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

MATEUS LOHN ANDRIANI

UM MÉTODO PARA A CONSTRUÇÃO DE TAXONOMIAS UTILIZANDO A DBPEDIA

Florianópolis
2017

Mateus Lohn Andriani

UM MÉTODO PARA A CONSTRUÇÃO DE TAXONOMIAS UTILIZANDO A DBPEDIA

Dissertação submetida ao Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento da Universidade Federal de Santa Catarina para a obtenção do Grau de Mestre em Engenharia do Conhecimento
Orientador: Prof. Dr. José Leomar Todesco.
Coorientador: Prof. Dr. Roberto Carlos dos Santos Pacheco

Florianópolis
2017

Ficha de identificação da obra elaborada pelo autor
através do Programa de Geração Automática da Biblioteca Universitária
da UFSC.

Andriani, Mateus Lohn
Um método para a construção de taxonomias utilizando a DBpedia / Mateus Lohn Andriani ; orientador, José Leomar Todesco; coorientador, Roberto Carlos dos Santos Pacheco - SC, 2017.
149 p.

Dissertação (mestrado) - Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento, Florianópolis, 2017.

Inclui referências.

1. Engenharia e Gestão do Conhecimento. 2. Geração de taxonomias. 3. Reconhecimento de entidades. 4. DBpedia. 5. Plataforma Lattes. I. Todesco, José Leomar. II. Pacheco, Roberto Carlos dos Santos. III. Universidade Federal de Santa Catarina. Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento. IV. Título.

Mateus Lohn Andriani

UM MÉTODO PARA A CONSTRUÇÃO DE TAXONOMIAS UTILIZANDO A DBPEDIA

Esta Dissertação foi julgada adequada para obtenção do Título de Mestre, e aprovada em sua forma final pelo Programa de Pós Graduação em Engenharia e Gestão do Conhecimento.

Florianópolis, 15 de fevereiro de 2017.

_____
Prof. Roberto Carlos dos Santos Pacheco, Dr.
Coordenador do Curso

Banca Examinadora:

_____
Prof. José Leomar Todesco, Dr.
Orientador
Universidade Federal de Santa Catarina

_____
Prof. Roberto Carlos dos Santos Pacheco, Dr.
Coordenador do Curso

_____
Prof. Denilson Sell, Dr.
Universidade Federal de Santa Catarina

_____
Prof. Alexandre Leopoldo Gonçalves, Dr.
Universidade Federal de Santa Catarina

_____
Prof. Jordan Paulesky Juliani, Dr.
Universidade do Estado de Santa Catarina

Esse trabalho é dedicado a todas as pessoas que acreditam em um objetivo e não medem esforços para conquistá-lo.

AGRADECIMENTOS

Apesar de constar somente meu nome como autor desse trabalho, sem o apoio de diversas pessoas o mesmo não teria sido realizado. Nesse espaço, realizo o devido agradecimento a todas essas pessoas (e peço desculpas caso tenha me esquecido de alguém).

Agradeço a minha família pelo incentivo e pelo apoio incondicional ao estudo. Aos meus pais (Francisco e Graça), minha irmã (Mariana), minha avó (Maria Matilde) e minha namorada (Helena), todo o agradecimento pelo carinho, conforto e reconhecimento do esforço prestado em torno deste trabalho. Sem eles, esta dissertação não teria surgido. Agradeço também aos meus amigos, pelo apoio e solidariedade diante das dificuldades, tanto pessoais quanto acadêmicas.

Um imenso agradecimento ao meu orientador, Tite, por todo o apoio para que esse mestrado fosse realizado, da concepção da ideia até sua conclusão, materializada neste trabalho. Também agradeço ao meu coorientador, Roberto Pacheco, pela atenção e apoio na coorientação. A mesma apreciação eu realizo aos membros da banca Denilson Sell, Alexandre Leopoldo Gonçalves e Jordan Paulesky Juliani, pelo carinho por aceitarem fazer parte dessa banca e disponibilizarem seus conhecimentos para que se possa melhorar a qualidade do trabalho.

Também gostaria de expressar gratidão à equipe do Instituto Stela, que me permitiu e auxiliou a trilhar todo o caminho desse mestrado. Muitas das ideias presentes aqui foram concebidas a partir dos trabalhos desenvolvidos no instituto.

Gostaria de realizar um agradecimento especial a Flavio Ceci, que foi responsável por plantar a semente que originou todo esse caminho trilhado no EGC e auxiliou ativamente no desenvolvimento de alguns trabalhos realizados neste percurso. Também agradeço a Rudger Taxweiler, que teve participação ativa na implementação do projeto, através do auxílio no acesso às informações da Plataforma Lattes utilizadas no experimento.

Agradeço também a todos os demais professores e colegas do EGC, por me guiarem durante esses anos no mestrado e terem me auxiliado a desenvolver as competências aderentes ao programa de forma integrada.

Por fim, agradeço também a todas as pessoas que estiveram envolvidas diretamente ou indiretamente na produção desse trabalho.

“As nuvens mudam sempre de posição, mas são sempre nuvens no céu. Assim devemos ser todo dia, mutantes, porém leais com o que pensamos e sonhamos; lembre-se, tudo se desmancha no ar, menos os pensamentos”.

(Paulo Beleki)

RESUMO

O processo de criação de taxonomias demanda esforço de especialistas de domínio, engenheiros de taxonomias, investimento financeiro e tempo. Devido às limitações existentes em fornecer estes recursos em sua integralidade em diversas organizações, muitos projetos que envolvem a construção de taxonomias não atingem o êxito esperado. Este trabalho pretende auxiliar na construção de taxonomias através da proposição de um método automatizado para a sua construção. Para a construção deste método, foi adotada uma série de procedimentos metodológicos, que se iniciou com o levantamento do referencial teórico sobre taxonomias e sua construção. Em sequência, foi realizada uma busca sistemática no domínio de construção automatizada de taxonomias, buscando encontrar abordagens e procedimentos já existentes neste campo de estudo. A partir desta revisão, foi elaborado um método para a geração de taxonomias a partir de repositórios de informações textuais com o apoio de bases de conhecimento, que fornecem as relações hierárquicas para verificação das relações taxonômicas entre os termos. Uma implementação deste método em formato de software foi realizada, utilizando uma amostra de currículos da área de conhecimento das Ciências Agrárias cadastrados na Plataforma Lattes como repositório de informações. A versão em português da DBpedia foi adotada como base de conhecimento neste experimento. Esta implementação também adota um processo de reconhecimento de entidades para a descoberta dos termos relevantes que podem ser cadastrados nas taxonomias. As propostas de taxonomias geradas pela implementação foram comparadas estatisticamente com o tesauro AGROVOC, referência na área da agricultura. Com a análise, verificou-se que 60% a 80% dos termos encontrados nas taxonomias geradas pela implementação também estão presentes no AGROVOC, sendo esta oscilação pertinente aos parâmetros de filtragem informados na entrada do método, o repositório de informações textuais utilizado e a base de conhecimento empregada para validação das relações hierárquicas.

Palavras-chave: Geração de taxonomias. Reconhecimento de entidades. Plataforma Lattes. DBpedia. AGROVOC.

ABSTRACT

The process of creating taxonomies demands effort from domain experts, taxonomy engineers, financial investment and time. Due to the limitations of providing these resources in their entirety in several organizations, many projects that involve the construction of taxonomies do not achieve the expected success. This work intends to assist in the construction of taxonomies through the proposition of an automated method for its construction. For the construction of this method, a series of methodological procedures was adopted, which began with the survey of the theoretical reference on taxonomies and their construction. In sequence, a systematic search was made in the field of automated taxonomy construction, seeking to find approaches and procedures that already exist in this field of study. From this review, a method was developed for the generation of taxonomies from textual information repositories with the support of knowledge bases, which provide the hierarchical relationships for the verification of the taxonomic relations between the terms. An implementation of this method in software format was performed, using a sample of curricula from the Agrarian Sciences knowledge area registered in the Plataforma Lattes as a repository of information. The DBpedia's Portuguese language version was adopted as knowledge base in this experiment. This implementation also adopts a process of entity recognition for the discovery of the relevant terms that can be registered in the taxonomies. The taxonomy proposals generated by the implementation were compared statistically with the AGROVOC thesaurus, reference in the area of agriculture. With the analysis, it was verified that 60% to 80% of the terms found in the taxonomies generated by the implementation are also present in AGROVOC, being this oscillation pertinent to the filter parameters informed in the method entry, the textual information repository used and the knowledge base used to validate hierarchical relationships.

Keywords: Taxonomy generation. Entity Recognition. Plataforma Lattes. DBpedia. AGROVOC.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE ABREVIATURAS E SIGLAS

AOL – *America Online*
API – *Application Programming Interface* (ver glossário)
ABNT – Associação Brasileira de Normas Técnicas
CAPES – Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
CNPq – Conselho Nacional de Desenvolvimento Científico e Tecnológico
CRM – Gestão de Relacionamento com o Cliente (originado do termo em inglês *Customer Relationship Management*)
C&T – Ciência e Tecnologia
CT&I – Ciência, Tecnologia e Inovação
DMOZ – *Directory Mozilla* (também referenciado como ODP – Open Directory Project)
DTD – *Document Type Definition*
GC – Gestão do Conhecimento
EC – Engenharia do Conhecimento
EGC – Engenharia e Gestão do Conhecimento
FINEP – Financiadora de Estudos e Projetos
IBGE – Instituto Brasileiro de Geografia e Estatística
ISO – Organização Internacional de Normalização (originado do termo em inglês *International Organization for Standardization*)
KBS – Sistemas Baseados em Conhecimento (originado do termo em inglês *Knowledge-Based System*)
MCT – Ministério da Ciência, Tecnologia e Inovação
MEC – Ministério da Educação
MSN – *The Microsoft Network*
ODP – *Open Directory Project*
OLAP – Processamento Analítico Online (originado do termo em inglês *On-Line Analytical Processing*)
RDF – Resource Description Framework
SPARQL – *SPARQL Protocol and RDF Query Language*
SVM – Máquina de Vetores de Suporte (originado do termo em inglês *Support Vector Machine*)
TICs – Tecnologias de Informação e Comunicação
XML - *eXtensible Markup Language*

SUMÁRIO

1 INTRODUÇÃO

A sociedade atual está passando por uma mudança de paradigma. A visão de mundo fundamentada na produção em formato industrial está sendo substituída por uma perspectiva orientada ao conhecimento. Como reflexo disso, o conhecimento torna-se progressivamente um recurso estratégico para o crescimento das organizações. (KIM; CENFETELLI; BENBASAT, 2012).

