistas da informação e da computação, linguistas, estatísticos, filósofos, economistas, entre outros. O grande atrativo é a possibilidade de extrair informações da massa de dados disponível no ambiente da web e nos repositórios digitais. Tais informações, tão heterogêneas quanto a quantidade de línguas existentes, podem fornecer valiosos insumos para guiar as decisões de pessoas comuns ou profissionais especializados.

O desenvolvimento da TI tem impulsionado fortemente a área que, por sua vez, estimula o aperfeiçoamento de novos recursos tecnológicos num círculo virtuoso. A natureza evolutiva e dinâmica da linguagem aliada à ubiquidade da web, torna o campo de estudo muito ativo e aberto para receber contribuições variadas. Um exemplo disso é a chegada da web semântica que trouxe à tona novamente alguns desafios propostos há décadas pela Inteligência Artificial, mas que ainda não foram adequadamente resolvidos. A natureza filosófica do significado tem provocado debates interessantes e certamente ainda será tema de futuras discussões.

No âmbito deste trabalho, o PNL ocupa um lugar bastante relevante. É ele que fornecerá as ferramentas necessárias para transformar e extrair informações dos recursos não estruturados. Além disso, propiciará a integração da linguagem natural com a linguagem lógica, utilizada na WS, na elaboração de um repositório híbrido rico em informações ontológicas e linguísticas. Enfim, as possibilidades linguísticas aliadas à precisão das ontologias devem contribuir

para incentivar o desenvolvimento de novos mecanismos de busca que lidam tanto com a informação estruturada quanto com a textual.

4.5 A Recuperação da Informação

A geração contemporânea da internet, Geração Y, tende a achar que Recuperação da Informação (RI) começou com a Web. Entretanto, a RI foi uma resposta aos desafios de acesso a conteúdos em vários formatos. A área inicia-se com os registros de bibliotecas e com os trabalhos científicos e logo se espalha por outros formatos que eram utilizados por profissionais da informação. Somente na metade do século XX, Mooers (1950) propõe o termo *Information Retrieval* que se traduz para o Português como Recuperação da Informação (RI).

É inegável o avanço da RI nos últimos anos em função da Web, da popularização de interfaces gráficas e do barateamento de dispositivos de armazenagem. Além disso, a contínua otimização dos motores de busca, que proporciona uma experiência razoavelmente satisfatória aos usuários, tem tornando a Web como a fonte padrão e preferida para se encontrar informação. Especialmente após o lançamento do motor de busca Google[49] por Brin & Page (1998) que tenta responder aos desafios de construir um sistema que reúna documentos da Web e os mantenha atualizados no ritmo de crescimento desse ambiente.

De acordo com Baeza-Yates & Ribeiro-Neto (1999), a RI lida com a representação, armazenamento, organização e acesso aos itens de informação. A representação e organização da informação deve prover facilidade de acesso à informação que interessa ao usuário. Manning, Raghavan & Schütze (2009) complementam afirmando que o foco da RI é encontrar material (usualmente documentos) de natureza não estruturada (usualmente texto) que satisfaça à necessidade de informação contida em grandes coleções (usualmente armazenada em computadores).

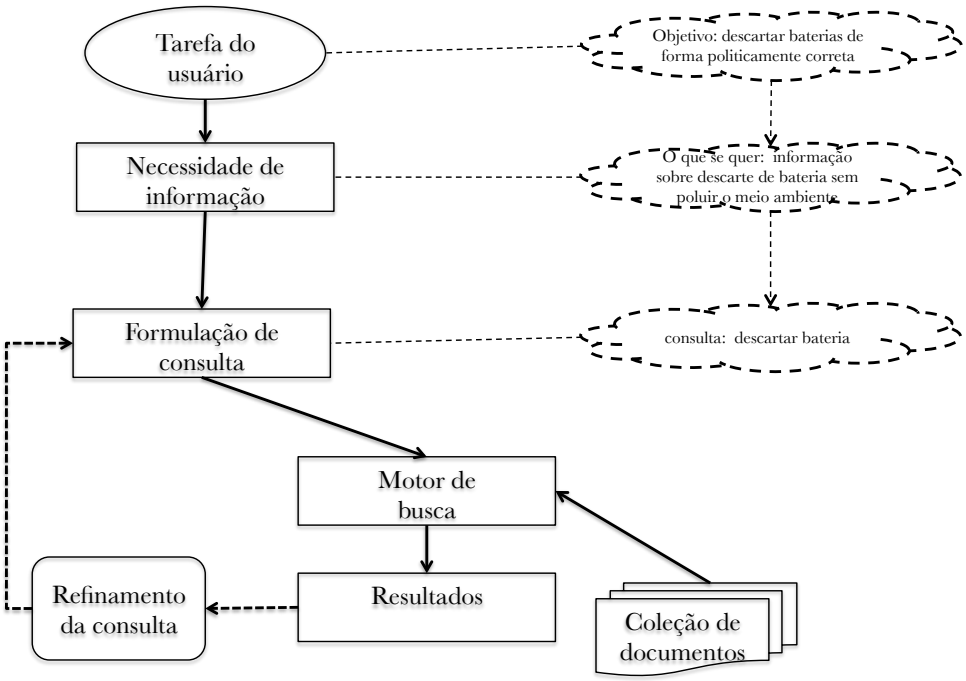
A segunda noção estabelece uma aproximação com a linguagem natural, que é a principal forma de interação entre pessoas que, de acordo com a necessidade de informação, formulam consultas e as submetem aos Sistemas de Recuperação da Informação (SRI). Tais SRI as transformam para uma linguagem compreensível e a comparam com os itens que descrevem os documentos armazenados e que serão recuperados Baeza-Yates & Ribeiro-Neto (1999) apoiam essa visão estabelecendo uma diferença entre recuperação de informação e de dados; a primeira lida com textos em linguagem natural que não são sempre bem estruturados e semanticamente precisos; a segunda, com dados — bases de dados relacionais — com estrutura e semântica bem definidas.

[49] www.google.com

4.5.1 Conceitos fundamentais

Manning, Raghavan & Schütze (2009) estabelece a estrutura fundamental apresentada na Figura 26 para realizar a tarefa de recuperar informação. Para tal, definem: i) coleção é um conjunto de documentos; ii) objetivo é recuperar documentos com informações relevantes à necessidade de informação do usuário e ajudá-lo a completar a tarefa.

Figura 26: Modelo Clássico de Busca



Fonte: Elaborado pelo Autor

Conforme a Figura 26, o processo inicia-se com a tarefa específica do usuário de traduzir sua necessidade de informação numa consulta na linguagem disponível do SRI. Nesse caso, como descartar baterias de forma politicamente correta. O objetivo é achar informação sobre como descartar baterias sem poluir o meio ambiente. Essa consulta é submetida ao motor de buscas que reúne os documentos mais representativos e os apresenta como resultado. O usuário analisa os documentos recuperados e verifica se atendem a sua necessidade, ou não. Por exemplo, o termo bateria pode representar um instrumento musical ou uma pilha. Nesta etapa, pode-se adicionar ou substituir termos para melhorar a recuperação. Portanto, o passo pode ser repetido e refinado interativamente até satisfazer à necessidade de informação do usuário.

Como se vê, o processo possui uma forte dependência da capacidade do usuário de expressar sua necessidade de informação suficientemente detalhada na consulta. Savoy & Gaussier (2010) alegam que o processo de busca deve ser visto mais como uma abordagem de "tentativa e erro" do que um paradigma de "pergunta e resposta". Afirmam também que isso decorre de três aspectos da linguagem natural, tais como polissemia, sinonímia e, em menor

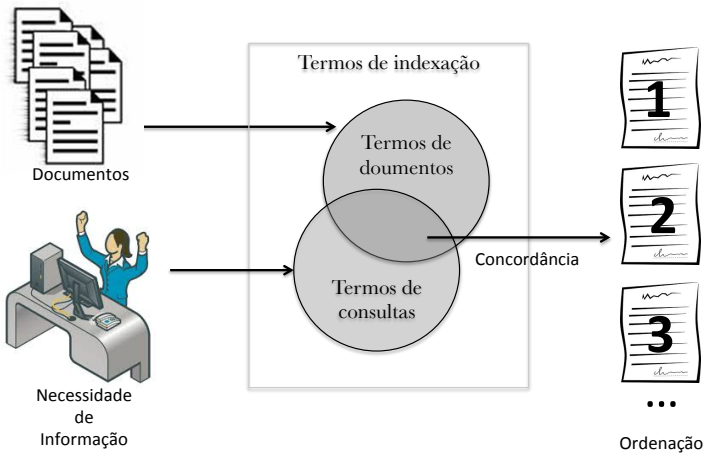
grau, erros de ortografia e outras variantes como, por exemplo, os termos “facto” e “fato” adotados por falantes do Português europeu e brasileiro, respectivamente.

Baeza-Yates & Ribeiro-Neto (1999) propõem a distinção entre a tarefa do usuário e a visão lógica do documento. Ambos influenciam diretamente a efetividade do SRI. A primeira implica a especificação de termos que contenham a semântica da necessidade do usuário e que preencham os requisitos da necessidade de informação do usuário na exploração (*browsing*) de documentos recuperados. A segunda compreende uma sequência de transformações que objetivam representar documentos por um conjunto de termos índices ou palavras-chave. Tal ação se justifica porque, apesar de textos completos representarem a visão lógica mais completa, têm alto custo computacional. Por outro lado, um pequeno conjunto de categorias fornece a visão lógica mais concisa do documento, mas o uso conduz a uma recuperação de qualidade inferior.

4.5.2 Modelos de RI

De acordo com Baeza-Yates & Ribeiro-Neto (1999), modelagem no contexto da RI é um processo complexo que almeja produzir uma função que atribua pontuação aos documentos com relação a uma dada consulta. Ainda, esse processo consiste na concepção de uma estrutura lógica para representar documentos e consultas; e na definição de uma função de classificação de documentos que permita a quantificação da similaridade entre eles.

Figura 27: Processo de Recuperação da Informação



Fonte: Adaptado de Baeza-Yates & Ribeiro-Neto (1999)

A Figura 27 mostra os passos que abrangem o processo geral da RI. De forma geral, um SRI deve considerar o problema de prever quais documentos o usuário achará relevante. Rijsbergen (1979) assegura que uma linguagem de indexação é aquela utilizada para descrever documentos e consultas. De um lado, existe uma coleção de documentos que está representada por meio de seus correspondentes termos de indexação. Do outro, há o usuário que formula uma consulta por meio de termos de indexação. Logo, a intersecção entre os índices de

documentos e da consulta recupera os documentos classificados em função da relevância com a consulta do usuário. Formalmente, os autores declaram que um modelo RI é uma quádrupla $[D, Q, \mathcal{F}, R(q_i, d_j)]$, na qual i) D é um conjunto de visão lógica de documentos na coleção; ii) Q é um conjunto de visão lógica de consultas do usuário; iii) $\mathcal{F}$ é uma estrutura de modelagem de documentos e de consultas; iv) $R(q_i, d_j)$ é uma função de classificação na qual, qi é a visão lógica do documento e dj é a visão lógica da consulta.

Para definir modelos de RI, há que se responder à pergunta: como a informação contida nos documentos e as consultas são representadas? A resposta é que não há o melhor modelo. Cada um deve ser adequado à necessidade. Diferentes modelos têm sido propostos para auxiliar na busca de informação. Uma unanimidade é que o processo de indexação é fundamental para garantir a qualidade da recuperação. Apresentam-se a seguir os principais modelos utilizados na recuperação da informação.

4.5.2.1 Modelo Booleano

A forma mais intuitiva de RI é a leitura de um documento para verificar se ele tem relevância, ou não, com a necessidade de informação do usuário. Obviamente, o processo é inviável para uma quantidade razoável de documentos. No ambiente Web, a tarefa é impossível devido ao volume de documentos envolvidos no processo. Portanto, uma forma de evitar a leitura de textos para uma consulta qualquer é a indexação prévia de documentos.

Uma tradição na CI é a preparação da representação do conteúdo temático dos documentos. Isso pode ser feito com o emprego de um ou mais termos de indexação — ou conjunto de palavras-chave — para indicar de que se trata o conteúdo ou sintetizar um documento. Esse processo de indexação é usualmente realizado com apoio de vocabulário controlado e executado manualmente por profissionais da informação (LANCASTER, 1993).

Assim, recuperar um documento em um repositório previamente indexado significa construir consultas que coincidam com os termos de indexação do documento. Para construir consultas, o usuário transforma sua necessidade de informação em uma expressão lógica utilizando os termos da indexação e operadores booleanos *AND, OR* e *NOT*. De acordo com Manning & Schütze (1999), o modelo de recuperação booleano é um modelo para RI no qual se pode criar qualquer consulta em que os termos são combinados com operadores booleanos. Nesse contexto, cada documento é visto simplesmente como um conjunto de palavras-chave.

Savoy & Gaussier (2010) alertam que existem alguns obstáculos desse modelo. O primeiro é que os documentos recuperados não constituem uma lista ordenada. Isto é, não refletem o grau de relevância do documento com a consulta submetida. Segundo, os operadores booleanos (binários) são geralmente muito limitados para indexação de documentos e consultas. Como se trata apenas da existência — ou não — do termo, não há como determinar se um termo é essencial ou marginal na consulta. Terceiro, um sistema de busca booleano é incapaz de recuperar documentos que preenchem parcialmente os requisitos da consulta, ou seja, o contexto é ignorado. Quarto, a escolha dos termos certos para a consulta impacta na

qualidade da lista recuperada. Isso significa que se certos termos aparecem na consulta, a quantidade de documentos retornados pode ser tão grande que pode tornar a tarefa de detectar os relevantes muito difícil.

4.5.2.2 Modelo Vetorial

Em função dos obstáculos apresentados, algumas propostas introduzem melhorias no processo de RI, como a representação de documentos e consultas em um espaço vetorial multidimensional. A seção 5.4.2.3 discute em detalhes a indexação de documentos e consultas que são utilizadas nos modelos de RI. O resultado é um conjunto de termos de indexação ponderados. Nesse tipo de modelo, a cada termo indexado corresponde uma dimensão. Os elementos desses vetores indicam a presença do termo ou da importância relativa de termos no documento ou na consulta.

O efeito prático desse tipo de representação é a utilização, como termos de indexação, do texto completo de documentos e na formulação de consultas em linguagem natural, de acordo com Savoy & Gaussier (2010). Isso transforma qualquer parte do texto em ponto de acesso e dispensa a conversão do conteúdo textual em expressões lógicas booleanas. O que evidencia um caráter mais amigável de interação com o usuário. Além disso, a ponderação dos termos demonstra que os termos de um texto não possuem a mesma importância para o conteúdo e, consequentemente, não possuem a mesma utilidade para descrever um documento.

Um estudo de Salton & Buckley (1988) compara as abordagens que empregam a ponderação dos termos de indexação para RI. Na fase de planejamento do SRI, deve-se considerar quais termos são incluídos na representação de documentos e de consultas, bem como determinar ponderações capazes de distinguir termos importantes daqueles menos cruciais na identificação do conteúdo. Observam que o uso de termos ponderados fornece melhores resultados do que aqueles que podem ser obtidos com representações de textos mais elaboradas. Concluem afirmando que a escolha do sistema de ponderação de termos é fundamental para o resultado da RI.

Um destaque interessante é a utilização dos princípios de Luhn, discutidos na seção 4.4.6.2, na construção do índice TF/IDF (*term frequency/inverse document frequency*), tratado na seção 4.4.5, que, conforme Manning, Raghavan & Schütze (2009), é um dos mais populares esquemas de indexação em RI. IDF é computado sobre toda coleção, enquanto TF o é por documento. Quando comparados, TF e IDF, apresentam comportamentos que se equilibram entre si. Isso resulta, como Luhn verificou em seus experimentos, que os termos com valores IDF intermediários mostram pesos máximos em TF/IDF e são os mais interessantes para classificação.

Baeza-Yates & Ribeiro-Neto (1999) alertam que existe uma simplificação nos modelos desse tipo, pois assumem que os termos de indexação são mutuamente independentes, o que, na prática, não se verifica. Contudo, as vantagens desse tipo de modelagem são a melhoria no conjunto de respostas, a recuperação de documentos baseada na correspondência

parcial entre documentos e consultas, a ordenação dos resultados de acordo com o grau de similaridade com os termos da consulta e a normalização da extensão do documento que é naturalmente considerada no processo de classificação.

4.5.2.3 Modelo Probabilístico

Rijsbergen (1979) lembra que o instrumento básico utilizado nas abordagens anteriores para separar documentos relevantes dos não relevantes é uma função de correspondência (*matching*), seja em ambiente estruturado ou não. A razão para escolher qualquer função nunca foi feita de maneira explícita, de fato, a maioria é baseada em argumentos intuitivos em conjunto com o princípio da Navalha de Ockham[50]. Portanto, buscam-se, nos modelos probabilísticos, argumentos essencialmente teóricos com fundamentação sólida para determinar uma função de correspondência e como deve ser seu uso.

Em qualquer tentativa de recuperar documentos, o usuário espera encontrar um grupo de documentos que representa o conjunto ideal que responde a sua pesquisa. Esse grupo se divide em documentos relevantes e não relevantes. De preferência, com muito mais documentos relevantes para compor o seu conjunto ideal de documentos recuperados que satisfaçam sua necessidade de informação.

De acordo com Savoy & Gaussier (2010), a recuperação no contexto da RI é vista como um processo de classificação. Nesse tipo de modelo um conjunto inicial de documentos é de alguma forma recuperado; o usuário inspeciona esses documentos procurando pelos relevantes; o sistema de RI usa essa informação para refinar a descrição de um conjunto ideal de respostas; pela repetição do processo, espera-se que a descrição do conjunto ideal de respostas seja melhorada.

Por conjunto ideal, Baeza-Yates & Ribeiro-Neto (1999) acrescentam que para uma dada consulta do usuário, existe um grupo de documentos que melhor responde à necessidade de informação. Os autores ainda mencionam que a principal vantagem do modelo probabilístico é o ordenamento por relevância dos documentos medida pela probabilidade, isto é, a criação automática de uma lista ordenada dos resultados. Contudo, os autores consideram desvantagens: i) a necessidade de "adivinhar" uma estimativa inicial da proporção de documentos relevantes; ii) não considera o fator TF (*term frequency*); e iii) a falta de normalização do tamanho do conteúdo dos documentos.

Manning, Raghavan & Schütze (2009) afirmam que nos modelos booleano e vetorial, a correspondência é calculada com uma definição formal, mas semanticamente imprecisa. Dada uma consulta, um SRI tem uma compreensão incerta da necessidade de informação. Por outro lado, um sistema tem um palpite incerto se o documento tem conteúdo relevante para a necessidade de informação. A teoria da probabilidade fornece o fundamento para o raciocínio sob a incerteza.

[50] "Se em tudo o mais forem idênticas as várias explicações de um fenômeno, a mais simples é a melhor" — http://pt.wikipedia.org/wiki/Navalha_de_Occam.

Os autores também afirmam que vários são os modelos de RI que possuem fundamentação na probabilidade, contudo o ***Binary Independency Model*** (BIM) proposto por Robertson & Spärck Jones (1976) permanece com grande influência nos modelos probabilísticos. O BIM pressupõe que os documentos são vetores binários — há o termo ou não — para possibilitar que todos os documentos e consultas tenham a mesma representação; o modelo não reconhece associação entre os termos, isto é, os termos são independentemente distribuídos tanto nos documentos relevantes, quanto nos irrelevantes — isso é um pressuposto de modelos baseados em classificadores ***Naive Bayes***[51]. Essa premissa equivale ao pressuposto do modelo vetorial, no qual cada termo é uma dimensão ortogonal a todas as outras.

Não se discutirão aqui os detalhes dos modelos por não ser o foco do trabalho, mas apenas mostrá-lo com uma opção. Para os interessados no tema, recomenda-se a leitura de Rijsbergen (1979), Baeza-Yates & Ribeiro-Neto (1999), Manning, Raghavan & Schütze (2009) e Savoy & Gaussier (2010).

4.5.3 Recuperação da Informação na Web Semântica

A Web quebrou diversos paradigmas da humanidade. A superação da barreira do espaço físico e a imediatez da informação são as mais evidentes, contudo ela não tem precedentes de outras tantas formas: o volume das informações disponíveis, a escalabilidade para contemplá-las, a falta de coordenação em relação à criação de conteúdo, diversidade de culturas e de níveis intelectuais na composição deste conteúdo. Todos esses ingredientes formam um conteúdo informacional que difere do tradicional e que precisa de abordagens distintas para gerenciá-lo.

Manning, Raghavan & Schütze (2009) professam que a massa de informação é inútil, caso essa riqueza informacional não possa ser descoberta e consumida por outros usuários. Tentativas para tornar a informação nesse ambiente passível de ser descoberta podem ser vistas com: i) motores de busca para índices de texto completos como o Excite[52] que busca palavras-chave por meio de índices invertidos e mecanismos de classificação ordenada; e II) taxonomias contendo páginas da web em categorias tal como em Yahoo![53] que permitem ao usuário a navegação através de estrutura hierárquica de categorias.

Os autores ainda afirmam que a qualidade e a relevância dos resultados de busca na internet deixam muito a desejar em função das idiossincrasias na criação de conteúdos nesse ambiente. Por isso, o que dizer da autoridade do texto nas páginas da web? A democratização da criação do conteúdo exprime um novo nível de granularidade de opiniões em, virtualmente, qualquer assunto. Logo, a web revela verdades, falsidades, contradições e suposições em escala global.

[51] Em termos simples, um classificador Naive Bayes assume que, dada uma variável, o valor de uma característica particular não se relaciona com a presença ou fata de qualquer outra característica. Fonte: http://en.wikipedia.org/wiki/Naive_Bayes_classifier

[52] http://www.excite.com/

[53] https://br.yahoo.com/?p=us

Da perspectiva tradicional da RI para a web, há uma mudança significativa no perfil de usuários. Os profissionais treinados na arte de construir consultas sobre coleções bem estruturadas e bem conhecidas são substituídos por pessoas comuns que tendem a desconhecer, ou a não se preocupar com a heterogeneidade dos conteúdos, a linguagem para consultas ou qualquer embasamento teórico de como funciona um sistema de informação. A consequência disso é o aumento da complexidade da infraestrutura envolvida em todo o processo de gestão da informação.

4.5.3.1 A Web como um grafo

A web tradicional consiste de páginas HTML e *hyperlinks* que conectam umas às outras. Nesse sentido, uma página pode ser vista como um nó — ou vértice —, enquanto cada *hyperlink* com o sentido da página A para B pode ser visto como uma aresta — ou arco. Ao conjunto de arestas entre os pares de nós dá-se o nome de grafo[54] da web.

Broder et al. (2000) declaram que existem várias razões para a compreensão da web como grafo, tais como: i) projetar estratégias de rastreamento ou varredura (*crawling*) da web; ii) compreender a sociologia da criação do conteúdo na web; iii) analisar o comportamento de algoritmos da web que utilizam informações de link, por exemplo a distribuição e evolução de valores *PageRank* de Brin & Page (1998) que simulam o conceito subjetivo de importância de páginas da web; iv) prever a evolução de estruturas da web; v) prever o aparecimento de novos fenômenos importantes no grafo da web.

A visão da web como grafo pode se aproveitar do fundamento teórico da matemática e propor soluções e caminhos para achar o "fio da meada" da aparentemente caótica web. Essa perspectiva abre o caminho para a mudança do foco primário da web de "documentos" para "dados". Obviamente, essa transformação é extraordinária em termos de como a web se servirá da total integração das associações semânticas com o dia a dia do usuário.

4.5.3.2 Modelos Semânticos para Recuperação da Informação

Para a concretização da visão da WS, a combinação de documentos textuais com marcações semântica deve estar disponível. Esses serão os documentos da WS (DWS) que estarão acessíveis às pessoas e aos agentes computadorizados. Para serem recuperados, deverão estar visíveis aos motores de busca que utilizam bases indexadas que não contemplam as informações semânticas nesses documentos.

Alonso, Banerjee & Drake (2006) declaram que quando uma informação não estruturada se incorpora a um recurso estruturado, ela pode ser integrada aos processos organizacionais diários, tais como pesquisa e conformidade. Para tal, os recursos na rede de informação abrangem todos os dados e os metadados necessários para torná-los significativos. Esses recursos podem ser dados estruturados, semi ou não estruturados, armazenados em qualquer

[54] Ao leitor interessado na teoria dos grafos, recomenda-se a leitura de Feofiloff, Kohayakawa & Wakabayashi (2011)

lugar como bases de dados, sistemas de arquivos locais, servidores de e-mails e criados por qualquer aplicação. Assim, a meta é possibilitar que uma organização veja todos os seus ativos em formato compreensível e rico em informação semântica.

Para os autores, uma rede de informação é autodescritiva: módulos de aplicação podem descobrir quais fontes existem, quais dados elas possuem, qual é o ciclo de vida dos dados e como dado deve ser interpretado. Os componentes principais são: i) gerenciador de repositórios, metadados e serviços; ii) buscadores e rastreadores semânticos; iii) apresentadores e visualizadores da informação e; iv) motores de inferência.

Apesar de ter sido proposta há mais uma década, Singh & Jain (2014) afirmam que a WS ainda permanece como um conceito futuro que ainda não está propriamente apresentado ao usuário comum. Há algumas razões para isso, talvez a WS ainda não foi completamente desenvolvida e as partes que já estão, ainda são tão pobres que não podem ser usadas no mundo real; ou nenhum software ou hardware ideal foi fornecido.

Por outro lado, tradicionais SRI baseiam-se, geralmente, em linguagem natural, isto é, em indexação de especialistas e em ferramentas de PLN para extrair palavras-chave e frases de documentos. Tais documentos estão na WWW que difere fundamentalmente da WS por não possuir elementos que expressem a semântica neles contida. Dessa dificuldade nasce a Recuperação da Informação Híbrida (RIH) que se propõe a recuperar os documentos híbridos que possuem texto padrão e marcadores semânticos(SHAH et al., 2002).

Finin et al. (2005) declaram que numa das visões da WS, a web será parecida com o que se tem hoje, exceto que os documentos serão enriquecidos com anotações compreensíveis por máquinas; na outra, o conteúdo da WS existirá em documentos separados que referenciam e descrevem os documentos convencionais; na terceira, construir motores de indexação e recuperação especialmente projetados para trabalhar com DWS. Na última, a inferência disponível em tecnologias da WS pode melhorar a busca de informações tradicionais.

Uren et al. (2007) afirmam que as buscas semânticas são classificadas em: i) por entidades — para obter informação sobre coisas representadas em classes ou instâncias de objetos; ii) por relações — para conhecer como as entidades se relacionam; e iii) por parametrização — para realizar combinação entre entidades e relações para atender necessidade específica.

Os autores também acrescentam que as principais interfaces para consultas semânticas disponíveis se classificam em: i) por palavras-chave — recebe uma, ou mais, palavra-chave como entrada e apresenta entidades semânticas correspondentes à entrada; ii) por formulário — utiliza formulários, menus ou listas *drop-down* que orientam o usuário na elaboração de consultas; iii) por visão/navegação — utiliza navegação pela representação visual da estrutura da ontologia para construir consultas; e iii) por linguagem natural — utiliza ferramentas linguísticas para traduzir consultas do usuário em LN para linguagem especializadas, como SPARQL.

De acordo com Shukla & Singh (2013), durante os próximos dez anos, as tecnologias da web incrementarão a habilidade de inserção de estruturas semânticas nos documentos

e de construção de vocabulários estruturados e ontologias para definir termos, conceitos e relacionamentos. A WS impulsionará avanços extraordinários na visibilidade e na exploração de informação, especialmente na habilidade de um sistema interpretar documentos e inferir significados sem intervenção humana. Logo, isso será um bônus para a expertise de profissionais da informação na organização de dados e na estruturação da informação.

A fim de explorar uma estreita integração de busca e inferência, Finin et al. (2005) propõe o seguinte esquema:

- a estrutura deve comportar tanto o processamento de recuperação quanto o de inferência;
- a recuperação deve contemplar o uso de termos, marcadores semânticos ou ambos como elementos de indexação.
- a busca na web deve se apoiar na atual ampla cobertura oferecida por motores de busca baseados em texto;
- inferência e recuperação devem ser firmemente acopladas, isto é, melhorias na recuperação devem acarretar melhores inferências e vice-versa.

Vallet, Fernández & Castells (2005) corroboram a visão de Mayfield & Finin (2003), na qual a busca semântica deve ser vista como um complemento à busca de palavras-chave já que ontologias e metadados são em menor número. A inferência, por sua vez, é útil para preencher os espaços deixados pelo conhecimento incompleto e pela informação faltante, por exemplo, a transitividade de uma relação localizadoEm em localidade geográfica como em "Brasília está localizada no Brasil", por transitividade pode ser inferido que também está na América do Sul.

4.5.4 Avaliação em RI

A interação do usuário com um SRI leva em consideração vários fatores que são mensuráveis, tais como velocidade de indexação, tamanho de índices, armazenamento, quantidade de documentos, latência (tempo) da busca etc. Contudo, considera-se a satisfação do usuário como o objetivo primordial de qualquer SRI, pois ela está relacionada com funcionamento geral do sistema. Mais especificamente correlaciona-se com a relevância de resultados que satisfazem a necessidade de informação do usuário. Mas, como avaliá-la?

Manning, Raghavan & Schütze (2009) ensinam que uma abordagem padrão de avaliação de um SRI gravita em torno da noção de documentos relevantes e não relevantes com relação à necessidade de informação do usuário. Nesse contexto, uma necessidade de informação é traduzida em uma consulta ou pesquisa. Por sua vez, a relevância é relacionada à avaliação dos documentos recuperados em relação à necessidade de informação do usuário, e não com

as palavras usadas na consulta. Portanto, elas refletem o quão bem formulada é a pergunta, ou consulta, para encontrar respostas na coleção de documentos.

Nessa perspectiva, Salton & Buckley (1988) afirmam que duas medidas são normalmente utilizadas para avaliar a habilidade de um sistema em recuperar os itens relevantes e rejeitar os não relevantes de uma coleção. Elas são a revocação e precisão, respectivamente.

Tabela 10: Tabela de contingência 2X2

Resultado	Relevância	
	Presente	Ausente
Positivo	VerdadeiroPositivo (VP)	FalsoPositivo (FP)
Negativo	Falso Negativo (FN)	VerdadeiroNegativo (VN)

Fonte: Adaptado de Salton & Buckley (1988)

A Tabela 10 mostra todos os resultados com as possíveis classificações em um processo de recuperação de documentos. Na Estatística, utiliza-se a tabela de contingência para se obter medidas estatísticas que avaliam o desempenho de um teste binário de classificação. Logo, o julgamento de um SRI pode utilizar do ferramental estatístico para avaliar um problema de classificação. Com essa distribuição na tabela é possível verificar os seguintes resultados: i) se o resultado for positivo, o documento pode ser relevante de fato (VP); ii) ou pode ser erroneamente classificado como não relevante (F); iii) quando for negativo, o documento pode ser classificado erroneamente como relevante (FP); iv) ou pode ser realmente não relevante. Desse modo, combinando esses resultados é possível verificar a:

- Precisão: fração de documentos recuperados que são relevantes à necessidade de informação do usuário — proporção de itens selecionados que são corretos.

$$P = \frac{VP}{(VP + FP)} \tag{4.7}$$

- Revocação: fração de documentos relevantes na coleção que são recuperados — proporção de itens corretos que são selecionados

$$R = \frac{VP}{(VP + FN)} \tag{4.8}$$

Em princípio, um sistema é preferível se produz alta revocação recuperando tudo que é relevante e também alta precisão rejeitando tudo que for não relevante. A função de revocação se aplica melhor utilizando alta frequência de variados termos que ocorrem em muitos documentos da coleção. Tais termos possibilitam a extração de mais documentos que deve incluir uma grande quantidade de relevantes. O fator de precisão apresenta melhor resultado utilizando termos mais específicos que são capazes de isolar os poucos itens relevantes da massa de não relevantes. Na prática, o ideal é utilizar termos que são gerais o

bastante para alcançar um nível razoável de revocação sem, ao mesmo tempo, produzir baixa precisão (SALTON; BUCKLEY, 1988).

Frequentemente, quer-se julgar um SRI por sua acurácia, que é a proporção de todas as classificações corretas:

$$A = \frac{(VP + VN)}{(VP + FP + FN + VN)} \tag{4.9}$$

Manning, Raghavan & Schütze (2009) alertam que essa medida representa a efetividade do sistema, pois o SRI é visto como um classificador de duas classes (relevante e não relevante). Contudo, em função da diferença de volume entre as categorias, geralmente 99,9% são não relevantes e não interessam ao usuário, mas influenciam positivamente no resultado final da acurácia. Por exemplo, se todos os 100 documentos de uma coleção são irrelevantes, o resultado será 1. Da mesma forma, se todos forem relevantes, o resultado também será 1.

Por outro lado, as medidas de P e R concentram a avaliação de verdadeiros positivos, que é o objetivo final do usuário: recuperar o que se precisa. Entretanto, essas duas medidas de características incompatíveis devem ser balanceadas para atingirem o ponto máximo em um processo de RI. Isso quer dizer que, em um bom SRI, o usuário quer obter a quantidade máxima de documentos relevantes, mas tolera um percentual mínimo de documentos não relevantes. Esse equilíbrio entre precisão e revocação é obtido pela medida-F que é a média harmônica ponderada entre a precisão e a revocação. A utilização da média harmônica se deve à característica da medida de ser mais conservadora e frequentemente estar próxima ao mínimo de dois números.

A manipulação dos pesos na medida-F enfatiza tanto a precisão quanto a revocação. A forma comumente apresentada é aquela que enfatiza igualmente as duas medidas. Por isso, ela é chamada de medida-F balanceada e representada por:

$$F_1 = \frac{2PR}{(P + R)} \tag{4.10}$$

Considerando que as medidas P, R e F_1 são computadas em um conjunto desordenado de documentos seria interessante que essa noção pudesse ser estendida também aos SRI atuais que recuperam de forma ordenada por relevância. Isso é muito comum nos atuais motores de busca, como o Google por exemplo, que apresentam como resultado os k mais relevantes. Contudo, para os propósitos deste trabalho essas medidas são suficientes. O leitor interessado pode ver uma discussão detalhada em Manning, Raghavan & Schütze (2009).

4.5.5 Estado da arte

As atividades que envolvem a WS têm sido estudadas e muitas propostas vem sendo apresentadas na tentativa de criar uma web distribuída de dados legíveis por máquinas. Desde a proposta da WS, muitos problemas têm sido solucionados e outros, mais complexos, ainda

permanecem como objeto de diversas abordagem para que a visão da WS seja concretizada em sua plenitude. Isso é o que se discute brevemente nesta seção.

Portais Semânticos, Maedche et al. (2001), Castells et al. (2004a), Castells et al. (2004b) e Contreras et al. (2004), fornecem essencialmente funcionalidades de buscas simples que se caracterizam como recuperação de dados semânticos. As buscas retornam instâncias de ontologias ao invés de documentos e também não fornecem métodos de classificação por relevância desses resultados. Em alguns sistemas, links para documentos que referenciam as instâncias são adicionados na interface do usuário, próximas a cada instância retornada na resposta à consulta Contreras et al. (2004), mas nem instâncias, nem documentos são ordenados. Esse tipo de solução pode ser suficiente para pequenas bases de conhecimento, mas não seria apropriada repositórios de grande escala, nos quais uma busca poderia retornar centenas ou milhares de resultados.

O problema da classificação por relevância é tratado em Rocha, Schwabe & Aragao (2004) que sugerem uma solução que fornece uma lista ordenada como resposta à pesquisa submetida pelo usuário. Para tal, propõe-se uma rede semântica na qual relações têm rótulos semânticos e pesos numéricos. Os termos da consulta são mapeados para os nós da rede semântica e a ordem dos resultados será computada de acordo com a relevância fornecida pelos pesos associados.

Guha & McCool (2003) e Guha, McCool & Miller (2003) assumem que o dado na WS é modelado como um grafo dirigido e rotulado, do qual cada nó corresponde a um recurso e cada aresta é rotulada com um tipo de propriedade, tal como um modelo de dados RDFS. Com essa premissa, foi desenvolvido o sistema, chamado TAP na *Stanford University, Knowledge Systems Lab.*, que recupera um grupo inicial de nós a partir de termos utilizados na busca. A consulta é expandida pela especificação das propriedades desejadas e pela exploração do conjunto inicial de nós.

Na visão de Popov et al. (2004), o resultado da combinação de técnicas de extração de informação, de ontologias semanticamente leves, da representação de conhecimento e da RI poderia abordar o problema da marcação e recuperação semânticas automáticas. Para tal, propuseram uma estrutura de gerenciamento de informação e do conhecimento — *Knowledge Information Management* (KIM). Esse sistema focaliza na criação de metadados a partir de documentos e fornece, além da estrutura, serviços para marcações, indexação e recuperação semânticas de documentos. O processo de recuperação desempenha a indexação de textos completos e realiza a busca por termos e metadados.

Portanto, TAP e KIM são propostas para construir bases de conhecimento de alta qualidade e anotar coleções de documentos em grande escala. Motores de buscas que usam ontologias desempenham bem o papel em intranets organizacionais, mas não são convincentes quando aplicados na web como um todo.

Nesse sentido, Vallet, Fernández & Castells (2005) e Castells, Fernandez & Vallet (2007) complementam KIM e TAP fornecendo um algoritmo de classificação (ranking) de resultados

especialmente projetado para um modelo de recuperação baseado em ontologia utilizando um esquema de indexação semântica fundamentada em técnicas de marcação ponderada.

Buscando extrapolar os limites das ontologias de organizações específicas, Fernandez et al. (2008) investigam a combinação e escala de espaços informacionais fornecidos pela WS e WWW. Os autores avançam em direção aos projetos de tecnologias de recuperação semântica para a web por: i) estabelecer a ponte entre os usuários e os dados semânticos e; ii) fazer a ponte entre os dados da WS e a informação não estruturada, textual, disponível na web.

Aproveitando o potencial do LOD, Hogan et al. (2011) propõem um Motor de Busca para Web Semântica (MBWS) para pesquisar e navegar em dados RDF na Web. Dada a flexibilidade da WS, objetos recuperados podem representar pessoas, companhias, cidades, proteínas ou qualquer coisa publicada, sem que haja categorização pré-definida como nos MB tradicionais. Ainda, tal sistema deve ser escalável para lidar com grandes quantidades de dados e robusto o suficiente para lidar com heterogeneidade, ruído, imprecisão e possíveis conflitos de dados coletados de várias fontes.

A exploração de metadados associados com documentos na WS melhora a precisão de sistemas de RI. Silva, Girardi & Drumond (2009) introduzem um modelo genérico de RI para a WS que utiliza metadados em todas as fases do processo: representação, correspondência (*matching*) e medida de similaridade. O modelo utiliza representação semântica, ao invés de palavras-chave. Os documentos são descritos por meio de conceitos e instâncias agrupadas em casos semânticos representando os interesses do usuário. Para atingir maior precisão nos resultados, os modelos de correspondência e de similaridade comparam os mesmos casos semânticos de consultas e documentos.

Nas tentativas de interligar a WS com a WWW, vários processos tem sido tentados e se intensificado mais recentemente. Apesar do crescimento das bases estruturadas em níveis que já viabilizam várias pesquisas, Heath & Bizer (2011) mencionam que o descompasso entre o texto e os dados estruturados ainda aparece como barreira para a popularização da WS e na utilização de ferramentas voltadas para esse ambiente.