As áreas de estudo da Engenharia e Gestão do Conhecimento originaram-se dessa mudança de paradigma, objetivando estruturar e auxiliar na demanda de gerenciar todo o conhecimento gerado, armazenado e disponibilizado. A Gestão do Conhecimento está cada vez mais interligada a cultura das grandes organizações. Suas práticas não são somente um diferencial na concorrência global, mas também são fundamentais para sua sobrevivência. A Engenharia de Conhecimento, por sua vez, suplementa o trabalho realizado na Gestão do Conhecimento. As técnicas desenvolvidas na área da Engenharia do Conhecimento permitem que o conhecimento seja representado através de diversos sistemas organizacionais, auxiliando na sua disseminação. Para Preece e outros (2001), essas técnicas favorecem e fomentam a prática da Gestão do Conhecimento nas organizações.

A construção das soluções de Engenharia do Conhecimento transformou-se com o tempo. Foram desenvolvidas diferentes estruturas para modelagem e representação do conhecimento, tais como *Role-Limiting Methods* e *Generic Tasks* (STUDER; BENJAMINS; FENSEL, 1998). Atualmente, as ontologias fazem parte das principais estruturas utilizadas para armazenamento do conhecimento. Elas são construídas em um modelo fundamentado em classes e instâncias, que por sua vez estão conectados através de hierarquias e associações entre si. Tais artefatos podem ser utilizados para compor aplicações baseadas em web semântica, compartilhamento de conhecimento e recuperação de informação. (ZOUAQ; NKAMBOU (2009); SELLAMI; CAMPS; AUSSENAC-GILLES, 2013).

De acordo com Knox e Logan (2003), as taxonomias são exemplos de ontologias. Para os autores, elas constituem elementos estruturais que compõem a base de ontologias, provendo suporte para organizar hierarquicamente os conceitos mapeados.

A construção de taxonomias, assim como a de ontologias em geral, possui um alto grau de complexidade. É necessária uma equipe de especialistas no domínio de análise, assim como o suporte de

engenheiros de conhecimento para que as informações estruturadas representem adequadamente o conhecimento.

Nesta pesquisa propõe-se o desenvolvimento de um método que visa auxiliar no processo de construção de taxonomias. O método realiza a análise do domínio em estudo a partir de documentos textuais, retirando a partir destes os conceitos que serão utilizados. Pretende-se aproveitar a classificação categórica presente na Wikipedia, disponibilizada em formato de bases de conhecimento a partir do projeto DBpedia, para organizar os conceitos hierarquicamente. Se necessário, as taxonomias obtidas ainda podem ser aprimoradas posteriormente de forma manual por especialistas de domínio.

1.1 DEFINIÇÃO DO PROBLEMA

O processo de modelagem e construção de taxonomias – e de ontologias, de forma geral – possui uma natureza complexa. Geralmente, elas são construídas por uma única pessoa ou por um pequeno grupo, que trabalham juntos por uma série de horas ou dias. Esse processo é caracterizado pela grande quantidade de tempo que necessita para ser executado, que por sua vez pode aumentar consideravelmente de acordo com o tamanho do domínio de análise. (CHILTON et al., 2013; IJNTEMA, 2012).

Em tais cenários de grande quantidade de dados, o processo de construção manual pode-se tornar insustentável, trazendo grandes chances de tornar a taxonomia construída consideravelmente desatualizada em relação ao domínio. Para Ochoa e outros (2013), um dos gargalos na construção de soluções de engenharia do conhecimento é justamente o próprio processo de construção e população das estruturas que irão armazenar o conhecimento.

Técnicas existentes na área de recuperação de informação e fontes de informação e conhecimento já existentes podem auxiliar nas tarefas que envolvem a construção de taxonomias. Bases de conhecimento colaborativas (caso de bases disponibilizadas em formato *wiki*, por exemplo) contém informações construídas e validadas manualmente, tornando desnecessária a realização de uma nova série de validações específicas para relações descobertas por procedimentos automatizados. Técnicas desenvolvidas no campo de estudo da descoberta de conhecimento permitem analisar grandes quantidades de texto e extrair novos conhecimentos a partir do domínio analisado. Algoritmos para o processamento de linguagem natural e análise estatística são exemplos de métodos que realizam extração de

informação de forma automatizada e que podem ser aplicados para fornecer insumos para a construção de taxonomias.

Considerando o cenário apresentado nos parágrafos anteriores, essa dissertação possui a seguinte pergunta de pesquisa: como facilitar o processo de construção de taxonomias, aproveitando-se de artefatos, estruturas e técnicas já existentes?

1.2 OBJETIVOS DO TRABALHO

Os objetivos deste trabalho dividem-se em objetivo geral e objetivos específicos.

1.2.1 Objetivo geral

O objetivo geral deste trabalho é desenvolver um método para a extração de taxonomias a partir de documentos, hierarquizadas segundo uma base de conhecimento.

1.2.2 Objetivos específicos

Os objetivos específicos do projeto de pesquisa são os seguintes:

- Apresentar o cenário atual da área de construção de taxonomias de forma automatizada;
- Propor um método automatizado para a construção de taxonomias, a partir do referencial teórico levantado na revisão e na análise;
- Implementar em formato de *software* uma versão do método que demonstre a sua viabilidade;
- Avaliar a funcionalidade do método, comparando os resultados obtidos com uma taxonomia de referência;

1.3 JUSTIFICATIVA E RELEVÂNCIA DO TEMA

Sem um corpo técnico com conhecimentos específicos para trabalhar com taxonomias, as organizações enfrentam dificuldades para construí-las. De acordo com a empresa de consultoria Gartner, cerca de sete em cada dez das organizações que investiram em iniciativas de gestão do conhecimento não alcançaram o retorno esperado devido aos recursos dedicados serem insuficientes (KNOX; LOGAN, 2003; R; RAO, 2011).

Uma taxonomia pode ser expandida para uma ontologia de domínio ao se acrescentar as associações que são transversais à árvore de conceitos. Assim como uma taxonomia, o custo de construir uma ontologia de forma manual é grande e algumas vezes inviável, considerando os planejamentos estratégicos e financeiros das organizações em geral (BREWSTER; WILKS, 2004). Um dos casos que se enquadra neste cenário é o desenvolvimento da Gene Ontology, referência na área de Ciências Biológicas e Médicas, que teve um custo avaliado em mais de 16 milhões de dólares (GOOD et al, 2006). O fornecimento de uma taxonomia de base para a construção de uma ontologia, dessa forma, elimina parte desse custo, pois sua evolução posterior para uma ontologia é mais simples do que construir uma ontologia a partir de um rascunho em branco.

Além do fator financeiro, o tempo influencia sobre a decisão de realizar projetos de GC que contenham em seu escopo a utilização de ontologias. Suárez-Figueroa, Gómez-Pérez e Villazón-Terrazas (2009) relataram que em um projeto sem a adoção de metodologias para o desenvolvimento de ontologias, foram necessários dez meses para concluir o seu desenvolvimento. Em um projeto semelhante, mas no qual foi adotada uma metodologia de desenvolvimento, foram despendidos seis meses. Um método automatizado que forneça propostas de taxonomias para o domínio em estudo pode auxiliar a diminuir ainda mais este tempo.

1.4 ESCOPO DO TRABALHO

A pesquisa realizada nesse trabalho se delimitará a construir um método de construção de taxonomias que trate da identificação dos termos e da elaboração da hierarquia taxonômica. Não serão abordados relacionamentos de sinonímia ou de termos relacionados neste método.

Por sua vez, a implementação do método proposto neste trabalho será desenvolvida sobre uma amostra de currículos disponibilizados na Plataforma Lattes. Outros repositórios de conteúdo textual não serão avaliados, embora o método possa ser empregado também nestes cenários.

A avaliação do método será realizada por meio da comparação dos resultados obtidos do experimento com uma taxonomia de referência. A avaliação subjetiva e a validação efetiva com especialistas de domínio não serão abordadas neste trabalho.

1.5 CARACTERIZAÇÃO DA PESQUISA

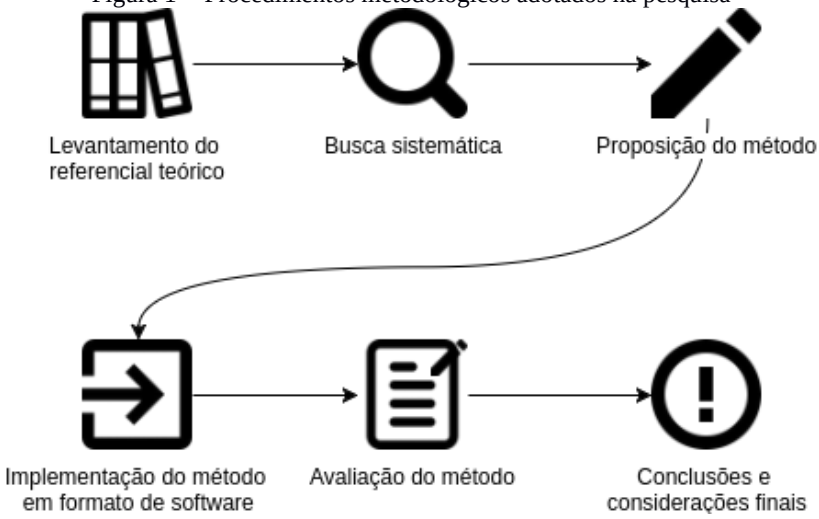
A visão de mundo desse projeto de pesquisa se aproxima mais do paradigma funcionalista. O paradigma funcionalista, de acordo com Morgan (2007, p.16), pressupõe "que a sociedade tem existência concreta e real e um caráter sistêmico orientado para produzir um estado de coisas ordenado e regulado" e possui a missão de compreender a sociedade de maneira que produza conhecimento empírico útil.

A pesquisa é caracterizada como tecnológica, pois possui o objetivo de projetar um método para construção de taxonomias com base no conhecimento científico já produzido. Essa definição de pesquisa tecnológica é suportada por Cupani (2006), ao considerar o método desenvolvido um artefato tecnológico.

O procedimento planejado para este projeto de pesquisa possui caráter exploratório e de levantamento bibliográfico. Segundo Gil (2002), a pesquisa exploratória é caracterizada pelo levantamento bibliográfico de referências com o intuito de verificar sua aplicabilidade dentro da pesquisa proposta. Já o caráter bibliográfico é definido por Gil (2002) como o tipo de pesquisa que é realizado sobre obras disponibilizadas em bibliotecas e repositórios físicos e/ou digitais.

Seguindo as diretrizes acima, a metodologia utilizada para a execução da pesquisa é composta pelos passos descritos na Figura 1.

Figura 1 - Procedimentos metodológicos adotados na pesquisa



Fonte: Elaboração do autor, 2017.

Os procedimentos metodológicos envolvendo o levantamento do referencial, a busca sistemática, a proposição do método, a implementação do método em formato de *software*, a avaliação do método e as conclusões e considerações finais serão abordados em detalhes no capítulo 3

1.6 ADERÊNCIA AO PROGRAMA DE ENGENHARIA E GESTÃO DO CONHECIMENTO (PPGEGC)

Esta pesquisa está inserida dentro da linha de pesquisa Teoria e Prática em Engenharia do Conhecimento do Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento, por lidar diretamente com a disciplina de engenharia de taxonomias e ontologias, que são instrumentos-base para a representação do conhecimento dentro da EC.

Um dos principais objetivos da engenharia de conhecimento consiste em responder a demandas levantadas pela área de gestão do conhecimento, em problemas que oportunizam a aquisição, representação e difusão de conhecimento percebido como propriedade de agentes artificiais em interação (suporte) a agentes humanos. Dessa forma, é necessário construir e manter estruturas de conhecimento, para que seu usufruto possa ser atingido em sua plenitude.

A relevância das ontologias dentro do programa pode ser confirmada através dos trabalhos de conclusão de curso já realizados por diversos alunos do EGC. O Quadro 1 menciona algumas dessas teses e dissertações.

Quadro 1 - Dissertações e teses do EGC que contextualizam a importância do objeto de pesquisa dentro do programa

Dissertação/Tese	**Autor(a)**	**Ano**	**Objetivo**
(D) Um modelo semiautomático para a construção e manutenção de ontologias a partir de bases de documentos não estruturados	CECI, Flavio	2010	Desenvolver um modelo semiautomático que promova suporte ao processo de manutenção de ontologias.
(D) Modelo de ontologia para	CARDENAS , Yuri Gomes	2014	Desenvolver um modelo baseado em ontologias para

representação de jogos digitais de disseminação do conhecimento.			representação de jogos digitais de disseminação de conhecimento.
(T) Ontologia de referência para periódico científico digital	FACHIN, Gleisy Regina Bories	2011	Propor um conjunto de metadados para periódico científico digital, possibili-tando a interoperabilidade por meio do uso de ferramenta ontológica.
(D) Processo para recuperar produtos de inteligência competitiva a partir da memória organizacional: proposta de uma taxonomia para o sistema Mindpuzzle	ROTHER, Rodrigo Garcia	2009	Desenvolver um processo para recuperação de produtos de inteligência competitiva em uma memória organizacional usando uma taxonomia como base.

Analisando o quadro acima, é possível conferir a importância das ontologias dentro do programa – tanto o desenvolvimento quanto a aplicação são foco de pesquisas já realizadas e publicadas.

Este trabalho diferencia-se pelo objetivo de propor um método de extração de taxonomias independentemente de domínios, tendo por elementos base a fonte documental e a hierarquização realizada por base de conhecimento. Trata-se, portanto, de uma contribuição às pesquisas futuras do EGC nos temas de concepção, criação e manutenção automáticas de taxonomias e ontologias.

1.7 ESTRUTURA DO TRABALHO

O conteúdo desse trabalho está segmentado em seis capítulos, com a adição de um conjunto de apêndices que reúne dados estatísticos da execução da implementação do método. São eles:

- Capítulo 1 – Introdução: descreve o problema de pesquisa, objetivos propostos, justificativa, metodologia da pesquisa e aderência do objeto de pesquisa ao PPGEGC;
- Capítulo 2 – Fundamentação teórica: apresenta o levantamento teórico que serviu para o embasamento dessa pesquisa. Os tópicos abrangidos são os conceitos gerais da área de Engenharia do Conhecimento, organização e representação do

conhecimento (com foco na construção de taxonomias), reconhecimento de entidades, bases de conhecimento e a Plataforma Lattes – escolhida como cenário de aplicação do método proposto. A seção também apresenta uma busca sistemática que retrata os esforços realizados na área de construção automatizada de taxonomias;

- Capítulo 3 – Procedimentos metodológicos: descreve os procedimentos de pesquisa adotados nesse trabalho. Esses procedimentos detalham tanto os procedimentos iniciais de pesquisa quanto as soluções adotadas para os problemas encontrados durante o desenvolvimento e implementação do método;
- Capítulo 4 – Método proposto: apresenta e explica o método desenvolvido nesse trabalho. Também descreve os parâmetros e a sua influência dentro dos resultados do método;
- Capítulo 5 – Experimentos e resultados: detalha os experimentos realizados para a verificação do funcionamento do método proposto. O tesauro utilizado como referência para comparação dos resultados (o AGROVOC) é conceitualizado. Por fim, são descritas as estatísticas geradas para cada execução do método e, em sequência, o funcionamento do método é discutido a partir de uma análise sobre os dados obtidos a partir da execução de baterias de testes com diferentes parâmetros;
- Capítulo 6 – Considerações finais e trabalhos futuros: apresenta os pareceres finais dessa pesquisa após a obtenção dos resultados. Também relata os possíveis trabalhos futuros que podem ser derivados a partir dessa pesquisa;
- Apêndices A, B, C e D – Resultados de baterias de testes: permite ao leitor acompanhar as estatísticas obtidas a partir da execução do método para diferentes séries de parâmetros.

2 FUNDAMENTAÇÃO TEÓRICA

Esta seção apresenta os tópicos essenciais para a compreensão do trabalho. Serão abordadas as definições basilares da EGC, interligando o conceito de conhecimento presente na área com o tópico de organização e representação do conhecimento presente na ciência da informação. Com este embasamento, são apresentadas as definições e classificações de taxonomias, que são foco deste trabalho.

A seção também abrange uma busca sistemática que visa retratar o estado atual da área de construção de taxonomias. As publicações encontradas através desse método foram analisadas com o objetivo de extrair informações como abordagem de construção e quais instrumentos automatizados foram aplicados.

Também são contemplados no texto os componentes de apoio para o desenvolvimento do método. São apresentadas as definições de bases de conhecimento, incluindo alguns exemplos encontrados na literatura. Do mesmo modo, são abordados alguns tópicos relevantes para o trabalho sobre a Plataforma Lattes e o tesauro AGROVOC. Esses componentes estão presentes nesta fundamentação objetivando a utilização na construção e verificação do funcionamento do método proposto.

2.1 ENGENHARIA E GESTÃO DO CONHECIMENTO

Atualmente, o conhecimento é visto como ativo de valor para a manutenção e progresso da sociedade. Alguns autores sugerem que o papel que o conhecimento exerce atualmente é semelhante ao das máquinas durante a revolução industrial (LASTRES; FERRAZ, 1999).

O destaque que o conhecimento possui demandou a criação de áreas específicas para o seu tratamento, armazenamento e utilização de forma eficiente. De forma geral, as áreas de engenharia e gestão do conhecimento possuem como foco pesquisar, conceber, desenvolver e aplicar sistemas de apoio para tarefas intensivas em conhecimento (SANTOS; ZANCANARO; NAKAYAMA, 2015).

As próximas subseções aprofundam a temática da EGC dentro do objeto de estudo tratado por este trabalho. São apresentadas inicialmente as fundamentações de dado, informação e conhecimento dentro da área, com objetivo de apresentar as definições adotadas para estes conceitos neste trabalho. Em sequência, são apresentadas as áreas de concentração da EGC abordadas nesta pesquisa.

2.1.1 Dado, informação e conhecimento

Os conceitos de dado, informação e conhecimento – embora sejam utilizados por vezes como sinônimos na literatura – possuem semânticas diferentes, embora relacionadas, no contexto da engenharia e gestão do conhecimento.

O conceito de dado, de acordo com Ponchirolli e Fialho (2005), é o que possui menor significado quanto à abrangência semântica de seu conteúdo. Um dado é um registro sobre um determinado evento, que pode ser facilmente obtido e catalogado dentro de um sistema. Ele pode ser registrado em forma gestual, gráfica, escrita ou oral.

Isoladamente, um dado não tem significado por não possuir contexto por si só, nem significado implícito. No entanto, através de um conjunto de dados é possível realizar correlações e comparativos, que podem gerar informação.

Informação, por sua vez, é um conceito que possui um significado semântico maior quando comparado ao de dado. Uma informação pode ser definida como um conjunto de dados que, ao serem processados, possuem significado e um contexto para um sistema (PONCHIROLLI; FIALHO, 2005).

De acordo com Nonaka e Takeuchi (2008), a informação é um fluxo de mensagens que fornece um ponto de vista para a interpretação de eventos ou objetos (que podem ser vistos como dados). Ela constitui um espaço intermediário entre o dado e a geração de conhecimento, sendo dessa forma um meio material ou necessário para extrair e construir o conhecimento.

Para Fialho e outros (2006), a informação é um conjunto de dados, dispostos de forma que haja um significado capaz de representar padrões e ativar significados durante o seu processamento pelo cérebro. Essa disposição deve explicitar a correlação entre os vários fatos e suas implicações para os indivíduos. "A informação é a base do conhecimento. " (FIALHO et al, 2006).

Conhecimento, por sua vez, é todo o conjunto de dados e informação reunido por pessoas para uso em uma ação, com o objetivo de realizar tarefas e criar mais informação (SCHREIBER et al., 1999).

Ainda de acordo com Schreiber e outros (1999), o conhecimento possui dois aspectos distintos: a existência de propósito, já que o mesmo é utilizado em busca do alcance de um objetivo, e uma capacidade generativa, já que uma das principais funções do conhecimento é gerar novo conhecimento.