Algumas iniciativas como em Navigli, Velardi & Gangemi (2003) e Reymonet, Thomas & Aussenac-Gilles (2007) propõem modelos de lexicalização de ontologias, mas sem integrar os níveis léxico e ontológico. O modelo proposto em Buitelaar et al. (2011) e aperfeiçoado em McCrae et al. (2012), Unger et al. (2013) e Cimiano, Unger & McCrae (2014) reflete a necessidade e urgência de estabelecer uma ligação entre o conhecimento do mundo — conceitos — e do termo, ao mesmo tempo que estabelece a diferença estrita entre eles.

Dada a escala dos volumes na web, não há como desenvolver soluções sem a parceria das máquinas. Portanto, para automatizar a criação do léxico, Walter, Unger & Cimiano (2013) e Walter, Unger & Cimiano (2014) utilizaram bases de dados estruturadas para fornecer a semântica e o corpus para se extraírem as variantes léxicas e morfológicas. A ideia é induzir a criação de um léxico a partir do conhecimento representado em ontologias para alimentar o modelo originalmente proposto.

A WS, na visão de seu criador, nasceu como uma extensão da WWW e, portanto, seria natural que a integração entre elas fosse perfeita. Na prática, há um problema na comunicação entre as partes, pois as duas estão separadas por diferentes linguagens e níveis de especialidade dos usuários. Para lidar com esse problema, as propostas mais recentes têm procurado formas de integrar o nível léxico e ontológico de maneira a que os usuários — pessoas e máquinas — tanto do mundo da linguagem natural, quanto da linguagem formal, sejam capazes de utilizar os sistemas e de cooperarem entre si.

Finalmente, com a integração da WS e WWW será possível obter informações estruturadas e não estruturadas apropriadas ao perfil do usuário. A nova geração de SRI será capaz de fazer, indistintamente, buscas ora em bases de conhecimento formais — contendo estruturas ontológicas incompreensíveis para pessoas — ora em bases textuais — igualmente incompreensíveis para programas "inteligentes" de computador — e fornecer resultados com qualidade. Assim, somente com essa comunicação livre e irrestrita entre os dois mundos, o potencial prometido da WS estará disponível ao usuário comum.

4.5.6 Considerações

Existe uma quantidade crescente de estruturas, na web, como resultado das linguagens modernas desenvolvidas, anotações de usuários, aparecimento de ferramentas de PLN mais robustas e um crescente volume de *linked data*. Essas estruturas mantêm a promessa de melhorar significativamente o acesso à informação, por meio do aumento da profundidade da análise dos sistemas de hoje. Atualmente, estamos apenas no começou da exploração das possibilidades e entendimento de como essas pistas semânticas podem ser utilizadas de forma mais proveitosa.

Parece existirem duas tendências que devem trazer grandes desafios à área: i) a necessidade de incluir os atuais e emergentes recursos de conhecimento (DBpedia, Freebase, CYC) como modelo semântico fundamental, promovendo um alcance e detalhamento da informação factual sem precedentes; e ii) a necessidade de incluir anotações que contenham pistas importantes para combinar as necessidades e perfis específicos com as ferramentas de busca.

Desde a proposição da WS, é visível que muito se caminhou, contudo o usuário comum parece estar à margem dos benefícios que poderia usufruir. A discussão para compreender o futuro da área na CI é promissora, destacando-se que o nascimento de bases de conhecimento em grande escala forma um componente crucial para a busca semântica, fornecendo uma estrutura unificada com incontáveis entidades e relações. Além disso, a adição de marcações factuais ou em grandes porções de texto podem fornecer melhores pistas sobre a especialidade da busca e da adequação da informação. Ainda, novas interfaces de usuário são fundamentais para desencadear a estruturação da web por meio das marcações semânticas, ou por meio da exploração de novas formas de sugestões interativas com o usuário.

Finalmente, sob a ótica da CI, o crescimento contínuo de bases com marcações semânticas na WS aponta um caminho que há muito se busca na organização e representação do

conhecimento, que é a estruturação automática de acervos informacionais. O aparecimento de novas tecnologias fornece uma oportunidade para se começar a trabalhar na direção de uma nova geração de sistemas inteligentes e para se formularem novas teorias e técnicas para manipular mais e melhor a nossa principal matéria-prima: a informação.

4.6 Considerações Finais acerca da Revisão de Literatura

O conteúdo da revisão contempla os temas que são tratados nesta pesquisa. Apesar de abordar muitos aspectos ligados à tecnologia, em função da temática da tese, procurou-se apresentá-la no contexto e com fundamentos da CI. O tratamento automático da linguagem natural já possui longa tradição na RI que fornece um dos principais alicerces teóricos para o desenvolvimento de pesquisas que lidam com a Ciência da Web. Essa convergência de teorias e técnicas apoiam os estudos em WS, um assunto ainda com muito potencial para ser explorado fora da CC e com total aderência aos propósitos da CI.

A web semântica nasce como uma extensão da web tradicional. Isso significa que o livro global voltado para humanos começa a receber atualizações e novos capítulos com informações apropriadas para outra audiência. O leiaute preparado para pessoas passa a compartilhar informações legíveis pelas máquinas. A linguagem natural, rica e ambígua, divide espaço com a linguagem voltada para máquinas que tem a tarefa de explicitar significados e eliminar ambiguidades. As ferramentas de inferência não são mais propriedade somente do intelecto humano, mas de agentes computadorizados que auxiliam na tarefa de organizar e recuperar um corpo de informação continuamente crescente.

O processamento de linguagem natural reúne ferramental para criar meios de acompanhar o crescimento da informação digital, pois não há como gerir o montante de informação sem o apoio de tratamentos automatizados. Nesse sentido, o PLN, que já está em fase madura de desenvolvimento, fornece tecnologias para representar e recuperar informações para o usuário comum. Com o aparecimento da WS, ele contribui na tarefa de extrair o significado implícito de textos para indução automática de estruturas ontológicas que conectam as palavras do dia a dia aos termos abstratos da lógica nesse novo ambiente.

O PLN como fundamento da WS é a manifestação da declaração de Wittgenstein de que o significado está no uso. As pesquisas recentes, utilizando *corpus* para povoar ou construir ontologias, mostram que as tecnologias do PLN procuram vincular as propriedades das palavras distribuídas no texto ao papel organizador dos conceitos nas ontologias. O argumento que se sustenta é que a linguagem natural é o principal método para transportar o significado e a WS possui grande parte de seu potencial conteúdo baseado em documentos textuais. Dessa forma, o sucesso da WS está relacionado à utilização do crescente corpo de textos com uma fonte de informação fundamental.

Resumidamente, as estruturas ontológicas e léxicas compartilham o mesmo ambiente, mas faltam abordagens para conectá-las automaticamente. Enquanto as ontologias compartilham

conceituações, observa-se que, geralmente, apenas rótulos em linguagem natural são anexados a elas. O que faz com que não aproveitem a riqueza da informação linguística. Assim, deve haver uma contrapartida sobre como as classes, propriedades e indivíduos são verbalizados em linguagem natural. Essa intersecção de abordagens lógicas e linguísticas deve contribuir para a melhoria da precisão na RI que ocupa um lugar de relevância no contexto da CI.

Assim, de forma a esquematizar o aporte teórico oriundo dessa revisão de literatura, o Quadro 4 apresenta os temas estudados relacionados aos respectivos objetivos específicos propostos nessa investigação.

Quadro 4: Objetivos específicos X Revisão de Literatura

OBJETIVOS ESPECÍFICOS	ÁREA DE ESTUDO
Identificar tecnologias da Web Semântica apropriadas para a representação do conhecimento em determinado domínio;	Web Semântica, seção 4.1 — provê a descrição de como os últimos avanços tecnológicos introduzem o significado ao mundo das máquinas. Além disso, apresenta as principais ferramentas utilizadas nesta pesquisa e como combiná-las de forma a atender ao objetivo específico.
Produzir ontologia para o domínio de risco financeiro;	Ontologia e PLN, seções 4.2 e 4.4, respectivamente — proporcionam elementos teóricos e práticos sobre ontologias e como elaborá-las utilizando ferramentas do PLN para acessar automaticamente coleções de textos para extrair, manipular e adequar os termos que comporão o vocabulário comum da ontologia.
Gerar uma base de léxicos em Português brasileiro que contenha os aspectos morfológicos, sintáticos e semânticos dos itens lexicais vinculados aos elementos ontológicos da indústria financeira;	Web Semântica e PLN, seções 4.1 e 4.4, respectivamente — abrangem as metodologias e as ferramentas necessárias para extrair rótulos da ontologia, buscar seus sinônimos e suas classificações nas fontes textuais e adequá-los ao padrão legível às máquinas.
Propor um modelo de recuperação de informação que utilize automaticamente a base de léxicos como forma de tratar a ambiguidade;	Web Semântica, PLN e RI, seções 4.1, 4.4 e 4.5, respectivamente — fornecem o suporte teórico para a elaboração de modelos de recuperação da informação, o processamento de consultas por meio do PLN e a adição de informações semânticas ao corpus para que seja possível produzir um índice léxico-semântico.
Testar o modelo e mensurar a precisão para comparar os resultados entre o modelo aqui proposto e um modelo clássico de recuperação da informação no domínio de risco financeiro.	PLN e RI, seções 4.4 e 4.5, respectivamente — adicionam subsídios para selecionar técnicas de avaliação de sistemas de recuperação da informação e de PLN para o propósito de verificar como as manipulações de textos podem afetar o desempenho desses sistemas.

Fonte: Elaborado pelo Autor

Finalmente, ressalta-se que todas as leituras aqui revisadas contribuiram, em maior ou menor grau, para o desenvolvimento desse estudo. Os temas abordados são norteados pelas questões de pesquisa que são a base para a elaboração do objetivo geral que, por sua vez, se desdobra no detalhamento expresso pelos obetivos específicos que também direcionam e limitam o escopo da revisão.

Parte III

Metodologia e Resultados

5 Procedimentos metodológicos

Este capítulo discute os caminhos e instrumentos comumente utilizados para alcançar os objetivos de pesquisa. Tais procedimentos são frequentemente compreendidos em coleta de dados, sistematização de processos e consolidação de resultados que visam descrever a realidade. Além disso, podem-se inquirir os limites da ciência por meio de hipóteses que problematizam criticamente os atuais perímetros de determinado conhecimento. Tanto aspectos teóricos como práticos são sempre discutíveis, pois nada é definitivo em ciência, especialmente nas Ciências Sociais. Essa é a face especialmente bela da pesquisa.

Marconi & Lakatos (2004) afirmam que os requisitos metodológicos para a pesquisa científica são escrutínio — passível de investigação ou exploração —, refutabilidade ou verificabilidade — apto à contestação por meio de argumentos ou submissão à verificação empírica — , confirmabilidade — sujeito à aquisição de resultado compatível com a observação dentro dos limites estabelecidos — e simplicidade metodológica — tecnicamente suscetível à teoria e às provas empíricas.

Segundo Braga (2007), a escolha da metodologia adequada tem a função de atribuir o caráter científico e de atestar qualidade e validade ao estudo feito e aos resultados conquistados. Ela é o roteiro que descreve o caminho e as ferramentas que serão julgadas pelos pares. Amatuzzi, Amatuzzi & Leme (2003) também afirmam que a escolha da metodologia na CI está relacionada ao tipo de pesquisa, à abordagem e, principalmente, à questão da pesquisa que deve ser tão clara quanto possível, pois orienta o melhor processo de pesquisa e indica o melhor desenho de resposta.

A adoção de procedimentos metodológicos, portanto, define o caminho percorrido na construção desta investigação que, como qualquer trabalho científico, é colocada à prova para se tornar objeto de crítica ou de validação pela comunidade científica. Nesse sentido, esta pesquisa possui três partes fundamentais que compreendem o levantamento de material teórico, a escolha de ferramentas tecnológicas e a caracterização de amostra e proposta de modelagem para lexicalização de ontologias e recuperação de informação léxico-semântica.

5.1 Tipo de pesquisa

Conforme a abrangência mencionada desta pesquisa, julga-se que:

- Quanto à natureza, trata-se de uma pesquisa aplicada, haja vista o direcionamento para a solução de problemas específicos por meio de aplicação prática.

- Do ponto de vista da abordagem, classifica-se como pesquisa qualitativa, pela interpretação de conteúdos e pelo aprofundamento em fenômenos e processos mais ou menos delimitados, mas com possibilidade de investigação intensa.

- Em relação à técnica, enquadra-se como pesquisa experimental, porque as condições da investigação são controladas em laboratório de forma a validar as hipóteses e possibilitar a replicação da pesquisa nas mesmas condições.

5.2 Método de Abordagem

As etapas consideradas para a consecução desta pesquisa foram agrupadas em duas partes distintas: i) a primeira concentra-se na fundamentação teórica que sustenta os argumentos aqui considerados, com foco nos conceitos fundamentais relacionados à Ciência da Informação; ii) A segunda parte desenvolve o método, com base no referencial teórico, demonstrando o caráter operacional da pesquisa para conceber uma proposta de modelo de lexicalização de ontologias e recuperação de informação léxico-ontológica em textos na língua portuguesa.

5.2.1 Fontes de pesquisa

O levantamento de material para estudo e referência é um esforço contínuo durante o desenvolvimento de uma pesquisa. Isso significa que não há um tempo determinado no qual se faz a coleta de material relevante, mas uma busca constante e dinâmica que pode sempre encontrar novas referências que suportam o aporte teórico. Entretanto, como a própria característica da pesquisa demanda, foram selecionados alguns pontos de partida para que outros pesquisadores possam reproduzir os caminhos aqui percorridos.

5.2.1.1 Busca da informação

Adotou-se o portal da Capes[1] como principal fonte de informação sobre o tema, pois concentra uma grande de quantidade de outros portais. Tal procedimento facilitou a operação que, muitas vezes, é tediosa, como quando se deve entrar em cada base de dados para submeter determinada consulta. No caso da Capes, basta entrar com o termo a ser pesquisado e determinar a área de concentração para limitar e especificar a abrangência da busca.

Além disso, outras buscas foram feitas diretamente, via navegador, em sites de instituições de ensino, repositórios públicos — como citeSeerx[2] —, Google Acadêmico[3] e outros recursos disponíveis na Web como complemento às bases mencionadas. Uma forma comum de

[1] http://www.periodicos.capes.gov.br/
[2] http://citeseerx.ist.psu.edu/index
[3] http://scholar.google.com.br/

levantamento bibliográfico é verificar as referências de uma publicação consultada como meio de se descobrirem outras fontes relevantes para o tema da pesquisa.

Importante ressaltar que esta pesquisa foi influenciada pelo estágio de doutoramento, durante um ano no *Institute for Web Science and Technologies* (WeST) da Universidade de Koblenz-Landau, Alemanha. Primeiramente, por ser uma instituição dedicada à pesquisa com foco em computação. Em segundo, pela possibilidade de estar em contato com pesquisadores em tempo integral, muita informação foi passada de forma verbal. Por exemplo, em reuniões semanais em que cada doutorando fala suscintamente do trabalho da semana e trocam-se experiências. Ou nos seminários, também semanais, nos quais se debatia sobre o desenvolvimento do trabalho do apresentador que, na maioria das vezes, era um estudante do instituto.

5.2.2 Parte I – Fundamentação teórica

5.2.2.1 Levantamento bibliográfico

O levantamento bibliográfico abrangeu as áreas de Ciência da Informação, Ciência da Computação, Estatística, Linguística e Filosofia.

O tema lexicalização de ontologias é inédito na Ciência da Informação e não foram encontradas referências, nesse campo, nas bases consultadas. Todavia, a busca para motivar a pesquisa sob a ótica da área não foi abandonada, por isso se investigaram obras de autores clássicos e dos atuais, que fundamentem os recentes avanços sobre o assunto. Os principais assuntos são o PNL, a RI, as ontologias e a organização e representação da informação e do conhecimento. Todos esses assuntos mantêm estreita ligação com outras áreas do conhecimento.

Os assuntos atinentes aos desafios relacionados à linguagem natural e à subjetividade da interpretação humana foram providos pela Linguística e o PNL. Os levantamentos concentraram-se na defesa do texto ou do *corpus* como base de dados confiável e nos aspectos do processamento da linguagem natural sob a perspectiva de linguistas. Ainda, relativo ao PNL, muitas técnicas são emprestadas da estatística, que fornece as principais ferramentas para coleta de dados, identificação de padrões e testes de validação.

A Filosofia é o berço das discussões acerca das ontologias e apoiou a fundamentação teórica sobre o assunto. Alguns aspectos da filosofia da linguagem também foram levantados, com o objetivo de delimitar a abordagem do processamento da linguagem natural e obter a garantia filosófica na utilização de repositórios textuais para construção de ontologias. A Lógica foi, também, revisitada com objetivo de aplicação nas regras de inferência e nos axiomas das ontologias.

A Ciência da Computação tem fornecido muitos insumos aos profissionais da informação. Essa área concentra os pesquisadores mais atuantes no tema desta pesquisa. Vale mencionar que, pela natureza da área, os trabalhos são focados em implementações de algoritmos

computacionais e aplicações relacionadas à automação de procedimentos tradicionalmente manuais. Os assuntos revisados foram o processamento da linguagem natural, a extração da informação, tecnologias da Web Semântica e extração de padrões léxico-sintáticos.

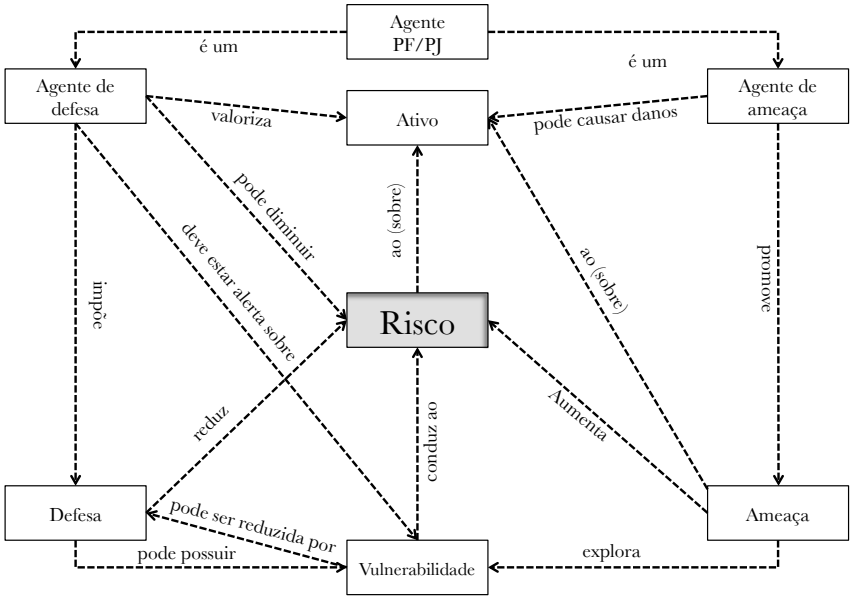
5.2.3 Parte II – Desenvolvimento

5.2.3.1 O *corpus* e a ontologia

Como ponto de partida, reutilizou-se uma ontologia que representa, em inglês, o domínio de finanças e está disponível em http://fadyart.com/Finance.owl. Para o propósito desta pesquisa e como forma de delimitar o tema, apenas a parte relacionada à área de risco financeiro foi considerada. Isso resultou em uma ontologia com 91.692 triplas, 39 classes e 101 propriedades.

Contudo, a aplicação direta dessa ontologia à realidade brasileira não é plausível, dada a singularidade do mercado financeiro e da regulamentação existente no país. Portanto, ela foi adaptada para contemplar essa característica e construiu-se manualmente uma nova versão a partir dos conceitos declarados em Gresser et al. (2010). Essa versão possui similaridade com conceitos e práticas orientados pelo comitê gestor de risco financeiro, composto pelo Banco Central do Brasil (BC) e pelas principais instituições financeiras nacionais.

Figura 28: Visão macro do domínio de Risco Financeiro



Fonte: Adaptado de Gresser et al. (2010, pág. 12)

A Figura 28 representa a visão macro do domínio de Risco Financeiro. Os vários conceitos estão interligados por relações que contrapõem as forças entre a ameaça e a proteção aos ativos de um ente. Cada dimensão da figura dá origem a conceitos cada vez mais específicos. Esse conjunto de conceitos deve ser interpretado seguindo as setas que estabelecem o tipo de

relação entre um conceito e outro. Por exemplo, “Agente PF/PJ” é um “Agente de defesa” que impõe medidas de “Defesa” para diminuir o “Risco” sobre o “Ativo”.

O conceito de gestão do risco interage com outras áreas, de forma que é impossível isolá-lo sem considerar outros conceitos correlatos, os quais foram agrupados conceitualmente desta forma:

a) Risco finaceiro— engloba as características de ameaça e de defesa;

b) Governança —contempla a formulação de políticas e regras para a regulação e proteção de ativos;

c) *Shareholders/stakeholders* — distribuem-se em proprietários, participantes, serviços e órgãos reguladores;

d) Ativo — o bem a ser protegido, que também pode estar associado à infraestrutura classificada em sistemas e redes, bem como à infraestrutura crítica, que apoiam o negócio. Sistema e rede aqui estão no sentido lato e não se restringem à noção de informática;

e) Tecnologia – contemplada por soluções e tendências tecnológicas.

Dessa forma, a ontologia final no domínio de Risco Financeiro e Corporativo em Português foi elaborada a partir da combinação das ontologias existentes em inglês adaptadas à necessidade brasileira. O que resultou em 2.178 triplas — que compreendem o conjunto de sujeito, predicado e objeto —; 65 classes — conceitos — e 47 propriedades — relações entre conceitos. Elas foram reformuladas manualmente utilizando, como fonte, 1.092 termos em Português, extraídos da intranet da área de risco financeiro da Caixa Econômica Federal — *WikiRisk*. Há duas bases que conformaram a ontologia:

- A primeira é composta por 984 termos, que foram validados por profissionais da área, os quais adicionaram uma definição em linguagem natural de cada termo, citando as fontes institucionais, como no exemplo a seguir, que normatizam o ambiente informacional da empresa.

> PLANO DE CONTINGÊNCIA — Plano que consiste em definir e testar ações que permitem dar continuidade às operações da CAIXA que não podem ser interrompidas. Fonte: MN CR 115

- A segunda possui 109 termos, que foram sugeridos e definidos por meio do portal colaborativo da empresa. Nesse caso, somente usuários do portal legitimaram os termos que, portanto, não possuem validação credenciada na documentação formal da instituição. Exemplo:

> Grau de Confiança — Percentual de confiança estatística atribuída ao cálculo do VaR.

- Há, ainda, uma terceira fonte, capturada no sítio[4] do Banco Central do Brasil, constituída de um glossário de termos, com as respectivas definições; utilizada pelos profissionais do risco para padronizar e validar termos e definições. Exemplo:

> Risco operacional — Risco de haver erro humano ou falha de equipamentos, programas de informática ou sistema de telecomunicações imprescindíveis ao funcionamento de determinado sistema.

Outro componente importante para desenvolvimento de pesquisas em PLN é a disponibilidade de bases de conhecimento léxico. No momento da realização da investigação, encontraram-se recursos disponíveis para auxiliar o desenvolvimento da pesquisa em processamento da linguagem natural. O idioma inglês é, sem dúvida, o mais abundante recurso para a área. Entretanto, há iniciativas no meio científico para promover a consolidação de bases de conhecimento léxico para o Português e com utilização, sob a licença, de domínio público.

Apesar de as definições serem extraídas da intranet e, portanto, inacessíveis para acesso externo, optou-se por essa alternativa devido à facilidade de identificação, recuperação, validação, padronização e confiabilidade dessa informação. Ela não é propriedade da empresa e pode ser encontrada em sites de informações financeiras, como o do Banco Central do Brasil, que atendem às necessidades de usuários.

A Priceton WordNet, proposta por Fellbaum (1998), é uma base comumente citada em trabalhos científicos da área. Trata-se de um recurso léxico para o inglês construído manualmente. Ele está estruturado em grupos de itens léxicos sinônimos chamados de *synsets* que contemplam as possíveis lexicalizações de conceitos em linguagem natural. Há também outras formas de relacionamento semântico como hiponímia ou meronímia. Além disso, cada *synset* possui indicações sobre se o termo é substantivo, verbo, advérbio ou adjetivo, isto é, a *part-of-speech*. Há ainda a definição textual, ou *gloss*, de cada *synset*.

Parte da iniciativa da *Open Multilingual Wordnet* (OMW), na qual wordnets em várias línguas são relacionadas à Pricenton WordNet (PWN), foi realizado por Paiva, Rademaker & Melo (2012) que resultou na OpenWN-PT. Num primeiro passo, a informação da versão em inglês é projetada para o Português, utilizando dicionários para mapear os membros em inglês de um *synset*, para possíveis candidatos à tradução. No segundo, páginas da Wikipédia são ligadas aos *synsets* mais importantes da PWN, segundo a posição na hierarquia e ao número de relações nas quais estão envolvidos. Essa ligação permite o mapeamento entre os

[4] http://www.bcb.gov.br/glossario.asp?Definicao=463&idioma=P&idpai=GLOSSARIO

títulos de artigos da Wikipédia em Português e as relações semânticas que são herdadas da PWN.

Outro trabalho na direção de universalizar a PWN foi proposto por Bond & Foster (2013). Eles criaram uma wordnet aberta de alta qualidade relacionada às correspondências dos sentidos entre as diversas línguas. Ela compreende 26 línguas, inclusive o Português, com mais de 10.000 *synsets*, que correspondem à cobertura entre 42% e 100% dos conceitos mais comuns de cada língua. O princípio é estabelecer a ligação entre o inglês e outras línguas que já possuem alguma base léxica que possa ser relacionada à PWN. Para o Português brasileiro, foi feita a combinação entre wordnets e wiktionary[5] para fornecer a tradução.

Oliveira (2013), no espírito de expansão de bases léxicas para o Português, criou a Onto.PT, uma ontologia léxica que se propõe a minimizar as principais limitações de recursos similares existentes. O que significa que o novo recurso deve ser público, construído automaticamente, criado desde o início para a língua portuguesa e estruturado semanticamente em relação ao sentido das palavras. A Onto.PT tem estrutura semelhante à PWN, isto é, agrupada por *synsets* que são lexicalizações de um conceito e relações semânticas entre eles.

Outro recurso utilizado foi o DBnary, proposto recentemente por Sérasset (2014). Trata-se de recurso lexical multilíngue que inclui correspondência entre idiomas e definições das palavras agrupadas conforme o sentido. Para tal, o autor empregou como fonte 25 diferentes línguas de edições do Wiktionary[6], que é um projeto colaborativo para produzir dicionários completos e gratuitos para toda e qualquer língua. O recurso foi disponibilizado para a comunidade como *linked data* que significa que qualquer pesquisador trabalhando com WS tem acesso direto. De certa forma, esse trabalho também é comparável à iniciativa da *Open Multilingual Wordnet*, haja vista o alinhamento entre diferentes idiomas para as entradas lexicais.

O corpus, em Português e em linguagem natural, é composto por documentos selecionados da Web publicados por instituições brasileiras tais como Banco Central[7], BNDES[8], CVM[9] etc. Essas informações são de domínio público, portanto, não há restrição referente ao acesso a elas. Logo, o conjunto de documentos, que regulamentam os procedimentos relacionados à boa prática na gestão de risco de entidades financeiras do Brasil, tais como ofícios-circulares, resoluções, leis, instruções normativas, compõe a matéria-prima da base de dados textual.

A coleção de textos sobre risco financeiro possui 2.978 documentos em língua portuguesa. Os documentos estão em vários formatos conhecidos pela maioria de usuários. São eles: *Microsoft Office Word* (doc e docx) e *PowerPoint* (ppt), *Portable Document Format* (pdf) e *HyperText Markup Language* (html). Além desses, utilizou-se a Wikipédia em Português que contém 1.385.451 documentos em *eXtensible Markup Language* (XML). A justificativa é a

[5]http://pt.wiktionary.org/wiki/Wikcion%C3%Alrio:P%C3%Algina_principal
[6]www.wiktionary.org
[7]http://www.bcb.gov.br
[8]http://www.bndes.gov.br/
[9]http://www.cvm.gov.br/

possibilidade de encontrar variadas formas de lexicalização de classes ou propriedades da ontologia que proporcionem melhor generalização de padrões.

Um último recurso utilizado como fonte para elaboração de modelo capaz de etiquetar automaticamente as categorias gramaticais de textos em Português foi o *corpus* Floresta Sintáctica[10], proposto por Freitas, Rocha & Bick (2008). O Floresta é um projeto de criação e disponibilização de um *corpus* sintaticamente anotado para o Português que, atualmente, se subdivide em 4 partes:

- Floresta Virgem: contém cerca de 1.6 milhões de palavras ou 95 mil frases coletadas dos corpora CETENFolha[11] e CETEMPúblico[12] e anotadas automaticamente;
- Amazônia: possui 3.8 milhões de palavras ou cerca de 194 mil frases retiradas do Overmundo[13], um site colaborativo voltado para a cultura brasileira. A Amazônia também é um *corpus* sintático que não foi revisado por linguistas;
- Selva: contém cerca de 300 mil palavras divididas entre diferentes modalidades, gêneros, domínios do Português brasileiro e europeu. A Selva foi criada com a intenção de ser parcialmente revista;
- Bosque: parte revisada linguisticamente da Floresta. Abrange 190 mil palavras, 9.368 frases, retiradas dos primeiros 1000 extratos dos corpora CETENFolha e CETEMPúblico.

Para essa pesquisa, utilizou-se o *corpus* Floresta que está incorporado à ferramenta NLTK. São 9.266 frases correspondente à "Floresta Sintá(c)tica Corpus"versão 7.4 — parte Bosque. A descrição das principais etiquetas (tag) realizada pelo parser são apresentadas no Quadro 5[14].

5.2.3.2 Recursos computacionais

Dada a natureza e a complexidade do PNL e da RI, alguns recursos computacionais foram essenciais para a execução deste trabalho. A premissa foi utilizar software livre ou de código aberto *free open-source*[15] e já estáveis no mercado. O objetivo foi minimizar o impacto com programação e com adequação de sistemas e equipamentos, haja vista que o foco era a apresentação do modelo e não o desenvolvimento de sistemas. Portanto, segue uma descrição sucinta dos softwares e hardwares utilizados.

Para elaboração da ontologia, utilizou-se o Protégé[16], versão 4.3. Trata-se de plataforma

[10] http://www.linguateca.pt/Floresta/
[11] parte do *corpus* NILC/São Carlos, retirado de textos do jornal brasileiro Folha de São Paulo, de 1994
[12] retirados do jornal português PÚBLICO
[13] http://www.overmundo.com.br
[14] A descrição e todas as etiquetas disponíveis podem ser encontrada em http://www.linguateca.pt/floresta/BibliaFlorestal/.
[15] Tecnicamente há diferenças entre software livre e código aberto. Pragmaticamente, ambos podem ser usados gratuitamente e o código está disponível para alteração. Contudo, possuem licenças de uso diferentes. Veja: http://pt.wikipedia.org/wiki/C%C3%B3digo_aberto e http://pt.wikipedia.org/wiki/Software_livre.
[16] http://protege.stanford.edu/

Quadro 5: Etiquetas da Floresta Sintá(c)tica

Símbolo		Categoria
n		nome, substantivo
prop		nome próprio
adj		Adjectivo
n-adj		flutuação entre substantivo e adjectivo
v	v-fin	verbo finito
	v-inf	Infinitivo
	v-pcp	Particípio
	v-ger	Gerúndio
art		Artigo
pron	pron-pers	pronome pessoal
	pron-det	pronome determinativo
	pron-indp	pronome independente (com comportamento semelhante ao nome)
adv		Advérbio
num		Numeral
prp		Preposição
intj		Interjeição
Conj	conj-s	conjunção subordinativa
	conj-c	conjunção coordenativa

Fonte: http://www.linguateca.pt/floresta/BibliaFlorestal/

livre para criação e edição de ontologias com interface de usuário personalizável. Possui ferramentas de visualização que permitem navegação interativa em relações ontológicas. Além disso, fornece extensa variedade de plug-in para atender às necessidades diversas na elaboração de ontologias como máquinas de inferência, visualizações, edição de regras, comparação e fusão de ontologias. No desenvolvimento dessa, o suporte na lista de discussão mantida pela Universidade de Stanford, criadora do Protégé, representou grande auxílio nas dificuldades operacionais com o programa.

A linguagem de programação utilizada foi Python 2.7, por ser relativamente fácil para elaborar programas e por possuir uma comunidade de usuários muito ativa, inclusive no Brasil. Outro fator para a escolha, foi a disponibilidade da biblioteca *Natural Language ToolKit* (NLTK) por Bird, Klein & Loper (2009) que oferece muitos recursos para o processamento da linguagem natural em Python. Isso possibilitou a análise da língua portuguesa e a exploração de conceitos linguísticos que permeiam os objetivos desta pesquisa.

Para elaboração de agrupamentos, ou comumente chamados de *clusters*, utilizou-se a biblioteca SciKit-Learn[17], de Pedregosa et al. (2011), que é um módulo Python, composto por extensa gama de algoritmos para aprendizado de máquina, para problemas supervisionados e não-supervisionados. Além da integração com a linguagem em que se desenvolve esta pesquisa, fornece implementações do estado da arte de conhecidas técnicas na área. Em particular, a preparação de dados, a execução de abordagens para criação de agrupamentos

[17]http://scikit-learn.org/stable/

e desempenho satisfatório foram fatores que determinaram a escolha dessa ferramenta.

A biblioteca RDFlib[18] foi empregada para lidar com os dados em RDF e OWL em ambiente Python. Ela possui parsers e serializadores para dados em RDF/XML e turtle que foram utilizados neste trabalho. Ainda, fornece implementação SPARQL que possibilita consultas e atualizações nas bases RDF, via programação em Python. Apesar da facilidade para uso da ferramenta, ela apresentou desempenho satisfatório somente para pequenas bases. Os volumes maiores demandaram outras soluções para que o desenvolvimento do projeto fosse factível.

O Apache Jena Fuseki[19], versão 1.0.0, foi utilizado para superar o baixo desempenho verificado na RDFlib em grandes bases. A solução livre e de código aberto, baseada em Java, é amplamente utilizada para desenvolvimento de aplicações de Web Semântica. Mostrou-se eficiente para armazenar os dados RDF, boa interface para submissão de consultas Sparql e bom desempenho nas respostas. Apesar de não utilizar uma *Application Programming Interface* (API) específica para conectar ao servidor jena-fuseki, contornou-se o problema utilizando comandos via sistema operacional que atenderam às expectativas.

O Solr[20] é um projeto de código aberto, baseado em Lucene Java, que oferece plataforma de busca de alta performance. Trata-se de um servidor de buscas em documentos completos — *full-text* — com desempenho satisfatório. Possui vários recursos para tratamento de textos e possibilita a indexação e busca de documentos em formatos tais como pdf, doc, docx, xml, html etc. No desenvolvimento da pesquisa, utilizou-se a versão 4.6.0 que se mostrou eficiente na indexação e ainda contém vários recursos para elaboração de consultas. Além disso, a integração com Python foi relativamente simples via API.

O conjunto de ferramentas mostrou-se suficiente e possibilitou restringir a execução das tarefas ao ambiente Python/Solr/Jena-Fuseki e Protégé. O sistema foi modelado e testado em um Macbook Pro com processador 2.7 GHz Intel Core i7, 4Gb de memória e 500Gb de HD.

5.3 Abordagem para responder a questão I

Relembrando a **Questao I**: É possível construir automaticamente, por meio de processamento automático da linguagem natural, uma base de léxicos para o Português brasileiro, a partir de ontologia e de *corpus* vinculados, que possa ser lida por computadores e na qual se explicitem os componentes morfológicos, sintáticos e semânticos?

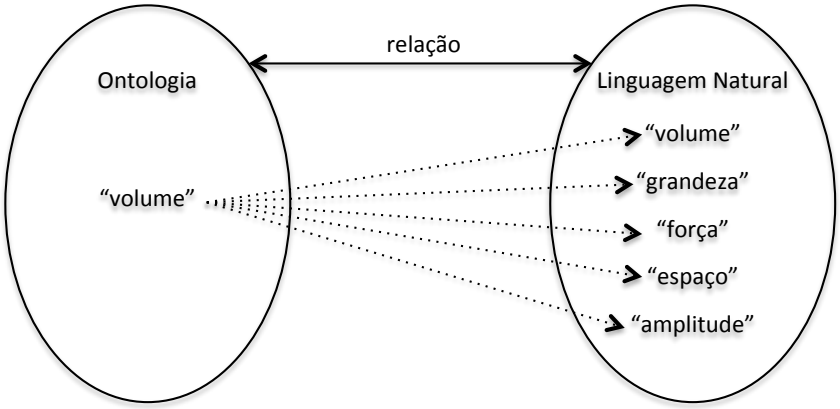
Da mesma forma como crescem os recursos estruturados como o *Linked Open Data*, cresce a necessidade de acessar essa informação em linguagem natural, normalmente utilizada por usuários finais. Aumenta, também, a necessidade de se conhecer como são verbalizados em LN os elementos disponíveis em ontologias. Embora se permita alguma informação linguística

[18] https://rdflib.readthedocs.org/en/latest/
[19] http://jena.apache.org/index.html
[20] http://lucene.apache.org/solr/

em SKOS ou RDFS, ainda é muito limitada e, portanto, restringe as opções para aplicações em linguagem natural.

Como ilustração, considere a tripla (:?sujeito :volume :?objeto). O sujeito e objeto são variáveis e a propriedade é fixa. O significado da propriedade "volume" pode ser verbalizada em LN de várias formas. Por exemplo, se o sujeito for som, o significado seria de "força" ou "amplitude"; se for rio, "espaço ocupado" ou "extensão"; se livro, "obra impressa com mais de 100 páginas"; tratando-se de instituição financeira, "vulto" ou "grandeza".

Figura 29: Relação Ontologia e Linguagem Natural



Fonte: Elaborado pelo Autor

A Figura 29 ilustra a relação entre rótulo anotado, *label*, em uma ontologia e possíveis formas de expressão do mesmo termo em linguagem natural, em que é possível observar que as relações expressas em ontologias podem possuir várias verbalizações. Do ponto de vista de aplicações de PLN, a vinculação do léxico às estruturas ontológicas evita o retrabalho em situações nas quais se necessite capturar esse conhecimento linguístico.

5.3.1 Modelo de lexicalização de ontologia

Para representar as informações linguísticas, adotam-se os princípios definidos por McCrae, Spohr & Cimiano (2011) na proposta do *Lexicon Model for Ontologies* (lemon), o qual se destina a criar um formato padrão de informações linguísticas em RDF. Esse modelo trata de especificações declarativas de um léxico legível por máquina que captura aspectos morfológicos, sintáticos e semânticos dos itens lexicais vinculados a uma ontologia. Isto é, indica-se explicitamente como os elementos do vocabulário de determinada ontologia são verbalizados em uma linguagem específica. Nesse caso, em Português que ainda não havia sido contemplado na visão disponível no momento da pesquisa.