A definição de conhecimento pode ser considerada parcialmente difusa. De acordo com Fialho e outros (2006), isso ocorre pelo conhecimento ser sensível ao contexto no qual está inserido. Os conceitos de dado e de informação também possuem essa relação com o contexto, o que torna o limite entre eles menos rígidos.

O quadro a seguir apresenta uma síntese comparativa entre os conceitos de dado, informação e conhecimento.

Quadro 2 - Comparativo entre dado, informação e conhecimento

Conceito	**Característica**	**Exemplo**
Dado	Não interpretado; Não preparado	SOS
Informação	Significado atrelado aos dados	Alerta de emergência
Conhecimento	Adiciona propósito e competência a informação; Possui potencial de gerar ação	Iniciar operação de resgate

Fonte: SCHREIBER et al, 1999, p. 4

2.1.2 Gestão do Conhecimento

A gestão do conhecimento é um dos instrumentos utilizados para compor a estratégia de uma empresa. Ela é a disciplina que "trata da prática de agregar valor à informação e distribuí-la, tendo como tema central o aproveitamento dos recursos existentes na empresa" (PONCHIROLLI; FIALHO, 2005, p. 130).

O termo surgiu na década de 1980 em pesquisas na área de inteligência artificial, que buscavam verificar como o aprendizado poderia ser simulado e auxiliado por meios tecnológicos (SIQUEIRA, 2005). Embora o termo tenha sido cunhado recentemente, a prática da gestão do conhecimento já ocorre na humanidade de forma intuitiva há milhares de anos.

Através da gestão do conhecimento, os bens intelectuais de uma organização são otimizados de forma que mais produtividade, valor e competitividade possam ser extraídos. O ponto chave de todo o processo é a ação sobre a informação, que será utilizada para a elaboração de estratégias e tomada de decisão, aprendizagem ou adaptação a mudanças. (PONCHIROLLI; FIALHO, 2005; SIQUEIRA, 2005).

O resultado dessas ações objetiva a formação de vantagens competitivas, o que destaca a organização perante as outras e aumentar o seu fator de concorrência. Entre as vantagens competitivas que podem ser formadas no processo, Siqueira (2005) destaca as seguintes:

- Conhecimentos e habilidades dos indivíduos e das equipes;
- Competências para construção e uso de sistemas de informação, bases de dados, software e equipamentos (que podem ser denominadas vantagens sobre sistemas físicos);
- Formação de processos organizados para otimização do fluxo do conhecimento (que podem ser denominadas vantagens sobre sistemas gerenciais);
- Elucidação de valores e normas da organização de forma mais clara, que constituem as diretrizes organizacionais.

A gestão do conhecimento não é uma disciplina que atua sozinha em uma organização. Suas soluções são elaboradas a partir da área da engenharia do conhecimento, que estrutura as demandas de gestão e fornece um sistema capaz de atender a solicitação. A atuação de ambas as áreas em conjunto é necessária para que o tratamento do conhecimento organizacional seja realizado de forma correta.

Um exemplo desta atuação em conjunto são os sistemas de memória organizacional, que empregam métodos de codificação de conhecimento desenvolvidos no campo da EC (MELGAR-SASIETA; BEPPLER; PACHECO, 2011). Portais corporativos, que captam todo o manancial de conteúdo informacional e de conhecimento organizacional para auxílio na tomada de decisão nos níveis estratégico, tático e operacional, também demonstram esta associação (TERRA; BAX, 2003).

2.1.3 Engenharia do conhecimento

A engenharia do conhecimento é uma disciplina que nasceu a partir das demandas geradas pela gestão do conhecimento.

Segundo Studer, Benjamins e Fensel (1998), o surgimento da área possui semelhanças com a motivação para o estabelecimento da engenharia de software. A criação em forma artesanal de sistemas baseados em conhecimento funcionava bem em protótipos acadêmicos, mas ao se transferir essa abordagem para sistemas corporativos ocorreram problemas de escalabilidade e de manutenção que tornaram o processo inaplicável.

Para Schreiber e outros (2002), a engenharia do conhecimento foi uma das disciplinas que avançaram bastante por conta da Era da Informação, em um fenômeno semelhante ao que ocorrera com a expansão das áreas de engenharias elétrica e mecânica por conta da Revolução Industrial.

A área tem por objetivo transformar o processo de criação de sistemas baseados em conhecimento em um processo de engenharia, com o aporte de metodologias, ferramentas e linguagens necessárias para transcrever um modelo fidedigno as necessidades do projeto (STUDER; BENJAMINS; FENSEL, 1998).

Junto com a criação da engenharia do conhecimento, houve uma mudança de paradigma da criação dos sistemas baseados em conhecimento. Inicialmente, os sistemas de conhecimento construídos utilizavam uma abordagem baseada na transferência de conhecimento. Com o estabelecimento da disciplina, a abordagem de transferência foi depreciada em virtude do surgimento da abordagem baseada em modelagem.

Nessa abordagem baseada em modelagem, o conhecimento normalmente passa por um processo de cinco estágios ao ser transformado de conhecimento humano em conhecimento estruturado para um sistema baseado em conhecimento (KENDAL; CREEN, 2007, p. 8):

- Aquisição do conhecimento: envolve a obtenção do conhecimento a partir das fontes de informação disponíveis, as quais abrangem pessoas, livros, vídeos, bases de dados e internet, por exemplo;
- Validação do conhecimento: abrange a validação do conhecimento por meio de casos de teste;
- Representação do conhecimento: trata do mapeamento do conhecimento obtido e validado e sua codificação na base de conhecimento;
- Inferência: envolve a criação de *links* na base de conhecimento, de forma que o sistema baseado em conhecimento que opere sobre a base consiga fazer decisões ou promover auxílio ao usuário;
- Explicação e justificativa: abrange o formato de processamento do conhecimento pelo programa, com o objetivo de alcançar a interpretação correta da necessidade do usuário e a apresentação da explicação de como o resultado foi obtido pelo sistema.

A engenharia e gestão do conhecimento utilizam instrumentos presentes na ciência da informação para auxílio na tarefa de estruturar, organizar e representar adequadamente o conhecimento. Estes tópicos são o tema da próxima seção.

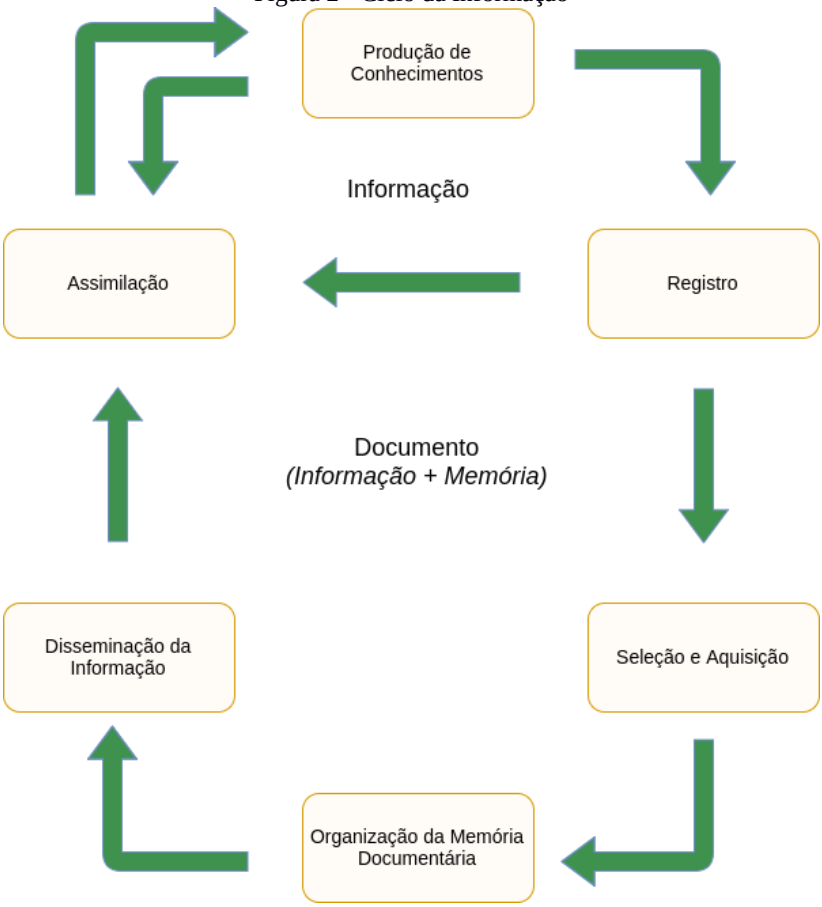
2.2 ORGANIZAÇÃO E REPRESENTAÇÃO DO CONHECIMENTO

De acordo com Brascher e Café (2008, p. 6), a organização do conhecimento compreende "todo tipo de método de indexação, resumo, catalogação, classificação, gestão de arquivos, bibliografia e a criação de bases de dados bibliográficas e textuais para a recuperação da informação". O principal objetivo da área é proporcionar a recuperação do conhecimento de forma eficaz pelos usuários. (ALVARENGA, 2003)

O domínio da área de organização e representação do conhecimento abrange o conhecimento como objeto de pesquisa, tendo como atividades principais sobre esse objeto de pesquisa a sua organização e representação. Dessas atividades resultam instrumentos, processos e produtos, que por sua vez auxiliam na organização e representação do conhecimento em outras áreas de ciência em geral. (FUJITA, 2008)

De acordo com Dodebei (2002 apud TRISTÃO; FACHIN; ALARCON, 2004), a organização do conhecimento possui uma dimensão cíclica. Esse ciclo é composto por seis fases: produção de conhecimentos, registro, seleção e aquisição, organização da memória documentária, disseminação de informação e assimilação. A Figura 2 apresenta este ciclo.

Figura 2 - Ciclo da Informação



Fonte: Adaptado de Dodebei (2002 apud TRISTÃO; FACHIN; ALARCON, 2004).

Embora a memória documentária compreenda a execução de todos esses passos, existe um atalho nesse ciclo que ignora o trabalho documental, tratando diretamente da aquisição de conhecimento a partir do momento no qual é feito seu registro em forma de informação.

A área de organização e representação do conhecimento utiliza-se essencialmente de sistemas para a tarefa da organizar o conhecimento – também denominados esquemas de organização do conhecimento por alguns autores. (CARLAN, 2010). Os sistemas de organização de

conhecimento são utilizados desde os tempos remotos, nas diversas áreas de conhecimento humano.

Hodge (2000) e Boccato (2011) apresentam uma classificação desses sistemas de acordo com sua função principal: sistemas de listas de termos, sistemas de categorização e classificação e sistemas de listas de relacionamentos.

Os sistemas de listas de termos possuem como função principal agrupar listas de termos, frequentemente também com as suas definições. Os sistemas de listas de termos compreendem (HODGE, 2000):

- Arquivos de autoridade: São listas de termos que apresentam os termos utilizados em determinado domínio e suas variações. Sua principal função é explicitar os termos preferenciais. Não possui uma organização complexa, podendo conter alguns níveis hierárquicos para facilitar a pesquisa em caso de listas muito grandes;
- Glossários: São listas de termos com suas definições. Normalmente estes termos pertencem a um ambiente ou campo de estudo específico. Raramente incluem variações ou sinônimos;
- Dicionários: São listas alfabéticas de termos, com suas definições. Diferenciam-se dos glossários por proverem também a etimologia dos termos, sinônimos e variações. O escopo de dicionários é mais generalizado, procurando atingir múltiplas disciplinas e áreas do conhecimento;
- *Gazeteers*/dicionários geográficos: São listas de termos referentes a localizações geográficas. Cada termo pode ser classificado de acordo com classes categorizadoras, como rios, escolas, igrejas, entre outras;

Já os sistemas de categorização e classificação possuem como princípio a criação e manutenção de conjuntos de elementos através de categorias e classes. Elas agrupam os elementos que possuem características em comum, constituindo sua própria unidade e distinguindo seu agrupamento de outros. (TRISTÃO; FACHIN; ALARCON, 2004). Os sistemas de categorização e classificação abrangem (HODGE, 2000; BOCATTO, 2011):

- Cabeçalhos de assunto: Estes esquemas de organização agrupam um conjunto de termos dentro de um vocabulário controlado para designar o conteúdo de uma coleção.

Geralmente apresentam uma estrutura hierárquica rasa, com poucos aprofundamentos;

- Esquemas de classificação, taxonomias e esquemas de categorização: Estes sistemas de organização do conhecimento provêm recursos para agrupar termos em tópicos mais abrangentes. Taxonomias têm sido usadas recorrentemente no *design* orientado a objetos e em sistemas de gestão do conhecimento para indicar grupos de objetos que possuem determinadas características.

Por fim, os sistemas de listas de relacionamento possuem ênfase em demonstrar as relações entre os conceitos. Os sistemas de listas de relacionamento englobam (HODGE, 2000; BOCATTO, 2011):

- Tesauros: Apresentam listas de conceitos que estão interligados com outros conceitos através de uma estrutura hierárquica, abrangendo também relações de sinonímia e de associação. Geralmente são desenvolvidos de forma direcionada ao domínio de uma disciplina ou de um produto em específico;
- Redes semânticas: São redes que conectam conceitos em formato de grafo. As conexões entre os conceitos abrangem as apresentadas nos tesauros, incluindo também relações de causa e efeito e outras detectadas através do processamento de linguagem natural.
- Ontologias: São modelos conceituais específicos. As ontologias são capazes de representar relacionamentos complexos entre os termos, regras e axiomas que estão ausentes das redes semânticas. Elas também incorporam a estrutura hierárquica presente nos tesauros.

Dentro do cenário atual da organização do conhecimento se destacam os sistemas de classificação, taxonomias e esquemas de categorização e tesauros, principalmente em decorrência do avanço nas últimas décadas nas áreas da *web* e redes digitais de troca de informação e nas demais TICs de forma geral. (TRISTÃO; FACHIN; ALARCON, 2004).

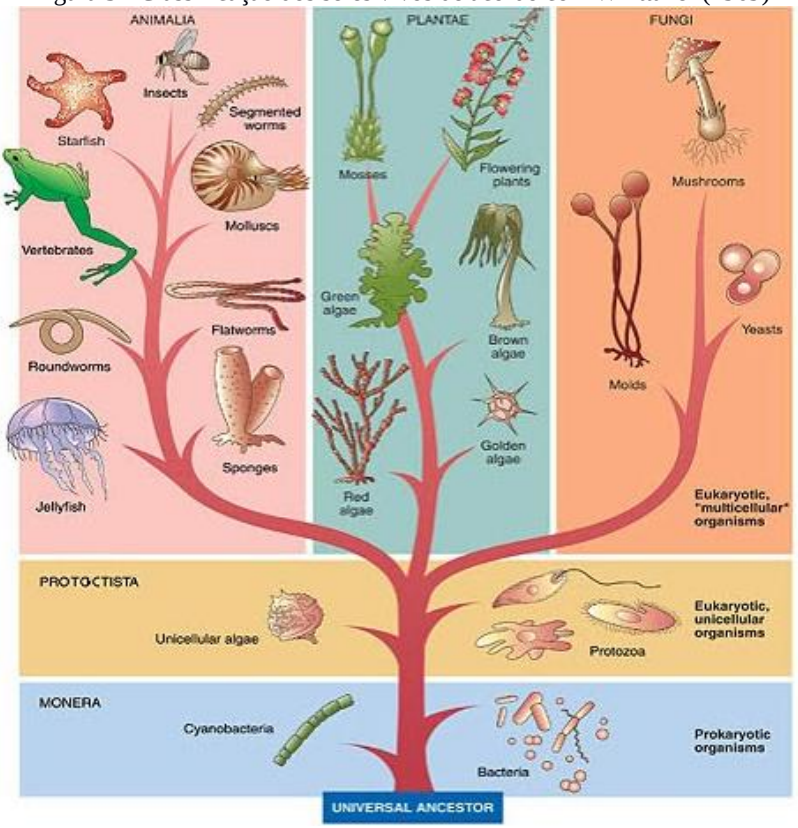
O foco desta pesquisa são as taxonomias, que são tema da próxima seção.

2.3 TAXONOMIAS

As taxonomias são elementos classificativos em estrutura de hierarquia, que estão disseminados em vários campos que possuem a necessidade de classificação de conhecimento.

Em sua etimologia, o termo taxonomia se origina dos termos gregos *taxis* (ordem) e *nomos* (lei, ou norma). Seu primeiro registro de utilização data de 1735, na publicação da primeira edição da obra *Systema Naturae*, pelo cientista e médico sueco Karl Von Linné. A obra acabou por tornar esse termo conhecido dentro do domínio da biologia (AGANETTE; ALVARENGA; SOUZA, 2010). A classificação taxonômica é a principal forma de categorização de seres vivos na biologia, sendo usada por cientistas e pesquisadores do mundo inteiro. A Figura 3 representa em forma gráfica um trecho desta classificação.

Figura 3 - Classificação dos seres vivos de acordo com Whittaker (1969)



Fonte: García, 2013

O termo não ficou restrito a área biológica. As taxonomias foram adotadas em outros campos onde existe a necessidade de classificação hierárquica. De acordo com Plosker (2005, p.58), "qualquer classificação ou divisão em grupos ordenados ou categorias" constitui uma taxonomia. Pela definição do autor, as taxonomias podem ser obtidas facilmente a partir de classificações hierárquicas realizadas no cotidiano.

Meijer, Frasincar e Hogenboom (2014, p.78, tradução do autor) propõem uma definição mais específica. Segundo os autores, "uma taxonomia é uma hierarquia de conceitos na qual são armazenadas as relações de abrangência e estreitamento entre diferentes conceitos".

Implicitamente, classificações hierárquicas possuem uma relação de abrangência e estreitamento, apresentando conceitos mais genéricos em níveis superiores e conceitos mais específicos em níveis inferiores. Entende-se que a definição de Meijer, Frasincar e Hogenboom, dessa forma, é uma definição mais explícita da ideia apresentada por Plosker, considerando a regra de abrangência e estreitamento.

A norma internacional ANSI/NISO Z39.19 (2005, p. 9, tradução do autor) também apresenta um conceito de taxonomia, que será utilizado no contexto desse trabalho. O conceito segue a mesma linha proposta pelos autores citados anteriormente, porém com uma definição ainda mais clara:

> Uma coleção de termos de um vocabulário controlado organizados em uma estrutura hierárquica. Cada termo numa taxonomia está em um ou mais relacionamentos de pais e filhos (mais abrangentes ou mais específicos) com outros termos da taxonomia.

A definição proposta pela norma também acrescenta a necessidade da construção da taxonomia através de um vocabulário controlado. Esta exigência está relacionada com a própria validade da taxonomia, que precisa conter um vocabulário controlado para expressar corretamente os níveis de abrangência ou especificidade dos conceitos dentro de um grupo de pessoas. Isso também ocorre em uma ontologia, que, em sua definição, já se apresenta como uma conceitualização compartilhada (GRUBER, 1993).

Para Knijff, Fransincar e Hogenboom (2013) e Tu, L. Chen e G. Chen (2015), a taxonomia é uma forma específica de ontologia – uma especificação formal e explícita de uma conceitualização compartilhada (GRUBER, 1993) – contendo somente relações do tipo “é um/é um tipo de”. Tais relações se originam da semântica entre os termos. Um exemplo disso ocorre ao classificar a conexão entre as palavras carro e veículo. Um carro é um veículo ou um tipo de veículo; no entanto, vale notar que nem todo veículo é um carro.

Vital (2007) denota que diversos autores relacionam o aparecimento e uso do termo em ambientes digitais ao surgimento de formas automatizadas de criação de informação. Essa adoção ocorre principalmente pelo fato da quantidade de informação catalogada de forma digital ter aumentado exponencialmente, sendo necessários

instrumentos cada vez mais inteligentes de filtragem para recuperá-las de forma eficaz.

Vital (2007, p. 47) ainda acrescenta que a criação de taxonomias, quando voltadas ao uso em ambientes organizacionais, está sujeita a um processo controlado:

> Taxonomias estão voltadas para a organização das informações em ambientes específicos, visando à recuperação eficaz. Para isso, estabelecem parâmetros em todo o ciclo de produção informacional, no qual profissionais distribuídos por espaços físicos distintos participam do processo de criação do conhecimento de forma organizada.