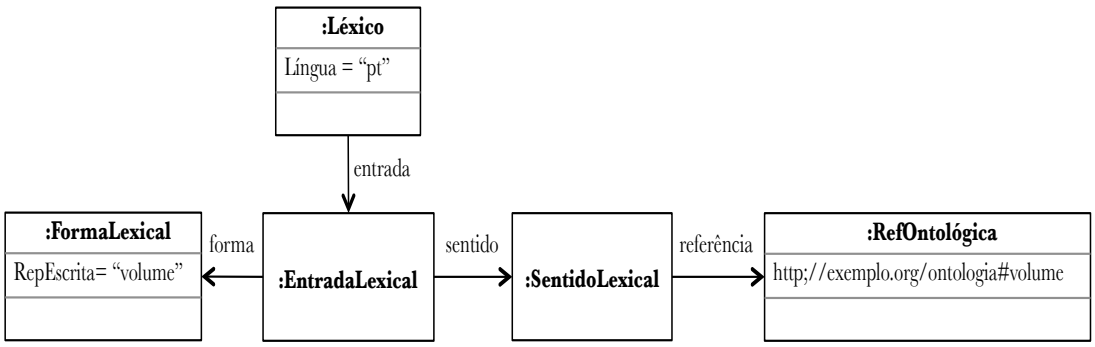
A arquitetura do projeto segue as premissas básicas de ser conciso tanto quanto possível, ser descritivo e não prescritivo, ser modular e ser representado nativamente em RDF. Essa abordagem permite manter poucas classes e definições, utilizar fontes externas para definições, se dividir em módulos para facilitar a utilização parcial do modelo e possibilitar o

compartilhamento na Web Semântica. As principais entidades, conforme McCrae et al. (2010), são:

- Léxico — acervo de todos os termos de um dado domínio;
- Entrada Lexical — padronização morfossintática de uma ou mais Formas Lexicais;
- Sentido léxico — objeto funcional, que liga uma entrada lexical a uma entidade ontológica, que proporciona a desambiguação do termo, i.e., a interpretação do sentido (significado) daquela entrada lexical;
- Referência — entidade ontológica representada por URI de um elemento da ontologia que se vincula a uma forma lexical;
- Forma Lexical — variante morfossintática de uma entrada lexical que inclui inflexões, declinações e variações sintáticas;
- Forma — representação da escrita padrão ou fonética de uma entrada lexical.

Como exemplo, a Figura 30 mostra essas entidades e suas relações:

Figura 30: Principais elementos



Fonte: Adaptado de McCrae et al. (2010, p. 4)

Dessa forma, um aspecto importante é que o sentido deve ser único para o par "Entrada Lexical/Referência Ontológica". Essa consideração implica que cada Entrada Lexical não é semanticamente desambiguada e, portanto, a Referência Ontológica fornece a semântica ao termo. Por exemplo, "O risco aumenta para todas as instituições financeiras" e "O risco é fino" possuem diferentes sentidos para o termo risco, apesar da mesma Entrada Léxica e de apresentarem o mesmo comportamento morfossintático.

Além disso, mais de um termo pode se relacionar com a mesma entidade ontológica, como em "risco" e "perigo". Os termos podem ser considerados sinônimos se, e somente se, se referirem à mesma entidade ontológica como no exemplo a seguir representado na notação turtle[21]:

[21]Para detalhamento da notação, recomenda-se a leitura de http://www.w3.org/TR/turtle/.

Exemplo 5.1: Modelo simples em lemon

```
@base <http://www.example.org/lexico> .
@prefix lemon: <http://lemon-model.net/lemon#> .
@prefix :<http://www.exemplo.org/> .
@prefix ontologia:<http://www.exemplo.org/ontology#> .

:meuLexico a lemon:Lexico;
  lemon:lingua "pt" ;
  lemon:entrada :risco, :perigo .

:risco a lemon:EntradaLexical ;
  lemon:formaCanonica [ lemon:RepEscrita "risco"@pt ] ;
  lemon:outraForma [ lemon:RepEscrita "riscos"@pt ] ;
  lemon:sentido [ lemon:referencia ontologia:risco ] .

:perigo a lemon:EntradaLexical ;
  lemon:formaCanonica [ lemon:RepEscrita "perigo"@pt ] ;
  lemon:outraForma [ lemon:RepEscrita "perigos"@pt ] ;
  lemon:sentido [ lemon:referencia ontologia:risco ] .
```

Dessa forma, tem-se o vocabulário expresso em formato compatível com os princípios da WS e, do mesmo modo, podem-se apresentar variações morfológicas e sintáticas. Tais variações são relevantes para aplicações em PLN que necessitem de representações linguísticas para fornecer melhores resultados. Ao mesmo tempo, a referência a uma ontologia permite que a interpretação fique restrita ao domínio com o qual se trabalha.

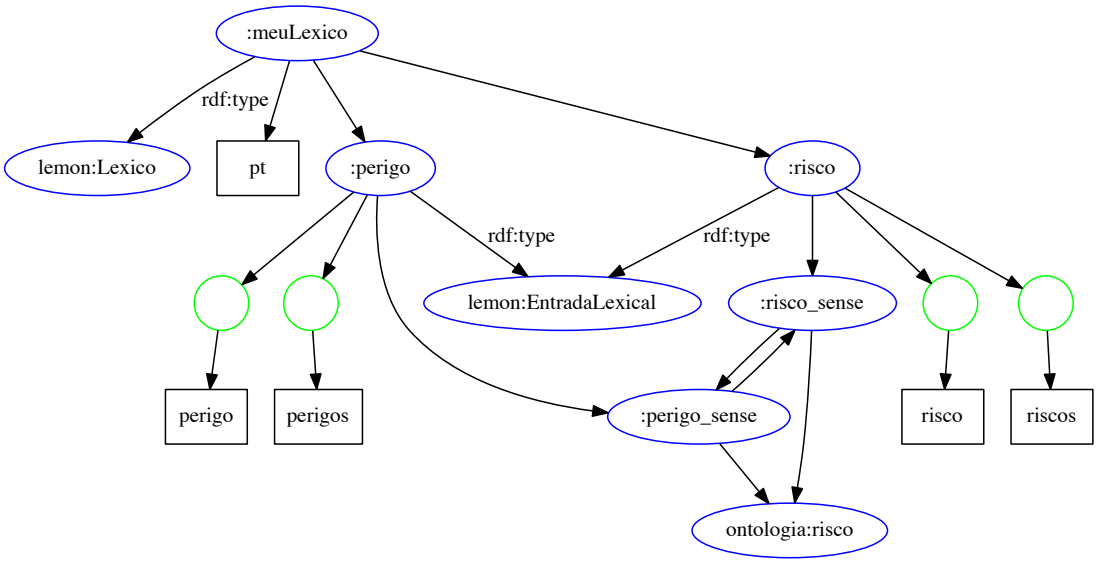
Além de ser legível por máquinas, a vantagem de se trabalhar com uma base léxica no padrão RDF é, a possibilidade de estabelecer relações léxico-semânticas. Na modelagem em formato *lemon*, são definidas as relações de equivalência, incompatibilidade, intensionalidade e extensionalidade. Por exemplo, a relação de equivalência entre "risco" e "perigo" pode ser elaborada na base e, com isso, os processos de manutenção e de depuração ficam centralizados. A Figura 31 mostra um trecho da base que ilustra a relação de equivalência entre os termos.

Outra relação importante para a pesquisa é a de hiponímia. Nela é possível percorrer a taxonomia de conceitos do geral para o mais específico. Na Figura 32, está ilustrada a inclusão do termo "risco operacional" que é hipônimo de "risco". Nota-se que ele adiciona mais especifidade para o significado herdando todas as características de seu hiperônimo.

O modelo *lemon* permite outras informações igualmente importantes para aplicações em PLN como a utilização da ontologia de descrição linguística proposta por Romary (2010) que adota o padrão ISOcat[22] *Data Category Registry* (DCR) que contém definições e descrições linguísticas. No caso desta pesquisa, adotou-se estratégia semelhante e utilizou-se a ontologia para o Português OpenWordnet-PT[23]. Dessa forma, pode-se indicar para cada termo, ou grupo de termos, a forma preferencial ou alternativa, a forma abstrata ou seu lexema — *stem*

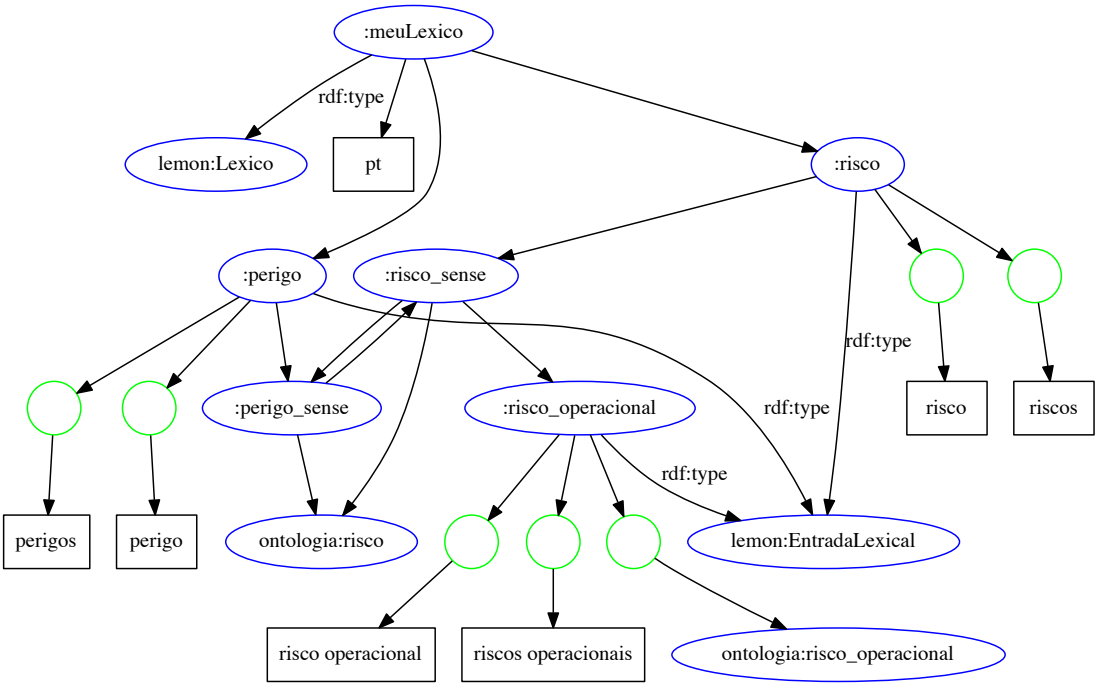
[22] http://www.isocat.org
[23] https://github.com/arademaker/openWordnet-PT

Figura 31: Ilustração da relação de equivalência entre Risco e Perigo



Fonte: Elaborado pelo Autor

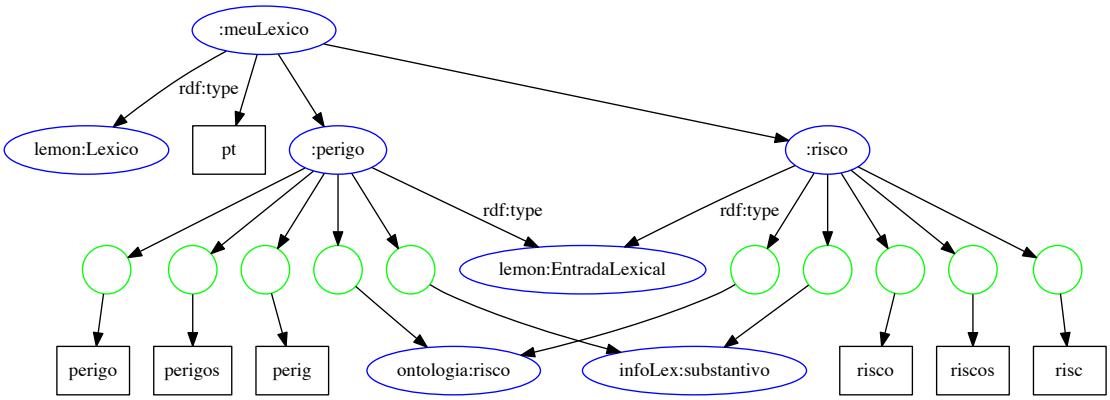
Figura 32: Ilustração da relação de hiponímia



Fonte: Elaborado pelo Autor

—, a classe sintática, a forma flexionada dos verbos entre outras informações linguísticas disponíveis.

Figura 33: Ilustração da Representação lexical em RDF



Fonte: Elaborado pelo Autor

A Figura 33 ilustra o padrão *lemon* dos termos "risco" e "perigo". Nele se observa indicação do idioma, da ontologia de referência, das formas canônica, alternativa e abstrata — lexema — e indicação da classe sintática, nesse caso, substantivo. O modelo é representado em RDF [24], segue os princípios da WS e, portanto, pode ser reutilizado por outras aplicações e é legível por máquinas.

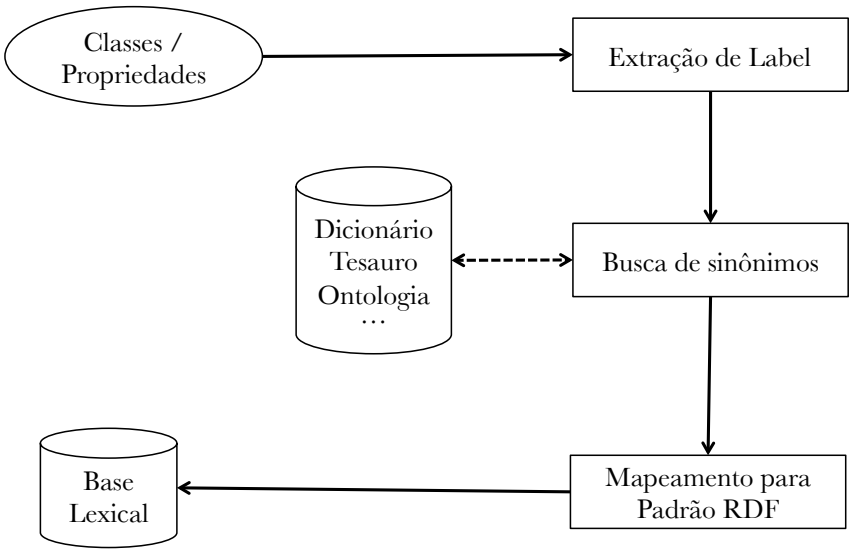
[24]

```
@base <http://www.example.org/lexico> .
@prefix lemon: <http://lemon-model.net/lemon#> .
@prefix :<http://www.exemplo.org/> .
@prefix ontoRisco:<http://www.exemplo.org/ontologia#> .
@prefix infoLex:<http://www.exemplo.org/linguistica#> .
:meuLexico a lemon:Lexico;
  lemon:lingua "pt" ;
  lemon:entrada :risco, :perigo .

:risco a lemon:EntradaLexical ;
  lemon:formaCanonica [ lemon:RepEscrita "risco"@pt ];
  lemon:formaAlternativa [ lemon:RepEscrita "riscos"@pt ] ;
  lemon:formaAbstrata [ lemon:RepEscrita "risc"@pt ] ;
  lemon:cat [lemon:partOfSpeech infoLex:substantivo];
  lemon:sentido [ lemon:referencia ontoRisco:risco ] .

:perigo a lemon:EntradaLexical ;
  lemon:formaCanonica [ lemon:RepEscrita "perigo"@pt ] ;
  lemon:formaAlternativa [ lemon:RepEscrita "perigos"@pt ] ;
  lemon:formaAbstrata [ lemon:RepEscrita "perig"@pt ] ;
  lemon:cat [lemon:partOfSpeech infoLex:substantivo];
  lemon:sentido [ lemon:referencia ontoRisco:risco ] .
```

5.3.2 Abordagem de lexicalização

Para atender à primeira questão desta pesquisa, propõe-se a construção semiautomática de uma base de léxicos em Português para o domínio de Risco Financeiro que daqui por adiante será denominada RiscoLex. Para tal, toma-se a ontologia de risco como base e o *corpus* correspondente, conforme a seção 5.2.3.1. A proposta é extrair os rótulos de classes e propriedades da ontologia, identificar e recuperar os respectivos sinônimos, as características morfossintáticas de cada termo, convertê-los em formato RDF e alimentar a base lexical no padrão *lemon*. A Figura 34 mostra os passos do processo de geração da RiscoLex.

Figura 34: Fluxo de criação da RiscoLex



Fonte: Elaborado pelo Autor

A abordagem inclui a proposição de uma ou mais entradas lexicais para cada classe e propriedade da ontologia. A primeira etapa trata da extração de rótulos da ontologia e informações adicionais, tais como sinônimos e características sintáticas, de recursos externos. A execução da tarefa segue os seguintes passos:

a) A partir da tripla (s, p, o) da ontologia, extraem-se todos os rótulos de s e p para criar uma lista de termos em LN.

b) Nos casos de rótulos em CamelCase (nascimentoLocal), separados por hífen (presidentes-do-Brasil) ou por underscore (instituições_financeiras), é necessário representá-los em LN encontrada em textos. Esse passo visa transformar formas como "paísDeOrigem" em "país de origem" ou "gerenciamento_de_risco" em "gerenciamento de risco".

c) Esses termos são procurados nos corpora para validação. Esse passo visa caracterizar termos frequentes e, consequentemente, preferidos no domínio e na língua portuguesa.

d) Em LN, é comum expressar-se o mesmo sentido com o uso de mais de uma palavra. Portanto, o objetivo é localizar a maior quantidade possível de sinônimos para os termos

da lista. Para essa tarefa, foram usadas as ontologias linguísticas para o Português descritas na seção 5.2.3.1.

e) Para tratar os termos polissêmicos automaticamente e coletar aqueles mais relevantes para o domínio, inspirou-se na abordagem de Lesk (1986). Nessa visão, o processo original de desambiguação emprega definições de dicionários para fornecer contexto para pequenas frases. Por exemplo, na frase "Eu preciso do banco para depositar dinheiro", a ambiguidade do termo "banco" pode representar problemas de interpretação. Afinal, o local de depositar dinheiro está alto e é preciso subir no assento individual, sem encosto e sem braços, para alcançá-lo ou trata-se de uma instituição financeira?

   Contudo, diferente da época, 1986, em que Lesk propôs o algoritmo, atualmente, há mais recursos linguísticos eletrônicos disponíveis como bases de dados que seguem o padrão da Wordnet e mais ferramentas de PLN. Nesse sentido, as definições são coletadas nas bases léxicas mencionadas. Substantivos, verbos e adjetivos são transformados para forma inflexionada, por exemplo: preciso para precisar ou banquinho para banco. As preposições e artigos são retiradas. O Quadro 6 ilustra essa abordagem na frase "Eu preciso do banco para depositar dinheiro".

Quadro 6: Abordagem de Lesk

Termo	Sentido	Contagem	Definição
Precisar	1	0	Requerer, reclamar em virtude de um direito
-	2	0	Estar necessitado, ter falta
Banco	**1 ***	**3**	**Instituição financeira para depósito, aplicação e guarda de dinheiro e valores**
-	2	0	Área administrada pelo Estado e destinada à preservação animal e vegetal
-	3	0	Assento com 3 ou 4 pernas com ou sem encosto
Depositar	**1 ***	**1**	**Por em depósito, entregar solenemente, confiar com solenidade jurídica**
-	2	0	Meter em loja, armazenar
Dinheiro	**1 ***	**2**	**Valor de uma dívida ou de depósito em dinheiro**
-	2	0	Antiga moeda romana
-	3	0	Metal de brilho amarelo

Fonte: Adaptado de Lesk (1986)

O autor utiliza as definições das palavras que compõem a frase e observa os termos que apresentam intersecção em outras definições. As definições em negrito têm o maior número de intersecções que estão sublinhadas. De acordo com a técnica, são esses os possíveis melhores significados no contexto da frase, isto é, o termo "banco"

relaciona-se com os sentidos em negrito dos termos “valor”, “depositar” e “dinheiro”. De fato, qualquer falante fluente do Português sabe que o termo “banco”, nessa frase, remete ao sentido de instituição financeira que é aquele detectado pelo algoritmo de Lesk.

Todavia, quando se tratam de rótulos, a abordagem de Lesk necessita de outra adaptação, pois o termo isolado como “banco” não fornece contexto suficiente. Assim, recorreu-se ao *corpus* de risco como fonte para se gerar agrupamentos de documentos utilizando as técnicas descritas na seção 4.4.4 e aplicadas detalhadamente em Schiessl (2007). Especialistas da área de risco analisaram esses agrupamentos e escolheram aquele que melhor representa o domínio. Nessa técnica, os agrupamentos contêm termos que melhor caracterizam o tema extraído de textos. Dessa forma, é possível fornecer contexto ao rótulo a ser desambiguado e coletar os sinônimos mais correlacionados, segundo a abordagem de Lesk.

O agrupamento fornece uma *Bag of Words* (BoW) das palavras mais relevantes para o domínio. Cada *synset* é comparado com os termos da BoW aplicando a métrica *String Matching* (*normalized Levenshtein*), Concordância de caracteres, proposta por Maedche & Staab (2002a) e revista em Cheatham & Hitzler (2013), que mede a “distância” entre um termo e outro.

Essa métrica baseia-se na *edit distance*, distância de edição, formulada por Levenshtein em 1965, que é um método bem estabelecido para quantificar a diferença entre duas sequências de caracteres. Ela mede o número mínimo de inserções, exclusões e substituições para transformar um termo em outro. Por exemplo, para transformar “risco” em “arisco” necessita-se de uma inserção da letra “a”, portanto (risco, arisco) = 1. A *String Matching* usa o mesmo princípio. Contudo, pondera pelo comprimento do menor termo. Essa ponderação limita o grau de similaridade entre 0 — nenhum ou nenhuma coincidência entre termos — e 1 — máximo ou termos idênticos.

Assim, sejam dois termos L_i e L_j:

$$SM(L_I; L_J) := max\left(0; \frac{min(|L_i|; |L_j|) - ed(L_i; L_j))}{min(|L_i|; |L_j|)}\right) \in [0; 1] \tag{5.1}$$

f) A etapa final é representar os termos coletados no padrão *lemon*. Por exemplo, a classe “empresa” recupera “companhia”. Ambas são representadas como:

Exemplo 5.2: Saída do processo

```
@base <http://www.example.org/lexico> .
@prefix lemon: <http://lemon-model.net/lemon#> .
@prefix :<http://www.exemplo.org/> .
@prefix ontoRisco:<http://www.exemplo.org/ontology#> .
@prefix infoLex:<http://www.exemplo.org/linguistica#> .

:meuLexico a lemon:Lexico;
```

```
  lemon:lingua "pt" ;
  lemon:entrada :empresa, :companhia .

:risco a lemon:EntradaLexical ;
  lemon:formaCanonica [ lemon:RepEscrita "empresa"@pt ] ;
  lemon:outraForma [ lemon:RepEscrita "empresas"@pt ] ;
  lemon:formaAbstrata [ lemon:RepEscrita "empres"@pt ] ;
  lemon:cat [lemon:partOfSpeech infoLex:noun];
  lemon:sentido [ lemon:referencia ontoRisco:risco ] .

:companhia a lemon:EntradaLexical ;
  lemon:formaCanonica [ lemon:RepEscrita "companhia"@pt ] ;
  lemon:outraForma [ lemon:RepEscrita "companhia"@pt ] ;
  lemon:formaAbstrata [ lemon:RepEscrita "companhia"@pt ] ;
  lemon:cat [lemon:partOfSpeech infoLex:noun];
  lemon:sentido [ lemon:referencia ontoRisco:risco ] .
```

Enfim, todos os elementos são recuperados e convertidos, por meio de programa em Python, para a forma padrão apresentada no exemplo 5.2. O processo se estabelece a relação de cada termo com a ontologia de referência e sua função sintática no contexto. Além disso, é possível inserir variantes como a forma plural, lexemas, comentários com exemplos de uso e outras informações relevantes para a aplicação.

5.4 Abordagem para responder a questão II

Retomando a **Questão II**: É possível melhorar a precisão em sistemas de recuperação da informação em Português brasileiro com o uso dessa base de léxicos para desambiguação?

No mundo da RI, a descrição de conteúdos e as técnicas de processamento de consultas baseiam-se em descritores, apesar dos avanços tecnológicos. Isso restringe a capacidade de capturar o conceito envolvido na necessidade do usuário e na significação do conteúdo. Essa limitação inclui a apreensão de sentidos figurados, por exemplo “A quebra do Lehman Brothers iniciou o terremoto no mundo todo”. O termo “terremoto” expressa a crise que abateu o mundo financeiro, não tremor de terra por todo o planeta.

Na falta de termos que traduzam o significado completo, a semântica preenche os espaços deixados pelos descritores. Na frase “O ebola já atravessou as fronteiras da África”, alguém razoavelmente informado infere que o ebola é uma febre hemorrágica em humanos, geralmente letal, originada no continente africano e que se espalha por outros continentes. A diferença entre o que está escrito e o que realmente significa está nas entrelinhas. Isto é, na interação entre leitor e escrita. Da mesma forma, descritores indicam o conteúdo por literalidade, mas podem ser complementados com a semântica daquilo que não está explícito.

O objetivo dessa proposta, na modelagem de SRI, é complementar à busca sintática, amplamente difundida, com a semântica. Intenta-se superar as limitações de modelos baseados em descritores com novas tecnologias fornecidas pela comunidade da WS. As ontologias fornecem a informação semântica para promover consultas mais expressivas e

resultados mais precisos, porém demandam linguagens especializadas. O desafio, portanto, é combinar a usabilidade de modelos baseados em descritores com a precisão das ontologias.

5.4.1 A RiscoLex e a RI

Os SRI tradicionais fundamentam-se na indexação por palavras-chave, ou descritores, para encontrar um documento, porém isso não é suficiente. O problema é que se o termo da consulta não coincide com as palavras-chave, o documento não será recuperado. Por exemplo, na consulta tem-se o termo "perigo", na representação do documento, o sinônimo "risco". Nenhum mecanismo baseado em cálculo de similaridade entre termos irá recuperar tal documento.

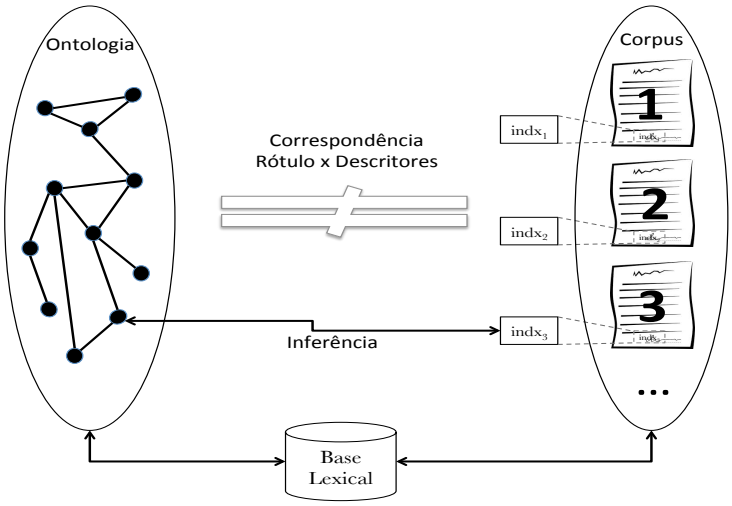
Existe um lapso entre termo e significado correspondente que não está explicitamente representado. Uma solução é acrescentar as possíveis variantes aos descritores para aumentar a cobertura da indexação. Esse procedimento coleta termos e adiciona mais significados ao corpus. Mas, representar o *corpus* exaustivamente pode aumentar a revocação e deteriorar a performance do sistema relacionado à precisão. A arte está no equilíbrio entre esses dois indicadores de desempenho.

O emprego de fontes externas em formato padrão minimiza o custo com indexação. Tal atividade é fundamental para a recuperação da informação, contudo é, geralmente, estática. O ideal é apoiar bases de indexação com recursos reutilizáveis, generalizáveis, automatizáveis e dinâmicos. As bases lexicais e as ontologias são esses recursos e adicionam a possibilidade de ampliar o alcance dos descritores de documentos de duas formas: fornecendo sinônimos e deduzindo termos relacionados por inferência.

A Figura 35 apresenta a visão geral do processo. Um *corpus* e uma ontologia representam o mesmo domínio para usuários distintos: máquinas e pessoas. Os rótulos disponíveis nas entidades ontológicas não possuem, geralmente, correspondência com os descritores de documentos. Por exemplo, o rótulo da classe PessoaFísica seria expresso nos descritores como "Pessoa Física". Nesse caso, a RiscoLex, vinculada a ontologia, provê o lema e sinônimos aos descritores. Caso o descritor não esteja na RiscoLex, nem na ontologia, ele pode ser incorporado semiautomaticamente em ambos para enfatizar a natureza dinâmica da área do conhecimento.

Além disso, a cobertura da indexação pode ser ampliada por meio de inferência que fornece significados semânticos explicitamente. A inclusão de ontologia para apoiar o SRI fornece mais significados por meio de motores de inferência. Nesse caso, ela pode ser vista como uma extensão dinâmica dos descritores de documentos. Por exemplo, da classe Especialista pode-se inferir que os membros também pertencem às classes *Stakeholder* e "pessoaFísica". Por isso, herdam todos os atributos e restrições destas, sem que sejam expressos explicitamente, apenas por meio de axiomas. Dessa forma, a resolução automática de hiperônimos, como "banco" e "instituição financeira", e de outras formas de dependência entre palavras acrescentam mais exatidão à representação de um dado documento.

Figura 35: Conceitos e termos



Fonte: Elaborado pelo Autor

5.4.2 Modelo de Recuperação da Informação Semântica

Na visão do Modelo de Recuperação da Informação Semântica (MoRIS), assume-se que a ontologia foi elaborada e associada às fontes de informações textuais que contemplam os conceitos que se quer representar. Admite-se também que, apesar desta pesquisa estar restrita ao domínio de risco financeiro, o modelo pode ser aplicado a qualquer área, desde que se disponham de informações estruturadas e não estruturadas que possam representar os conceitos compreendidos pelo domínio.

A ferramenta Solr foi usada para armazenar e indexar o corpus. Ela facilita a criação de motor de busca para websites, bases de dados e arquivos. Além disso, permite respostas rápidas às buscas graças ao tipo de indexação utilizado, o índice invertido. Nessa estrutura, o dado é representado da seguinte forma: documento é a unidade de busca e de indexação. O índice consiste de um ou mais documentos que contêm um ou mais campos. Analogamente às bases de dados, documentos representam as linhas das tabelas, enquanto campos representam as colunas.

A customização do sistema emprega a linguagem Python. Usa-se o módulo NLTK e Scikit Learn para tarefas de PLN e para interagir com o Solr, a biblioteca PySolr, que é um cliente Python que permite adicionar, remover e indexar documentos, submeter consultas, processar resultados e alterar parâmetros — no momento de execução do programa — à instância Solr. Dessa forma, além de permanecer com soluções em software livre, as tarefas de ajustes do sistema se mantêm separadas do ambiente de armazenamento e indexação do corpus.

O domínio é representado por ontologias e corpora. De um lado, entidades ontológicas representam conceitos e motores de inferência deduzem automaticamente informações que não estão explícitas. De outro, descritores descrevem conteúdos de documentos e pessoas interagem em LN para inferir significados não expressos. Ambos se complementam na tarefa

de prover informação, mas em formatos diferentes ou mesmo incompatíveis.

Influenciado por Fernández et al. (2011) e Kara et al. (2012), o modelo adota a estrutura da RI baseada em descritores com a inclusão de um módulo semântico. Embora as buscas com linguagens especializadas, como SPARQL, apresentem melhor desempenho, demandam conhecimento restrito à pequena parcela de usuários técnicos. Por outro lado, graças aos populares motores de busca como Google e Yahoo!, o usuário típico já se habituou à interface de buscas por palavras-chave, ao desempenho e à escalabilidade testadas no âmbito da Web dos tradicionais SRI.

Propõe-se, então, o método de consultas baseado em descritores para possibilitar a submissão de pesquisas ao corpus, à ontologia e à RiscoLex de forma transparente para o usuário. O domínio do risco financeiro e corporativo está circunscrito à ontologia de domínio, a qual complementa a informação não explícita no *corpus* por meio de inferência. Todavia, não há buscas explícitas na ontologia. As pesquisas são respondidas pelo *corpus* por meio de índices invertidos de SRI tradicionais.

Os documentos e entidades ontológicas são indexados conjuntamente. Essa opção de modelagem facilita a interação com o usuário final, pois continua realizando pesquisas da mesma forma que faz nos motores de busca tradicionais. Além disso, apresenta o resultado final, no mínimo, tão bom quanto à abordagem tradicional. Isto é, caso a consulta não encontre correspondentes na base de conhecimento, o sistema recupera a informação relacionada aos descritores dos documentos.

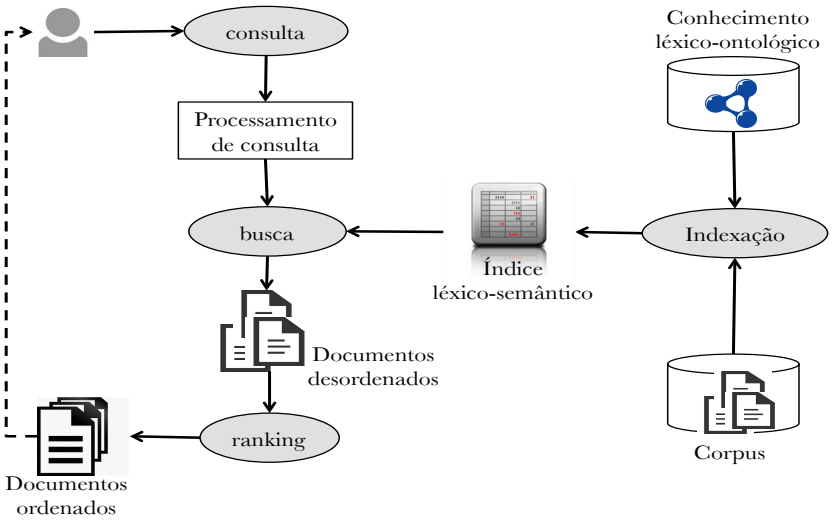
A Figura 36 ilustra o processo da RI com a adição do módulo semântico. O usuário interage de forma tradicional para submeter a consulta. O processamento da consulta realiza a padronização dos termos para a busca. O conhecimento léxico-ontológico é constituído pela ontologia e pela RiscoLex. O *corpus* caracteriza a base que contém documentos a serem recuperados. A indexação conjunta das bases envolvidas fornece o índice léxico-semântico, que é utilizado na recuperação e no ranking de documentos recuperados para apresentação ao usuário.

O núcleo do MoRIS compreende quatro etapas: consulta, busca, indexação e ranking. Os detalhes são apresentados a seguir.

5.4.2.1 Consulta

O usuário interage com o sistema de forma tradicional, ou seja, realiza a pesquisa em LN. A interface para consulta semântica é baseada em descritores para proporcionar melhor usabilidade. No passo seguinte, a consulta é processada com ferramenta de PLN para normalizá-la. O processo envolve a retirada de stopwords ou palavras não discriminatórias, a padronização de termos às respectivas formas canônicas. Por exemplo, considerando a consulta “Bancos situados em Brasília”, o sistema analisa e traduz para “banco situar brasília”.

Figura 36: Visão Geral do MoRIS



Fonte: Adaptado de Fernández et al. (2011, p. 438)

5.4.2.2 Busca

A transformação da consulta determina os conceitos envolvidos na busca para compará-los com os índices. O índice semântico coleta todos os documentos que contenham os conceitos expressos na consulta. O grau de correspondência entre consulta e documentos é fornecido por meio de cálculo de similaridade entre termos da consulta e do índice. O motor de busca oferece opções de pesquisa exata — recupera somente termos morfologicamente iguais — ou por semelhança — termos aproximados. Exemplo: na pesquisa exata pelo termo “contato”, os documentos com a variação do Português europeu, “contacto”, não seriam recuperados.

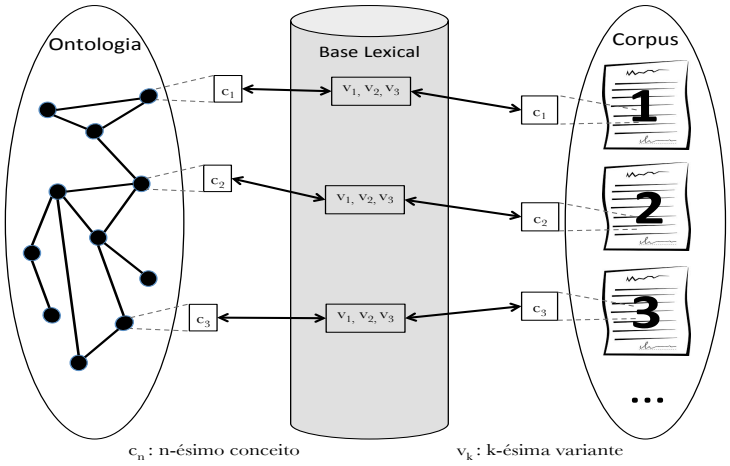
5.4.2.3 Indexação

O processo de indexação é a principal diferença do modelo com os SRI tradicionais. A informação disponível no *corpus* é vinculada ao espaço semântico por meio de anotação explícita nos documentos com dados léxico-semânticos. O *corpus* e o conhecimento léxico-ontológico, ontologia e RiscoLex respectivamente, são indexados em conjunto para produzir o índice léxico-semântico. Esse índice contém as palavras-chave extraídas dos documentos e as informações léxico-ontológicas da RiscoLex.

O índice invertido é a forma de indexação mais utilizada na RI. Para usuários não técnicos, a expressão pode soar estranha, mas o conceito é simples e engenhoso. Metaforicamente, pode-se imaginar um livro como um *corpus* e as páginas como documentos. O índice invertido seria o índice remissivo no final do livro. Ao invés de procurar o termo página por página, localiza-se o termo no índice remissivo que contém o número de página(s) em que ele se encontra. Similarmente à busca por índice remissivo, a recuperação do documento também é mais rápida.

Portanto, no índice invertido, as palavras-chave são associadas aos documentos nos quais estão presentes. Na indexação léxico-semântica, a construção do índice invertido é semelhante ao processo de anotação semântica. Isto é, o índice contém as palavras-chave e as entidades semânticas correspondentes que estão explicitamente anotadas nos documentos. Essa relação entre entidade semântica e documento é chamada de anotação que é considerada tanto para recuperação quanto para ranking de documentos.

Figura 37: Anotação semântica



Fonte: Elaborado pelo Autor

A Figura 37 mostra a relação entre conceitos — c_n — e variantes léxicas — v_n — na base de conhecimento e documentos que contêm esses conceitos. De um lado, a RiscoLex possui a entrada de um termo relacionado com suas variantes lexicais, classe sintática e vinculação com a ontologia. Por outro, as informações léxicas são vinculadas aos documentos nos quais estão presentes.

Nessa visão, a RiscoLex faz a ligação entre os conceitos expressos nos documentos e na ontologia. Por exemplo: o conceito “risco” está associado ao termo “perigo” e classificado como substantivo. Os documentos que tenham os termos “risco” e “perigo” classificados como substantivos são candidatos à anotação semântica. Além disso, os termos maiores são preferíveis, pois se admite que termos mais longos contêm significado mais específico, por exemplo, “risco de crédito” é mais específico que “risco”. Nesse caso, o primeiro será o selecionado para anotação.