A participação de diversos profissionais faz-se necessária pelas competências distintas que possuem. Uma taxonomia de domínio que possa atender um domínio organizacional por completo depende do conhecimento especializado de diversos profissionais. Somente com a junção de competências é possível selecionar os conceitos de forma que estes reúnam todas as informações relevantes para a organização.

Para Conway e Sligar (2002), o objetivo das taxonomias nesses ambientes corporativos é fornecer um mapeamento entre conceitos que permita a conexão dos usuários com o conhecimento certo de forma rápida. Em caso de informações semiestruturadas (armazenadas de forma sistematizada em servidores da organização) ou não estruturadas (armazenadas em discos rígidos de estações de trabalho e em servidores locais), as taxonomias provêm uma ligação entre conceitos e conteúdos.

Conway e Sligar ainda definem as taxonomias corporativas em três tipos: taxonomias descritivas, taxonomias de navegação e taxonomias de gerenciamento de dados.

As taxonomias descritivas permitem a realização de recuperação de informação. Elas são constituídas por um conjunto de vocábulos controlados, que podem ser utilizados como marcadores dentro dos metadados dos arquivos. Mesmo com atualizações em nomes dos marcadores ou no conteúdo dos arquivos, é possível manter a associação. Um exemplo são os nomes comerciais de produtos, que podem sofrer atualizações até a versão final - no entanto, o significado continua sendo o mesmo.

Esse tipo de taxonomia, quando construída em formato de tesauro, possibilita o uso do recurso de expansão de consulta. O formato de tesauro caracteriza a taxonomia pela existência de termos principais ou mais comuns, adicionando também os sinônimos e termos

relacionados. A expansão de consulta é um dos recursos que enriquecem resultados de buscas, sendo uma das ferramentas que melhoram a recuperação de informação.

Já as taxonomias de navegação possuem como foco disponibilizar conhecimento através da navegação. Em vez de auxiliar na tarefa de expansão de consulta, esse tipo de taxonomia apresenta uma árvore baseada nos modelos de conhecimento dos engenheiros de conhecimento sobre como a informação é organizada.

Taxonomias de navegação eram a base de diretórios de conteúdo, tais como Yahoo, Cadê e MSN, comuns na *web* no início do século. Esses sites acabaram perdendo espaço para buscadores baseados em palavras-chave e foram desativados, mas projetos *open-source* ainda resgatam a essência. O portal DMOZ[1], hospedado pela AOL, é um *site* mantido por voluntários que agrupa sites organizados de acordo com uma taxonomia de navegação.

Em uma taxonomia de navegação é possível encontrar termos que não seriam empregados em uma taxonomia descritiva. A expressão "Matricule-se", por exemplo, dentro do item Disciplinas, pode constituir um elemento de navegação de uma taxonomia para um sistema acadêmico, apesar de não constituir em sua essência uma disciplina. O termo, no entanto, é compreensível para o usuário final.

Em geral, as taxonomias de navegação são desenvolvidas especialmente para um *site*, sistema ou portal, possuindo regras menos restritivas que as taxonomias descritivas.

Por fim, as taxonomias de gerenciamento de dados servem para definir um dicionário de suporte para transações de negócio. Elas não são necessariamente hierárquicas, e contem termos relacionados aos tipos de dados armazenados nas organizações, com o intuito de facilitar trabalhos posteriores com a mesma informação.

Um exemplo de utilização desse tipo de taxonomia é em sistemas CRM, nos quais os usuários armazenam informações relativas a clientes. A existência de uma taxonomia de gerenciamento de dados permite com que uma equipe de trabalho realize o processamento analítico dos dados posteriormente – em uma ferramenta de OLAP, por exemplo – com maior garantia de seleção dos campos de dados corretos no sistema.

Em alguns cenários, as taxonomias de gerenciamento de dados podem ser as próprias taxonomias descritivas, economizando recursos no momento da elaboração e trazendo uma experiência de uso melhor para os usuários.

---

1 Disponível em: <http://www.dmoz.org/>

Quanto à estrutura da taxonomia, Blackburn (2006) classifica-as como de assunto, de unidade de negócio e funcional.

As taxonomias de assunto são compostas por termos de um vocabulário controlado organizados em ordem alfabética. Os termos mais genéricos ficam nos níveis mais abrangentes, e a árvore de termos é expandida em direção a termos mais específicos.

É comum essas taxonomias conterem uma seção de termos relacionados ou preferenciais, que guiem o usuário a partir do termo que ele conhece até o termo preferencial presente na taxonomia. Isto ocorre visto a dificuldade de estabelecer uma taxonomia de assuntos com termos que são universalmente reconhecidos.

Esse tipo de taxonomia assemelha-se ao perfil de taxonomias descritivas proposto por Conway e Sligar. Sua base estrutural é também a de um tesauro simples.

As taxonomias de unidade de negócio são compostas pelas hierarquias presentes em organogramas. Os documentos são categorizados de acordo com a unidade de negócio que os gerencia, cada qual é representada por um termo na árvore.

O maior problema desse tipo de taxonomia, segundo Blackburn (2006), é a dificuldade de enquadrar documentos organizacionais em termos específicos da taxonomia, principalmente quando o documento circula por mais de um departamento. Outro problema são as mudanças na hierarquia organizacional, que por sua vez refletem no organograma e nos documentos associados.

Por outro lado, por respeitar o próprio organograma da organização, o uso da estrutura da taxonomia torna-se transparente para os usuários, sendo necessário menos tempo para o seu aprendizado.

Por fim, as taxonomias funcionais são organizadas de acordo com as funções, atividades e transações que produzem os documentos organizacionais. Estas funções estão transcritas dentro dos processos de negócio da organização.

Esse tipo de taxonomia é mais sustentável ao longo do tempo em organizações, pois, enquanto mudanças no organograma da organização ocorrem, processos de negócio e atividades tendem a se manter os mesmos. Além disso, os conceitos relacionados a processos de negócio tendem a ser mais definidos. Isso permite alterações mais pontuais na estrutura para adicionar o suporte a novos conceitos.

Taxonomias funcionais não permitem em sua essência a associação de documentos de projeto e de coleções de documentos relacionadas a uma entidade em específico, devido ao fato desses documentos serem originados a partir de um conjunto de diferentes

atividades. Segundo Blackburn (2006), uma forma de controlar esse problema é a utilização de referencias cruzadas de metadados. Esses metadados podem ser utilizados com o objetivo de criar uma listagem virtual da coleção de documentos associados a uma entidade em específico.

O Quadro 3 ilustra os benefícios e as desvantagens de se utilizar cada um dos tipos estruturais de taxonomia.

Quadro 3 - Prós e contras da utilização de cada um dos tipos estruturais de taxonomia.

Tipos	**Prós**	**Contras**
Assunto	Abordagem mais reconhecida pelos usuários; Possibilidade de aproveitar hierarquias já existentes.	Requer conhecimento da terminologia empregada ou o uso de tesauros de apoio;
Unidade de negócio	Familiar para os usuários (que já conhecem o organograma da organização);	Mudanças no organograma da organização implicam em mudanças na taxonomia; Documentos compartilhados são difíceis de enquadrar em uma única categoria;
Funcional	Suporta melhor mudanças organizacionais;	Dificulta a associação de coleções de documentos.

Fonte: Adaptado de Blackburn (2006)

O uso de taxonomias mostra-se útil em ferramentas de busca de informação, classificação, navegação, dentre outras aplicações. (MEIJER; FRASINCAR; HOGENBOOM, 2014). Além disso, as taxonomias desempenham o papel de facilitador de comunicação entre pessoas envolvidas no processo de busca ou categorização de uma informação. O fato de propor uma hierarquia fornece navegabilidade

fácil na estrutura, o que facilita a recuperação de informação. (SANTOS; CORRÊA, 2011).

2.3.1 Abordagens para a construção de taxonomias

Segundo Punera (2007), a construção de taxonomias tradicionalmente é feita de forma manual. Mesmo taxonomias relativamente recentes, como as taxonomias Yahoo! e DMOZ, foram construídas por funcionários e voluntários.

Essa construção muitas vezes é realizada sem a utilização de algum processo base consolidado. O material utilizado para a coleta de conceitos é levantado, analisado e posteriormente organizado em forma hierárquica.

Alguns autores propõem a elaboração de taxonomias com base em um método fixo ou em uma norma. Dessa forma, a falta de passos específicos inerentes à construção de taxonomias de forma manual não prejudica a qualidade final do artefato construído. Entretanto, a utilização de tais metodologias-base não garante maior velocidade na construção desses artefatos, diminuindo assim a motivação de se utilizar este ferramental.

A utilização de um processo manual, apesar da precisão do resultado final, consome tempo, exige recursos e não consegue abranger totalmente o domínio de análise, em casos como o mapeamento do corpus da *World Wide Web*, por exemplo. (PUNERA, 2007). Com isso, a criação de taxonomias de forma automatizada é objeto de estudo de diversos autores, com o intuito de reduzir seu custo.

Essas abordagens podem ser classificadas de acordo com o nível de interação entre o usuário e a abordagem. Métodos que exigem a intervenção humana são classificados como semiautomatizados. Já abordagens que não dependem de parâmetros ou de validação durante o processo são classificadas como automatizadas.

2.3.2 Busca sistemática sobre construção de taxonomias

Para a averiguação do estado da área de construção de taxonomias nos últimos anos, foi realizada uma busca sistemática para a constatação das abordagens utilizadas para a construção de taxonomias. O Quadro 4 lista as publicações que foram selecionadas após a análise.

Quadro 4 - Publicações selecionadas da busca sistemática para análise

Artigo	**Autor(es)**	**Ano**	**Periódico**

An algebraic taxonomy for locus computation in dynamic geometry.	Abánades, M. Á., Botana, F., Montes, A., & Recio, T.	2014	Computer-Aided Design, 56, 22–33
Rich Mobile Applications: Genesis, taxonomy, and open issues.	Abolfazli, S., Sanaei, Z., Gani, A., Xia, F., & Yang, L. T.	2014	Journal of Network and Computer Applications, 40, 345–362.
Seamless application execution in mobile cloud computing: Motivation, taxonomy, and open challenges.	Ahmed, E., Gani, A., Khurram Khan, M., Buyya, R., & Khan, S. U.	2015	Journal of Network and Computer Applications, 52, 154–172.
Application optimization in mobile cloud computing: Motivation, taxonomies, and open challenges.	Ahmed, E., Gani, A., Sookhak, M., Hamid, S. H. A., & Xia, F.	2015	Journal of Network and Computer Applications, 52, 52–68.
Automatic extraction of ontological relations from Arabic text.	Zamil, M. G. H. A., & Al-Radaideh, Q.	2014	Journal of King Saud University - Computer and Information Sciences, 26(4), 462–472.
A taxonomy for decision support capabilities of enterprise content management systems.	Alalwan, J. A.	2013	The Journal of High Technology Management Research, 24(1), 10–17.
Providing metrics and automatic enhancement for hierarchical taxonomies.	Beydoun, G., García-Sánchez, F., Vincent-Torres, C. M., Lopez-Lorca, A. A., & Martínez-Béjar, R.	2013	Information Processing & Management, 49(1), 67–82.
Marine ecology service reuse through taxonomy-oriented SPL development.	Buccella, A., Cechich, A., Pol'la, M., Arias, M., del	2014	Computers & Geosciences, 73, 108–121.

	Socorro Doldan, M., & Morsan, E.		
Towards automatic conflict detection in home and building automation systems.	Carreira, P., Resendes, S., & Santos, A. C.	2014	Pervasive and Mobile Computing, 12, 37–57.
A Method for Refining a Taxonomy by Using Annotated Suffix Trees and Wikipedia Resources.	Chernyak, E., & Mirkin, B.	2014	Procedia Computer Science, 31, 193–200.
Adversarial attacks against intrusion detection systems: Taxonomy, solutions and open issues.	Corona, I., Giacinto, G., & Roli, F.	2013	Information Sciences, 239, 201–225.
Domain taxonomy learning from text: The subsumption method versus hierarchical clustering.	De Knijff, J., Frasincar, F., & Hogenboom, F.	2013	Data & Knowledge Engineering, 83, 54–69.
Fuzzy nearest neighbor algorithms: Taxonomy, experimental analysis and prospects.	Derrac, J., García, S., & Herrera, F.	2014	Information Sciences, 260, 98–119.
Towards a unified taxonomy and architecture of cloud frameworks.	Dukaric, R., & Juric, M. B.	2013	Future Generation Computer Systems, 29(5), 1196–1210.
A survey of Cloud monitoring tools: Taxonomy, capabilities and objectives.	Fatema, K., Emeakaroha, V. C., Healy, P. D., Morrison, J. P., & Lynn, T.	2014	Journal of Parallel and Distributed Computing, 74(10), 2918–2933.
Succeeding metadata based annotation scheme and visual tips for the automatic assessment of video aesthetic quality in car commercials.	Fernández-Martínez, F., Hernández García, A., & Díaz de María, F.	2015	Expert Systems with Applications, 42(1), 293–305.
A survey of fingerprint classification Part I: Taxonomies on feature extraction methods and	Galar, M., Derrac, J., Peralta, D., Triguero, I.,	2015	Knowledge-Based Systems, 81, 76–97.

learning models.	Paternain, D., Lopez-Molina, C., … Herrera, F.		
Review and taxonomies of assembly and disassembly path planning problems and approaches.	Ghandi, S., & Masehian, E.	2015	Computer-Aided Design, 67, 58–86.
Classification, representation, and automatic extraction of deformation features in sheet metal parts.	Gupta, R. K., & Gurumoorthy, B.	2013	Computer-Aided Design, 45(11), 1469–1484.
Network attacks: Taxonomy, tools and systems.	Hoque, N., Bhuyan, M. H., Baishya, R. C., Bhattacharyya, D. K., & Kalita, J. K.	2014	Journal of Network and Computer Applications, 40, 307–324.
Crowdsourcing: A taxonomy and systematic mapping study.	Hosseini, M., Shahri, A., Phalp, K., Taylor, J., & Ali, R.	2015	Computer Science Review.
Application partitioning algorithms in mobile cloud computing: Taxonomy, review and future directions.	Liu, J., Ahmed, E., Shiraz, M., Gani, A., Buyya, R., & Qureshi, A.	2015	Journal of Network and Computer Applications, 48, 99–117.
A semantic approach for extracting domain taxonomies from text.	Meijer, K., Frasincar, F., & Hogenboom, F.	2014	Decision Support Systems, 62, 78–93.
Consensus under a fuzzy context: Taxonomy, analysis framework AFRYCA and experimental case of study.	Palomares, I., Estrella, F. J., Martínez, L., & Herrera, F.	2014	Information Fusion, 20, 252–271.
Open Directory Project based universal taxonomy for Personalization of Online (Re)sources.	Ševa, J., Schatten, M., & Grd, P.	2015	Expert Systems with Applications, 42(17-18), 6306–6314.
Automatic multi-way domain	Tu, D., Chen,	2015	Neurocomputing,

concept hierarchy construction from customer reviews.	L., & Chen, G.		147, 472–484.
An extensible pattern-based library and taxonomy of security threats for distributed systems.	Uzunov, A. V., & Fernandez, E. B.	2014	Computer Standards & Interfaces, 36(4), 734–747.
A taxonomy of privacy-preserving record linkage techniques.	Vatsalan, D., Christen, P., & Verykios, V. S.	2013	Information Systems, 38(6), 946–969.

A maior parte dos trabalhos selecionados no estudo (64,3%) apresenta um processo de construção de taxonomias manual. Os autores efetuam a revisão teórica da área de aplicação da taxonomia, com o intuito de identificar os conceitos que devem ser colocados na estrutura. No trabalho de Fernández-Martínez, Hernández García, e Díaz de María (2015), existe a fundamentação da taxonomia proposta em uma taxonomia já construída anteriormente. As áreas de aplicação dos artefatos produzidos variam entre a área de informática, manufatura e modelos de resolução de problemas de negócio.

Já 14,3% mencionam um método ou abordagem para a construção de taxonomias claramente. É o caso dos trabalhos de Buccella e outros (2014), Hosseini e outros (2015), Ševa, Schatten e Grd (2015) e Uzunov e Fernandez (2014).

No trabalho de Buccella e outros (2014), a taxonomia é baseada em um padrão definido pelas ISOs 19101-2 e 19119. A construção da taxonomia segue a normatização definida pelas ISOs, estendendo a taxonomia a partir dos conceitos já existentes e estruturados nas normas.

Hosseini e outros (2015) realizam uma análise do domínio a partir de um processo de revisão sistemática e elaboram uma classificação hierárquica sobre os diferentes aspectos do *crowdsourcing*. Unindo todas as tabelas classificativas apresentadas, é possível construir uma taxonomia hierárquica.

Já Ševa, Schatten e Grd (2015) defendem a construção de taxonomias baseadas em diretórios de conteúdo: os autores criaram taxonomias a partir de dos repositórios da DBpedia, WikiData e ODP, dos quais o último é foco de trabalho do artigo coletado na pesquisa.

Uzunov e Fernandez (2014) elaboram uma abordagem mista, construindo uma taxonomia a partir do método artesanal (embora fundamentado em classes de taxonomias já existentes), mas propondo outro método em seguida para a criação de taxonomias especializadas a

partir da taxonomia inicial. No final, essas especializações podem ser reunidas para a montagem de uma única taxonomia aglutinadora.

No contexto da pesquisa realizada nesta dissertação, as abordagens mais relevantes são as automatizadas e as semiautomatizadas. Elas constituem proporções de 14,3% e 7,1%, respectivamente, das publicações coletadas na análise.

Nesse contexto, o quadro apresenta trabalhos como o de Zamil e Al-Radaideh (2014), que propõe a geração automatizada de taxonomias a partir do algoritmo de detecção de padrões de Hearst. Os conceitos coletados são filtrados e passam por um processo de *stemming*. O processo de *stemming* é um tratamento linguístico, no qual são removidos alguns padrões de sufixos das palavras, que em geral definem singularidade/pluralidade e conjugações verbais. Por fim, são detectados padrões dentro das sentenças existentes nos textos, através da utilização da suíte de ferramentas GATE, que permite realizar o *tagging* dos termos que compõem as sentenças, definindo-os como substantivos, adjetivos, verbos, etc.

Já De Knijff, Frasincar e Hogenboom (2013) propõem uma estratégia baseada no método da subordinação e da clusterização. Os termos são inicialmente identificados e filtrados, através de um interpretador do tipo *Part-of-Speech*. Através de cálculos estatísticos, são definidos os conceitos que serão utilizados para a construção da taxonomia. O processamento da hierarquia é o foco do trabalho dos autores: são apresentadas duas abordagens para o levantamento da estrutura, que são mais apropriadas de acordo com os requisitos da taxonomia. Para taxonomias com mais profundidade (mais níveis), é recomendado o uso do método de clusterização para a construção da taxonomia; já visando taxonomias mais rasas e mais fidedignas com a criação manual de taxonomias, é recomendada a utilização do método de subordinação. Essas recomendações foram extraídas a partir do teste e validação de ambas as abordagens.

Meijer, Frasincar, e Hogenboom (2014) apresentam uma abordagem para a construção de taxonomias baseada em um *framework* automatizado, que possui quatro etapas de processamento: extração de termos, filtragem de termos, desambiguação de sentidos e construção da hierarquia. Para avaliar a qualidade da taxonomia, os autores utilizam medidas comparativas em relação a taxonomias já existentes, retirando métricas como *recall*, precisão e outros coeficientes.

Tu, L. Chen e G. Chen (2015) descrevem uma abordagem baseada em métodos estatísticos e técnicas de processamento de linguagem natural. Essa abordagem também possui quatro fases, a

exemplo das outras descritas anteriormente: pré-processamento, extração das palavras-chaves, extração das entidades candidatas e cálculo da distância computacional (estas duas em uma única fase) e por fim a construção da hierarquia. Para organizar a abordagem e justificar a sua eficácia, os autores organizam uma série de definições conceituais sobre o processo de construção de taxonomias elaborado.

2.4 RECONHECIMENTO DE ENTIDADES NOMEADAS

O reconhecimento de entidades nomeadas é uma tarefa importante dos sistemas de extração de informação. Em síntese, ela consiste na detecção de elementos informativos em textos, tais como nomes de pessoas, empresas, cidades, datas e valores monetários. (ANDRIANI et al., 2014).