O processo de anotação semântica é, portanto, fundamental para vincular documentos ao espaço semântico conformado pela ontologia de domínio. O PLN é a ferramenta principal para identificar, comparar e anotar os documentos. Contudo, buscando minimizar possíveis efeitos da ambiguidade, ela é complementada por validação humana. Os passos da anotação semântica são os seguintes:

a) extrair todas as entidades ontológicas e variantes léxicas para uma lista;

b) analisar documentos para retirar símbolos sem relevância para o conteúdo textual, como tags HTML ou logos;

c) analisar o texto para extrair os termos com POS e lexemas correspondentes;

d) identificar n-grams ou padrões POS, exemplo “substantivo + substantivo”, “nome próprio + nome próprio”;

e) eliminar stopwords;

f) comparar com rótulos da ontologia para obter o subgrupo a ser anotado;

g) ndicar a classe gramatical do termo a ser anotado;

h) indicar a similaridade do termo com o significado do domínio, em caso de termos homógrafos. Exemplo: “Artigo” está relacionado à mercadoria para o domínio e não para documento escrito.

i) confirmar da anotação por especialista de domínio;

j) adicionar a anotação aos documentos do corpus.

Essa anotação beneficia tanto na recuperação quanto no ranking. A primeira acrescenta pontos de acesso que, consequentemente, aumentam a revocação e, ainda, os termos anotados restringem o significado ao domínio delimitado pela ontologia. Isso favorece a redução da ambiguidade das palavras e melhoram a precisão. A segunda utiliza tradicionais algoritmos de ranking para atribuir maior importância às palavras-chave combinadas com as anotações semânticas.

5.4.2.4 Ranking

O algoritmo mais popular de ranking para SRI é o tf-idf – seção 4.4.5, pág. 106. Heuristicamente, essa abordagem está conectada à Lei de Zipf — seção 4.4.6.1, pág. 108. O valor do tf-idf aumenta proporcionalmente ao número de vezes que um termo aparece no documento, mas compensado pela frequência do mesmo termo no corpus. A ideia é dar maior relevância aos termos que possuem maior poder de discriminação, ao invés dos mais comuns como artigos e preposições.

A recuperação semântica é alcançada pela introdução de fator de ponderação entre descritores e respectivas anotações semânticas. O processo resulta em modificação do algoritmo tf-idf para obter valores mais altos considerando a informação ontológica. Ou seja, as ocorrências de descritores com anotação semântica correspondente são ponderadas de forma a aumentar a relevância de documentos com essas características.

Finalmente, o usuário obtém lista de documentos ordenada, de acordo com os valores do tf-idf, e ponderada, também, pela informação semântica. Caso não haja anotação semântica,

o resultado é uma lista ordenada de documentos de forma clássica, isto é, pelas palavras-chave. Esse enfoque garante que o ranking seja, no mínimo, igual ao padrão de SRI tradicionais.

5.4.3 Avaliação

Julga-se que o principal propósito da pesquisa não é testar exaustivamente um *corpus* de risco financeiro, mas mostrar as vantagens e apontar obstáculos na construção de um sistema de recuperação de informação semântica. Portanto, para avaliar o desempenho do sistema, foram escolhidos 785 documentos relacionados às definições de conceitos do domínio para viabilizar a verificação manual dos resultados obtidos pelo processo de recuperação.

Assim, a operação de indexação ocorreu duas vezes: na primeira, os documentos foram indexados em seu estado bruto, sem pré-processamento para representar o processo de busca tradicional; na segunda, contemplaram-se os documentos pré-processados por mecanismos de processamento de linguagem natural e contendo anotações semânticas para compor o índice léxico-semântico.

Formularam-se cinco consultas para testar o sistema. A ideia foi partir de consultas mais simples, tais como as realizadas com termos simples do tipo "banco", até as mais complexas, como a utilização de termos com ambiguidade inclusive no domínio, como "organização", "seguro", "recompensa". Todas as consultas foram submetidas às duas bases para comparação do resultado entre a busca tradicional e a busca semântica. O conjunto das 5 questões está listado e discutido na seção 5.5.3.

Como forma de observar o comportamento do sistema com a aplicação do PLN, ainda foram testadas as mesmas consultas, utilizando a lematização dos termos. O objetivo foi constatar a influência na precisão em função da aplicação dessa técnica. Como se trata de algo custoso, do ponto de vista de processamento e de conhecimento técnico, o resultado mostra os efeitos da adoção de tal procedimento.

5.5 Resultados e discussão

5.5.1 Questão I

O primeiro ponto a see destacado é a criação da primeira base léxica, RiscoLex, em Português brasileiro no padrão lemon que se diferencia das demais pela interpretação da linguagem circunscrita ao domínio bem definido. Além disso, a ontologia, como recurso para interpretação da linguagem natural, coloca a base léxica no centro do processo de interpretação. Nesse sentido, o nível de granularidade da representação, à qual o significado da linguagem natural é capturado, não é direcionado pela língua, mas pela distinção semântica feita em uma ontologia. Assim, essas distinções são relevantes apenas no contexto de um domínio específico.

Outro ponto a ser ressaltado é a construção da primeira ontologia de gestão de risco em Português. A dificuldade de construção desse tipo de recurso é frequentemente relatada em trabalhos acadêmicos. Aqui não foi diferente. Apesar de utilizar outros recursos como ponto de partida, a especificidade do assunto demandou a construção dessa ontologia como se fosse nova. A diversidade entre os mercados financeiros internacional e nacional obrigou a repensar os conceitos e suas relações, para que estivessem de acordo com o mercado brasileiro e, em especial, a empresa pública. Essa adequação exigiu grande esforço para representar tal conhecimento.

A validação dos rótulos da ontologia com o *corpus* apresentou um resultado aquém do esperado. Apenas 50,77% referentes às classes — ou sujeitos — foram encontrados no *corpus* e das propriedades, somente 20%. O pequeno número constatado reflete a má escolha de termos para constar nos rótulos das classes e propriedades. Esse processo indica que a seleção de termos sinônimos deve melhorar a representação do conhecimento em relação ao material escrito disponível. A participação de especialistas da área é preponderante para a escolha mais adequada desses termos.

Para a coleta de sinônimos — *synsets*—, a primeira opção foi a utilização da base Onto.PT. Apesar da elaboração semelhante à PWN, todas as relações também estão traduzidas para o Português. Inicialmente, essa potencial vantagem para a língua mostrou-se desfavorável, pois todos os outros recursos disponíveis mantêm essas palavras-chave em Inglês. Tal uniformidade justifica-se para manter um padrão de acesso às variadas bases, como ocorreu neste trabalho.

A OpenWN-BR apresentou-se como a principal fonte de consulta para a obtenção dos *synsets*. Uma vez que possui boa cobertura, está totalmente integrada com a PWN, a estrutura é idêntica e utiliza as mesmas palavras-chaves. Diferentemente da Onto.PT, não fornece acesso online, mas está disponível publicamente para *download*. A partir disso, basta carregar em um servidor para acesso às informações. Outro inconveniente é o acesso exclusivo via SPARQL, que demanda conhecimento especializado para elaborar e realizar consultas.

A base DBnary também foi utilizada. Nela foram recuperadas definições de termos e classificações gramaticais quando disponíveis. Contudo, as dificuldades foram o acesso exclusivo via SPARQL, a baixa cobertura e a falta de identificação para variantes do Português europeu e brasileiro. Isso proporcionou a ocorrência de resultados inesperados como, por exemplo, para o termo "economias", recuperam-se "absorvente_feminino", "absorvente_íntimo" e "penso_higiénico que correspondem à variante europeia. Para a brasileira, o mesmo termo significa o plural de "economia" ou referência informal à "poupança", mas esses significados não estão disponíveis na base.

Em complemento às bases anteriores, a atualização da biblioteca NLTK 3.0 em julho de 2014, durante o processo de construção desta pesquisa, foi, portanto, uma grata surpresa. Com ela, foi oferecida a *Open Multilingual Wordnet* contemplando o Português. Por ser uma biblioteca nativa da linguagem Python, utilizada em todo trabalho, facilitou muito no quesito programação. Ela possui o Português alinhado ao Inglês, isto é, para descobrir alguma

relação busca-se o termo inglês que retorna o resultado em Português. O problema é a baixa cobertura e definições somente em Inglês.

Na etapa de verificação dos sinônimos coletados com os termos utilizados nos documentos que compõem o corpus, houve vários percalços. Primeiro, em função de incompatibilidade nos formatos de arquivos com a ferramenta. Isso requereu a transformação de documentos nos formatos pdf, html, ppt e xls para o txt que demandou esforço considerável. Segundo, a dificuldade da ferramenta para o reconhecimento de caracteres latinos, com diacríticos característicos da língua portuguesa, requereu inúmeras intervenções para solucionar erros apresentados durante a execução de programas.

Para a elaboração dos agrupamentos de documentos, alguns cuidados foram tomados com relação à especificidade do vocabulário da área. Para tanto, foi realizada a identificação de *collocations* ou, em Português, termo composto que é a reunião de palavras que expressam um sentido. Exemplo, "risco de crédito" possui sentido específico e não pode ser separado em "risco", "de" e "crédito". Nessa tarefa, a identificação de locuções deve vir antes da eliminação das *stopwords*. Caso contrário, a eliminação da preposição "de" impossibilitaria a identificação de "risco de crédito".

O procedimento de agrupamento gerou três grupos. O grupo escolhido foi considerado por especialistas aquele que mais possui palavras relacionadas ao Risco Financeiro. Em seguida, utilizou-se o mais representativo deles como a fonte para a criação da *bag of words* (BoW) do risco. O agrupamento passou por diversos processamentos para que se encontrasse a BoW apropriada para a comparação com os sinônimos.

Quadro 7: Etapas de pré-processamento

Sem processamento	
Total number of terms (tokens)	90.533
Total number of unique terms (Vocabulary)	11.227
Termos compostos identificados	
Total number of terms (tokens)	89.613
Total number of unique terms (Vocabulary)	11.251
Termos em minúsculo	
Total number of terms (tokens)	89.613
Total number of unique terms (Vocabulary)	9.130
Sem *stopwords*	
Total number of terms (tokens)	42.394
Total number of unique terms (Vocabulary)	8.511
Lematização – *Snowball stemmer*	
Total number of terms (tokens)	42.394
Total number of unique terms (Vocabulary)	5.494
Lematização – RSLPS *stemmer*	
Total number of terms (tokens)	42394
Total number of unique terms (Vocabulary)	5056

Fonte: Elaborado pelo Autor

Os resultados em etapas do pré-processamento – descritos na seção 4.4.4 – estão apresentados no Quadro 7 que mostra a inserção de recursos a cada passo e como isso interefere na redução da quantidade termos a serem utilizados para o estudo.

A cada etapa do processamento observa-se a redução de termos que farão parte da BoW. Há uma redução 53%, i.e., de 90.533 para 42.394 termos no agrupamento. Destaca-se que a utilização dos algoritmos[25] de lematização não produziu bons resultados, pois a redução de termos como "empresa" em "empres" agrupa também a forma lematizada de "empresário" ou "empresarial" e, nesse trabalho, não se deseja que representem a mesma coisa. Portanto, adotou-se a forma mais usual na língua.

Desse modo, a versão final para comparação com os sinônimos contém termos mais característicos que diferenciam os agrupamentos e os homogeneíza internamente. A representação gráfica está na Figura 38:

Figura 38: Bag of Words de Risco Financeiro

Fonte: Elaborado pelo Autor

Para o cálculo de similaridade entre *synsets* e termos do agrupamento, optou-se utilizar a forma não lematizada, pois os termos inseridos na RiscoLex devem estar na forma mais usual. A adaptação de Lesk para esse procedimento mostrou-se efetiva, pois evidencia os termos que fazem parte do domínio. Por exemplo, na recuperação de sinônimos para o termo "sinistro", obteve-se "infernal" e "diabólico". Na comparação entre esses sinônimos e a BoW tem-se:

- (sinistro ; sinistro) = 1.000
- (sinistro ; ministro) = 0.889
- (infernal ; not found) = 0.000
- (diabólico ; not found) = 0.000

[25]Os algoritmos Snowball e RSLPS estão disponíveis para o Português na NLTK. Para os leitores interessados nos detalhes técnicos, recomenda-se a leitura da documentação em http://www.nltk.org/api/nltk.stem.html.

Observa-se, portanto, que os termos "infernal" e "diabólico" apresentaram nenhuma similaridade de acordo com a medida *String Matching*, o que significa que eles não foram encontrados na BoW. Isso era exatamente o que se esperava, pois para o mercado de risco, sinistro é algum evento inesperado que causa perda financeira. Logo, os significados desses dois termos não estão associados ao domínio. Por outro lado, a identificação do termo "ministro" pelo algoritmo indica atenção e que a avaliação humana nessa etapa é importante.

Da mesma forma, para o termo "exposição", recuperaram-se os termos "exibição" e "vulnerabilidade":

- (exposição ; exposição) = 1.0
- (exposição ; reposição) = 0.8
- (exibição ; not found) = 0
- (vulnerabilidade ; vulnerabilidade) = 1.0

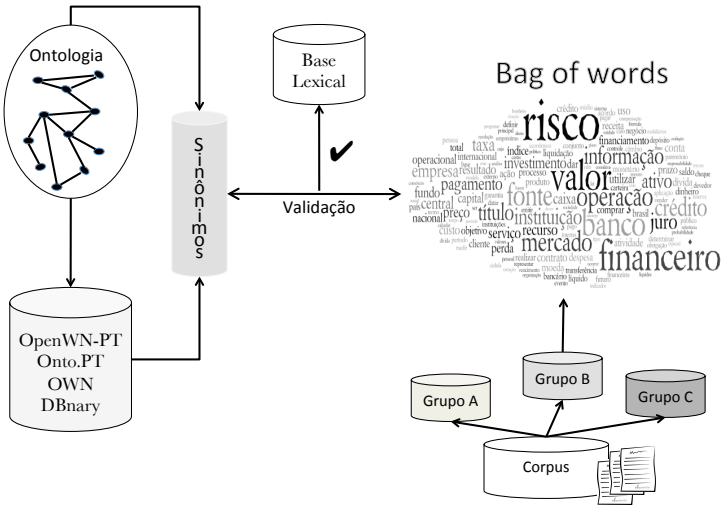
Nesse caso também se identifica que somente os sinônimos relacionados ao domínio foram encontrados. Igualmente, o termo "reposição" foi recuperado por ser similar a "exposição", isto é, 0.8. A forma de eliminar esses casos indesejados foi identificar somente os termos com correspondência exata, isto é, com similaridade igual a 1.

Por esses resultados, nota-se que a técnica ajuda na identificação dos *synsets* candidatos a comporem a RiscoLex. Desse modo, a lista de sinônimos foi analisada e nenhum caso de identificação de termos não relacionados ao domínio foi verificado. Entretanto, na análise manual, alguns termos que fazem parte do domínio e não foram encontrados no agrupamento, como os sinônimos de "ataque" – "assalto" e "atentado". Nesse caso, foram adicionados à RiscoLex. A explicação para o fato é que a técnica de agrupamento utilizada tende a eliminar tanto os termos frequentes quanto os raros, obedecendo à Lei de Zipf e à Regra de Luhn, discutidas nas seções 4.4.6.1 e 4.4.6.2, respectivamente.. Numa verificação desses termos na coleção, observou-se que a utilização de ambos é menos frequente.

Logo, os rótulos de 65 classes e 47 propriedades, isto é, 112 entradas foram pesquisadas e buscaram-se as variações léxicas ou sinônimos nos dicionários e ontologias léxicas para compor a RiscoLex. Ao todo foram encontrados e validados 122 novos termos. A versão final, portanto, foi aumentada em 109% que totaliza 234 termos para compor a RiscoLex que se encontra no anexo B. A Figura 39 ilustra o processo descrito para a validação dos termos e composição da RiscoLex.

De um lado, de acordo com a Figura 39, a ontologia fornece os rótulos para iniciar a busca por sinônimos nas bases de apoio. Por outro lado, o *corpus* é segmentado e o grupo que contém os termos que melhor representam o domínio é transformado numa BoW. Então, sinônimos e termos da BoW são validados pela medida de similaridade. Os termos candidatos que alcançam similaridade igual 1, isto é, são idênticos, compõem a RiscoLex

Figura 39: Validação do léxico do domínio



Fonte: Elaborado pelo Autor

automaticamente. Para os termos não encontrados na BoW, é feita uma análise manual para verificar se, de fato, se relacionam com o domínio. Em caso positivo, são integrados à RiscoLex.

Finalmente, esse resultado demonstra a eficiência da abordagem semiautomática para a criação de base léxica a partir de ontologia e corpus. Esse novo recurso pode beneficiar a pesquisa em processamento de linguagem natural, que ainda carece de recursos em língua portuguesa, a despeito dos recentes avanços da área. Embora o presente experimento seja realizado em área bem delimitada, a técnica é viável para qualquer domínio, desde que haja ontologia e *corpus* correspondentes à área.

5.5.2 Questão II

Ressalta-se que a configuração do Solr apresentou algumas dificuldades. O programa foi construído para armazenar, preferencialmente, documentos em XML. Contudo, os documentos do acervo estudado estão nos formatos PDF, HTML, DOC e TXT. A identificação e extração dos conteúdos dos documentos nesses formatos requereu a utilização de integração com outra ferramenta chamada Apache Tika[26]. A configuração dessa ferramenta com o Solr demandou investigação por fóruns e tutoriais para que funcionasse como pretendido.

Ainda no quesito configuração do ambiente, foram testadas três bibliotecas para a conexão à ferramenta Solr via programação Python. Optou-se pela pysolr[27] por possibilitar a melhor interface com a ferramenta de busca. Além dessa, as bibliotecas mysolr[28] e solrpy[29] também

[26] http://tika.apache.org/
[27] https://github.com/toastdriven/pysolr
[28] https://pypi.python.org/pypi/mysolr/
[29] https://pypi.python.org/pypi/solrpy

proporcionam o acesso, mas apresentaram algumas dificuldades nas tarefas de manutenção da base como a adição, remoção e indexação de documentos e, por isso, foram descartadas.

Para o experimento, considerou-se que, de acordo com os objetivos da investigação, não se necessita de quantidade exaustiva de documentos, mas de um conjunto de dados no qual se possa verificar as vantagens e desvantagens da metodologia, bem como a viabilidade de escrutinar todo o *corpus* manualmente para verificar e validar os resultados apresentados pelos procedimentos automáticos. Consequentemente, foram selecionados 785 documentos contendo as principais definições do domínio de risco. Esses documentos foram indexados com e sem anotação semântica para possibilitar a comparação entre as abordagens.

Uma etapa importante no pré-processamento do *corpus* é a identificação dos termos compostos frequentes, o que contribui para o equilíbrio da revocação e da precisão discutidas na seção 4.5.4. Termos compostos como "patrimônio líquido" ou "Tesouro Nacional" foram transformados em patrimônio_líquido e Tesouro_Nacional, respectivamente, para que fossem indexados como um único termo. Nessa tarefa, há que se tomar a precaução de verificar os falsos positivos como "por exemplo" ou "pode ser". Essas combinações são muito frequentes em quase todos os textos em Português, mas no caso deste estudo não são relevantes e por isso não foram consideradas.

Os termos identificados no *corpus* que correspondem aos rótulos da ontologia recebem peso para aumentar a relevância do documento e torná-lo mais "visível" ao motor de buscas. O Solr vem com a opção de ponderação baseada no algoritmo *tf-idf*, tratado na seção 4.4.5, que apresenta uma lista ordenada na qual se apresentam as informações detalhadas de pontuações atribuídas a cada documento recuperado para o cálculo de relevância. Segue um excerto do resultado da busca para "comite basileia":

Exemplo 5.3: Exemplo de informações para cálculo de relevância

```
...
<lst name="debug">
<str name="rawquerystring">comite basileia~</str>
<str name="querystring">comite basileia~</str>
<str name="parsedquery">fullText:comite fullText:basileia~2</str>
<str name="parsedquery_toString">fullText:comite fullText:basileia~2</str>
<lst name="explain">
<str name="d71fb4b1e20ecd73e19571e03a933794">
0.26190314 = (MATCH) product of: 0.5238063 = (MATCH) sum of: 0.5238063 = MATCH) sum of: 0.27503976
     = (MATCH) weight(fullText:ébasilia^0.875 in 504) [DefaultSimilarity], result of: 0.27503976 =
     score(doc=504,freq=2.0 = termFreq=2.0), product of: 0.31383118 = queryWeight, product of: 0.875
      = boost 4.9576335 = idf(docFreq=14, maxDocs=785) 0.07234585 = queryNorm 0.87639403 =
     fieldWeight in 504, product of: 1.4142135 = tf(freq=2.0), with freq of: 2.0 = termFreq=2.0
     4.9576335 = idf(docFreq=14, maxDocs=785) 0.125 = fieldNorm(doc=504) 0.24876651 = (MATCH) weight
     (fullText:ébasilia.^0.75 in 504) [DefaultSimilarity], result of: 0.24876651 = score(doc=504,
     freq=1.0 = termFreq=1.0), product of: 0.3286082 = queryWeight, product of: 0.75 = boost
     6.056246 = idf(docFreq=4, maxDocs=785) 0.07234585 = queryNorm 0.7570307 = fieldWeight in 504,
     product of: 1.0 = tf(freq=1.0), with freq of: 1.0 = termFreq=1.0 6.056246 = idf(docFreq=4,
     maxDocs=785) 0.125 = fieldNorm(doc=504) 0.5 = coord(1/2)
</str>
...
```

Como se observa no exemplo 5.3, o aplicativo fornece uma extensa lista de estatísticas que especificam o ranking dos documentos para determinar a ordem em que eles serão exibidos. A lista apresentada refere-se a um documento e, portanto, cada um possui todas essas estatísticas disponíveis para o caso de verificação dos resultados. Para o leitor interessado no detalhamento desses números, recomenda-se consultar o site do aplicativo[30].

Além disso, a ferramenta agrega um fator de coordenação que, no caso de múltiplos termos na consulta, quanto mais termos encontrados, maior o peso atribuído. Esse princípio considera que termos combinados reduzem a ambiguidade e especificam o significado. Assim, uma busca pelo termo "risco" recupera 158 documentos. Para a expressão "risco de crédito", retornam-se 18 documentos que contêm a expressão exata. No entanto, para "risco" e "crédito" com fator de ponderação para os termos, recuperam-se 193 documentos.

Analisando o resultado verifica-se que os 18 primeiros documentos estão relacionados ao termo "risco de crédito", os restantes relacionados aos termos "risco" ou "crédito". Tal fato demonstra a vantagem de se utilizar a ponderação estendida, com o fator de coordenação como critério de desambiguação.

De acordo com a RiscoLex, o termo "risco" compreende o mesmo espaço semântico de "perigo" e "ameaça". O primeiro é a forma mais frequente e os outros são as variantes que compreendem o mesmo significado, porém menos utilizadas nesse contexto. Considerando o exemplo extraído do corpus:

✎ ... evento ou situação que implique uma ameaça significativa para a missão, operação, integridade ou recurso da ...

Uma pesquisa sintática, isto é, por palavras-chave, pelo termo "ameaça" recuperaria somente os textos que contenham o termo explícito. Contudo, quando se coloca a anotação relacionada ao termo fazendo referência à ontologia tem-se:

✎ ... evento ou situação que implique uma <ameaça><ontologia: risco> significativa para a missão, operação, integridade ou recurso da ...

Na busca pelo mesmo termo, ameaça, tem-se agora o espaço semântico provido pela RiscoLex que significa também pesquisar pelos termos "risco", "perigo" e "ameaça". A Tabela 11 mostra a comparação entre a busca pelos termos isolados e a opção com notação semântica. A busca sintática procura somente o termo explícito, enquanto a semântica procura por qualquer um deles. Por exemplo, o termo perigo, na busca sintática, só recupera um documento, na semântica, 159 documentos:

Como era de se esperar, o termo "risco", a forma mais frequente, está presente na maioria dos documentos. Suas variantes são utilizadas em apenas 3 documentos. O que significaria, no caso de uma busca sintática pelo termo "ameaça", a ausência de 99% dos

[30] https://wiki.apache.org/solr/SolrRelevancyFAQ

Tabela 11: Busca sintática X semântica

Tipo da Busca	Termo	Docs encontrados
Sintática	ameaça	2
	perigo	1
	risco	158
Semântica	ameaça	159

Fonte: Elaborado pelo Autor

textos semanticamente relacionados. Portanto, o resultado semântico é o conjunto de textos contendo qualquer um dos termos semanticamente relacionados que estão presentes na RiscoLex. O primeiro benefício que essa técnica evidencia é o incremento da revocação.

Como se enfatiza na literatura, a revocação e precisão estão inversamente correlacionadas e, portanto, o que se busca é o equilíbrio para que se alcance máxima revocação e precisão. Como exemplo, observa-se um comportamento ambíguo comum para termos que variam a função sintática de acordo com o uso. Toma-se, por exemplo, o substantivo "bem" que na RiscoLex abarca o mesmo conceito de "propriedade", "posse", "ativo" e "recurso", isto é, algo que se é o dono ou se possui.

A busca retorna 171 documentos que supostamente contemplam o significado pretendido no domínio. Contudo, é necessária uma análise mais detalhada como se segue:

(a) (substantivo) ...sobre **<bens>** imóveis. Neste caso, o **<bem>** hipotecado permanece em poder do . . .

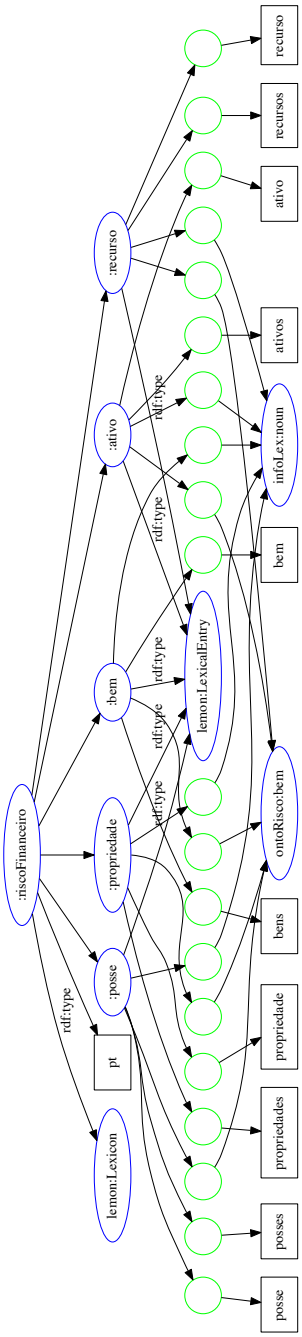
(b) (conjunção aditiva – locução comparativa) ...estrutura de capitais, situação econômico-financeira, < **<bem>** como> sua posição no mercado . . .

O termo "bem" no exemplo (b) não é pertinente ao substantivo "posse", mas à locução comparativa "assim como" e, logo, não deveria ser recuperado por não ser semanticamente correlacionado com a intenção da busca pelo substantivo "bem". A Figura 40 ilustra o campo semântico do termo para o domínio de acordo com a RiscoLex.

Para solucionar o problema de falso positivo, é necessária a identificação sintática dos termos que serão anotados no corpus. Assim, procedeu-se ao processo de *parsing* dos textos que determina a categoria sintática de cada termo de acordo com sua função na frase. Para tanto, utilizou-se o *corpus* Floresta Sintá(c)tica para a criação de um modelo capaz de executar tal tarefa automaticamente.

Portanto, para a elaboração do modelo apto a etiquetar automaticamente as categorias sintáticas do texto, utilizou-se o *corpus* na proporção de 90% para treinamento e 10% para teste. Essa proporção garantiu um resultado satisfatório, que foi uma taxa de acerto de 81%, considerando que se trata de um domínio que possui muitos termos específicos ou em Inglês. A melhoria dessa taxa pode ser desenvolvida com a entrada manual desses termos na base de treinamento ou com a criação de um *corpus* do domínio anotado sintaticamente.

Figura 40: Campo semântico do termo “bem”



Fonte: Elaborado pelo Autor

Assim, o texto foi analisado pelo modelo e transformado para a versão contendo as categorias sintáticas como ilustra o exemplo:

(a) . . . sobre[prp] <**bens**[n]> imóveis[n] .[.] Neste[n] caso[conj-s] ,[,] o[art] <**bem**[n]> hipotecado[n] permanece[v-fin] em[prp] poder[v-inf] do[n]. . .

(b) . . . sua[pron-det] estrutura[n] de[prp] capitais[n] ,[,] situação[n] econômico-financeira[n]

,[,] < <**bem**[adv]> como[adv]> sua[pron-det] posição[n] no[pron-pers] mercado[n]...

Verifica-se, então, que o modelo classifica corretamente o termo "bem" e sua forma plural "bens" na frase (a) como substantivos e na (b) como advérbio. Após esse passo, pode se efetuar a marcação semântica nos casos em que o termos e respectivas categorias sintáticas coincidem com os termos da RiscoLex. Assim, somente o termo "bem", classificado como substantivo (n), é candidato à anotação semântica, enquanto as ocorrências do mesmo termo classificado como advérbio (adv), que não pertence ao mesmo campo semântico desejado, são ignoradas.

Após a identificação das categorias sintáticas, a consulta pelo termo "bem" recupera 17 documentos a menos, isto é, 154 referências. Em uma análise manual mais detalhada, constatou-se que os documentos não recuperados continham a expressão "bem como" na qual "bem" é advérbio e, portanto, não foram recuperados. Verificando pela expressão "bem como", encontram-se 20 referências, contudo 2 possuem o termo "recurso" e 1, "ativo" como se mostra um trecho do texto que contém "bem como" e "ativo" a seguir:

✎ ... estabelecendo[v-ger] papéis[n] e[conj-c] responsabilidades[n] ,[,] < <**bem**>[adv] como[adv]> as[art] dos[n] prestadores[n] de[prp] serviços[n] terceirizados[n] ... danos[n] a[prp] <**ativos**[n]> físicos[n] próprios[n] ou[conj-c] em[prp] uso[n] pela[n] instituição[n] ...

Essa referência é recuperada acertadamente por conta do termo "ativos", classificado com substantivo, e que denota o mesmo conceito de "bem". Isso mostra que a diferenciação sintática dos termos auxilia na remoção da ambiguidade causada por policategorização e, consequentemente, na melhoria da precisão.

Um terceiro procedimento para lidar com ambiguidade por homografia também é executado. Trata-se da identificação semântica de termos que têm a mesma classificação sintática, mas significados diferentes. Tome-se como exemplo a busca pelo termo "produto" que é sinônimo de "artigo". Esse termo também é frequente no domínio de risco, mas usualmente refere-se ao dispositivo textual inserido em leis, normas etc. Considera-se o seguinte trecho anotado sintaticamente:

✎ ... A[art] Constituição[n] de[prp] 1988[num] prevê[n] ainda[adv] ,[,] em[prp] seu[pron-det] **artigo**[n] 192[n] ,[,] a[art] elaboração[n] de[prp] Lei[n] Complementar[n] do[n] Sistema[n] Financeiro[n] Nacional[prop] ...

O termo "artigo" nessa expressão possui significado diferente de "produto", apesar de também ser classificado como "substantivo".

Como mencionado na seção 5.2.3.1, a RiscoLex foi gerada a partir da WordNet, que fornece relações entre termos que auxiliam na tarefa de delimitação do significado como a presença das relações de hiperonímia e hiponímia entre os conceitos expressos pelos termos. Dessa forma, é possível quantificar o grau de similaridade entre o termo e o significado

pretendido e, portanto, automatizar a seleção de termos candidatos à anotação semântica. Como exemplo, considere os sentidos de "mercadoria" e de "gênero literário" extraídos da base para o termo "artigo":

Quadro 8: Significados de artigo

Significado	Mercadoria	Gênero Literário
Definição	Produto oferecido para venda	Prosa não-ficcional formando uma parte independente de uma publicação
Similaridade – (artigo ; produto)	1,0000	0,07692

Fonte: Elaborado pelo Autor

O Quadro 8 mostra os significados, as definições e o cálculo de similaridade baseado na hierarquia dos significados dos conceitos de cada termo. A similaridade é calculada baseada na "distância" entre os termos que representam conceitos de forma hierárquica. O grau de similaridade varia entre 1 para idênticos e 0 para nenhuma relação.

Assim, a aplicação para a similaridade semântica é a identificação de termos com significados relacionados de forma que seja possível quantificá-la. Dado um termo específico, pode-se percorrer toda coleção para identificar outros que estão no mesmo campo semântico. Conhecer os termos semanticamente relacionados mostra-se bastante útil para indexar um corpus, de forma que uma busca por um conceito abrangente tal como "mercadoria" também recupere documentos com termos específicos como "artigo".

Os conceitos estão conectados por relações semânticas que denotam significados mais abrangentes ou mais específicos às palavras, formando uma rede complexa.

A Figura 41 ilustra um fragmento da hierarquia de conceitos da Wordnet. As caixas representam os conceitos e a linhas indicam as relações de hiperonímia ou hiponímia, ou seja, a relação de conceitos subordinados ou superordenados. As figuras com cor cinza estão relacionadas a conceito de "abstração", enquanto as outras, "entidade física". Observa-se que o conceito "artigo" é comum a ambos. O conceito mais geral é "entidade" e os mais específicos, "artigo" e "produto". Ainda, "entidade" é hiperônimo comum aos significados em questão.

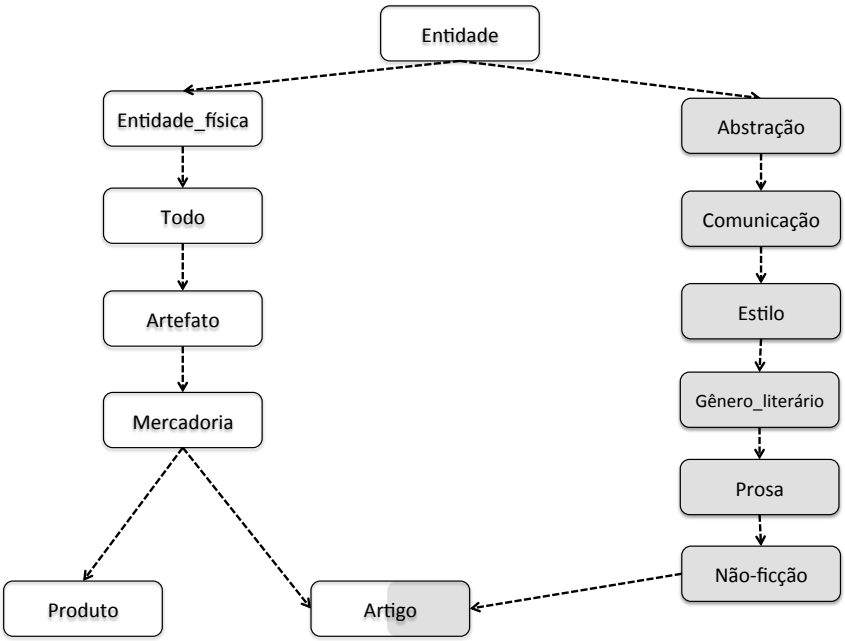
A fórmula para o cálculo é conhecida por ***path-based similarity*** ou similaridade baseada no caminho, Banerjee & Pedersen (2002), como se segue:

$$CS(c_1, c_2) = \frac{1}{comprimentoCaminho(c_1, c_2)} \tag{5.2}$$

na qual c_1 e c_2 representam os conceitos e comprimentoCaminho é a contagem de passos que separam os dois conceitos na hierarquia da Wordnet.

A ideia é que menores distâncias significam proximidade semântica. No exemplo, o valor $1,0000$ representa que os termos "produto" e "artigo" expressam o mesmo conceito sob a

Figura 41: Hierarquia de Significados



Fonte: Elaborado pelo Autor

perspectiva de "mercadoria". O que já se esperava, pois são termos considerados sinônimos nesse âmbito. Já com o sentido de "gênero literário", o valor é menor, apenas $0,0769$, ou pela fórmula, $\frac{1}{13}$. A biblioteca NLTK possui várias implementações para calcular a similaridade baseada na Wordnet.

Retomando o exemplo:

> ✎ ... A[art] Constituição[n] de[prp] 1988[num] prevê[n] ainda[adv] ,[,] em[prp] seu[pron-det] **artigo**[n] 192[n] ,[,] a[art] elaboração[n] de[prp] Lei[n] Complementar[n] do[n] Sistema[n] Financeiro[n] Nacional[prop] ...

O termo "artigo" chama a atenção por não ter o significado pretendido pelo usuário que é o de "produto". Caracterizando o aspecto dinâmico da manutenção de SRI, sempre que se verifica, usuário ou equipe de manutenção, uma ocorrência dessa natureza, convém marcar o termo para averiguação manual. Esse procedimento gera uma lista de candidatos à desambiguação e, consequentemente, ao refinamento da anotação semântica.

De acordo com essa técnica, tem-se a indicação pela medida de similaridade que o termo "artigo" tem significados diferentes, apesar da mesma função sintática. São retornados os valores 1,000 para estar semanticamente relacionado à "mercadoria" e 0,0769 ao "gênero literário". Esses casos, em que ocorrem mais de uma medida, são separados para avaliação pelo especialista que decide se será feita a marcação ou não. Nesse exemplo, a opção é por não assinalar a ocorrência do termo com a anotação semântica.

Uma ressalva à utilização da Wordnet como base de apoio para calcular a similaridade semântica é a baixa cobertura para o idioma Português. Espera-se que a pesquisa crescente em

PLN incentive a expansão dessa base, que se mostra útil para o tratamento da ambiguidade com o processamento da linguagem natural. Contudo, mesmo com o tamanho aquém do esperado, mostrou-se uma boa alternativa para identificar candidatos a termos semanticamente relacionados e, consequentemente, para a melhoria da recuperação da informação.

Finalmente, sabe-se que a supervisão humana incrementa a exatidão das anotações. A tarefa, contudo, não é factível para alguns milhões de anotações a que se pode chegar nas bases de dados textuais. O processamento automático de anotação descrito pode apresentar uma lista de termos em que se tem alguma incerteza em relação à anotação, para averiguação de especialistas de domínio. Essa lista constitui-se em ferramenta de depuração, para que se resolvam casos de polissemia ou de anotações semânticas equivocadas, que não correspondem ao conceito ou à categoria sintática, bem como a inexistência de termos importantes para o domínio na base de conhecimento.

5.5.3 Avaliação

Para avaliar a performance da proposta, foram indexadas duas bases com os mesmos documentos. A primeira, que representa a busca tradicional (bt), isto é, um índice construído a partir dos textos sem nenhum processamento para servir de ponto de partida para comparação. A segunda contém o índice que foi construído, a partir dos textos com as anotações semânticas para representar a busca semântica (bs).

A avaliação da proposta é baseada na revocação e da precisão discutidas na seção 4.5.4. Para tal, deve-se estabelecer a relevância de cada documento de acordo com o interesse do usuário, nesse caso, à consulta formulada ao sistema. Como se trata de uma base não classificada previamente, é necessária a avaliação das referências relevantes ao tema das consultas. Portanto, para determinar a relevância, a base foi avaliada por 5 especialistas[31], orientados pelas consultas do Quadro 9, para que também julgassem os documentos não recuperados.

Quadro 9: Consultas da avaliação

P-1 Encontrar todos os documentos relacionados a ameaça (consulta: ameaça)
P-2 Encontrar todos os documentos relacionados a risco operacional (consulta: risco operacional)
P-3 Encontrar todos os documentos relacionados a risco de crédito (consulta: risco de crédito)
P-4 Encontrar todos os documentos relacionados a bens (consulta: bem)
P-5 Encontrar todos os documentos relacionados a crime (consulta: crime)

Fonte: Elaborado pelo Autor

Todas as consultas do Quadro 9 evidenciam características que podem influenciar os resultados fornecidos por motores de busca. A complexidade do ponto de vista do tratamento linguístico vai aumentando para demonstrar a melhoria da utilização de um sistema de

[31]Assim distribuídos: 2 de Ciência da Computação, 2 de Finanças, 1 de Estatística e de Ciência da Informação

recuperação da informação semântica. A seguir, são apresentados os resultados, as complexidades linguísticas que interferem no desempenho dos motores de busca tradicionais e como a abordagem proposta lidou com elas.

Tabela 12: Avaliação de resultados

Consulta	bt	DRR[32]	P %	R %	F %	bs	DRR	P %	R %	F %	DRB[33]
P-1	2	2	100,00	1,26	2,48	159	159	100,00	100,00	100,00	159
P-2	175	14	8,00	93,33	14,74	15	15	100,00	100,00	100,00	15
P-3	720	18	2,50	100,00	4,88	19	18	94,74	100,00	97,30	18
P-4	30	10	33,33	6,54	10,93	154	153	99,35	100,00	99,67	153
P-5	2	2	100,00	25,00	40,00	7	7	100,00	87,50	93,33	8

Fonte: Elaborado pelo Autor

Os resultados podem ser vistos na Tabela 12 que apresenta, para cada consulta, a quantidade de documentos recuperados pela busca tradicional (bt) de documentos considerados relevantes: os valores da precisão (P), da revocação (R) e da medida (F). Para a busca semântica (bs), os mesmos indicadores e na última coluna, a quantidade de documentos existentes na base, considerados relevantes pelos especialistas que a escrutinaram

A consulta P-1 apresenta diferença considerável entre bt e bs. A razão para isso deve-se à preferência pelo termo "risco" nos documentos que compõem a base e, portanto, refletida na baixa revocação, apenas $1,26$. O índice tradicional não é capaz de recuperar o termo "risco", pois não está expresso na consulta. Contudo, a bs é capaz de reconhecer outros termos que estão presentes na RiscoLex e que, dessa forma, foram incorporados ao índice. Assim, tanto "ameaça" quanto "risco" e "perigo" são indexados e, por isso, considerados como pontos de acesso.

A consulta P-2 apresenta uma boa revocação, mas uma precisão baixa na bt. Isto é, recuperam-se muitos documentos, porém majoritariamente irrelevantes. Tal fato se deve à falta de identificação de termos compostos, o que resulta na busca dos termos isolados "risco" ou "operacional". Além disso, não há o tratamento para plural. Assim, um documento com "riscos operacionais" também não foi recuperado na bt. Por outro lado, a bs localiza um único termo "risco operacional" que proporciona a exata recuperação de todos os documentos que contém o termo composto, inclusive aquele que possui o termo no plural.

A terceira consulta P-3 evidencia a importância do tratamento de *stopwords* e de termos compostos, bem como a correta identificação de diacríticos característicos do Português. As buscas tradicionais geralmente eliminam os acentos gráficos e, portanto, as palavras "crédito" e "credito" — do verbo "creditar" conjugado na primeira pessoa do presente do indicativo – terão a mesma forma, se retirado o acento agudo. No domínio analisado, ambas as palavras são encontradas e, dessa forma, o reconhecimento e transformação para a forma canônica antes da anotação é conveniente. Logo, a alta revocação e baixa precisão na bt é reflexo

[32]Documento Relevante Recuperado
[33]Documento Relevante na Base

da falta de processamento dos tópicos mencionados. Na bs, observam-se altos índices de revocação e precisão como se esperava.

Nota-se também um documento a mais em bs que se configura de difícil tratamento automático. Essa referência consta no documento desta forma:

✎ ... além dos <riscos de crédito> e de mercado, introduziu-se o risco operacional...

A normalização da forma plural fez com que o documento fosse recuperado, contudo considerado irrelevante pelos especialistas, por tratar-se de um texto em que o termo aparece somente em uma lista de vários riscos no contexto do risco operacional. Esse é um caso típico no qual somente o julgamento humano define a relevância do termo e não há como tratá-lo automaticamente.

A penúltima consulta P-4 mostra a melhoria no resultado quando se contemplam as categorias sintáticas como forma de delimitar o sentido dos termos. A bt recupera os documentos que possuem o termo "bem", mas não contempla seu plural "bens", com 16 referências na base. Destaca-se que, desse resultado, 20 documentos apresentam a locução comparativa "bem como" e, com exceção de três referências que contêm também outros termos relevantes e com mesmo significado, não deveriam ser recuperados. No caso da bs, são recuperados 154 documentos que contemplam também os sinônimos "ativo", "propriedade", "posse" e "recurso". Assim, observa-se um documento recuperado a mais na bs que se deve a uma anotação equivocada verificada em decorrência desse resultado. Além disso, das 20 referências com o termo "bem como", 17 são descartadas, por não conterem os sinônimos no texto.

A última consulta P-5 demanda mais atenção aos detalhes, pois envolve também o processo de inferência que abrange a busca por hipônimos. O sinônimo de "crime" é "violação". O sentido adotado é na perspectiva do cidadão ordinário que é um ato passível de punição pela lei, isto é, não abrange todos os aspectos técnicos que a palavra exprime para o Direito. Assim, possui os seguintes hipônimos para o domínio: "ataque", "assalto", "falsificação", "roubo", "rapto" e "infração".

Ressalta-se que, geralmente, uma busca percorre do conceito mais geral para os mais específicos e, assim, no caminho extensional do conceito, um usuário padrão não leva em conta as complexidades e relações linguísticas existentes entre sua consulta e o resultado que quer receber. Quando se busca um termo geral, ele fica satisfeito em receber documentos que respondam a sua necessidade de informação, sem que se perceba que os resultados abrangem conceitos mais específicos, como, por exemplo, "roubo" é uma extensão do conceito "crime". Os hipônimos devidamente anotados desempenham esse papel especificador numa consulta semântica.

Contudo, o contrário, ou seja, no caminho intensional do conceito, aumenta-se a abrangência e colateralmente a vagueza na busca que aumenta a revocação, mas deteriora a precisão que não é o objetivo da maioria dos motores de busca. O exemplo a seguir ilustra

a taxonomia de conceitos do termo “crime” de acordo com a WordNet. Da esquerda para direita, do sentido mais geral para o mais específico:

entidade→abstração→característica psicológica→evento→ato→atividade→transgressão→crime

Observa-se que quanto mais genérico, mais abrangente seria a busca. Por exemplo, a busca por “atividade” recuperaria certamente documentos com nenhuma relação com “crime”. O que não é desejável nesse exemplo. Dessa forma, nessa abordagem de busca semântica, somente são consideradas as relações hiponímicas que visam melhorar a especificidade da busca e, consequentemente, a precisão.

Observa-se, na Tabela 12, que a bt recupera 2 documentos com 100% de precisão, mas com revocação baixa, apenas 25%. A bs, contudo, também apresenta a precisão ótima, mas com a revocação bem superior, $87,5\%$, em função da busca também localizar os hipônimos anotados na base.

Um fato que chama a atenção é a não identificação de todos os documentos relevantes na base, pois existe um, além dos 7 que foram recuperados. Na verificação manual, identificou-se um documento que possui o termo “atentado”, que também é um hipônimo de “crime”, mas não foi anotado, nem identificado para ser carregado para a RiscoLex.

Esse caso poderia ser facilmente resolvido com a introdução da anotação e da inserção do termo na RiscoLex. Entretanto, tal feito é importante para enfatizar que a verificação humana é parte relevante do processo para depuração da base de pesquisa e das bases léxico-ontológicas — OntoRisco e RiscoLex — que resultam em melhoria substancial para a recuperação da informação. Além disso, evidencia a natureza dinâmica da manutenção de bases de conhecimento que necessitam de manutenção periódica.

Por fim, considera-se que as consultas submetidas ao modelo mostram que a busca semântica supera o desempenho da tradicional e validam a metodologia empregada para responder a segunda questão da investigação . O processamento que envolve a preparação das bases onto-léxicas e a anotação do *corpus* do domínio que será indexado são complexos, mas proporcionam avanço relevante na tarefa de fornecer a informação mais adequada ao usuário. O procedimento, embora adicione complexidade em sua elaboração, pode ser reproduzido em qualquer outro domínio, com otimização dos resultados observados.

6 Conclusões e recomendações

A pesquisa demonstrou a utilização de tecnologias da Web Semântica e o processamento de informação textual para a construção de base léxico-semântica adequada ao padrão adotado pela W3C. Tal recurso destina-se a apoiar o modelo proposto de recuperação de informação semântica que utiliza informações linguísticas e semânticas na construção de um índice léxico-semântico que melhoram a precisão no processo de RI.

Apesar do desenvolvimento dos recursos tecnológicos nas últimas décadas, verificou-se que o progresso alcançado para línguas como o Inglês ainda carece de profissionais e ferramentas para produzir recursos linguísticos e ferramentas tecnológicas adequadas para o Português. Muitos procedimentos já amplamente difundidos no meio acadêmico e profissional necessitam de intervenções pontuais para tratar os problemas característicos de nossa língua, como acentos, caracteres e letras não existentes em outros idiomas.

Além disso, alguns recursos elaborados para o idioma Português, no meio acadêmico, não estão disponíveis para serem utilizados livremente pela comunidade científica o que, certamente, representa uma dificuldade a ser superada. Enquanto o mundo anglo-saxão trabalha para a máxima divulgação e disponibilização livre de inovações linguísticas, ainda temos poucas iniciativas nesse sentido. Um exemplo é a presença do idioma de Camões na *Linked Open Data cloud*, que fica atrás do Alemão e do Francês e possuem menos falantes nativos no mundo que falantes de Português apenas no Brasil.

Quando se olha pela perspectiva da Ciência da Informação, a pesquisa nesse campo está ainda incipiente, pois a literatura é majoritariamente produzida na Ciência da Computação. Há muito espaço para investigações científicas que impulsionariam a CI e a auxiliariam na concretização da visão da Web Semântica, também sob a ótica dos cientistas da informação. Esse é, também, um dos pontos em que esta pesquisa vem contribuir de forma a preencher a lacuna entre o desenvolvimento de recursos computacionais e a gestão e organização da informação, na perspectiva da CI.

Nesse sentido, a contribuição da Filosofia, da Linguística e da Estatística completam o cabedal teórico utilizado para atingir os objetivos propostos neste trabalho. A escolha das bases de dados, de tecnologias da Web Semântica e de ferramentas computacionais ancoraram-se nessa convergência de teorias e técnicas para possibilitar a proposição da metodologia, dos modelos e de resultados que justificaram a pesquisa.

Como várias pesquisas que utilizam a PNL para o seu desenvolvimento, as bases de dados em Português demandaram tratamento pontual e significaram parcela considerável de tempo para a pesquisa. Os problemas relacionados ao idioma apresentam-se em todas as fases, desde a importação de arquivos de textos que não são convertidos de forma adequada

para utilização em algoritmos, até a apresentação final dos textos recuperados. Apesar de a ferramenta NLTK estar disponível para o nosso idioma, muitas intervenções via programação foram necessárias.

Um ponto positivo foi a qualidade dos textos que são produzidos por profissionais especializados e, por isso, não se verificam erros ortográficos ou utilização da linguagem informal. Por outro lado, a presença de termos em Inglês e Português, que são muito frequentes nesse domínio, dificulta a automatização completa dos procedimentos. Isso demanda a construção de regras para "entender" o idioma diferente do qual se está processando.

Durante a fase de importação e tratamento das bases utilizadas, não foram encontrados recursos linguísticos e ontológicos disponíveis para atender a indústria financeira brasileira, apesar de sua importância para o progresso do país. Do ponto de vista econômico, isso se caracteriza como um potencial de negócio muito interessante. Do acadêmico, uma grande fonte de financiamento para pesquisas, uma vez que os principais usuários desse tipo de informação detêm grandes somas de recursos financeiros. Essa é uma parceria que pode beneficiar ambos os lados.

Assim, outra contribuição deste trabalho é a produção de recursos em Português para o segmento financeiro. O primeiro é a construção da base léxica inédita, RiscoLex, que contém informações morfológicas, sintáticas e semânticas de termos para o risco financeiro. O segundo é a elaboração da ontologia, OntoRisco, também para o mesmo domínio. Essa foi inspirada em outras elaboradas para o Inglês, mas devido às características do mercado nacional e especialmente das entidades financeiras públicas, a ontologia teve que ser praticamente redesenhada.

Em relação às limitações que permearam esta investigação, pode-se destacar a dificuldade de construção da ontologia pelo processo manual. As discussões para definir e estabelecer o relacionamento entre os conceitos por especialistas, com diferentes formações, tornam o desenvolvimento lento e, às vezes, sem consenso. Obviamente, alguma dose de arbitrariedade é impositiva em algumas circunstâncias, contudo é um processo que não pode ser prescindido e que resulta na representação do conhecimento do domínio mais próximo à realidade dos profissionais da área.

Outro ponto é a qualidade dos corpora encontrados para a extração de sinônimos para compor a RiscoLex. Como mencionado, a metodologia tem sido desenvolvida majoritariamente por cientistas da computação que primam pela automatização e pelo desempenho de algoritmos computacionais. Encontraram-se inconsistências como, por exemplo, o termo "mercadoria" ser indicado, ao mesmo tempo, como sinônimo e hiperônimo de "artigo", o que demonstra que a depuração dessas bases por linguistas, lexicógrafos, terminólogos e outros especialistas da área é uma necessidade.

O quadro 10 reflete os objetivos específicos contrastados com as principais dificuldades no decorrer da investigação. A intenção é mostrá-las de forma sistematizada para facilitar o trabalho de outros pesquisadores que venham percorrer os mesmos caminhos deste estudo.

Quadro 10: Objetivos específicos X dificuldades encontradas

OBJETIVOS ESPECÍFICOS	DIFICULDADES ENCONTRADAS
Identificar tecnologias da Web Semântica apropriadas para a representação do conhecimento em determinado domínio;	i) O nível técnico das publicações, em geral, demandou conhecimento em especializado; ii) Apesar de haver muito material publicado, geralmente está em Inglês e com a premissa que o leitor tenha autonomia no mundo da informática. Por exemplo, entender SPARQL sem noção de SQL é uma tarefa árdua.
Produzir ontologia para o domínio de risco financeiro;	i) A reutilização de ontologias voltados para o mercado financeiro internacional não se mostrou efetiva devido às diferenças entre o mercado internacional e brasileiro. ii) A elaboração manual de ontologias exige esforço, tempo e colaboração de especialistas de domínio para resultados consistentes; iii) As fontes textuais contendo as definições que fornecem o vocabulário para a ontologia devem ser validadas com parcimônia para resultados mais efetivos.
Gerar uma base de léxicos em Português brasileiro que contenha os aspectos morfológicos, sintáticos e semânticos dos itens lexicais vinculados aos elementos ontológicos da indústria financeira;	i) As fontes de dados, como dicionários, tesauros e ontologias léxicas, em Português ainda são escassas e possuem problemas de cobertura; ii) Algumas, principalmente aquelas construídas automaticamente, mostraram problemas com a qualidade das associações entre termos; iii) Nem todas estavam disponíveis para acesso público ou para *downloads*; iv) Diferentes formatos dessas fontes demandaram intervenção via programação para que pudessem ser integradas.
Propor um modelo de recuperação de informação que utilize automaticamente a base de léxicos como forma de tratar a ambiguidade;	i) A padronização dos formatos pdf, doc, html etc para compor o corpus demandou tempo e recurso, mas foi inevitável para que o Solr funcionasse como esperado; ii) O processo de anotação semântica do corpus necessitou de várias intervenções manuais, pois a decisão humana não pode ser prescindida; iii) A programação para integrar base léxica, corpus e motor de busca não é trivial e requereu tempo e técnica para que fosse completada; iv) A configuração das ferramentas que compõem o modelo exigiram buscas em tutoriais e fóruns para que funcionassem de acordo com o previsto.
Testar o modelo e mensurar a precisão para comparar os resultados entre o modelo aqui proposto e um modelo clássico de recuperação da informação no domínio de risco financeiro.	i) A verificação manual para validar os resultados demandou tempo e colaboração entre especialistas envolvidos para discussões e conclusões de pontos investigados; ii) A extração das estatísticas do motor de busca requereu programação via Python para que fossem utilizadas.

Fonte: Elaborado pelo Autor

Apesar das dificuldades apontadas, conclui-se que os objetivos específicos relacionados aos recursos em Português para o risco financeiro foram satisfatoriamente alcançados. Não se objetiva a cobertura exaustiva desses recursos, mas a proposição de modelo dinâmico e crível na construção automatizada, na medida do possível, de fontes de informação léxico-semânticas que apoiem o domínio. Assim, os procedimentos seguramente podem ser melhorados, tanto no aperfeiçoamento dos métodos computacionais, quanto no aumento na cobertura do vocabulário relacionado ao domínio, bem como adoção para outras áreas do conhecimento.

A integração da RiscoLex ao modelo recuperação da informação também encontrou alguns obstáculos, mas, de forma geral, as ferramentas disponíveis para a construção desse tipo de modelo já estão bastante maduras, de modo a tornar a tarefa de implementação menos penosa. Novamente, a aplicação da PNL nesse processo necessita de mediações pontuais,

por conta da particularidade do Português, o que evidencia a importância e a urgência de pesquisas, como esta, para aperfeiçoar e aumentar a oferta de ferramentas computacionais e as bases de dados adequadas ao nosso idioma.

Aqui também conclui-se que o objetivo foi alcançado, pois a proposta do índice léxico-semântico reúne componentes linguísticos e ontológicos que mostraram melhoria na precisão. Nesse sentido, a principal contribuição é preencher a lacuna existente entre a busca baseada em palavras-chave e o significado dos termos submetidos na consulta. A adoção de índice híbrido proporciona, no mínimo, o desempenho da busca tradicional por palavras-chave e não requer que o usuário seja especialista em qualquer linguagem especializada para realizar tais buscas.

Por tudo isso, considera-se que o objetivo desta pesquisa foi conseguido, pois a proposta era criar uma base de léxicos — RiscoLex — em Português brasileiro, contendo informações morfológicas, sintáticas e semânticas apropriadas para leitura por máquinas — RDF —, permitindo a vinculação entre dados estruturados — OntoRisco — e não estruturados – *corpus* textual – e integrá-la a um modelo de recuperação da informação — MoRIS — com o objetivo de aumentar a precisão.

Quadro 11: Objetivos específicos X resultados

OBJETIVOS ESPECÍFICOS	RESULTADOS
Identificar tecnologias da Web Semântica apropriadas para a representação do conhecimento em determinado domínio;	Adoção do RDF e OWL para elaborar os recursos léxico-semânticos e SPARQL para extrair informações das ontologias léxicas.
Produzir ontologia para o domínio de risco financeiro;	OntoRisco – exibida na seção 5.2.3.1 e presente no anexo A.
Gerar uma base de léxicos em Português brasileiro que contenha os aspectos morfológicos, sintáticos e semânticos dos itens lexicais vinculados aos elementos ontológicos da indústria financeira;	RiscoLex – exposta na seção 5.3 e disponível no anexo B.
Propor um modelo de recuperação de informação que utilize automaticamente a base de léxicos como forma de tratar a ambiguidade;	MoRIS – apresentado na seção 5.4.2.
Testar o modelo e mensurar a precisão para comparar os resultados entre o modelo aqui proposto e um modelo clássico de recuperação da informação no domínio de risco financeiro.	Os resultados foram discutidos na seção 5.5.3 e resumidos na Tabela 12.

Fonte: Elaborado pelo Autor

O Quadro 11 sintetiza os objetivos específicos e os resultados encontrados com o desenvolvimento deste trabalho.

A metodologia proposta mostrou-se útil e factível na tarefa de produzir uma base léxica, a partir de ontologias e textos, que visa aumentar a disponibilidade de recursos léxico-semânticos em Português. Tais recursos podem servir de insumo para inserir um módulo semântico nos motores de busca que, dessa forma, recuperam informações mais contextualizadas sob a expectativa do usuário de determinado domínio.

No âmbito acadêmico, considera-se que a pesquisa obteve êxito, embora se acredite que ainda se possa melhorar a experiência do usuário de forma a que ela seja transparente e que possa ser aplicada em outros domínios. Nesse sentido, percebeu-se que um estudo de usuário contribuiria para a depuração dos recursos léxico-semânticos. Essa melhoria pode ser proporcionada pela captura da preferência no uso de termos para a formulação de consultas. A adição desse vocabulário, tanto na ontologia quanto na base léxica, poderia proporcionar melhor interação entre usuário e sistema.

Da perspectiva da aplicação em empresas, constatou-se que o perfil da equipe que elabora e mantém o sistema deve ser multiplidisciplinar. A integração de variadas *expertises* é componente crítico para que o modelo contemple tanto o componente genérico quanto o especialista, para que se possa atender a maior quantidade possível de usuários. A linguagem técnica de especialistas que elaboram boletins e normas, convive com a usual do usuário consumidor dessas informações e ambas devem estar consideradas e relacionadas para que o significado seja preciso no âmbito computacional.

Do ponto de vista pessoal, o desenrolar desta investigação mostrou faces prazerosas e angustiantes da mesma moeda. A lida diária com pequenos e grandes obstáculos ensina que a persistência e a disciplina são ingredientes para o sucesso em qualquer tarefa. A troca de experiências com culturas e especialidades diferentes coloca o pesquisador completamente fora de sua zona de conforto, o que, no primeiro momento, se traduz em ansiedade extrema, mas, ao fim e ao cabo, as recompensas pessoais e intelectuais são incomensuráveis.

Como sugestão de trabalhos futuros, o estudo de usuários como forma de coletar insumos para enriquecer o vocabulário e, ao mesmo tempo, contemplar as idiosincrasias e jargões de domínios a serem explorados poderia contribuir para melhorar a assertividade de bases léxicas em conjunto com SRI semânticos. Ainda, a adoção de outros índices de representação, além de tf-idf, abordando a indexação léxico-semântica seria de grande utilidade. Noutra abordagem, as bases produzidas pela lexicalização de ontologias poderiam ser instrumentos para melhorar a construção de resumos automáticos ou de elaboração automática de textos. Por fim, a adição de lexicógrafos, terminólogos e linguistas na construção de bases léxicas poderia contribuir sobremaneira na interpretação e na adequação de fenômenos linguísticos ao ambiente das ontologias.

Finalmente, parafraseando dois seres humanos notáveis, Einstein e Wittgestein, expresso o maior aprendizado na conclusão desta obra: os limites do meu mundo são os limites da minha linguagem, que se ampliam a cada dia e expandem a minha mente que, portanto, jamais retornará ao tamanho de ontem.

Referências

ALLEMANG, D.; HENDLER, J. **Semantic Web for the Working Ontologist**: **Effective Modeling in RDFS and OWL**. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2008. ISBN 0123735564, 9780123735560. Citado 5 vezes nas páginas 47, 52, 55, 56 e 59.

ALONSO, O.; BANERJEE, S.; DRAKE, M. Gio: a semantic web application using the information grid framework. In: ACM. **Proceedings of the 15th international conference on World Wide Web**. [S.l.], 2006. p. 857–858. Citado na página 126.

AMATUZZI, M. L. L.; AMATUZZI, M. M.; LEME, L. E. G. Metododia da pesquisa: o desenho da pesquisa. **Acta Ortop. Brasil**, v. 11, n. 1, p. 58–61, 2003. Disponível em: <http://ref.scielo-.org/bp25mr>. Acesso em: 27 jul. 2012. Citado na página 139.

ANTONIOU, G.; VAN HARMELEN, F. Web ontology language: Owl. In: STAAB, S.; STUDER, R. (Ed.). **Handbook of ontologies**. 2. ed. Berlin: Springer, 2009. p. 95 – 110. Disponível em: <http://www.springerlink.com/index/10.1007/978-3-540-92673-3>. Acesso em: 27 set. 2011. Citado 3 vezes nas páginas 63, 65 e 67.

AUGER, A.; BARRIÈRE, C. Pattern-based approaches to semantic relation extraction: A state-of-the-art. In: AUGER, A.; BARRIÈRE, C. (Ed.). **Terminology**. Canada: John Benjamins Publishing Co., 2008. v. 14, n. 1, p. 1–19. Citado na página 89.

BAADER, F.; HORROCKS, I.; SATTLER, U. Description logics. In: STAAB, S.; STUDER, R. (Ed.). **Handbook of ontologies**. 2. ed. Berlin: Springer, 2009. p. 21 – 43. Disponível em: <http://www.springerlink.com/index/10.1007/978-3-540-92673-3>. Acesso em: 12 mai. 2012. Citado na página 63.

BAEZA-YATES, R.; RIBEIRO-NETO, B. **Modern Information Retrieval**. 1. ed. Boston, MA: Addison-Wesley Longman Publishing Co., Inc., 1999. ISBN 020139829X. Citado 9 vezes nas páginas 35, 86, 106, 111, 119, 121, 123, 124 e 125.

BANERJEE, S.; PEDERSEN, T. An adapted lesk algorithm for word sense disambiguation using wordnet. In: **Computational linguistics and intelligent text processing**. [S.l.]: Springer, 2002. p. 136–145. Citado na página 175.

BERNERS-LEE, T. **The World Wide Web**: **A very short personal history**. 1998. Disponível em: <http://www.w3.org/People/Berners-Lee/ShortHistory.html>. Citado na página 51.

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The semantic web. **Scientific American**, v. 284, n. 5, p. 34–43, 2001. ISSN 0036-8733. Disponível em: <http://www.jeckle.de/files-/tblSW.pdf>. Acesso em: 17 set. 2011. Citado 3 vezes nas páginas 33, 52 e 72.

BERNERS-LEE, T. et al. **Semantic web road map**. 1998. Citado 2 vezes nas páginas 51 e 53.

BIRD, S.; KLEIN, E.; LOPER, E. **Natural Language Processing with Python**. 1st. ed. O'Reilly Media, Inc., 2009. ISBN 0596516495, 9780596516499. Disponível em: <http://www-.nltk.org/book/>. Citado 3 vezes nas páginas 98, 99 e 147.

BOND, F.; FOSTER, R. Linking and extending an open multilingual wordnet. In: **ACL (1)**. [S.l.: s.n.], 2013. p. 1352–1362. Citado na página 145.

BORST, W. N. **Construction of engineering ontologies for knowledge sharing and reuse**. Tese (Doutorado) — Institute for Telematica and Information Technology, University of Twente, Enschede, The Netherlands, 1997. Disponível em: <http://doc.utwente.nl/17864/1/t0000004-.pdf>. Acesso em: 21 set. 2011. Citado na página 74.

BRAGA, K. S. Aspectos relevantes para a seleção de metodologia adequada à pesquisa social em Ciência da Informação. In: MUELLER, S. P. M. (Org.). **Métodos para a pesquisa em Ciência da Informação**. Brasília: Thesaurus, 2007. cap. 1, p. 17–38. Citado na página 139.

BRÄSCHER, M. **Tratamento automático de ambiguidades na recuperação da informação**. 290 f. Tese (Doutorado em Ciência da Informação) — Faculdade de Ciência da Informação, Universidade de Brasília, Brasília, DF, 1999. Citado 4 vezes nas páginas 34, 35, 39 e 89.

BRÄSCHER, M. A ambigüidade na recuperação da informação. IASI, 2002. Citado na página 97.

BREWSTER, C. A. et al. The ontology: Chimaera or pegasus. In: PROCEEDINGS OF THE DAGSTUHL SEMINAR MACHINE LEARNING FOR THE SEMANTIC WEB, 13–18 de maio de 2005, Dagstuhl, Alemanha. **Anais**... Dagstuhl, Alemanha: Schloss Dagstuhl, 2005. Disponível em: <http://eprints.aston.ac.uk/83/1/dagstuhl05.pdf>. Citado na página 33.

BRIN, S.; PAGE, L. The anatomy of a large-scale hypertextual web search engine. **Computer networks and ISDN systems**, Elsevier, v. 30, n. 1, p. 107–117, 1998. Citado 2 vezes nas páginas 119 e 126.

BRODER, A. et al. Graph structure in the web. **Computer networks**, Elsevier, v. 33, n. 1, p. 309–320, 2000. Citado na página 126.

BUITELAAR, P. et al. Ontology lexicalisation: The lemon perspective. In: **WS 2 Workshop Extended Abstracts, 9th International Conference on Terminology and Artificial Intelligence, TIA**. Paris: [s.n.], 2011. p. 33–36. Citado 3 vezes nas páginas 39, 40 e 132.

BUSH, V. As we may think. The Atlantic On-line. **The Atlantic Monthly**, n. 1, p. 101–108, 1945. Disponível em: <http://www.theatlantic.com/unbound/flashbks/computer/bushf.htm>. Citado na página 47.

CAMPOS, M. L. de A. Modelização de domínios de conhecimento: uma investigação dos princípios fundamentais. **Ciência da Informação**, v. 33, n. 1, p. 22–32, jan./abr. 2004. Disponível em: <http://goo.gl/B735P>. Acesso em: 16 ago. 2012. Citado na página 81.

_____. O papel das definições na pesquisa em ontologia. **Perspectivas em Ciência da Informação**, v. 15, n. 1, p. 220–238, jan/abr. 2010. Disponível em: <http://www.scielo.br/pdf-/pci/v15n1/13.pdf>. Acesso em: 18 ago. 2012. Citado na página 74.

CASTEL, F. Ontological computing. **Commun. ACM**, ACM, New York, v. 45, n. 2, p. 29–30, fev. 2002. ISSN 0001-0782. Disponível em: <http://doi.acm.org/10.1145/503124.503141>. Citado na página 74.

CASTELLS, P.; FERNANDEZ, M.; VALLET, D. An adaptation of the vector-space model for ontology-based information retrieval. **Knowledge and Data Engineering, IEEE Transactions on**, IEEE, v. 19, n. 2, p. 261–272, 2007. Citado na página 131.

CASTELLS, P. et al. Semantic web technologies for economic and financial information management. In: **The semantic web: Research and applications**. [S.l.]: Springer, 2004. p. 473–487. Citado na página 131.

_____. Neptuno: Semantic web technologies for a digital newspaper archive. In: **The Semantic Web: Research and Applications**. [S.l.]: Springer, 2004. p. 445–458. Citado na página 131.

CHARNIAK, E. Jack and janet in search of a theory of knowledge. In: MORGAN KAUFMANN PUBLISHERS INC. **Proceedings of the 3rd international joint conference on Artificial intelligence**. [S.l.], 1973. p. 337–343. Citado na página 35.

CHAUÍ, M. **Convite à Filosofia**. 13. ed. São Paulo: Editora Ática, 2003. 424 p. ISBN 85-08-08935-X. Citado na página 73.

CHEATHAM, M.; HITZLER, P. String similarity metrics for ontology alignment. In: ALANI, H. et al. (Ed.). **12th International Semantic Web Conference - Part II. Lecture Notes in Computer Science**. Sidney, Australia: Springer, 2013. v. 8219. Disponível em: <http://knoesis-.wright.edu/pascal/pub/strings-iswc13.pdf>. Citado na página 156.

CHOMSKY, N. Three models for the description of language. **IEEE on Information theory**, v. 2, n. 3, p. 113–124, sep. 1956. Disponível em: <www.chomsky.info/articles/195609–.pdf>. Acesso em: 22 nov. 2012. Citado na página 86.

CIMIANO, P.; UNGER, C.; MCCRAE, J. **Ontology-Based Interpretation of Natural Language**. [S.l.]: Morgan & Claypool Publishers, 2014. 1–178 p. Citado 2 vezes nas páginas 97 e 132.

CIMIANO, P.; VÖLKER, J.; BUITELAAR, P. Ontology construction. In: INDURKHYA, N.; DAMERAU, F. J. (Ed.). **Handbook of Natural Language Processing, Second Edition**. 2. ed. FLorida: CRC Press, Taylor and Francis Group, 2010. cap. 24, p. 577 – 604. Citado 3 vezes nas páginas 39, 84 e 92.

CONTRERAS, J. et al. A semantic portal for the international affairs sector. In: **Engineering Knowledge in the Age of the Semantic Web**. [S.l.]: Springer, 2004. p. 203–215. Citado na página 131.

CORAZZON, R. **Ontology: Theory and History from a Philosophical Perspective**. Outubro 2012. On-line. Disponível em: <http://www.ontology.co/>. Citado na página 74.

DACONTA, M. C.; SMITH, K. T.; OBRST, L. J. **The Semantic Web**: a guide to the future of xml, web services, and knowledge management. Indianapolis: Wiley, 2003. 281 p. Citado 3 vezes nas páginas 55, 56 e 59.

DAHLBERG, I. A referent-oriented analytical concept theory of interconcept. **International Classification**, v. 5, n. 3, p. 142 – 150, 1978. Citado 4 vezes nas páginas 77, 84, 90 e 93.

_____. Teoria do conceito. **Ciência da Informação**, Rio de Janeiro, v. 7, n. 2, p. 101–107, 1978. Disponível em: <goo.gl/sgPlv>. Acesso em: 22 out. 2011. Citado 3 vezes nas páginas 39, 75 e 80.

_____. Teoria da classificação, ontem e hoje. In: CONFERÊNCIA BRASILEIRA DE CLASSIFICAÇÃO BIBLIOGRÁFICA, 12–17 de setembro de 1972, Rio de Janeiro. **Anais**... Rio de Janeiro: IBICT, 1979. v. 1, p. 352–370. Acesso em: 01 nov. 2011. Citado na página 93.

DALE, R. Classical approaches to natural language processing. In: INDURKHYA, N.; DAMERAU, F. J. (Ed.). **Handbook of natural language processing**. 2nd. ed. Boca Raton, FL, USA: CRC Press, 2010, (Machine Learning & Pattern Recognition Series, v. 2). cap. 1, p. 3–7. Citado na página 92.

DUCHARME, B. **Learning SPARQL**. Sebastopol: O'Reilly Media, Inc, 2011. Citado na página 68.

FEITOSA, A. L. G. **A Integração entre Sistemas Legislativos, Terminologia e Web Semântica na organização e representação da informação legislativa**. 405 f. Tese (Doutorado em Ciência da Informação) — Universidade de Brasília, Brasília, DF, mar. 2005. Citado na página 70.

FELLBAUM, C. **Wordnet: An electronic lexical database**. Cambridge, Massachusetts / London, England: The MIT Press, 1998. 423 p. ISBN 0-262-06197-X. Citado na página 144.

FEOFILOFF, P.; KOHAYAKAWA, Y.; WAKABAYASHI, Y. Uma introdução sucinta à teoria dos grafos. **Disponıvel em http://www. ime. usp. br/~ pf/teoriadosgrafos**, 2011. Citado na página 126.

FERNÁNDEZ, M. et al. Semantically enhanced information retrieval: An ontology-based approach. **Web Semantics: Science, Services and Agents on the World Wide Web**, v. 9, n. 4, p. 434 – 452, 2011. ISSN 1570-8268. {JWS} special issue on Semantic Search. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1570826810000910>. Citado 2 vezes nas páginas 160 e 161.

FERNANDEZ, M. et al. Semantic search meets the web. In: IEEE. **Semantic Computing, 2008 IEEE International Conference on**. [S.l.], 2008. p. 253–260. Citado na página 132.

FININ, T. et al. Information retrieval and the semantic web. In: IEEE. **System Sciences, 2005. HICSS'05. Proceedings of the 38th Annual Hawaii International Conference on**. [S.l.], 2005. p. 113a–113a. Citado 2 vezes nas páginas 127 e 128.

FIORIN, J. (Org.). **Introdução à Linguística**. São Paulo: Contexto, 2002. (Introdução à linguística, I e II). Citado na página 48.

FREITAS, C.; ROCHA, P.; BICK, E. Um mundo novo na floresta sintá (c) tica–o treebank do português. **Calidoscópio**, v. 6, n. 3, p. 142–148, 2008. Disponível em: <http://www.linguateca-.pt/Floresta/>. Citado na página 146.

FREITAS, M. C. de. **Elaboração automática de ontologias de domínio: discussão e resultados**. 142 p. Tese (Doutorado) — Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, RJ, 2007. Disponível em: <http://www.linguateca.pt/Repositorio/MCFreitas_Certificada-.rar>. Acesso em: 04 abr. 2012. Citado 2 vezes nas páginas 94 e 96.

GARRÃO, M. U. Linguística de córpus: o lugar da fusão entre semântica e pragmática. **Calidoscópio**, v. 4, n. 3, p. 135–140, set/dez. 2006. Citado na página 95.

GÓMEZ, P. C. Do we need statistics when we have linguistics? **D.E.L.T.A. - Documentação de Estudos em Linguística Teórica e Aplicada**, v. 18, n. 2, p. 233–271, 2002. Disponível em: <http://www.scielo.br/pdf/delta/v18n2/v18n2a03.pdf>. Acesso em: 30 ago. 2012. Citado na página 86.

GOMÉZ-PERÉZ, A.; BENJAMINS, V. R. Overview of knowledge sharing and reuse components: Ontologies and problem-solving methods. In: BENJAMINS, V. R. et al. (Ed.). **Proceedings of the IJCAI-99 Workshop on Ontologies and Problem-Solving Methods (KRR5)**. Stockholm, Sweden: [s.n.], 1999. p. 1–15. Disponível em: <http://citeseerx.ist.psu-.edu/viewdoc/summary?doi=10.1.1.39.249>. Acesso em: 02 abr. 2011. Citado na página 77.

GRESSER, J.-Y. et al. D2. 1 draft ontology of financial risks & dependencies within & outside the financial sector vol. 1-building ontologies. 2010. Citado na página 142.

GRUBER, T. R. A translation approach to portable ontology specifications. **Knowledge Acquisition**, v. 5, n. 2, p. 199–220, jun 1993. Disponível em: <http://tomgruber.org/writing-/ontolingua-kaj-1993.pdf>. Acesso em: 20 set. 2011. Citado 2 vezes nas páginas 33 e 77.

_____. What is an ontology? 1993. Disponível em: <http://www-ksl.stanford.edu/kst/what-is-an-ontology.html>. Acesso em: 18 ago. 2011. Citado na página 74.

GUARINO, N. Understanding, building and using ontologies. **International Journal of Human Computer Studies**, Citeseer, v. 46, p. 293–310, 1997. Disponível em: <http://goo.gl-/DNhbk>. Citado na página 33.

_____. Formal ontology in information systems. In: **Formal Ontology in Information Systems**: **Proceedings of the 1st International Conference June 6-8**, **1998**, **Trento**, **Italy**. 1st. ed. Amsterdam, The Netherlands: IOS Press, 1998. p. 3 – 15. ISBN 9051993994. Citado 2 vezes nas páginas 41 e 78.

GUARINO, N.; GIARETTA, P. Ontologies and knowledge bases: towards a terminological clarification. In: MARS, N. (Ed.). **Towards very large knowledge bases**: knowledge building and knowledge sharing. Amsterdan: IOS Press, 1995. p. 25–32. Disponível em: <http://www-.loa.istc.cnr.it/Papers/KBKS95.pdf>. Acesso em: 23 set. 2011. Citado na página 73.

GUARINO, N.; OBERLE, D.; STAAB, S. What is an ontology? In: STAAB, S.; STUDER, R. (Ed.). **Handbook of ontologies**. 2. ed. Berlin: Springer, 2009. p. 1 – 17. Disponível em: <http://www.springerlink.com/index/10.1007/978-3-540-92673-3>. Acesso em: 27 set. 2011. Citado 4 vezes nas páginas 40, 74, 93 e 94.

GUARINO, N.; WELTY, C. A. An overview of ontoclean. In: STAAB, S.; STUDER, R. (Ed.). **Handbook of ontologies**. 2. ed. Berlin: Springer, 2009. p. 201–220. Disponível em: <http:/-/www.springerlink.com/index/10.1007/978-3-540-92673-3>. Acesso em: 29 set. 2011. Citado 3 vezes nas páginas 75, 76 e 77.

GUHA, R.; MCCOOL, R. Tap: a semantic web platform. **Computer Networks**, v. 42, n. 5, p. 557 – 577, 2003. ISSN 1389-1286. The Semantic Web: an evolution for a revolution. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1389128603002251>. Citado na página 131.

GUHA, R.; MCCOOL, R.; MILLER, E. Semantic search. In: ACM. **Proceedings of the 12th international conference on World Wide Web**. [S.l.], 2003. p. 700–709. Citado na página 131.

HATZIVASSILOGLOU, V. Do we need linguistics when we have statistics? a comparative analysis of the contributions of linguistic cues to a statistical word grouping system. In: KLAVANS, J. L.; RESNIK, P. (Ed.). **The Balancing Act**: Combining symbolic and statistical approaches to language. New York: MIT Press, 1996. cap. 4, p. 67–94. Disponível em: <http://www.cs.columbia.edu/~vh/Papers/1994/BalancingAct.pdf>. Acesso em: 30 ago. 2012. Citado na página 85.

HEARST, M. A. Automatic acquisition of hyponyms from large text corpora. In: LINGUISTICS, A. for C. (Ed.). **Proceedings of the 14th Conference on Computational linguistics**. Stroudsburg: Association for Computational Linguistics, 1992. (COLING '92, v. 2), p. 23–28. Disponível em: <http://acl.ldc.upenn.edu/C/C92/C92-2082.pdf>. Acesso em: 11 nov. 2011. Citado na página 96.

_____. Untangling text data mining. In: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS ON COMPUTATIONAL LINGUISTICS, 37., 1999, College Park. **Proceedings**... Stroudsburg: Association for Computational Linguistics, 1999. p. 3–10. Citado na página 91.

HEATH, T.; BIZER, C. **Linked Data**: Evolving the web into a global data space. 1st. ed. [S.l.]: Morgan & Claypool Publishers, 2011. (Synthesis Lectures on the Semantic Web: Theory and Technology). Citado 3 vezes nas páginas 34, 54 e 132.

HIRST, G. Ontology and the lexicon. In: STAAB, S.; STUDER, R. (Ed.). **Handbook of ontologies**. 2. ed. Berlin: Springer, 2009. p. 269 – 287. Disponível em: <http://www.springerlink.com/index/10.1007/978-3-540-92673-3>. Acesso em: 26 jun. 2012. Citado 2 vezes nas páginas 84 e 85.

HOGAN, A. et al. Searching and browsing linked data with swse: The semantic web search engine. **Web Semantics: Science, Services and Agents on the World Wide Web**, v. 9, n. 4, p. 365 – 401, 2011. ISSN 1570-8268. {JWS} special issue on Semantic Search. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1570826811000473>. Citado na página 132.

HORROCKS, I. Reasoning with expressive description logics: Theory and practice. In: **In: Andrei Voronkov, (ed) Proc. 18th Int. Conf. on Automated Deduction (CADE-18**. [S.l.]: Springer, 2002. p. 1–15. Citado na página 64.

JURAFSKY, D.; MARTIN, J. H. **Speech and Language Processing (2Nd Edition)**. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 2009. ISBN 0131873210. Citado 6 vezes nas páginas 97, 100, 102, 104, 105 e 106.

KAO, A.; POTEET, S. Text mining and natural language processing: introduction for the special issue. **SIGKDD Explor. Newsl.**, ACM, v. 7, n. 1, p. 1–2, 2005. Citado na página 92.

KARA, S. et al. An ontology-based retrieval system using semantic indexing. **Information Systems**, v. 37, n. 4, p. 294 – 305, 2012. ISSN 0306-4379. Semantic Web Data Management. Disponível em: <http://www.sciencedirect.com/science/article/pii/S030643791100113X>. Citado na página 160.

KHATTREE, R.; NAIK, D. N. Book. **Multivariate Data Reduction and Discrimination with SAS Software**. [S.l.]: SAS Publishing, 2000. Citado 2 vezes nas páginas 116 e 117.

KRÖTZSCH, M.; SIMANCIK, F.; HORROCKS, I. A description logic primer. 01 2012. Disponível em: <http://arxiv.org/abs/1201.4089>. Citado na página 64.

LANCASTER, F. W. **Information Retrieval Systems**: Characteristics, testing and evaluation. New York: Wiley, 1968. 233 p. Citado na página 86.

_____. **Indexação e resumos: teoria e prática**. Brasília: Briquet de Lemos, 1993. 347 p. Citado na página 122.

LASSILA, O.; MCGUINNESS, D. The role of frame-based representation on the semantic web. **Linköping Electronic Articles in Computer and**, 2001. Disponível em: <http://www-.ep.liu.se/ea/cis/2001/005/cis01005.pdf>. Citado na página 79.

LESK, M. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In: **Proceedings of the 5th Annual International Conference on Systems Documentation**. New York, NY, USA: ACM, 1986. (SIGDOC '86), p. 24–26. ISBN 0-89791-224-1. Disponível em: <http://doi.acm.org/10.1145/318723.318728>. Citado na página 155.

LIMA-MARQUES, M. **Ontologias**: da filosofia à representação do conhecimento. Brasília: Thesaurus, 2006. 72 p. Citado 2 vezes nas páginas 33 e 73.

LUHN, H. P. The automatic creation of literature abstracts. **IBM Journal of Reasearch and Development**, p. 159–165, 1958. Citado 2 vezes nas páginas 113 e 114.

MAEDCHE, A.; STAAB, S. Measuring similarity between ontologies. In: **Knowledge engineering and knowledge management: Ontologies and the semantic web**. [S.l.]: Springer, 2002. p. 251–263. Citado na página 156.

_____. Mining ontologies from text. In: EUROPEAN WORKSHOP ON KNOWLEDGE ACQUISITION, MODELING AND MANAGEMENT, 12., 2000, Juan-les-Pin, France. **Proceedings**... Berlin: Springer-Verlag, 2002. (Lecture Notes in Computer Science, v. 1937), p. 189 – 202. Citado na página 33.

MAEDCHE, A. et al. Seal—a framework for developing semantic web portals. In: **Advances in Databases**. [S.l.]: Springer, 2001. p. 1–22. Citado na página 131.

MANNING, C. D.; RAGHAVAN, P.; SCHÜTZE, H. **Introduction to information retrieval**. Cambridge, England: Cambridge University Press, 2009. 544 p. Online Edition (c) 2009 Cambridge UP. Disponível em: <http://www.informationretrieval.org>. Citado 11 vezes nas páginas 36, 39, 113, 115, 119, 120, 123, 124, 125, 128 e 130.

MANNING, C. D.; SCHÜTZE, H. **Foundations of Statistical Natural Language Processing**. Cambridge: The MIT Press, 1999. 680 p. Citado 10 vezes nas páginas 87, 88, 91, 102, 105, 111, 115, 116, 117 e 122.

MARCONDES, C. H.; CAMPOS, M. L. de A. Ontologia e web semântica: o espaço da pesquisa em ciência da informação. **PontodeAcesso**, v. 2, n. 1, p. 117–136, jul 2008. Disponível em: <http://www.portalseer.ufba.br/index.php/revistaici/article/view/2669/1885>. Acesso em: 02 ago. 2012. Citado 3 vezes nas páginas 33, 70 e 81.

MARCONI, M. de A.; LAKATOS, E. M. **Metodologia Científica**. 4. ed. São Paulo: Atlas, 2004. 305 p. Citado na página 139.

MAYFIELD, J.; FININ, T. Information retrieval on the semantic web: Integrating inference and retrieval. In: **Proceedings of the SIGIR Workshop on the Semantic Web**. [S.l.: s.n.], 2003. Citado na página 128.

MCCLEARY, L.; VIOTTI, E. Semântica e pramática. Apostila do Curso de Licenciatura e Bacharelado em Letras-Libras na Modalidade à Distância, Universidade Federal de Santa Catarina. 2009. Citado 3 vezes nas páginas 89, 90 e 91.

MCCRAE, J. et al. **The lemon cookbook**. 2010. Disponível em: <http://lemon-model.net-/index.php>. Citado na página 150.

_____. Interchanging lexical resources on the semantic web. **Language Resources and Evaluation**, Springer, v. 46, n. 4, p. 701–719, 2012. Citado na página 132.

MCCRAE, J.; SPOHR, D.; CIMIANO, P. Linking lexical resources and ontologies on the semantic web with lemon. In: **The Semantic Web: Research and Applications**. [S.l.]: Springer, 2011. p. 245–259. Citado na página 149.

MELO, F. J. D. de; BRÄSCHER, M. **Fundamentos da linguística para a formação do profissional de informação**. Brasília, DF: Centro Editorial, 2011. 124 p. Citado 3 vezes nas páginas 83, 87 e 88.

MIRANDA, A. Book. **Ciência da Informação: teoria e metodologia de uma área em expansão**. 2. ed. Brasília: Elmira Simeão, organizadora., 2003. 212 p. Citado na página 98.

MOENS, M. F. Book. **Automatic indexing and abstracting of document texts**. [S.l.]: Kluwer Academic Publishers, 2000. Citado na página 114.

MOOERS, C. E. Coding, information retrieval, and the rapid selector. **American Documentation**, v. 1, n. 4, p. 225–229, 1950. Citado na página 119.

NARDI, D.; BRACHMAN, R. J. An introduction to description logics. In: BAADER, F. et al. (Ed.). **The description logic handbook: theory, implementation, and applications**. 1st. ed. Cambridge (Massachusetts): Cambridge University Press, 2003. p. 5 – 44. Citado na página 61.

NAVIGLI, R.; VELARDI, P.; GANGEMI, A. Ontology learning and its application to automated terminology translation. **IEEE Intelligent Systems**, v. 18, n. 1, p. 22–31, 2003. Citado na página 132.

NIVRE, J. Two notions of parsing. In: ARPPE, A. et al. (Ed.). **Inguiries into Words, Constraints and Contexts**. Festschrift for kimmo koskenniemi on his 60th birthday. Stanford, CA, USA: CSLI Publications/Stanford University, 2005, (CSLI Studies in Computational Linguistics ONLINE). cap. 11, p. 106–115. Disponível em: <http://web.stanford.edu/group-/cslipublications/cslipublications/koskenniemi-festschrift/kk-festschrift-all-2005.pdf>. Citado na página 105.

OGDEN, C. K.; RICHARDS, I. **The meaning of meaning**. London: Trubner & Co, 1923. Citado 3 vezes nas páginas 49, 50 e 54.

OLIVEIRA, H. G. **Onto. PT: Towards the Automatic Construction of a Lexical Ontology for Portuguese**. Tese (Doutorado) — Ph. D. thesis, University of Coimbra, 2013. Citado na página 145.

PAIVA, V. de; RADEMAKER, A.; MELO, G. de. Openwordnet-pt: An open brazilian wordnet for reasoning. In: **Proceedings of the 24th International Conference on Computational Linguistics**. [s.n.], 2012. Disponível em: <http://hdl.handle.net/10438/10274>. Citado na página 144.

PALMER, D. D. Text preprocessing. In: INDURKHYA, N.; DAMERAU, F. J. (Ed.). **Handbook of natural language processing**. 2nd. ed. Boca Raton, FL, USA: CRC Press, 2010, (Machine Learning & Pattern Recognition Series, v. 2). cap. 1, p. 9–30. Citado 2 vezes nas páginas 101 e 102.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in python. **The Journal of Machine Learning Research**, JMLR. org, v. 12, p. 2825–2830, 2011. Citado na página 147.

PICKLER, M. E. V. Web semântica: ontologias como ferramentas de representação do conhecimento. **Perspectivas em ciência da informação**, v. 12, n. 1, p. 65–83, 2007. Citado na página 70.

POPOV, B. et al. Kim – a semantic platform for information extraction and retrieval. **Natural Language Engineering**, v. 10, p. 375–392, 9 2004. ISSN 1469-8110. Citado na página 131.

PRUD'HOMMEAUX, E. **Semantic Web Specification at W3C - slide "Semantic Web "Layer Cake**. mar. 2004. Disponível em: <http://www.w3.org/2004/Talks/0412-RDF-functions-/slide4-0.html>. Citado na página 53.

RAYWARD, W. B. **Anticipating the Digital World: Paul Otlet and his Paper Internet**. Bartels Memorial Lecture, Leeds Metropolitan University, mai. 2002. Disponível em: <http:/-/courseweb.lis.illinois.edu/~wrayward/Otlet/Bartelslecture1.pdf>. Citado na página 47.

REYMONET, A.; THOMAS, J.; AUSSENAC-GILLES, N. Modelling ontological and terminological resources in owl dl. In: **Proceedings of ISWC**. [S.l.: s.n.], 2007. v. 7. Citado 2 vezes nas páginas 35 e 132.

RIJSBERGEN, C. J. V. **Information Retrieval**. 2. ed. London: Butterworth-Heinemann, 1979. 224 p. Disponível em: <http://www.dcs.gla.ac.uk/Keith/Preface.html>. Acesso em: 23 fev. 2012. Citado 5 vezes nas páginas 34, 115, 121, 124 e 125.

ROBERTSON, S. E.; SPÄRCK JONES, K. Relevance weighting of search terms. **Journal of the American Society for Information Science**, School of Library, Archive and Information Studies University College London London WC1E 6BT, England; Computer Laboratory University of Cambridge Cambridge CB2 3QG, England, v. 27, n. 3, p. 129–146, 1976. Disponível em: <http://dx.doi.org/10.1002/asi.4630270302>. Citado na página 125.

ROBREDO, J. Ciência da informação e web semântica: Linhas convergentes ou linhas paralelas? In: **Passeios pelo bosque da informação: estudos sobre a representação e organização da informação e do conhecimento**. Edição comemorativa dos 10 anos do grupo de pesquisa eroic. Brasília: IBICT, 2010. cap. 1, p. 12–47. Citado 2 vezes nas páginas 71 e 81.

_____. Do documento impresso à informação nas nuvens: reflexões. **PBCIB**, v. 6, n. 2, 2012. Citado na página 47.

ROCHA, C.; SCHWABE, D.; ARAGAO, M. P. A hybrid approach for searching in the semantic web. In: **Proceedings of the 13th International Conference on World Wide Web**. New York, NY, USA: ACM, 2004. (WWW '04), p. 374–383. ISBN 1-58113-844-X. Disponível em: <http://doi.acm.org/10.1145/988672.988723>. Citado na página 131.

ROMARY, L. Standardization of the formal representation of lexical information for nlp. **Dictionaries. An International Encyclopedia of Lexicography. Supplementary volume: Recent developments with special focus on computational lexicography**, 2010. Citado na página 151.

SALES, L. F.; CAMPOS, M. L. de A.; GOMES, H. E. Ontologias de domínio: um estudo das relações conceituais. **Perspectivas em Ciência da Informação**, v. 13, n. 2, p. 62–76, maio/ago. 2008. Disponível em: <http://www.eci.ufmg.br/pcionline/index.php/pci/article-/viewFile/219>. Acesso em: 20 mar. 2012. Citado 2 vezes nas páginas 78 e 80.

SALTON, G. **Automatic Text Processing: The Transformation, Analysis, and Retrieval of**. [S.l.]: Addison-Wesley, 1989. Citado na página 106.

SALTON, G.; BUCKLEY, C. Term-weighting approaches in automatic text retrieval. **Information processing & management**, v. 24, n. 5, p. 513 – 523, 1988. Citado 3 vezes nas páginas 123, 129 e 130.

SANTOS, P. Paul otlet: um pioneiro da organização das redes mundiais de tratamento e difusão da informação registrada. **Ciência da Informação**, Brasília, DF, Brasil, v. 36, n. 2, set. 2008. Disponível em: <http://revista.ibict.br/ciinf/index.php/ciinf/article/view/971/1637>. Citado na página 46.

SARACEVIC, T. Information Science. In: BATES, M. J.; MAACK, M. N. (Ed.). **Encyclopedia of Library and Information Science**. New York: Taylor an Francis, 2009. p. 2570–2586. Disponível em: <http://comminfo.rutgers.edu/~tefko/SaracevicInformationScienceELIS2009-.pdf>. Acesso em: 23 fev. 2012. Citado na página 82.

SAUSSURE, F. d. **Curso de Linguística Geral**. 22. ed. São Paulo: Cultrix, 2000. 279 p. Citado 2 vezes nas páginas 86 e 90.

SAVOY, J.; GAUSSIER, E. Information retrieval. In: INDURKHYA, N.; DAMERAU, F. J. (Ed.). **Handbook of Natural Language Processing**. 2. ed. Florida: CRC Press, Taylor and Francis Group, 2010. cap. 19, p. 455–484. Citado 7 vezes nas páginas 87, 88, 120, 122, 123, 124 e 125.

SCHIESSL, M. **Descoberta de Conhecimento em Textos aplicada a um Sistema de Atendimento ao Consumidor**. 106 p. Dissertação (Mestrado) — Universidade de Brasília, Brasília, 2007. Disponível em: <http://hdl.handle.net/10482/1414>. Citado na página 156.

SCHIESSL, M.; BRÄSCHER, M. Ontologia: ambiguidade e precisão. **Encontros Bibli: Revista Eletrônica de Biblioteconomia e Ciência da Informação**, v. 17, n. n. esp. 1, p. 125–141, 2012. Disponível em: <http://dx.doi.org/10.5007/1518-2924.2012v17nesp1p125>. Acesso em: 25 ago. 2012. Citado na página 33.

SÉRASSET, G. Dbnary: Wiktionary as a lemon based rdf multilingual lexical resource. **Semantic Web Journal-Special issue on Multilingual Linked Open Data**, 2014. Citado na página 145.

SHAH, U. et al. Information retrieval on the semantic web. In: ACM. **Proceedings of the eleventh international conference on Information and knowledge management**. [S.l.], 2002. p. 461–468. Citado na página 127.

SHAMSFARD, M.; BARFOROUSH, A. A. The state of the art in ontology learning: a framework for comparison. **Knowledge Engineering Review**, v. 18, n. 4, p. 293–316, Dec. 2003. Citado na página 33.

_____. Learning ontologies from natural language texts. **International Journal of Human-Computer Studies**, v. 60, n. 1, p. 17–63, 2004. Citado na página 33.

SHANNON, C. E. The mathematical theory of communication. **The Bell System Technical Journal**, v. 27, n. 3, 4, p. 379–423, 623–656, July, October, 1948. Disponível em: <http://cm-.bell-labs.com/cm/ms/what/shannonday/shannon1948.pdf>. Citado na página 86.

SHARMA, S. Book. **Applied Multivariate Techniques**. [S.l.]: John Wiley & Sons, Inc., 1996. Citado 2 vezes nas páginas 116 e 117.

SHUKLA, P. P.; SINGH, R. K. Semantic web: An introduction to information retrieval. AISSMS College of Engineering, 2013. Disponível em: <http://dspace.iconad.in:8080/jspui/handle-/123456789/789>. Citado na página 127.

SILVA, B. C. D. da et al. **Introdução ao Processamento das Linguas Naturais e Algumas Aplicações**. São Paulo, Agosto 2007. Citado 3 vezes nas páginas 87, 88 e 89.

SILVA, F.; GIRARDI, R.; DRUMOND, L. An information retrieval model for the semantic web. In: IEEE. **Information Technology: New Generations, 2009. ITNG'09. Sixth International Conference on**. [S.l.], 2009. p. 143–148. Citado na página 132.

SINGH, G.; JAIN, V. Information retrieval (ir) through semantic web (sw): An overview. **arXiv preprint arXiv:1403.7162**, 2014. Citado na página 127.

SMITH, B. Beyond concepts: ontology as reality representation. In: **Proceedings of the third international conference on formal ontology in information systems (FOIS 2004)**. [S.l.: s.n.], 2004. p. 73–84. Citado na página 73.

SMITH, D. W. Phenomenology. **The Stanford Encyclopedia of Philosophy**, 2003. Disponível em: <http://plato.stanford.edu/archives/win2003/entries/phenomenology>. Citado na página 73.

SNOW, R.; JURAFSKY, D.; NG, A. Y. Learning syntactic patterns for automatic hypernym discovery. In: SAUL, L. K.; WEISS, Y.; BOTTOU, L. (Ed.). **Advances in Neural Information Processing Systems 17**. Cambridge, MA: MIT Press, 2005. p. 1297–1304. Citado na página 96.

SOERGEL, D. The rise of ontologies or the reinvention of classification. **Journal of the American Society for Information Science**, v. 50, n. 12, p. 1119–1120, October 1999. Disponível em: <http://www.dsoergel.com/cv/B70.pdf>. Acesso em: 09 dez. 2011. Citado na página 81.

SOUZA, R. R.; ALVARENGA, L. A web semântica e suas contribuições para a ciência da informação. **Ciência da Informação**, Brasília, v. 33, n. 1, jun. 2004. Disponível em: <http:-//revista.ibict.br/index.php/ciinf/article/view/50/49>. Acesso em: 20 nov. 2011. Citado na página 70.

SOWA, J. F. **Knowledge representation**: logical, philosophical and computational foundations. 1. ed. Pacific Grove, CA, USA: Brooks/Cole Publishing Co., 2000. 594 p. ISBN 0-534-94965-7. Citado 2 vezes nas páginas 75 e 77.

SPÄRCK JONES, K. A statistical interpretation of term specificity and its application in retrieval. **Journal of documentation**, MCB UP Ltd, v. 28, n. 1, p. 11–21, 1972. Citado na página 110.

STUDER, R.; BENJAMINS, R. R.; FENSEL, D. Knowledge engineering: Principles and methods. **Data Knowledge Engineering**, v. 25, n. 1/2, p. 161–197, 1998. Acesso em: 25 mai. 2012. Citado 2 vezes nas páginas 74 e 75.

TERRA, E. Book. **Curso Prático de Gramática**. [S.l.]: Scipione, 2002. Citado na página 106.

UNGER, C. et al. A lemon lexicon for dbpedia. In: **Proceedings of 1st International Workshop on NLP and DBpedia, co-located with the 12th International Semantic Web Conference (ISWC 2013), October 21-25, Sydney, Australia**. [S.l.: s.n.], 2013. Citado na página 132.

UREN, V. et al. The usability of semantic search tools: a review. **The Knowledge Engineering Review**, Cambridge Univ Press, v. 22, n. 04, p. 361–377, 2007. Citado na página 127.

VALLET, D.; FERNÁNDEZ, M.; CASTELLS, P. The quest for information retrieval on the semantic web. **Upgrade 6 (6), Monograph: The Semantic Web**, p. 19–23, December 2005. Disponível em: <http://ir.ii.uam.es/~search/publications/upgrade05.pdf>. Citado 2 vezes nas páginas 128 e 131.

VICKERY, B. C. Ontologies. **Journal of Information Science**, v. 23, n. 4, p. 277–286, 1997. Disponível em: <http://mba.eci.ufmg.br/downloads/recol/277.pdf>. Acesso em: 04 dez. 2010. Citado na página 80.

VIEIRA, R.; LIMA, V. L. S. de. Linguística computacional: princípios e aplicações. In: CONGRESSO DA SOCIEDADE BRASILEIRA DE COMPUTAÇÃO, 21., 2001, Fortaleza. **Anais**... Fortaleza: SBC, 2001. v. 2, p. 47–88. Disponível em: <http://goo.gl/6HKaI>. Acesso em: 08 jun. 2011. Citado na página 92.

VITAL, L. P.; CAFÉ, L. M. A. Ontologias e taxonomias: diferenças. **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 16, n. 2, p. 115 - 130, abr/jun. 2011. Disponível em: <http://portaldeperiodicos.eci.ufmg.br/index.php/pci/article/view/200/866>. Acesso em: 26 out. 2011. Citado na página 33.

VÖLKER, J.; HITZLER, P.; CIMIANO, P. Acquisition of owl dl axioms from lexical resources. In: SHETH, A. et al. (Ed.). **The Semantic Web: Research and Applications**. [S.l.]: Springer Berlin / Heidelberg, 2007, (Lecture Notes in Computer Science, v. 5318). p. 670–685. Citado na página 84.

W3C. **OWL 2 Web Ontology Language**. jul 2014. Disponível em: <http://www.w3.org/TR/2012/REC-owl2-new-features-20121211/>. Citado 2 vezes nas páginas 62 e 66.

W3SCHOOL. fevereiro 2013. Disponível em: <http://www.w3schools.com/rdf/rdf_intro.asp>. Citado na página 63.

WAKEFIELD, T. Miscellaneous, **A Perfect Storm is Brewing**: **Better Answers are Possible by Incorporating Unstructured Data Analysis Techniques**. 2004. Disponível em: <http://www.datawarehouse.com/article/?articleid=4766>. Citado 2 vezes nas páginas 99 e 100.

WALTER, S.; UNGER, C.; CIMIANO, P. A corpus-based approach for the induction of ontology lexica. In: **Natural Language Processing and Information Systems**. [S.l.]: Springer, 2013. p. 102–113. Citado na página 132.

_____. Atoll – a framework for the automatic induction of ontology lexica. 2014. Citado na página 132.

WEISS, S. et al. Book. **Text Mining**: **Predictive Methods for Analyzing Unstructured Information**. [S.l.]: Springer, 2005. Citado na página 111.

WILKS, Y. Knowledge structures and language boundaries. In: MORGAN KAUFMANN PUBLISHERS INC. **Proceedings of the 5th international joint conference on Artificial intelligence-Volume 1**. [S.l.], 1977. p. 151–157. Citado na página 35.

WILKS, Y.; BREWSTER, C. Natural language processing as a foundation of the semantic web. **Foundations and Trends in Web Science**, Now Publishers Inc., v. 1, n. 3–4, p. 199–327, 2009. Citado 2 vezes nas páginas 35 e 63.

WITTGENSTEIN, L. J. J. **Investigações Filosóficas**. Tradução de José Carlos Bruni. São Paulo: Nova Cultural, 1999. (Os Pensadores). Citado 2 vezes nas páginas 90 e 91.

WIVES, L. K. **Técnicas de descoberta de conhecimento em textos aplicadas à inteligência competitivaTécnicas de descoberta de conhecimento em textos aplicadas à inteligência competitiva**. Tese (Thesis (PhD)), 2001. Citado na página 111.

WOODFIELD, T. Manual (technical documentation), **Mining Textual Data Using SAS® Text Miner for SAS®9 Course Notes**. 2004. Citado 3 vezes nas páginas 98, 99 e 117.

YANG, Y.; PEDERSEN, J. O. A comparative study on feature selection in text categorization. In: . [S.l.]: Morgan Kaufmann Publishers, 1997. p. 412–420. Citado 2 vezes nas páginas 111 e 116.

ZIPF, G. K. **Human Behavior and the Principle of Least Effort**. Cambridge, MA: Addison-Wesley, 1949. Citado 3 vezes nas páginas 86, 112 e 113.

Anexos

ANEXO A – Ontologia do Risco Financeiro (OntoRisco)

```
<?xml version="1.0"?>
<!DOCTYPE rdf:RDF [
   <!ENTITY owl "http://www.w3.org/2002/07/owl#" >
   <!ENTITY xsd "http://www.w3.org/2001/XMLSchema#" >
   <!ENTITY rdfs "http://www.w3.org/2000/01/rdf-schema#" >
   <!ENTITY rdf "http://www.w3.org/1999/02/22-rdf-syntax-ns#" >
   <!ENTITY risk_management "http://www.semanticweb.org/ontologies/2014/8/risk_management#" >
]>
<rdf:RDF xmlns="http://www.semanticweb.org/ontologies/2014/8/untitled-ontology-106#"
     xml:base="http://www.semanticweb.org/ontologies/2014/8/untitled-ontology-106"
     xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
     xmlns:risk_management="http://www.semanticweb.org/ontologies/2014/8/risk_management#"
     xmlns:owl="http://www.w3.org/2002/07/owl#"
     xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
     xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
    <owl:Ontology rdf:about="http://www.semanticweb.org/ontologies/2014/8/risk_management">
       <rdfs:label xml:lang="en">Financial Risk Management Ontology</rdfs:label>
       <rdfs:label xml:lang="pt">Ontologia de Gestão de Risco Financeiro</rdfs:label>
       <rdfs:isDefinedBy>Marcelo Schiessl</rdfs:isDefinedBy>
    </owl:Ontology>

    <!--
    ////////////////////////////////////////////////////////////////////////////////////
    //
    // Object Properties
    //
    ////////////////////////////////////////////////////////////////////////////////////
     -->

    <!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#attraction -->
    <owl:ObjectProperty rdf:about="&risk_management;attraction">
       <rdfs:label xml:lang="en">attraction</rdfs:label>
       <rdfs:label xml:lang="pt">atração</rdfs:label>
       <rdfs:domain rdf:resource="&risk_management;Risk"/>
       <rdfs:range rdf:resource="&risk_management;Threat_agent"/>
    </owl:ObjectProperty>

    <!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#defines -->
    <owl:ObjectProperty rdf:about="&risk_management;defines">
       <rdfs:label xml:lang="en">defines</rdfs:label>
       <rdfs:label xml:lang="pt">define</rdfs:label>
       <rdfs:range rdf:resource="&risk_management;Countermeasure"/>
       <rdfs:range rdf:resource="&risk_management;Organization"/>
       <rdfs:domain rdf:resource="&risk_management;Rule"/>
    </owl:ObjectProperty>
```

```
<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#givesRiseTo -->
<owl:ObjectProperty rdf:about="&risk_management;givesRiseTo">
   <rdfs:label xml:lang="en">gives rise to</rdfs:label>
   <rdfs:label xml:lang="pt">suscita</rdfs:label>
   <rdfs:range rdf:resource="&risk_management;Threat"/>
   <rdfs:domain rdf:resource="&risk_management;Threat_agent"/>
</owl:ObjectProperty>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#has -->
<owl:ObjectProperty rdf:about="&risk_management;has">
   <rdfs:label xml:lang="en">has</rdfs:label>
   <rdfs:label xml:lang="pt">possui</rdfs:label>
   <rdfs:domain rdf:resource="&risk_management;Asset"/>
   <rdfs:range rdf:resource="&risk_management;Vulnerability"/>
</owl:ObjectProperty>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#imposes -->
<owl:ObjectProperty rdf:about="&risk_management;imposes">
   <rdfs:label xml:lang="en">imposes</rdfs:label>
   <rdfs:label xml:lang="pt">impõe</rdfs:label>
   <rdfs:domain rdf:resource="&risk_management;&apos;Agente_de_defesa&apos;"/>
   <rdfs:range rdf:resource="&risk_management;Defesa"/>
</owl:ObjectProperty>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#includes -->
<owl:ObjectProperty rdf:about="&risk_management;includes">
   <rdfs:label xml:lang="en">includes</rdfs:label>
   <rdfs:label xml:lang="pt">Inclui</rdfs:label>
   <rdfs:range rdf:resource="&risk_management;Correction"/>
   <rdfs:domain rdf:resource="&risk_management;Countermeasure"/>
   <rdfs:range rdf:resource="&risk_management;Critical_financial_infraestructure"/>
   <rdfs:range rdf:resource="&risk_management;Detection"/>
   <rdfs:range rdf:resource="&risk_management;Machines"/>
   <rdfs:range rdf:resource="&risk_management;Prevention"/>
   <rdfs:range rdf:resource="&risk_management;Protection"/>
   <rdfs:domain rdf:resource="&risk_management;Response_team"/>
   <rdfs:range rdf:resource="&risk_management;Response_team"/>
</owl:ObjectProperty>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#isConductedBy -->
<owl:ObjectProperty rdf:about="&risk_management;isConductedBy">
   <rdfs:label xml:lang="en">is conducted by</rdfs:label>
   <rdfs:label xml:lang="pt">é conduzido por</rdfs:label>
   <rdfs:range rdf:resource="&risk_management;Risk"/>
   <rdfs:domain rdf:resource="&risk_management;Vulnerability"/>
   <owl:inverseOf rdf:resource="&risk_management;leadingTo"/>
</owl:ObjectProperty>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#leadingTo -->
<owl:ObjectProperty rdf:about="&risk_management;leadingTo">
   <rdfs:label xml:lang="en">leading to</rdfs:label>
   <rdfs:label xml:lang="pt">conduz a</rdfs:label>
   <rdfs:range rdf:resource="&risk_management;Loss"/>
   <rdfs:range rdf:resource="&risk_management;Risk"/>
   <rdfs:domain rdf:resource="&risk_management;Unfortunate_event"/>
   <rdfs:domain rdf:resource="&risk_management;Vulnerability"/>
```

```
</owl:ObjectProperty>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#materializeAs -->
<owl:ObjectProperty rdf:about="&risk_management;materializeAs">
   <rdfs:label xml:lang="en">materialize as</rdfs:label>
   <rdfs:label xml:lang="pt">materializar como</rdfs:label>
   <rdfs:range rdf:resource="&risk_management;Loss"/>
   <rdfs:domain rdf:resource="&risk_management;Risk"/>
   <rdfs:domain rdf:resource="&risk_management;Threat"/>
   <rdfs:range rdf:resource="&risk_management;Unfortunate_event"/>
</owl:ObjectProperty>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#mayBeAwareOf -->
<owl:ObjectProperty rdf:about="&risk_management;mayBeAwareOf">
   <rdfs:label xml:lang="en">may be aware of</rdfs:label>
   <rdfs:label xml:lang="pt">estar ciente de </rdfs:label>
   <rdfs:domain rdf:resource="&risk_management;&apos;Agente_de_defesa&apos;"/>
   <rdfs:range rdf:resource="&risk_management;Vulnerability"/>
</owl:ObjectProperty>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#mayDamage -->
<owl:ObjectProperty rdf:about="&risk_management;mayDamage">
   <rdfs:label xml:lang="en">may damage</rdfs:label>
   <rdfs:label xml:lang="pt">pode danificar</rdfs:label>
   <rdfs:range rdf:resource="&risk_management;Asset"/>
   <rdfs:domain rdf:resource="&risk_management;Threat_agent"/>
</owl:ObjectProperty>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#on -->
<owl:ObjectProperty rdf:about="&risk_management;on">
   <rdfs:label xml:lang="en">on</rdfs:label>
   <rdfs:label xml:lang="pt">sobre</rdfs:label>
   <rdfs:range rdf:resource="&risk_management;Asset"/>
   <rdfs:domain rdf:resource="&risk_management;Loss"/>
   <rdfs:domain rdf:resource="&risk_management;Threat"/>
</owl:ObjectProperty>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#pode_causar_danos_a -->
<owl:ObjectProperty rdf:about="&risk_management;pode_causar_danos_a">
   <rdfs:comment xml:lang="pt">pode causar danos a</rdfs:comment>
   <rdfs:range rdf:resource="&risk_management;Asset"/>
   <rdfs:domain rdf:resource="&risk_management;Threat_agent"/>
</owl:ObjectProperty>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#thatExploits -->
<owl:ObjectProperty rdf:about="&risk_management;thatExploits">
   <rdfs:label xml:lang="en">that exploits</rdfs:label>
   <rdfs:label xml:lang="pt">explora</rdfs:label>
   <rdfs:domain rdf:resource="&risk_management;Threat"/>
   <rdfs:range rdf:resource="&risk_management;Vulnerability"/>
</owl:ObjectProperty>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#thatIncreases -->
<owl:ObjectProperty rdf:about="&risk_management;thatIncreases">
   <rdfs:label xml:lang="en">that increases</rdfs:label>
   <rdfs:label xml:lang="pt">aumenta</rdfs:label>
   <rdfs:range rdf:resource="&risk_management;Risk"/>
```

```
    <rdfs:domain rdf:resource="&risk_management;Threat"/>
</owl:ObjectProperty>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#thatMayBeReducedBy -->
<owl:ObjectProperty rdf:about="&risk_management;thatMayBeReducedBy">
    <rdfs:label xml:lang="en">that may be reduced by</rdfs:label>
    <rdfs:label xml:lang="pt">pode ser reduzido por</rdfs:label>
    <rdfs:range rdf:resource="&risk_management;Countermeasure"/>
    <rdfs:range rdf:resource="&risk_management;Defesa"/>
    <rdfs:domain rdf:resource="&risk_management;Vulnerability"/>
    <owl:inverseOf rdf:resource="&risk_management;thatMayPosess"/>
</owl:ObjectProperty>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#thatMayPosess -->
<owl:ObjectProperty rdf:about="&risk_management;thatMayPosess">
    <rdfs:label xml:lang="en">that may posess</rdfs:label>
    <rdfs:label xml:lang="pt">pode possuir</rdfs:label>
    <rdfs:domain rdf:resource="&risk_management;Defesa"/>
    <rdfs:range rdf:resource="&risk_management;Vulnerability"/>
</owl:ObjectProperty>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#to -->
<owl:ObjectProperty rdf:about="&risk_management;to">
    <rdfs:label xml:lang="en">to</rdfs:label>
    <rdfs:label xml:lang="pt">para</rdfs:label>
    <rdfs:range rdf:resource="&risk_management;Asset"/>
    <rdfs:domain rdf:resource="&risk_management;Risk"/>
</owl:ObjectProperty>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#toReduce -->
<owl:ObjectProperty rdf:about="&risk_management;toReduce">
    <rdfs:label xml:lang="en">to reduce</rdfs:label>
    <rdfs:label xml:lang="pt">para reduzir</rdfs:label>
    <rdfs:domain rdf:resource="&risk_management;Countermeasure"/>
    <rdfs:range rdf:resource="&risk_management;Risk"/>
</owl:ObjectProperty>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#values -->
<owl:ObjectProperty rdf:about="&risk_management;values">
    <rdfs:label xml:lang="en">values</rdfs:label>
    <rdfs:label xml:lang="pt">valoriza</rdfs:label>
    <rdfs:domain rdf:resource="&risk_management;&apos;Agente_de_defesa&apos;"/>
    <rdfs:range rdf:resource="&risk_management;Asset"/>
</owl:ObjectProperty>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#wishesToMinimize -->
<owl:ObjectProperty rdf:about="&risk_management;wishesToMinimize">
    <rdfs:label xml:lang="en">wishes to minimize</rdfs:label>
    <rdfs:label xml:lang="pt">deseja diminuir</rdfs:label>
    <rdfs:domain rdf:resource="&risk_management;&apos;Agente_de_defesa&apos;"/>
    <rdfs:range rdf:resource="&risk_management;Risk"/>
</owl:ObjectProperty>

<!--
///////////////////////////////////////////////////////////////////////////////////////
//
// Data properties
```

```
//
///////////////////////////////////////////////////////////////////////////////
 -->

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#ISO3166_CountryCodes -->
<owl:DatatypeProperty rdf:about="&risk_management;ISO3166_CountryCodes">
   <rdfs:label xml:lang="pt">ISO3166 código de país</rdfs:label>
   <rdfs:comment xml:lang="pt">relaciona os códigos de países internacionalmente definidos pela
        ISO 3166 que é um conjunto de três normas geográficas para codificar nomes de países e
        dependências, e das suas principais subdivisões administrativas</rdfs:comment>
   <rdfs:subPropertyOf rdf:resource="&risk_management;internationalCodes"/>
</owl:DatatypeProperty>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#ISOCurrency -->
<owl:DatatypeProperty rdf:about="&risk_management;ISOCurrency">
   <rdfs:label xml:lang="pt">ISO4217 moeda</rdfs:label>
   <rdfs:comment xml:lang="pt">Identifica as moedas de cada paísde acordo com a ISO4217-Currency
         codes que é o padrão internaciona para códigos de moedas. Ex: o dólar americano é
        representado por USD, o real brasileiro por BRL.</rdfs:comment>
   <rdfs:subPropertyOf rdf:resource="&risk_management;internationalCodes"/>
</owl:DatatypeProperty>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#abstract -->
<owl:DatatypeProperty rdf:about="&risk_management;abstract">
   <rdfs:label xml:lang="pt">resumo</rdfs:label>
   <rdfs:range rdf:resource="&xsd;string"/>
</owl:DatatypeProperty>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#areaServed -->
<owl:DatatypeProperty rdf:about="&risk_management;areaServed">
   <rdfs:label xml:lang="pt">localidade de atuação</rdfs:label>
   <rdfs:subPropertyOf rdf:resource="&risk_management;institutionData"/>
   <rdfs:range rdf:resource="&xsd;string"/>
</owl:DatatypeProperty>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#dateOfBirth -->
<owl:DatatypeProperty rdf:about="&risk_management;dateOfBirth">
   <rdfs:label xml:lang="pt">data de nascimento</rdfs:label>
   <rdfs:subPropertyOf rdf:resource="&risk_management;personData"/>
   <rdfs:range rdf:resource="&xsd;dateTime"/>
</owl:DatatypeProperty>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#dateOfDeath -->
<owl:DatatypeProperty rdf:about="&risk_management;dateOfDeath">
   <rdfs:label xml:lang="pt">data de morte</rdfs:label>
   <rdfs:subPropertyOf rdf:resource="&risk_management;personData"/>
   <rdfs:range rdf:resource="&xsd;dateTime"/>
</owl:DatatypeProperty>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#description -->
<owl:DatatypeProperty rdf:about="&risk_management;description">
   <rdfs:label xml:lang="pt">descrição</rdfs:label>
   <rdfs:range rdf:resource="&xsd;string"/>
   <rdfs:subPropertyOf rdf:resource="&owl;topDataProperty"/>
</owl:DatatypeProperty>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#foundationDate -->
```

```
<owl:DatatypeProperty rdf:about="&risk_management;foundationDate">
   <rdfs:label xml:lang="pt">data de fundação</rdfs:label>
   <rdfs:subPropertyOf rdf:resource="&risk_management;institutionData"/>
   <rdfs:range rdf:resource="&xsd;dateTime"/>
</owl:DatatypeProperty>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#foundationPlace -->
<owl:DatatypeProperty rdf:about="&risk_management;foundationPlace">
   <rdfs:label xml:lang="pt">lugar de fundação</rdfs:label>
   <rdfs:subPropertyOf rdf:resource="&risk_management;institutionData"/>
   <rdfs:range rdf:resource="&xsd;string"/>
</owl:DatatypeProperty>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#foundedBy -->
<owl:DatatypeProperty rdf:about="&risk_management;foundedBy">
   <rdfs:label xml:lang="pt">fundado por</rdfs:label>
   <rdfs:subPropertyOf rdf:resource="&risk_management;institutionData"/>
   <rdfs:range rdf:resource="&xsd;string"/>
</owl:DatatypeProperty>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#hasAddress -->
<owl:DatatypeProperty rdf:about="&risk_management;hasAddress">
   <rdfs:label xml:lang="pt">tem endereço</rdfs:label>
   <rdfs:domain rdf:resource="&risk_management;Person"/>
   <rdfs:domain rdf:resource="&risk_management;business"/>
   <rdfs:range rdf:resource="&rdf;PlainLiteral"/>
</owl:DatatypeProperty>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#hasName -->
<owl:DatatypeProperty rdf:about="&risk_management;hasName">
   <rdf:type rdf:resource="&owl;FunctionalProperty"/>
   <rdfs:label xml:lang="pt">tem nome</rdfs:label>
   <rdfs:domain rdf:resource="&risk_management;Person"/>
   <rdfs:domain rdf:resource="&risk_management;business"/>
   <rdfs:range rdf:resource="&rdf;PlainLiteral"/>
</owl:DatatypeProperty>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#industry -->
<owl:DatatypeProperty rdf:about="&risk_management;industry">
   <rdfs:label xml:lang="pt">indústria</rdfs:label>
   <rdfs:subPropertyOf rdf:resource="&risk_management;institutionData"/>
   <rdfs:range rdf:resource="&xsd;string"/>
</owl:DatatypeProperty>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#institutionData -->
<owl:DatatypeProperty rdf:about="&risk_management;institutionData">
   <rdfs:label xml:lang="pt">dados institucionais</rdfs:label>
   <rdfs:domain rdf:resource="&risk_management;business"/>
</owl:DatatypeProperty>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#internationalCodes -->
<owl:DatatypeProperty rdf:about="&risk_management;internationalCodes">
   <rdfs:label xml:lang="pt">códigos internacionais</rdfs:label>
   <rdfs:comment xml:lang="pt">relaciona aos códigos padronizados internacionalmente</rdfs:
        comment>
</owl:DatatypeProperty>
```

```
<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#keyPerson -->
<owl:DatatypeProperty rdf:about="&risk_management;keyPerson">
   <rdfs:label xml:lang="pt">pessoa-chave</rdfs:label>
   <rdfs:subPropertyOf rdf:resource="&risk_management;institutionData"/>
   <rdfs:range rdf:resource="&xsd;string"/>
</owl:DatatypeProperty>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#local_de_morte -->
<owl:DatatypeProperty rdf:about="&risk_management;local_de_morte">
   <rdfs:subPropertyOf rdf:resource="&risk_management;personData"/>
   <rdfs:range rdf:resource="&xsd;string"/>
</owl:DatatypeProperty>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#locationCity -->
<owl:DatatypeProperty rdf:about="&risk_management;locationCity">
   <rdfs:label xml:lang="pt">cidade de localização</rdfs:label>
   <rdfs:subPropertyOf rdf:resource="&risk_management;institutionData"/>
   <rdfs:range rdf:resource="&xsd;string"/>
</owl:DatatypeProperty>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#netIncome -->
<owl:DatatypeProperty rdf:about="&risk_management;netIncome">
   <rdfs:label xml:lang="pt">receita líquida</rdfs:label>
   <rdfs:subPropertyOf rdf:resource="&risk_management;institutionData"/>
   <rdfs:range rdf:resource="&xsd;int"/>
</owl:DatatypeProperty>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#numEmployees -->
<owl:DatatypeProperty rdf:about="&risk_management;numEmployees">
   <rdfs:label xml:lang="pt">número de empregados</rdfs:label>
   <rdfs:subPropertyOf rdf:resource="&risk_management;institutionData"/>
   <rdfs:range rdf:resource="&xsd;int"/>
</owl:DatatypeProperty>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#numberOfLocations -->
<owl:DatatypeProperty rdf:about="&risk_management;numberOfLocations">
   <rdfs:label xml:lang="pt">número de localidades</rdfs:label>
   <rdfs:subPropertyOf rdf:resource="&risk_management;institutionData"/>
   <rdfs:range rdf:resource="&xsd;int"/>
</owl:DatatypeProperty>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#personData -->
<owl:DatatypeProperty rdf:about="&risk_management;personData">
   <rdfs:label xml:lang="pt">dados pessoais</rdfs:label>
   <rdfs:comment xml:lang="pt">relaciona os dados pessoais de pessoas</rdfs:comment>
   <rdfs:domain rdf:resource="&risk_management;Person"/>
   <rdfs:subPropertyOf rdf:resource="&owl;topDataProperty"/>
</owl:DatatypeProperty>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#placeOfBirth -->
<owl:DatatypeProperty rdf:about="&risk_management;placeOfBirth">
   <rdfs:label xml:lang="pt">local de nascimento</rdfs:label>
   <rdfs:subPropertyOf rdf:resource="&risk_management;personData"/>
   <rdfs:range rdf:resource="&xsd;string"/>
</owl:DatatypeProperty>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#position -->
```

```
<owl:DatatypeProperty rdf:about="&risk_management;position">
   <rdfs:label xml:lang="pt">cargo</rdfs:label>
   <rdfs:subPropertyOf rdf:resource="&risk_management;personData"/>
   <rdfs:range rdf:resource="&xsd;string"/>
</owl:DatatypeProperty>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#shortDescription -->
<owl:DatatypeProperty rdf:about="&risk_management;shortDescription">
   <rdfs:label xml:lang="pt">descrição resumida</rdfs:label>
   <rdfs:range rdf:resource="&xsd;string"/>
</owl:DatatypeProperty>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#symbol -->
<owl:DatatypeProperty rdf:about="&risk_management;symbol">
   <rdfs:label xml:lang="pt">símbolo</rdfs:label>
   <rdfs:subPropertyOf rdf:resource="&risk_management;institutionData"/>
   <rdfs:range rdf:resource="&xsd;string"/>
</owl:DatatypeProperty>

<!--
///////////////////////////////////////////////////////////////////////////////////
//
// Classes
//
///////////////////////////////////////////////////////////////////////////////////
 -->

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#&apos;Agente_de_defesa&apos;
     -->
<owl:Class rdf:about="&risk_management;&apos;Agente_de_defesa&apos;">
   <rdfs:subClassOf rdf:resource="&risk_management;Agente"/>
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#&apos;Pessoa_abstrata&apos; -->

<owl:Class rdf:about="&risk_management;&apos;Pessoa_abstrata&apos;">
   <rdfs:label xml:lang="pt">Pessoa Abstrata</rdfs:label>
   <rdfs:subClassOf rdf:resource="&risk_management;Agente"/>
   <owl:disjointWith rdf:resource="&risk_management;Person"/>
   <owl:disjointWith rdf:resource="&risk_management;business"/>
   <rdfs:comment xml:lang="pt">Ente que não é classificado como pessoa física ou jurídica, como
       programas de computador.</rdfs:comment>
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Academia -->
<owl:Class rdf:about="&risk_management;Academia">
   <rdfs:label xml:lang="en">Academia</rdfs:label>
   <rdfs:label xml:lang="pt">Academia</rdfs:label>
   <rdfs:subClassOf rdf:resource="&risk_management;Reasearch"/>
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Acionistas -->
<owl:Class rdf:about="&risk_management;Acionistas">
   <rdfs:subClassOf rdf:resource="&risk_management;Shareholder"/>
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Agente -->
```

```
<owl:Class rdf:about="&risk_management;Agente">
   <rdfs:comment xml:lang="pt">Entidade que engloba os tipos de pessoas: física, jurídica e
       abstrata.</rdfs:comment>
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Artigo -->
<owl:Class rdf:about="&risk_management;Artigo">
   <owl:equivalentClass rdf:resource="&risk_management;Product"/>
   <rdfs:subClassOf rdf:resource="&risk_management;Asset"/>
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Assalto -->
<owl:Class rdf:about="&risk_management;Assalto">
   <rdfs:subClassOf rdf:resource="&risk_management;Crime"/>
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Asset -->
<owl:Class rdf:about="&risk_management;Asset">
   <rdfs:label xml:lang="en">Asset</rdfs:label>
   <rdfs:label xml:lang="pt">Ativo</rdfs:label>
   <owl:equivalentClass rdf:resource="&risk_management;Bem"/>
   <owl:equivalentClass rdf:resource="&risk_management;Posse"/>
   <owl:equivalentClass rdf:resource="&risk_management;Propriedade"/>
   <owl:equivalentClass rdf:resource="&risk_management;Recurso"/>
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Atentado -->
<owl:Class rdf:about="&risk_management;Atentado">
   <rdfs:subClassOf rdf:resource="&risk_management;Unfortunate_event"/>
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Attack -->
<owl:Class rdf:about="&risk_management;Attack">
   <rdfs:label xml:lang="en">Attack</rdfs:label>
   <rdfs:label xml:lang="pt">Ataque</rdfs:label>
   <rdfs:subClassOf rdf:resource="&risk_management;Crime"/>
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Banco -->
<owl:Class rdf:about="&risk_management;Banco">
   <rdfs:subClassOf rdf:resource="&risk_management;Regulator"/>
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Behavioral_method -->
<owl:Class rdf:about="&risk_management;Behavioral_method">
   <rdfs:label xml:lang="en">Behavioral method</rdfs:label>
   <rdfs:label xml:lang="pt">Método comportamental</rdfs:label>
   <rdfs:subClassOf rdf:resource="&risk_management;Rule"/>
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Bem -->
<owl:Class rdf:about="&risk_management;Bem"/>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Black_swan -->
<owl:Class rdf:about="&risk_management;Black_swan">
   <rdfs:label xml:lang="en">Black Swan</rdfs:label>
   <rdfs:label xml:lang="pt">Evento inesperado</rdfs:label>
```

```
    <rdfs:subClassOf rdf:resource="&risk_management;Unfortunate_event"/>
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Central_Bank -->
<owl:Class rdf:about="&risk_management;Central_Bank">
    <rdfs:label xml:lang="en">Central Bank</rdfs:label>
    <rdfs:label xml:lang="pt">Banco Central</rdfs:label>
    <rdfs:subClassOf rdf:resource="&risk_management;Banco"/>
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#
     Comitê_Gestor_de_Risco_Financeiro -->
<owl:Class rdf:about="&risk_management;Comitê_Gestor_de_Risco_Financeiro">
    <rdfs:label xml:lang="pt">Comitê Gestor de Risco Financeiro</rdfs:label>
    <rdfs:subClassOf rdf:resource="&risk_management;Regulator"/>
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Concentration -->
<owl:Class rdf:about="&risk_management;Concentration">
    <rdfs:label xml:lang="en">Concentration</rdfs:label>
    <rdfs:label xml:lang="pt">Concentração</rdfs:label>
    <rdfs:subClassOf rdf:resource="&risk_management;Risk_factor"/>
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Consultant -->
<owl:Class rdf:about="&risk_management;Consultant">
    <rdfs:label xml:lang="en">Consultant</rdfs:label>
    <rdfs:label xml:lang="pt">Consultoria</rdfs:label>
    <rdfs:subClassOf rdf:resource="&risk_management;Reasearch"/>
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Correction -->
<owl:Class rdf:about="&risk_management;Correction">
    <rdfs:label xml:lang="en">Correction</rdfs:label>
    <rdfs:label xml:lang="pt">Correção</rdfs:label>
    <rdfs:subClassOf rdf:resource="&risk_management;Risk_mitigation"/>
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Countermeasure -->
<owl:Class rdf:about="&risk_management;Countermeasure">
    <rdfs:label xml:lang="en">Countermeasures</rdfs:label>
    <rdfs:label xml:lang="pt">Contramedidas</rdfs:label>
    <rdfs:subClassOf rdf:resource="&risk_management;Defesa"/>
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Counterparty_risk -->
<owl:Class rdf:about="&risk_management;Counterparty_risk">
    <rdfs:label xml:lang="en">Counterparty_risk</rdfs:label>
    <rdfs:label xml:lang="pt">Risco de contraparte</rdfs:label>
    <rdfs:subClassOf rdf:resource="&risk_management;Risk"/>
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Credit_risk -->
<owl:Class rdf:about="&risk_management;Credit_risk">
    <rdfs:label xml:lang="en">Credit risk</rdfs:label>
    <rdfs:label xml:lang="pt">Risco de crédito</rdfs:label>
    <rdfs:subClassOf rdf:resource="&risk_management;Risk"/>
```

```
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Crime -->
<owl:Class rdf:about="&risk_management;Crime">
   <owl:equivalentClass rdf:resource="&risk_management;Violação"/>
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#
     Critical_financial_infraestructure -->
<owl:Class rdf:about="&risk_management;Critical_financial_infraestructure">
   <rdfs:label xml:lang="en">Critical financial infraestructure</rdfs:label>
   <rdfs:label xml:lang="pt">Infraestrutura crítica financeira</rdfs:label>
   <rdfs:subClassOf rdf:resource="&risk_management;Infrastructure"/>
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Crititical_market_participant
     -->
<owl:Class rdf:about="&risk_management;Crititical_market_participant">
   <rdfs:label xml:lang="en">Critical market participant</rdfs:label>
   <rdfs:label xml:lang="pt">Participante de mercado crítico</rdfs:label>
   <rdfs:subClassOf rdf:resource="&risk_management;Participant"/>
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Default -->
<owl:Class rdf:about="&risk_management;Default">
   <rdfs:label xml:lang="en">Default</rdfs:label>
   <rdfs:label xml:lang="pt">Default</rdfs:label>
   <rdfs:subClassOf rdf:resource="&risk_management;Unfortunate_event"/>
   <rdfs:comment xml:lang="pt">É o descumprimento de qualquer cláusula importante de um contrato
        que vincula devedor e credor. Calote</rdfs:comment>
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Defesa -->
<owl:Class rdf:about="&risk_management;Defesa"/>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Detection -->
<owl:Class rdf:about="&risk_management;Detection">
   <rdfs:label xml:lang="en">Detection</rdfs:label>
   <rdfs:label xml:lang="pt">Detecção</rdfs:label>
   <rdfs:subClassOf rdf:resource="&risk_management;Risk_mitigation"/>
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Directive -->
<owl:Class rdf:about="&risk_management;Directive">
   <rdfs:label xml:lang="en">Directive</rdfs:label>
   <rdfs:label xml:lang="pt">Diretriz</rdfs:label>
   <rdfs:subClassOf rdf:resource="&risk_management;Rule"/>
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Dono -->
<owl:Class rdf:about="&risk_management;Dono">
   <rdfs:subClassOf rdf:resource="&risk_management;Shareholder"/>
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Empregados -->
<owl:Class rdf:about="&risk_management;Empregados">
   <rdfs:subClassOf rdf:resource="&risk_management;Person"/>
```

```
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Environmental_and_social_risk
     -->
<owl:Class rdf:about="&risk_management;Environmental_and_social_risk">
   <rdfs:label xml:lang="en">Environmental and social risk</rdfs:label>
   <rdfs:label xml:lang="pt">Risco social e ambiental</rdfs:label>
   <rdfs:subClassOf rdf:resource="&risk_management;Risk"/>
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Equipamento -->
<owl:Class rdf:about="&risk_management;Equipamento">
   <rdfs:label xml:lang="en">Hardware</rdfs:label>
   <rdfs:label xml:lang="pt">Equipamento</rdfs:label>
   <rdfs:subClassOf rdf:resource="&risk_management;&apos;Pessoa_abstrata&apos;"/>
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Exposure -->
<owl:Class rdf:about="&risk_management;Exposure">
   <rdfs:label xml:lang="pt">Exposição</rdfs:label>
   <rdfs:subClassOf rdf:resource="&risk_management;Credit_risk"/>
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Failed_transaction -->
<owl:Class rdf:about="&risk_management;Failed_transaction">
   <rdfs:label xml:lang="en">Failed transaction</rdfs:label>
   <rdfs:label xml:lang="pt">Falha em transação </rdfs:label>
   <rdfs:subClassOf rdf:resource="&risk_management;Unfortunate_event"/>
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Falsificação -->
<owl:Class rdf:about="&risk_management;Falsificação">
   <rdfs:subClassOf rdf:resource="&risk_management;Crime"/>
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Fraqueza -->
<owl:Class rdf:about="&risk_management;Fraqueza">
   <rdfs:subClassOf rdf:resource="&risk_management;Risk_factor"/>
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Fraud -->
<owl:Class rdf:about="&risk_management;Fraud">
   <rdfs:label xml:lang="en">Fraud</rdfs:label>
   <rdfs:label xml:lang="pt">Fraude</rdfs:label>
   <rdfs:subClassOf rdf:resource="&risk_management;Unfortunate_event"/>
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Functional_leaders -->
<owl:Class rdf:about="&risk_management;Functional_leaders">
   <rdfs:label xml:lang="en">Functional leaders</rdfs:label>
   <rdfs:label xml:lang="pt">Líderes funcionais</rdfs:label>
   <rdfs:subClassOf rdf:resource="&risk_management;Participant"/>
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Hazard -->
<owl:Class rdf:about="&risk_management;Hazard">
   <rdfs:label xml:lang="en">Hazard</rdfs:label>
```

```
    <rdfs:label xml:lang="pt">Desastre</rdfs:label>
    <rdfs:subClassOf rdf:resource="&risk_management;Unfortunate_event"/>
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Industry_leader -->
<owl:Class rdf:about="&risk_management;Industry_leader">
    <rdfs:label xml:lang="en">Industry leader</rdfs:label>
    <rdfs:label xml:lang="pt">Líder da indústria </rdfs:label>
    <rdfs:subClassOf rdf:resource="&risk_management;Regulator"/>
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Infrastructure -->
<owl:Class rdf:about="&risk_management;Infrastructure">
    <rdfs:label xml:lang="en">Infrastruture</rdfs:label>
    <rdfs:label xml:lang="pt">Infraestrutura</rdfs:label>
    <rdfs:subClassOf rdf:resource="&risk_management;Asset"/>
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Infração -->
<owl:Class rdf:about="&risk_management;Infração">
    <rdfs:subClassOf rdf:resource="&risk_management;Crime"/>
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Instituição_governamental -->
<owl:Class rdf:about="&risk_management;Instituição_governamental">
    <rdfs:label xml:lang="pt">Instituição Governamental</rdfs:label>
    <rdfs:subClassOf rdf:resource="&risk_management;Regulator"/>
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Legal_risk -->
<owl:Class rdf:about="&risk_management;Legal_risk">
    <rdfs:label xml:lang="en">Legal risk</rdfs:label>
    <rdfs:label xml:lang="pt">Risco legal</rdfs:label>
    <rdfs:subClassOf rdf:resource="&risk_management;Risk"/>
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Limit -->
<owl:Class rdf:about="&risk_management;Limit">
    <rdfs:label xml:lang="en">Limit</rdfs:label>
    <rdfs:label xml:lang="pt">Limite</rdfs:label>
    <rdfs:subClassOf rdf:resource="&risk_management;Rule"/>
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Liquidity_risk -->
<owl:Class rdf:about="&risk_management;Liquidity_risk">
    <rdfs:label xml:lang="en">Liquidity risk</rdfs:label>
    <rdfs:label xml:lang="pt">Risco de liquidez</rdfs:label>
    <rdfs:subClassOf rdf:resource="&risk_management;Risk"/>
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Loss -->
<owl:Class rdf:about="&risk_management;Loss">
    <rdfs:label xml:lang="en">Loss</rdfs:label>
    <rdfs:label xml:lang="pt">Perda</rdfs:label>
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Machines -->
```

```
<owl:Class rdf:about="&risk_management;Machines">
   <rdfs:label xml:lang="en">Mashine</rdfs:label>
   <rdfs:label xml:lang="pt">Máquina</rdfs:label>
   <rdfs:subClassOf rdf:resource="&risk_management;Infrastructure"/>
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Major_operational_disruption
     -->
<owl:Class rdf:about="&risk_management;Major_operational_disruption">
   <rdfs:label xml:lang="en">Major operational disruption</rdfs:label>
   <rdfs:label xml:lang="pt">Interrupção de operaçao significativa</rdfs:label>
   <rdfs:subClassOf rdf:resource="&risk_management;Unfortunate_event"/>
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Operational_risk -->
<owl:Class rdf:about="&risk_management;Operational_risk">
   <rdfs:label xml:lang="en">Operational risk</rdfs:label>
   <rdfs:label xml:lang="pt">Risco operacional</rdfs:label>
   <rdfs:subClassOf rdf:resource="&risk_management;Risk"/>
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Organization -->
<owl:Class rdf:about="&risk_management;Organization">
   <rdfs:label xml:lang="en">Organization</rdfs:label>
   <rdfs:label xml:lang="pt">Organização</rdfs:label>
   <rdfs:subClassOf rdf:resource="&risk_management;business"/>
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Participant -->
<owl:Class rdf:about="&risk_management;Participant">
   <rdfs:label xml:lang="en">Participant</rdfs:label>
   <rdfs:label xml:lang="pt">Participante</rdfs:label>
   <rdfs:subClassOf rdf:resource="&risk_management;Stakeholder"/>
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Perigo -->
<owl:Class rdf:about="&risk_management;Perigo">
   <rdfs:label xml:lang="en">Danger</rdfs:label>
   <rdfs:label xml:lang="pt">Perigo</rdfs:label>
   <owl:equivalentClass rdf:resource="&risk_management;Threat"/>
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Person -->
<owl:Class rdf:about="&risk_management;Person">
   <rdfs:label xml:lang="pt">Pessoa física</rdfs:label>
   <rdfs:subClassOf rdf:resource="&risk_management;Agente"/>
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Political_risk -->
<owl:Class rdf:about="&risk_management;Political_risk">
   <rdfs:label xml:lang="en">Political risk</rdfs:label>
   <rdfs:label xml:lang="pt">Risco político</rdfs:label>
   <rdfs:subClassOf rdf:resource="&risk_management;Risk"/>
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Posse -->
<owl:Class rdf:about="&risk_management;Posse"/>
```

```
<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Prevention -->
<owl:Class rdf:about="&risk_management;Prevention">
   <rdfs:label xml:lang="en">Prevention</rdfs:label>
   <rdfs:label xml:lang="pt">Prevenção</rdfs:label>
   <rdfs:subClassOf rdf:resource="&risk_management;Risk_mitigation"/>
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Principal_risk -->
<owl:Class rdf:about="&risk_management;Principal_risk">
   <rdfs:label xml:lang="en">Principal risk</rdfs:label>
   <rdfs:label xml:lang="pt">Risco principal</rdfs:label>
   <rdfs:subClassOf rdf:resource="&risk_management;Risk"/>
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Product -->
<owl:Class rdf:about="&risk_management;Product">
   <rdfs:label xml:lang="en">Product</rdfs:label>
   <rdfs:label xml:lang="pt">Produto</rdfs:label>
   <rdfs:subClassOf rdf:resource="&risk_management;Asset"/>
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Programa_de_computador -->
<owl:Class rdf:about="&risk_management;Programa_de_computador">
   <rdfs:label xml:lang="en">Software</rdfs:label>
   <rdfs:label xml:lang="pt">Programa de computador</rdfs:label>
   <rdfs:subClassOf rdf:resource="&risk_management;&apos;Pessoa_abstrata&apos;"/>
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Propriedade -->
<owl:Class rdf:about="&risk_management;Propriedade"/>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Proprietarios -->
<owl:Class rdf:about="&risk_management;Proprietarios">
   <rdfs:subClassOf rdf:resource="&risk_management;Person"/>
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Protection -->
<owl:Class rdf:about="&risk_management;Protection">
   <rdfs:label xml:lang="en">Protection</rdfs:label>
   <rdfs:label xml:lang="pt">Proteção</rdfs:label>
   <rdfs:subClassOf rdf:resource="&risk_management;Risk_mitigation"/>
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Rapto -->
<owl:Class rdf:about="&risk_management;Rapto">
   <rdfs:subClassOf rdf:resource="&risk_management;Crime"/>
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Reasearch -->
<owl:Class rdf:about="&risk_management;Reasearch">
   <rdfs:label xml:lang="en">Research</rdfs:label>
   <rdfs:label xml:lang="pt">Pesquisa</rdfs:label>
   <rdfs:subClassOf rdf:resource="&risk_management;Service"/>
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Recurso -->
```

```
<owl:Class rdf:about="&risk_management;Recurso"/>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Regulator -->
<owl:Class rdf:about="&risk_management;Regulator">
   <rdfs:label xml:lang="en">Regulator</rdfs:label>
   <rdfs:label xml:lang="pt">Órgão Regulador</rdfs:label>
   <rdfs:subClassOf rdf:resource="&risk_management;business"/>
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Relying_party -->
<owl:Class rdf:about="&risk_management;Relying_party">
   <rdfs:label xml:lang="en">Relying party</rdfs:label>
   <rdfs:label xml:lang="pt">Parte confiável</rdfs:label>
   <rdfs:subClassOf rdf:resource="&risk_management;Participant"/>
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Resilience -->
<owl:Class rdf:about="&risk_management;Resilience">
   <rdfs:label xml:lang="en">Resilience</rdfs:label>
   <rdfs:label xml:lang="pt">Resiliência</rdfs:label>
   <rdfs:subClassOf rdf:resource="&risk_management;Risk_factor"/>
   <rdfs:comment xml:lang="pt">mesmo que meta ou objetivo</rdfs:comment>
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Response_team -->
<owl:Class rdf:about="&risk_management;Response_team">
   <rdfs:label xml:lang="en">Response team</rdfs:label>
   <rdfs:label xml:lang="pt">Grupo de resposta</rdfs:label>
   <rdfs:subClassOf rdf:resource="&risk_management;Defesa"/>
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Risk -->
<owl:Class rdf:about="&risk_management;Risk">
   <rdfs:label xml:lang="en">Risk</rdfs:label>
   <rdfs:label xml:lang="pt">Risco</rdfs:label>
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Risk_factor -->
<owl:Class rdf:about="&risk_management;Risk_factor">
   <rdfs:label xml:lang="en">Risk factor</rdfs:label>
   <rdfs:label xml:lang="pt">Fator de risco</rdfs:label>
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Risk_mitigation -->
<owl:Class rdf:about="&risk_management;Risk_mitigation">
   <rdfs:label xml:lang="en">Risk mitigation</rdfs:label>
   <rdfs:label xml:lang="pt">Mitigação do risco</rdfs:label>
   <rdfs:subClassOf rdf:resource="&risk_management;Countermeasure"/>
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Rule -->
<owl:Class rdf:about="&risk_management;Rule">
   <rdfs:label xml:lang="en">Rule</rdfs:label>
   <rdfs:label xml:lang="pt">Regra</rdfs:label>
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Sectorial_risk -->
```

```
<owl:Class rdf:about="&risk_management;Sectorial_risk">
   <rdfs:label xml:lang="en">Sectorial risk</rdfs:label>
   <rdfs:label xml:lang="pt">Risco setorial</rdfs:label>
   <rdfs:subClassOf rdf:resource="&risk_management;Risk"/>
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Service -->
<owl:Class rdf:about="&risk_management;Service">
   <rdfs:label xml:lang="en">Service</rdfs:label>
   <rdfs:label xml:lang="pt">Serviço</rdfs:label>
   <rdfs:subClassOf rdf:resource="&risk_management;Stakeholder"/>
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Service_offering -->
<owl:Class rdf:about="&risk_management;Service_offering">
   <rdfs:label xml:lang="en">Service offering</rdfs:label>
   <rdfs:label xml:lang="pt">Oferta de serviço</rdfs:label>
   <rdfs:subClassOf rdf:resource="&risk_management;Asset"/>
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Settlement_risk -->
<owl:Class rdf:about="&risk_management;Settlement_risk">
   <rdfs:label xml:lang="en">Settlement risk</rdfs:label>
   <rdfs:label xml:lang="pt">Risco de compensação</rdfs:label>
   <rdfs:subClassOf rdf:resource="&risk_management;Operational_risk"/>
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Shareholder -->
<owl:Class rdf:about="&risk_management;Shareholder">
   <rdfs:subClassOf rdf:resource="&risk_management;business"/>
   <rdfs:comment xml:lang="pt">O shareholder significa acionista, ou seja, é uma pessoa que
        possui pelo menos uma ação de uma organização ou empresa.</rdfs:comment>
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Specialist -->
<owl:Class rdf:about="&risk_management;Specialist">
   <rdfs:label xml:lang="en">Specialist</rdfs:label>
   <rdfs:label xml:lang="pt">Especialistas</rdfs:label>
   <rdfs:subClassOf rdf:resource="&risk_management;Participant"/>
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Staff -->
<owl:Class rdf:about="&risk_management;Staff">
   <rdfs:label xml:lang="en">Staff</rdfs:label>
   <rdfs:label xml:lang="pt">Pessoal</rdfs:label>
   <rdfs:subClassOf rdf:resource="&risk_management;Participant"/>
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Stakeholder -->
<owl:Class rdf:about="&risk_management;Stakeholder">
   <rdfs:label xml:lang="en">Stakeholder</rdfs:label>
   <rdfs:label xml:lang="pt">Stakeholder</rdfs:label>
   <rdfs:subClassOf rdf:resource="&risk_management;Person"/>
   <rdfs:comment xml:lang="pt">O stakeholder é uma pessoa ou um grupo, que legitima as ações de
        uma organização e que tem um papel direto ou indireto na gestão e resultados dessa mesma
         organização. É formado pelos funcionários da empresa, gestores, gerentes, proprietários,
         fornecedores, concorrentes, ONGs, clientes, o Estado, credores, sindicatos e diversas
```

```
        outras pessoas ou empresas que estejam relacionadas com uma determinada ação ou projeto
        .</rdfs:comment>
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Supplier -->
<owl:Class rdf:about="&risk_management;Supplier">
   <rdfs:label xml:lang="en">Supplier</rdfs:label>
   <rdfs:label xml:lang="pt">Fornecedor</rdfs:label>
   <rdfs:subClassOf rdf:resource="&risk_management;Service"/>
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Technical_risk -->
<owl:Class rdf:about="&risk_management;Technical_risk">
   <rdfs:label xml:lang="en">Technical risk</rdfs:label>
   <rdfs:label xml:lang="pt">Risco técnico</rdfs:label>
   <rdfs:subClassOf rdf:resource="&risk_management;Risk"/>
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Technology_options -->
<owl:Class rdf:about="&risk_management;Technology_options">
   <rdfs:label xml:lang="en">Technology options</rdfs:label>
   <rdfs:label xml:lang="pt">Opções de tecnologia</rdfs:label>
   <rdfs:subClassOf rdf:resource="&risk_management;Infrastructure"/>
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Theft -->
<owl:Class rdf:about="&risk_management;Theft">
   <rdfs:label xml:lang="en">Theft</rdfs:label>
   <rdfs:label xml:lang="pt">Roubo</rdfs:label>
   <rdfs:subClassOf rdf:resource="&risk_management;Crime"/>
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Threat -->
<owl:Class rdf:about="&risk_management;Threat">
   <rdfs:label xml:lang="en">Threat</rdfs:label>
   <rdfs:label xml:lang="pt">Ameaça</rdfs:label>
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Threat_agent -->
<owl:Class rdf:about="&risk_management;Threat_agent">
   <rdfs:label xml:lang="en">Threat agent</rdfs:label>
   <rdfs:label xml:lang="pt">Agente de ameaça</rdfs:label>
   <rdfs:subClassOf rdf:resource="&risk_management;Agente"/>
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Unfortunate_event -->
<owl:Class rdf:about="&risk_management;Unfortunate_event">
   <rdfs:label xml:lang="en">Unfortunate event</rdfs:label>
   <rdfs:label xml:lang="pt">Sinistro</rdfs:label>
   <rdfs:subClassOf rdf:resource="&risk_management;Crime"/>
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Violação -->
<owl:Class rdf:about="&risk_management;Violação"/>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Vulnerability -->
<owl:Class rdf:about="&risk_management;Vulnerability">
```

```
    <rdfs:label xml:lang="en">Vulnerability</rdfs:label>
    <rdfs:label xml:lang="pt">Vulnerabilidade</rdfs:label>
</owl:Class>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#business -->
<owl:Class rdf:about="&risk_management;business">
    <rdfs:label xml:lang="pt">Pessoa jurídica</rdfs:label>
    <rdfs:subClassOf rdf:resource="&risk_management;Agente"/>
</owl:Class>

<!--
////////////////////////////////////////////////////////////////////////////////
//
// Individuals
//
////////////////////////////////////////////////////////////////////////////////
 -->

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Banco_Central_do_Brasil -->
<owl:NamedIndividual rdf:about="&risk_management;Banco_Central_do_Brasil">
    <rdf:type rdf:resource="&risk_management;Central_Bank"/>
    <rdfs:label xml:lang="pt">Banco Central do Brasil</rdfs:label>
</owl:NamedIndividual>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Banco_do_Brasil -->
<owl:NamedIndividual rdf:about="&risk_management;Banco_do_Brasil">
    <rdf:type rdf:resource="&risk_management;Banco"/>
    <rdfs:label xml:lang="pt">Banco do Brasil</rdfs:label>
</owl:NamedIndividual>

<!-- http://www.semanticweb.org/ontologies/2014/8/risk_management#Caixa_Econômica_Federal -->
<owl:NamedIndividual rdf:about="&risk_management;Caixa_Econômica_Federal">
    <rdf:type rdf:resource="&risk_management;Banco"/>
    <rdfs:label xml:lang="pt">Caixa Econômica Federal</rdfs:label>
    <risk_management:hasName xml:lang="pt">Caixa Econômica Federal</risk_management:hasName>
</owl:NamedIndividual>

<!--
////////////////////////////////////////////////////////////////////////////////
//
// Annotations
//
////////////////////////////////////////////////////////////////////////////////
 -->
<rdf:Description rdf:about="&owl;Thing">
    <rdfs:label xml:lang="en">Risk management</rdfs:label>
    <rdfs:label xml:lang="pt">Gestão de Risco</rdfs:label>
</rdf:Description>

<!--
////////////////////////////////////////////////////////////////////////////////
//
// General axioms
//
////////////////////////////////////////////////////////////////////////////////
 -->
<rdf:Description>
```

```
        <rdf:type rdf:resource="&owl;AllDisjointClasses"/>
        <owl:members rdf:parseType="Collection">
            <rdf:Description rdf:about="&risk_management;Correction"/>
            <rdf:Description rdf:about="&risk_management;Detection"/>
            <rdf:Description rdf:about="&risk_management;Prevention"/>
            <rdf:Description rdf:about="&risk_management;Protection"/>
        </owl:members>
    </rdf:Description>
</rdf:RDF>
<!-- Generated by the OWL API (version 3.4.2) http://owlapi.sourceforge.net -->
```

ANEXO B – Base Lexical (RiscoLex)

```
@base <http://www.example.org/lexico> .
@prefix lemon: <http://lemon-model.net/lemon#> .
@prefix :<http://www.exemplo.org/> .
@prefix ontoRisco:<http://www.semanticweb.org/ontologies/2014/8/risk_management#> .
@prefix infoLex: <http://www.semanticweb.org/ontologies/2014/9/openWordnet-PT#> .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :bem .

:bem a lemon:LexicalEntry ;
    lemon:canonicalForm [ lemon:writtenRep "bem"@pt ] ;
    lemon:otherForm [ lemon:writtenRep "bens"@pt ] ;
    lemon:cat [lemon:partOfSpeech infoLex:noun];
    lemon:sense [ lemon:reference ontoRisco:bem ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :propriedade .

:propriedade a lemon:LexicalEntry ;
    lemon:canonicalForm [ lemon:writtenRep "propriedade"@pt ] ;
    lemon:otherForm [ lemon:writtenRep "propriedades"@pt ] ;
    lemon:cat [lemon:partOfSpeech infoLex:noun];
    lemon:sense [ lemon:reference ontoRisco:bem ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :posse .

:posse a lemon:LexicalEntry ;
    lemon:canonicalForm [ lemon:writtenRep "posse"@pt ] ;
    lemon:otherForm [ lemon:writtenRep "posses"@pt ] ;
    lemon:cat [lemon:partOfSpeech infoLex:noun];
    lemon:sense [ lemon:reference ontoRisco:bem ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :ativo .

:ativo a lemon:LexicalEntry ;
    lemon:canonicalForm [ lemon:writtenRep "ativo"@pt ] ;
    lemon:otherForm [ lemon:writtenRep "ativos"@pt ] ;
    lemon:cat [lemon:partOfSpeech infoLex:noun];
    lemon:sense [ lemon:reference ontoRisco:bem ] .
```

```
:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :recurso .

:recurso a lemon:LexicalEntry ;
    lemon:canonicalForm [ lemon:writtenRep "recurso"@pt ] ;
    lemon:otherForm [ lemon:writtenRep "recursos"@pt ] ;
    lemon:cat [lemon:partOfSpeech infoLex:noun];
    lemon:sense [ lemon:reference ontoRisco:bem ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :evento_inesperado .

:evento_inesperado a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "evento inesperado"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:evento_inesperado ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :detecção .

:detecção a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "detecção"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:detecção ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :prevenção .

:prevenção a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "prevenção"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:prevenção ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :fornecedor .

:fornecedor a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "fornecedor"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:fornecedor ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :provedor .

:provedor a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "provedor"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:fornecedor ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
```

```
  lemon:entry :vendedor .

:vendedor a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "vendedor"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:fornecedor ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :comerciante .

:comerciante a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "comerciante"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:fornecedor ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :resiliência .

:resiliência a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "resiliência"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:resiliência ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :elasticidade .

:elasticidade a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "elasticidade"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:resiliência ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :perda .

:perda a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "perda"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:perda ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :derrota .

:derrota a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "derrota"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:perda ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :desgaste .

:desgaste a lemon:LexicalEntry ;
```

```
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "desgaste"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:perda ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :perdas .

:perdas a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "perdas"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:perda ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :prejuízo .

:prejuízo a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "prejuízo"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:perda ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :privação .

:privação a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "privação"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:perda ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :ataque .

:ataque a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "ataque"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:ataque ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :atentado .

:atentado a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "atentado"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:ataque ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :assalto .

:assalto a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "assalto"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:ataque ] .
```

```
:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :ataque .

:ataque a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "ataque"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:ataque ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :participante .

:participante a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "participante"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:participante ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :activo .

:activo a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "activo"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:participante ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :organização .

:organização a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "organização"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:organização ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :sociedade .

:sociedade a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "sociedade"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:organização ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :ordem .

:ordem a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "ordem"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:organização ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
```

```
  lemon:entry :associações .

:associações a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "associações"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:organização ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :união .

:união a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "união"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:organização ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :organização .

:organização a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "organização"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:organização ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :sistema .

:sistema a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "sistema"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:organização ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :organização .

:organização a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "organização"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:organização ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :estrutura .

:estrutura a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "estrutura"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:organização ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :estabelecimento .

:estabelecimento a lemon:LexicalEntry ;
```

```
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "estabelecimento"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:organização ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :constituição .

:constituição a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "constituição"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:organização ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :administração .

:administração a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "administração"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:organização ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :método_comportamental .

:método_comportamental a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "método comportamental"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:método_comportamental ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :fraude .

:fraude a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "fraude"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:fraude ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :hoax .

:hoax a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "hoax"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:fraude ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :trote .

:trote a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "trote"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:fraude ] .
```

```
:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :sinistro .

:sinistro a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "sinistro"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:sinistro ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :default .

:default a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "default"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:default ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :banco_central .

:banco_central a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "banco central"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:banco_central ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :Bancos_centrais .

:Bancos_centrais a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "Bancos centrais"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:banco_central ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :grupo_de_resposta .

:grupo_de_resposta a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "grupo de resposta"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:grupo_de_resposta ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :exposição .

:exposição a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "exposição"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:exposição ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
```

```
  lemon:entry :exibição .

:exibição a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "exibição"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:exposição ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :vulnerabilidade .

:vulnerabilidade a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "vulnerabilidade"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:exposição ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :contramedidas .

:contramedidas a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "contramedidas"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:contramedidas ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :concentração .

:concentração a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "concentração"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:concentração ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :atenção .

:atenção a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "atenção"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:concentração ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :densidade .

:densidade a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "densidade"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:concentração ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :infraestrutura .

:infraestrutura a lemon:LexicalEntry ;
```

```
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "infraestrutura"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:infraestrutura ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :regra .

:regra a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "regra"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:regra ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :regulamentos .

:regulamentos a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "regulamentos"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:regra ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :lei .

:lei a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "lei"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:regra ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :norma .

:norma a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "norma"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:regra ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :limite .

:limite a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "limite"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:limite ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :limitação .

:limitação a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "limitação"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:limite ] .
```

```
:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :fronteira .

:fronteira a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "fronteira"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:limite ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :fronteiras .

:fronteiras a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "fronteiras"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:limite ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :máximo .

:máximo a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "máximo"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:limite ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :delimitação .

:delimitação a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "delimitação"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:limite ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :perímetro .

:perímetro a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "perímetro"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:limite ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :margem .

:margem a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "margem"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:limite ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
```

```
  lemon:entry :máquina .

:máquina a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "máquina"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:máquina ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :especialistas .

:especialistas a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "especialistas"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:especialistas ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :serviço .

:serviço a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "serviço"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:serviço ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :instalação .

:instalação a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "instalação"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:serviço ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :serviço .

:serviço a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "serviço"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:serviço ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :incumbência .

:incumbência a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "incumbência"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:serviço ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :profissão .

:profissão a lemon:LexicalEntry ;
```

```
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "profissão"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:serviço ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :função .

:função a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "função"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:serviço ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :serviço .

:serviço a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "serviço"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:serviço ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :emprego .

:emprego a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "emprego"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:serviço ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :ofício .

:ofício a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "ofício"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:serviço ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :trabalho .

:trabalho a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "trabalho"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:serviço ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :ocupação .

:ocupação a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "ocupação"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:serviço ] .
```

```
:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :infraestrutura_crítica_financeira .

:infraestrutura_crítica_financeira a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "infraestrutura crítica financeira"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:infraestrutura_crítica_financeira ] .
:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :proteção .

:proteção a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "proteção"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:proteção ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :proteção .

:proteção a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "proteção"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:proteção ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :mecenato .

:mecenato a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "mecenato"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:proteção ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :defesa .

:defesa a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "defesa"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:proteção ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :desastre .

:desastre a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "desastre"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:desastre ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :flagelo .
```

```
:flagelo a lemon:LexicalEntry ;
   infoLex:partOfSpeech infoLex:noun ;
   lemon:canonicalForm [ lemon:writtenRep "flagelo"@pt ] ;
   lemon:sense [ lemon:reference ontoRisco:desastre ] .

:riscoFinanceiro a lemon:Lexicon;
 lemon:language "pt" ;
 lemon:entry :calamidade .

:calamidade a lemon:LexicalEntry ;
   infoLex:partOfSpeech infoLex:noun ;
   lemon:canonicalForm [ lemon:writtenRep "calamidade"@pt ] ;
   lemon:sense [ lemon:reference ontoRisco:desastre ] .

:riscoFinanceiro a lemon:Lexicon;
 lemon:language "pt" ;
 lemon:entry :catástrofe .

:catástrofe a lemon:LexicalEntry ;
   infoLex:partOfSpeech infoLex:noun ;
   lemon:canonicalForm [ lemon:writtenRep "catástrofe"@pt ] ;
   lemon:sense [ lemon:reference ontoRisco:desastre ] .

:riscoFinanceiro a lemon:Lexicon;
 lemon:language "pt" ;
 lemon:entry :tragédia .

:tragédia a lemon:LexicalEntry ;
   infoLex:partOfSpeech infoLex:noun ;
   lemon:canonicalForm [ lemon:writtenRep "tragédia"@pt ] ;
   lemon:sense [ lemon:reference ontoRisco:desastre ] .

:riscoFinanceiro a lemon:Lexicon;
 lemon:language "pt" ;
 lemon:entry :catástrofes .

:catástrofes a lemon:LexicalEntry ;
   infoLex:partOfSpeech infoLex:noun ;
   lemon:canonicalForm [ lemon:writtenRep "catástrofes"@pt ] ;
   lemon:sense [ lemon:reference ontoRisco:desastre ] .

:riscoFinanceiro a lemon:Lexicon;
 lemon:language "pt" ;
 lemon:entry :cataclismo .

:cataclismo a lemon:LexicalEntry ;
   infoLex:partOfSpeech infoLex:noun ;
   lemon:canonicalForm [ lemon:writtenRep "cataclismo"@pt ] ;
   lemon:sense [ lemon:reference ontoRisco:desastre ] .

:riscoFinanceiro a lemon:Lexicon;
 lemon:language "pt" ;
 lemon:entry :consultoria .

:consultoria a lemon:LexicalEntry ;
   infoLex:partOfSpeech infoLex:noun ;
```

```
    lemon:canonicalForm [ lemon:writtenRep "consultoria"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:consultoria ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :consultoria .

:consultoria a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "consultoria"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:consultoria ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :consultoria .

:consultoria a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "consultoria"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:consultoria ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :consultas .

:consultas a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "consultas"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:consultoria ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :participante_de_mercado_crítico .

:participante_de_mercado_crítico a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "participante de mercado crítico"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:participante_de_mercado_crítico ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :falha_em_transação .

:falha_em_transação a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "falha em transação"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:falha_em_transação ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :interrupção_de_operaçao_significativa .

:interrupção_de_operaçao_significativa a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "interrupção de operaçao significativa"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:interrupção_de_operaçao_significativa ] .
```

```
:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :roubo .

:roubo a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "roubo"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:roubo ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :roubo .

:roubo a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "roubo"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:roubo ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :rapto .

:rapto a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "rapto"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:roubo ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :furto .

:furto a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "furto"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:roubo ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :espoliação .

:espoliação a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "espoliação"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:roubo ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :assalto .

:assalto a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "assalto"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:roubo ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :roubo .
```

```
:roubo a lemon:LexicalEntry ;
   infoLex:partOfSpeech infoLex:noun ;
   lemon:canonicalForm [ lemon:writtenRep "roubo"@pt ] ;
   lemon:sense [ lemon:reference ontoRisco:roubo ] .

:riscoFinanceiro a lemon:Lexicon;
 lemon:language "pt" ;
 lemon:entry :rapina .

:rapina a lemon:LexicalEntry ;
   infoLex:partOfSpeech infoLex:noun ;
   lemon:canonicalForm [ lemon:writtenRep "rapina"@pt ] ;
   lemon:sense [ lemon:reference ontoRisco:roubo ] .

:riscoFinanceiro a lemon:Lexicon;
 lemon:language "pt" ;
 lemon:entry :saque .

:saque a lemon:LexicalEntry ;
   infoLex:partOfSpeech infoLex:noun ;
   lemon:canonicalForm [ lemon:writtenRep "saque"@pt ] ;
   lemon:sense [ lemon:reference ontoRisco:roubo ] .

:riscoFinanceiro a lemon:Lexicon;
 lemon:language "pt" ;
 lemon:entry :botim .

:botim a lemon:LexicalEntry ;
   infoLex:partOfSpeech infoLex:noun ;
   lemon:canonicalForm [ lemon:writtenRep "botim"@pt ] ;
   lemon:sense [ lemon:reference ontoRisco:roubo ] .

:riscoFinanceiro a lemon:Lexicon;
 lemon:language "pt" ;
 lemon:entry :espólio .

:espólio a lemon:LexicalEntry ;
   infoLex:partOfSpeech infoLex:noun ;
   lemon:canonicalForm [ lemon:writtenRep "espólio"@pt ] ;
   lemon:sense [ lemon:reference ontoRisco:roubo ] .

:riscoFinanceiro a lemon:Lexicon;
 lemon:language "pt" ;
 lemon:entry :pilhagem .

:pilhagem a lemon:LexicalEntry ;
   infoLex:partOfSpeech infoLex:noun ;
   lemon:canonicalForm [ lemon:writtenRep "pilhagem"@pt ] ;
   lemon:sense [ lemon:reference ontoRisco:roubo ] .

:riscoFinanceiro a lemon:Lexicon;
 lemon:language "pt" ;
 lemon:entry :agente_de_ameaça .

:agente_de_ameaça a lemon:LexicalEntry ;
   infoLex:partOfSpeech infoLex:noun ;
```

```
    lemon:canonicalForm [ lemon:writtenRep "agente de ameaça"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:agente_de_ameaça ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :pessoa_jurídica .

:pessoa_jurídica a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "pessoa jurídica"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:pessoa_jurídica ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :líder_da_indústria .

:líder_da_indústria a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "líder da indústria"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:líder_da_indústria ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :opções_de_tecnologia .

:opções_de_tecnologia a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "opções de tecnologia"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:opções_de_tecnologia ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :correção .

:correção a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "correção"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:correção ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :emenda .

:emenda a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "emenda"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:correção ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :retificação .

:retificação a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "retificação"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:correção ] .
```

```
:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :correção .

:correção a lemon:LexicalEntry ;
   infoLex:partOfSpeech infoLex:noun ;
   lemon:canonicalForm [ lemon:writtenRep "correção"@pt ] ;
   lemon:sense [ lemon:reference ontoRisco:correção ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :artefato .

:artefato a lemon:LexicalEntry ;
   infoLex:partOfSpeech infoLex:noun ;
   lemon:canonicalForm [ lemon:writtenRep "artefato"@pt ] ;
   lemon:otherForm [ lemon:writtenRep "artefatos"@pt ] ;
   lemon:sense [ lemon:reference ontoRisco:artefato ] ;
   lemon:sense :artefato_sense .

:artefato_sense lemon:reference ontoRisco:artefato .

:artefato_sense lemon:narrower :mercadoria .

:mercadoria a lemon:LexicalEntry ;
   lemon:canonicalForm [ lemon:writtenRep "mercadoria"@pt ] ;
   lemon:otherForm [ lemon:writtenRep "mercadorias"@pt ] ;
   lemon:cat [lemon:partOfSpeech infoLex:noun];
   lemon:sense [ lemon:reference ontoRisco:mercadoria ] ;
   lemon:sense :mercadoria_sense .

:mercadoria_sense lemon:reference ontoRisco:mercadoria .

:mercadoria_sense lemon:narrower :artigo, :produto .

:artigo a lemon:LexicalEntry ;
   lemon:canonicalForm [ lemon:writtenRep "artigo"@pt ] ;
   lemon:otherForm [ lemon:writtenRep "artigos"@pt ] ;
   lemon:cat [lemon:partOfSpeech infoLex:noun];
   lemon:sense [ lemon:reference ontoRisco:artigo ] .

:produto a lemon:LexicalEntry ;
   lemon:canonicalForm [ lemon:writtenRep "produto"@pt ] ;
   lemon:otherForm [ lemon:writtenRep "produtos"@pt ] ;
   lemon:cat [lemon:partOfSpeech infoLex:noun];
   lemon:sense [ lemon:reference ontoRisco:produto ] .

:gênero_literário a lemon:LexicalEntry ;
   lemon:canonicalForm [ lemon:writtenRep "gênero literário"@pt ] ;
   lemon:otherForm [ lemon:writtenRep "gêneros literários"@pt ] ;
   lemon:cat [lemon:partOfSpeech infoLex:noun];
   lemon:sense [ lemon:reference ontoRisco:gênero_literário ] ;
   lemon:sense :gênero_literário_sense .

:gênero_literário_sense lemon:reference ontoRisco:gênero_literário .

:gênero_literário_sense lemon:narrower :prosa .
```

```
:prosa a lemon:LexicalEntry ;
   lemon:canonicalForm [ lemon:writtenRep "prosa"@pt ] ;
   lemon:otherForm [ lemon:writtenRep "prosas"@pt ] ;
   lemon:cat [lemon:partOfSpeech infoLex:noun];
   lemon:sense [ lemon:reference ontoRisco:prosa ] ;
   lemon:sense :prosa_sense .

:prosa_sense lemon:reference ontoRisco:prosa .

:prosa_sense lemon:narrower :não-ficção .

:não-ficção a lemon:LexicalEntry ;
   lemon:canonicalForm [ lemon:writtenRep "não-ficção"@pt ] ;
   lemon:otherForm [ lemon:writtenRep "não-ficção"@pt ] ;
   lemon:cat [lemon:partOfSpeech infoLex:noun];
   lemon:sense [ lemon:reference ontoRisco:não-ficção ] ;
   lemon:sense :não-ficção_sense .

:não-ficção_sense lemon:reference ontoRisco:não-ficção .

:não-ficção_sense lemon:narrower :artigo .

:mercadoria_sense lemon:incompatible :não-ficção_sense .

:não-ficção_sense lemon:incompatible :mercadoria_sense .

:riscoFinanceiro a lemon:Lexicon;
 lemon:language "pt" ;
 lemon:entry :pessoal .

:pessoal a lemon:LexicalEntry ;
   infoLex:partOfSpeech infoLex:noun ;
   lemon:canonicalForm [ lemon:writtenRep "pessoal"@pt ] ;
   lemon:sense [ lemon:reference ontoRisco:pessoal ] .

:riscoFinanceiro a lemon:Lexicon;
 lemon:language "pt" ;
 lemon:entry :companhia .

:companhia a lemon:LexicalEntry ;
   infoLex:partOfSpeech infoLex:noun ;
   lemon:canonicalForm [ lemon:writtenRep "companhia"@pt ] ;
   lemon:sense [ lemon:reference ontoRisco:pessoal ] .

:riscoFinanceiro a lemon:Lexicon;
 lemon:language "pt" ;
 lemon:entry :pessoal .

:pessoal a lemon:LexicalEntry ;
   infoLex:partOfSpeech infoLex:noun ;
   lemon:canonicalForm [ lemon:writtenRep "pessoal"@pt ] ;
   lemon:sense [ lemon:reference ontoRisco:pessoal ] .

:riscoFinanceiro a lemon:Lexicon;
 lemon:language "pt" ;
 lemon:entry :elenco .
```

```
:elenco a lemon:LexicalEntry ;
   infoLex:partOfSpeech infoLex:noun ;
   lemon:canonicalForm [ lemon:writtenRep "elenco"@pt ] ;
   lemon:sense [ lemon:reference ontoRisco:pessoal ] .

:riscoFinanceiro a lemon:Lexicon;
 lemon:language "pt" ;
 lemon:entry :equipe_de_funcionários .

:equipe_de_funcionários a lemon:LexicalEntry ;
   infoLex:partOfSpeech infoLex:noun ;
   lemon:canonicalForm [ lemon:writtenRep "equipe de funcionários"@pt ] ;
   lemon:sense [ lemon:reference ontoRisco:pessoal ] .

:riscoFinanceiro a lemon:Lexicon;
 lemon:language "pt" ;
 lemon:entry :gente .

:gente a lemon:LexicalEntry ;
   infoLex:partOfSpeech infoLex:noun ;
   lemon:canonicalForm [ lemon:writtenRep "gente"@pt ] ;
   lemon:sense [ lemon:reference ontoRisco:pessoal ] .

:riscoFinanceiro a lemon:Lexicon;
 lemon:language "pt" ;
 lemon:entry :vulnerabilidade .

:vulnerabilidade a lemon:LexicalEntry ;
   infoLex:partOfSpeech infoLex:noun ;
   lemon:canonicalForm [ lemon:writtenRep "vulnerabilidade"@pt ] ;
   lemon:sense [ lemon:reference ontoRisco:vulnerabilidade ] .

:riscoFinanceiro a lemon:Lexicon;
 lemon:language "pt" ;
 lemon:entry :debilidade .

:debilidade a lemon:LexicalEntry ;
   infoLex:partOfSpeech infoLex:noun ;
   lemon:canonicalForm [ lemon:writtenRep "debilidade"@pt ] ;
   lemon:sense [ lemon:reference ontoRisco:vulnerabilidade ] .

:riscoFinanceiro a lemon:Lexicon;
 lemon:language "pt" ;
 lemon:entry :exposição .

:exposição a lemon:LexicalEntry ;
   infoLex:partOfSpeech infoLex:noun ;
   lemon:canonicalForm [ lemon:writtenRep "exposição"@pt ] ;
   lemon:sense [ lemon:reference ontoRisco:vulnerabilidade ] .

:riscoFinanceiro a lemon:Lexicon;
 lemon:language "pt" ;
 lemon:entry :risco .

:risco a lemon:LexicalEntry ;
 infoLex:partOfSpeech infoLex:noun ;
```

```
  lemon:canonicalForm [ lemon:writtenRep "risco"@pt ] ;
  lemon:otherForm [ lemon:writtenRep "riscos"@pt ] ;
  lemon:sense :risco_sense .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :perigo .

:perigo a lemon:LexicalEntry ;
  infoLex:partOfSpeech infoLex:noun ;
  lemon:canonicalForm [ lemon:writtenRep "perigo"@pt ] ;
  lemon:otherForm [ lemon:writtenRep "perigos"@pt ] ;
  lemon:sense :perigo_sense .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :ameaça .

:ameaça a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "ameaça"@pt ] ;
    lemon:otherForm [ lemon:writtenRep "ameaças"@pt ] ;
  lemon:sense :ameaça_sense .

:risco_sense lemon:reference ontoRisco:risco ;
    lemon:equivalent :perigo_sense, :ameaça_sense .

:perigo_sense lemon:reference ontoRisco:risco ;
   lemon:equivalent :risco_sense, :ameaça_sense .

:ameaça_sense lemon:reference ontoRisco:risco ;
   lemon:equivalent :risco_sense, :perigo_sense .

:risco_sense lemon:narrower :risco_de_credito .

:risco_de_credito a lemon:LexicalEntry ;
  infoLex:partOfSpeech infoLex:noun ;
  lemon:canonicalForm [ lemon:writtenRep "risco de crédito"@pt ] ;
  lemon:otherForm [ lemon:writtenRep "riscos de crédito"@pt ] ;
  lemon:sense [ lemon:reference ontoRisco:risco_de_credito ] .

:risco_sense lemon:narrower :risco_operacional .

:risco_operacional a lemon:LexicalEntry ;
  infoLex:partOfSpeech infoLex:noun ;
  lemon:canonicalForm [ lemon:writtenRep "risco operacional"@pt ] ;
  lemon:otherForm [ lemon:writtenRep "riscos operacionais"@pt ] ;
  lemon:sense [ lemon:reference ontoRisco:risco_operacional ] .

:risco_operacional lemon:narrower :risco_de_compensação .

:risco_de_compensação a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "risco de compensação"@pt ] ;
    lemon:otherForm [ lemon:writtenRep "riscos de compensação"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:risco_de_compensação ] .
```

```
:risco_sense lemon:narrower :risco_de_liquidez .

:risco_de_liquidez a lemon:LexicalEntry ;
   infoLex:partOfSpeech infoLex:noun ;
   lemon:canonicalForm [ lemon:writtenRep "risco de liquidez"@pt ] ;
   lemon:otherForm [ lemon:writtenRep "riscos de liquidez"@pt ] ;
   lemon:sense [ lemon:reference ontoRisco:risco_de_liquidez ] .

:risco_sense lemon:narrower :risco_setorial .

:risco_setorial a lemon:LexicalEntry ;
   infoLex:partOfSpeech infoLex:noun ;
   lemon:canonicalForm [ lemon:writtenRep "risco setorial"@pt ] ;
   lemon:otherForm [ lemon:writtenRep "riscos setoriais"@pt ] ;
   lemon:sense [ lemon:reference ontoRisco:risco_setorial ] .

:risco_sense lemon:narrower :risco_técnico .

:risco_técnico a lemon:LexicalEntry ;
   infoLex:partOfSpeech infoLex:noun ;
   lemon:canonicalForm [ lemon:writtenRep "risco técnico"@pt ] ;
   lemon:otherForm [ lemon:writtenRep "riscos técnicos"@pt ] ;
   lemon:sense [ lemon:reference ontoRisco:risco_técnico ] .

:risco_sense lemon:narrower :risco_político .

:risco_político a lemon:LexicalEntry ;
   infoLex:partOfSpeech infoLex:noun ;
   lemon:canonicalForm [ lemon:writtenRep "risco político"@pt ] ;
   lemon:otherForm [ lemon:writtenRep "riscos políticos"@pt ] ;
   lemon:sense [ lemon:reference ontoRisco:risco_político ] .

:risco_sense lemon:narrower :risco_principal .

:risco_principal a lemon:LexicalEntry ;
   infoLex:partOfSpeech infoLex:noun ;
   lemon:canonicalForm [ lemon:writtenRep "risco principal"@pt ] ;
   lemon:otherForm [ lemon:writtenRep "riscos principais"@pt ] ;
   lemon:sense [ lemon:reference ontoRisco:risco_principal ] .

:risco_sense lemon:narrower :risco_social_e_ambiental .

:risco_social_e_ambiental a lemon:LexicalEntry ;
   infoLex:partOfSpeech infoLex:noun ;
   lemon:canonicalForm [ lemon:writtenRep "risco social e ambiental"@pt ] ;
   lemon:otherForm [ lemon:writtenRep "riscos sociais e ambientais"@pt ] ;
   lemon:sense [ lemon:reference ontoRisco:risco_social_e_ambiental ] .

:risco_sense lemon:narrower :risco_legal .

:risco_legal a lemon:LexicalEntry ;
   infoLex:partOfSpeech infoLex:noun ;
   lemon:canonicalForm [ lemon:writtenRep "risco legal"@pt ] ;
   lemon:otherForm [ lemon:writtenRep "riscos legais"@pt ] ;
   lemon:sense [ lemon:reference ontoRisco:risco_legal ] .

:risco_sense lemon:narrower :risco_de_contraparte .
```

```
:risco_de_contraparte a lemon:LexicalEntry ;
   infoLex:partOfSpeech infoLex:noun ;
   lemon:canonicalForm [ lemon:writtenRep "risco de contraparte"@pt ] ;
   lemon:otherForm [ lemon:writtenRep "riscos de contraparte"@pt ] ;
   lemon:sense [ lemon:reference ontoRisco:risco_de_contraparte ] .

:riscoFinanceiro a lemon:Lexicon;
 lemon:language "pt" ;
 lemon:entry :Órgão_regulador .

:Órgão_regulador a lemon:LexicalEntry ;
   infoLex:partOfSpeech infoLex:noun ;
   lemon:canonicalForm [ lemon:writtenRep "Órgão regulador"@pt ] ;
   lemon:sense [ lemon:reference ontoRisco:Órgão_regulador ] .

:riscoFinanceiro a lemon:Lexicon;
 lemon:language "pt" ;
 lemon:entry :academia .

:academia a lemon:LexicalEntry ;
   infoLex:partOfSpeech infoLex:noun ;
   lemon:canonicalForm [ lemon:writtenRep "academia"@pt ] ;
   lemon:sense [ lemon:reference ontoRisco:academia ] .

:riscoFinanceiro a lemon:Lexicon;
 lemon:language "pt" ;
 lemon:entry :escola .

:escola a lemon:LexicalEntry ;
   infoLex:partOfSpeech infoLex:noun ;
   lemon:canonicalForm [ lemon:writtenRep "escola"@pt ] ;
   lemon:sense [ lemon:reference ontoRisco:academia ] .

:riscoFinanceiro a lemon:Lexicon;
 lemon:language "pt" ;
 lemon:entry :mitigação_do_risco .

:mitigação_do_risco a lemon:LexicalEntry ;
   infoLex:partOfSpeech infoLex:noun ;
   lemon:canonicalForm [ lemon:writtenRep "mitigação do risco"@pt ] ;
   lemon:sense [ lemon:reference ontoRisco:mitigação_do_risco ] .

:riscoFinanceiro a lemon:Lexicon;
 lemon:language "pt" ;
 lemon:entry :oferta_de_serviço .

:oferta_de_serviço a lemon:LexicalEntry ;
   infoLex:partOfSpeech infoLex:noun ;
   lemon:canonicalForm [ lemon:writtenRep "oferta de serviço"@pt ] ;
   lemon:sense [ lemon:reference ontoRisco:oferta_de_serviço ] .

:riscoFinanceiro a lemon:Lexicon;
 lemon:language "pt" ;
 lemon:entry :líderes_funcionais .

:líderes_funcionais a lemon:LexicalEntry ;
```

```
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "líderes funcionais"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:líderes_funcionais ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :diretriz .

:diretriz a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "diretriz"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:diretriz ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :diretiva .

:diretiva a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "diretiva"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:diretriz ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :directivas .

:directivas a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "directivas"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:diretriz ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :pesquisa .

:pesquisa a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "pesquisa"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:pesquisa ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :sindicância .

:sindicância a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "sindicância"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:pesquisa ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :investigação .

:investigação a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "investigação"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:pesquisa ] .
```

```
:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :inquérito .

:inquérito a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "inquérito"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:pesquisa ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :exploração .

:exploração a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "exploração"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:pesquisa ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :inquérito .

:inquérito a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "inquérito"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:pesquisa ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :pessoa_física .

:pessoa_física a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "pessoa física"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:pessoa_física ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :stakeholder .

:stakeholder a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "stakeholder"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:stakeholder ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :fator_de_risco .

:fator_de_risco a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "fator de risco"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:fator_de_risco ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
```

```
  lemon:entry :parte_confiável .

:parte_confiável a lemon:LexicalEntry ;
   infoLex:partOfSpeech infoLex:noun ;
   lemon:canonicalForm [ lemon:writtenRep "parte confiável"@pt ] ;
   lemon:sense [ lemon:reference ontoRisco:parte_confiável ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :é_conduzido_por .

:é_conduzido_por a lemon:LexicalEntry ;
   infoLex:partOfSpeech infoLex:collocation ;
   lemon:canonicalForm [ lemon:writtenRep "é conduzido por"@pt ] ;
   lemon:sense [ lemon:reference ontoRisco:é_conduzido_por ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :possui .

:possui a lemon:LexicalEntry ;
   infoLex:partOfSpeech infoLex:verb ;
   lemon:canonicalForm [ lemon:writtenRep "possui"@pt ] ;
   lemon:sense [ lemon:reference ontoRisco:possui ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :para .

:para a lemon:LexicalEntry ;
   infoLex:partOfSpeech infoLex:preposition ;
   lemon:canonicalForm [ lemon:writtenRep "para"@pt ] ;
   lemon:sense [ lemon:reference ontoRisco:para ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :define .

:define a lemon:LexicalEntry ;
   infoLex:partOfSpeech infoLex:verb ;
   lemon:canonicalForm [ lemon:writtenRep "define"@pt ] ;
   lemon:sense [ lemon:reference ontoRisco:define ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :pode_ser_reduzido_por .

:pode_ser_reduzido_por a lemon:LexicalEntry ;
   infoLex:partOfSpeech infoLex:collocation ;
   lemon:canonicalForm [ lemon:writtenRep "pode ser reduzido por"@pt ] ;
   lemon:sense [ lemon:reference ontoRisco:pode_ser_reduzido_por ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :pode_danificar .

:pode_danificar a lemon:LexicalEntry ;
```

```
    infoLex:partOfSpeech infoLex:collocation ;
    lemon:canonicalForm [ lemon:writtenRep "pode danificar"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:pode_danificar ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :inclui .

:inclui a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:verbo ;
    lemon:canonicalForm [ lemon:writtenRep "inclui"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:inclui ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :materializar_como .

:materializar_como a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:collocation ;
    lemon:canonicalForm [ lemon:writtenRep "materializar como"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:materializar_como ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :conduz_a .

:conduz_a a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:collocation ;
    lemon:canonicalForm [ lemon:writtenRep "conduz a"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:conduz_a ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :pode_possuir .

:pode_possuir a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:collocation ;
    lemon:canonicalForm [ lemon:writtenRep "pode possuir"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:pode_possuir ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :impõe .

:impõe a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:verb ;
    lemon:canonicalForm [ lemon:writtenRep "impõe"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:impõe ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :atração .

:atração a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "atração"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:atração ] .
```

```
:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :atratividade .

:atratividade a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "atratividade"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:atração ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :deseja_diminuir .

:deseja_diminuir a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:collocation ;
    lemon:canonicalForm [ lemon:writtenRep "deseja diminuir"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:deseja_diminuir ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :sobre .

:sobre a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:preposition ;
    lemon:canonicalForm [ lemon:writtenRep "sobre"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:sobre ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :valoriza .

:valoriza a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:verb ;
    lemon:canonicalForm [ lemon:writtenRep "valoriza"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:valoriza ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :suscita .

:suscita a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:verb ;
    lemon:canonicalForm [ lemon:writtenRep "suscita"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:suscita ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :estar_ciente_de .

:estar_ciente_de a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:collocation ;
    lemon:canonicalForm [ lemon:writtenRep "estar ciente de"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:estar_ciente_de ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
```

```
  lemon:entry :para_reduzir .

:para_reduzir a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:collocation ;
    lemon:canonicalForm [ lemon:writtenRep "para reduzir"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:para_reduzir ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :aumenta .

:aumenta a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:verb ;
    lemon:canonicalForm [ lemon:writtenRep "aumenta"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:aumenta ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :explora .

:explora a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:verb ;
    lemon:canonicalForm [ lemon:writtenRep "explora"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:explora ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :local_de_nascimento .

:local_de_nascimento a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "local de nascimento"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:local_de_nascimento ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :lugar_de_nascimento .

:lugar_de_nascimento a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "lugar de nascimento"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:local_de_nascimento ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :terra_natal .

:terra_natal a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "terra natal"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:local_de_nascimento ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :local_de_nascimento .

:local_de_nascimento a lemon:LexicalEntry ;
```

```
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "local de nascimento"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:local_de_nascimento ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :pessoa-chave .

:pessoa-chave a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "pessoa-chave"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:pessoa-chave ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :cargo .

:cargo a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "cargo"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:cargo ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :ofício .

:ofício a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "ofício"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:cargo ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :poder .

:poder a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "poder"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:cargo ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :cargo .

:cargo a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "cargo"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:cargo ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :posição .

:posição a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "posição"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:cargo ] .
```

```
:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :número_de_empregados .

:número_de_empregados a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "número de empregados"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:número_de_empregados ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :tem_nome .

:tem_nome a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "tem nome"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:tem_nome ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :descrição_resumida .

:descrição_resumida a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "descrição resumida"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:descrição_resumida ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :iso3166_código_de_país .

:iso3166_código_de_país a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "iso3166 código de país"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:iso3166_código_de_país ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :data_de_morte .

:data_de_morte a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "data de morte"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:data_de_morte ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :indústria .

:indústria a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "indústria"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:indústria ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
```

```
  lemon:entry :indústria .

:indústria a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "indústria"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:indústria ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :manufatura .

:manufatura a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "manufatura"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:indústria ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :empresa_industrial .

:empresa_industrial a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "empresa industrial"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:indústria ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :lugar_de_fundação .

:lugar_de_fundação a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "lugar de fundação"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:lugar_de_fundação ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :cidade_de_localização .

:cidade_de_localização a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "cidade de localização"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:cidade_de_localização ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :receita_líquida .

:receita_líquida a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "receita líquida"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:receita_líquida ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :tem_endereço .

:tem_endereço a lemon:LexicalEntry ;
```

```
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "tem endereço"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:tem_endereço ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :dados_pessoais .

:dados_pessoais a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "dados pessoais"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:dados_pessoais ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :data_de_fundação .

:data_de_fundação a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "data de fundação"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:data_de_fundação ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :número_de_localidades .

:número_de_localidades a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "número de localidades"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:número_de_localidades ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :resumo .

:resumo a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "resumo"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:resumo ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :sumário .

:sumário a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "sumário"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:resumo ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :revisão_rápida .

:revisão_rápida a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "revisão rápida"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:resumo ] .
```

```
:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :resumo_geral .

:resumo_geral a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "resumo geral"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:resumo ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :compêndio .

:compêndio a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "compêndio"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:resumo ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :abreviação .

:abreviação a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "abreviação"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:resumo ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :abreviatura .

:abreviatura a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "abreviatura"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:resumo ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :sinopse .

:sinopse a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "sinopse"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:resumo ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :currículo .

:currículo a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "currículo"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:resumo ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
```

```
  lemon:entry :fundado_por .

:fundado_por a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:collocation ;
    lemon:canonicalForm [ lemon:writtenRep "fundado por"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:fundado_por ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :iso4217_moeda .

:iso4217_moeda a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "iso4217 moeda"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:iso4217_moeda ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :dados_institucionais .

:dados_institucionais a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "dados institucionais"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:dados_institucionais ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :data_de_nascimento .

:data_de_nascimento a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "data de nascimento"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:data_de_nascimento ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :aniversário .

:aniversário a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "aniversário"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:data_de_nascimento ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :data_de_nascimento .

:data_de_nascimento a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "data de nascimento"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:data_de_nascimento ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :descrição .

:descrição a lemon:LexicalEntry ;
```

```
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "descrição"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:descrição ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :caraterização .

:caraterização a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "caraterização"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:descrição ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :ilustração .

:ilustração a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "ilustração"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:descrição ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :declaração .

:declaração a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "declaração"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:descrição ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :códigos_internacionais .

:códigos_internacionais a lemon:LexicalEntry ;
    infoLex:partOfSpeech infoLex:noun ;
    lemon:canonicalForm [ lemon:writtenRep "códigos internacionais"@pt ] ;
    lemon:sense [ lemon:reference ontoRisco:códigos_internacionais ] .

:riscoFinanceiro a lemon:Lexicon;
  lemon:language "pt" ;
  lemon:entry :crime, :violação .

:crime a lemon:LexicalEntry ;
    lemon:canonicalForm [ lemon:writtenRep "crime"@pt ] ;
    lemon:otherForm [ lemon:writtenRep "crimes"@pt ] ;
    lemon:cat [lemon:partOfSpeech infoLex:noun];
    lemon:sense [ lemon:reference ontoRisco:crime ] ;
    lemon:sense :crime_sense .

:crime_sense lemon:reference ontoRisco:crime ;
    lemon:canonicalForm [ lemon:writtenRep "violação"@pt ] ;
    lemon:otherForm [ lemon:writtenRep "violações"@pt ] ;
    lemon:cat [lemon:partOfSpeech infoLex:noun];
    lemon:sense [ lemon:reference ontoRisco:violação ] ;
    lemon:equivalent :violação_sense .
```

```
:violação lemon:canonicalForm [ lemon:writtenRep "violação"@en ] ;
  lemon:sense :violação_sense .

:violação_sense lemon:reference ontoRisco:crime ;
  lemon:equivalent :crime_sense .

:crime_sense lemon:narrower :roubo .

:roubo a lemon:LexicalEntry ;
   lemon:canonicalForm [ lemon:writtenRep "roubo"@pt ] ;
   lemon:otherForm [ lemon:writtenRep "roubos"@pt ] ;
   lemon:cat [lemon:partOfSpeech infoLex:noun];
   lemon:sense [ lemon:reference ontoRisco:roubo ] .

:crime_sense lemon:narrower :rapto .

:rapto a lemon:LexicalEntry ;
   lemon:canonicalForm [ lemon:writtenRep "rapto"@pt ] ;
   lemon:otherForm [ lemon:writtenRep "raptos"@pt ] ;
   lemon:cat [lemon:partOfSpeech infoLex:noun];
   lemon:sense [ lemon:reference ontoRisco:rapto ] .

:crime_sense lemon:narrower :ataque .

:ataque a lemon:LexicalEntry ;
   lemon:canonicalForm [ lemon:writtenRep "ataque"@pt ] ;
   lemon:otherForm [ lemon:writtenRep "ataques"@pt ] ;
   lemon:cat [lemon:partOfSpeech infoLex:noun];
   lemon:sense [ lemon:reference ontoRisco:ataque ] .

:crime_sense lemon:narrower :assalto .

:assalto a lemon:LexicalEntry ;
   lemon:canonicalForm [ lemon:writtenRep "assalto"@pt ] ;
   lemon:otherForm [ lemon:writtenRep "assaltos"@pt ] ;
   lemon:cat [lemon:partOfSpeech infoLex:noun];
   lemon:sense [ lemon:reference ontoRisco:assalto ] .

:crime_sense lemon:narrower :falsificação .

:falsificação a lemon:LexicalEntry ;
   lemon:canonicalForm [ lemon:writtenRep "falsificação"@pt ] ;
   lemon:otherForm [ lemon:writtenRep "falsificações"@pt ] ;
   lemon:cat [lemon:partOfSpeech infoLex:noun];
   lemon:sense [ lemon:reference ontoRisco:falsificação ] .

:crime_sense lemon:narrower :infração .

:infração a lemon:LexicalEntry ;
   lemon:canonicalForm [ lemon:writtenRep "infração"@pt ] ;
   lemon:otherForm [ lemon:writtenRep "infrações"@pt ] ;
   lemon:cat [lemon:partOfSpeech infoLex:noun];
   lemon:sense [ lemon:reference ontoRisco:infração ] .
```