O Quadro 5 apresenta um exemplo envolvendo os resultados obtidos a partir da aplicação de um processo de reconhecimento de entidades.

Quadro 5 - Exemplo de aplicação da técnica de reconhecimento de entidades

Texto original	O oficial da ONU Erkeus dirige-se para Bagdá.
Texto com os marcadores	O oficial da [ONU] [Erkeus] dirige-se para [Bagdá]
Entidades extraídas (com a classe entre parênteses)	ONU (organização), Erkeus (pessoa) e Bagdá (localização geográfica)

Fonte: ANDRIANI et al., 2014, p. 74.

De acordo com Nadeau e Sekine (2007), o termo “nomeadas” refere-se à restrição inicial do processo em extrair somente as entidades com designadores rígidos, como nomes próprios, substâncias químicas e espécies biológicas – tais como as exemplificadas no quadro anterior. No entanto, a comunidade que estuda e trabalha com reconhecimento de entidades nomeadas também inclui expressões temporais e números que envolvam a utilização de alguma unidade de medida (como unidades monetárias, por exemplo). No final, essas definições tornam-se mais ou menos abrangentes, dependendo do domínio de análise e o tipo de resultado que se pretende obter com a técnica.

Existem algumas classificações para as técnicas de reconhecimento de entidades. Uma delas agrupa essas técnicas de

acordo com o tipo de abordagem empregado (AMARAL, 2013; PELLUCI et al., 2011):

a) Abordagens baseadas em regras: também denominadas abordagens baseadas em conhecimento, utilizam heurísticas definidas como expressões regulares ou padrões linguísticos. Alguns desses padrões consistem em expressões regulares para detecção dos termos LTDA, S.A., *sites* e endereços de *e-mail*;
b) Abordagens baseadas em aprendizado de máquina: utilizam algoritmos de aprendizado de máquina para classificar as entidades nomeadas, tais como Naive-Bayes, SVM e árvores de decisão;
c) Abordagens híbridas: utilizam em conjunto princípios das abordagens baseadas em regras e das abordagens baseadas em aprendizado de máquina para realizar o processo de reconhecimento de entidades.

Já quanto ao tipo de aprendizado, Nadeau e Sekine (2007) classificam as técnicas de reconhecimento de entidades em três grupos. Estes grupos serão abordados nas próximas subseções.

2.4.1 Técnicas de aprendizado supervisionado

Segundo Nadeau e Sekine (2007), as técnicas de reconhecimento de entidades que pertencem a este grupo utilizam um corpus anotado sintaticamente para memorizar as entidades que podem ser extraídas. Esse corpus é considerado o modelo para o processo de reconhecimento de entidades e seu tamanho e qualidade influencia diretamente no resultado.

A aplicação de técnicas desse tipo em outros corpora não se restringe somente ao resgate das entidades memorizadas a partir do modelo, mas também a desambiguação de valores baseadas em critérios discriminativos.

2.4.2 Técnicas de aprendizado semi-supervisonado

De acordo com Nadeau e Sekine (2007), as técnicas de aprendizado semi-supervisonado também trabalham com o fornecimento de um conjunto de entidades delimitado, mas em escala muito menor do que nas técnicas de aprendizado supervisionado.

Tipicamente, estas técnicas tem como o requisito o fornecimento de um conjunto de entidades que estão disponíveis dentro do *corpus* de aplicação da técnica. Essas entidades são pesquisadas e analisadas

dentro do *corpus*, com o objetivo de se extrair padrões sobre contexto em comum entre as entidades descritas no conjunto. A técnica então realiza a identificação de novas entidades seguindo os padrões encontrados nos exemplos. As novas entidades alimentam a lista de entidades de exemplo, aprimorando o padrão de reconhecimento de entidades identificado.

2.4.3 Aprendizado não supervisionado

Por sua vez, as técnicas de aprendizado não supervisionado utilizam princípios baseados em clusterização, análise estatística e recursos léxicos para a extração de entidades. Ao contrário de outros tipos de técnicas de reconhecimento, nesse cenário não são fornecidas informações sobre sementes de entidades validas. (NADEAU; SEKINE, 2007)

Um dos exemplos desse tipo de técnica é a extração de entidades a partir de grupos clusterizados. Dentre as regras que auxiliam na captação de entidades dentro dos documentos, está a detecção de padrões nos termos. Evidências como a utilização de palavras capitalizadas e a ocorrência de um conjunto de termos em contextos específicos formam pistas para a extração de entidades.

2.5 BASES DE CONHECIMENTO

De acordo com Lin e Mendelzon (1999), uma base de conhecimento é um conjunto finito de sentenças especificadas em uma linguagem formal. Essa linguagem formal permite expressividade e compreensão computacional, permitindo com que agentes inteligentes possam interpretar seu significado.

As bases de conhecimento podem ser classificadas em colaborativas e linguísticas (ZESCH, T.; MÜLLER, C.; GUREVYCH, 2008):

- Bases colaborativas: são bases construídas colaborativamente principalmente por voluntários não profissionais. Seu custo de construção é menor, visto o trabalho colaborativo, assim como seu crescimento é rápido e sua atualização é constante. A comunidade é responsável pelo controle de qualidade do conteúdo;
- Bases linguísticas: são bases construídas por linguistas. A construção segue um modelo teórico ou evidencias de um

corpus. O tamanho da base, assim como o intervalo de atualização e o controle de qualidade são parâmetros controlados de forma editorial. Demandam um custo significativo para serem construídas.

As subseções seguintes descrevem em mais detalhes as principais bases de conhecimento colaborativas existentes que estão envolvidas com parte dos objetivos específicos dessa pesquisa.

2.5.1 DBpedia

De acordo com Bizer e outros (2009), a DBpedia[2] é uma base de conhecimento gerada a partir da informação contida na Wikipedia.

A Wikipedia é o exemplo mais notável de uma base de conhecimento. Apesar do texto dos artigos não estar fortemente estruturado, o *site* permite a adição de diversos atributos e classificações estruturadas em suas páginas.

A base contém cerca de 2,6 milhões de entidades, e contem para cada uma delas um identificador único, que pode ser utilizado como referência na *web*. Os termos possuem descrições em cerca de 30 idiomas, com informações sobre termos relacionados, classificações em quatro hierarquias de conceitos e *links* para outras bases de dados existentes na *web* sobre a mesma entidade.

Cerca de metade das informações contidas na DBpedia estão disponibilizadas sob a forma de uma ontologia inter-domínios, com granularidade de classes que varia desde pessoas, organizações ou localizações populares a jogadores de basquete e flores ornamentais (MENDES et al, 2011).

2.5.2 Freebase

A Freebase[3] é um sistema de base de dados construído com o intuito de formar um repositório público do conhecimento disponível no mundo. O formato é baseado em comunidades de conhecimento já existentes, tais como a Wikipedia e a The Semantic Web. (BOLLACKER et al, 2008)

A Freebase é uma base estruturada em formato de grafo, capaz de suportar uma grande diversidade de dados estruturados com

2 Disponível em <https://pt.wikipedia.org/>
3 Disponível em <https://www.freebase.com/>

escalabilidade e rapidez. Ela é composta principalmente pelos seguintes artefatos (BOLLACKER; COOK; TUFTS, 2007):

- Uma *data store* em formado de grafo;
- Uma *object store* para objetos de dados grandes;
- Uma API pública baseada no protocolo HTTP;
- Um sistema de entrada de conteúdo leve;

A Freebase foi desativada em 31 de março de 2015. Os autores do projeto sugeriram mudanças na Wikipedia com o intuito de fornecer as mesmas funcionalidades presentes na Freebase.

2.5.3 YAGO

A YAGO[4] é uma ontologia leve, extensível de grande cobertura. Segundo Suchanek, Kasneci e Weikum (2007), a YAGO, em seu lançamento, já possuía mais de um milhão de entidades e cinco milhões de relações, que abrangem relações taxonômicas (do tipo é um/é um tipo de) e relações não taxonômicas (como os verbos encontrados em ontologias).

As relações existentes na YAGO são automaticamente extraídas da Wikipedia, através de extratores de informação baseados em regras fixas e processamento heurístico. Esses extratores retiram o conteúdo a partir das *infoboxes* e categorias existentes na *wiki*, que são fortemente estruturadas e possuem uma linguagem fixa quando comparadas ao texto livre contido nos artigos. (KANESCI et al,2008).

Por fim, essas mesmas informações extraídas a partir da Wikipedia são validadas através do conteúdo existente na WordNet. Dentre as verificações realizadas estão as de raciocínio lógico, hierarquia de classes e condições de classificação de entidades em classes. Segundo Kanesci e outros (2008), devido a isso, a taxa de acurácia da YAGO chega a ser acima de 95%.

2.6 PLATAFORMA LATTES

A Plataforma Lattes é um sistema de informação que integra as bases de dados de currículos, grupos de pesquisa e instituições catalogados pelo CNPq. Seu nome é uma homenagem ao pesquisador Cesare M. Giulo Lattes, que foi um dos pesquisadores responsáveis pela descoberta do *méson pi*, relativo à partícula subatômica que garante a

4 Disponível em: <https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/>.

coesão do núcleo de um átomo (MENA-CHALCO, J. P.; CESAR JUNIOR, 2009).

A plataforma foi concebida com o intuito de integrar os sistemas de informações das agências federais (BALANCIERI, 2004). Esses dados eram disponibilizados anteriormente de forma isolada, o que dificultava o acesso (PACHECO; KERN, 2001, p. 58):

> A Plataforma Lattes de sistemas de informações em ciência e tecnologia surgiu a partir da necessidade de integração de informações mantidas por CNPq, Capes, Fapesp, Finep e outros sistemas do Ministério de Ciência e Tecnologia. Atendendo a uma reivindicação antiga da comunidade científica, o sistema Currículo Lattes permitiu integrar informações de aplicativos não integrados que, apenas no CNPq, envolviam Bcurr, minicurrículo e Genos – precursor do Currículo Lattes.

Por consolidar informações em um grande espectro da área de ciência, tecnologia e inovação, o ambiente é utilizado por pesquisadores, grupos, instituições de ensino e pesquisa, CNPq, Capes e outros órgãos governamentais.

> Em termos evolutivos, a Plataforma partiu da agregação de informações disponíveis nas diversas bases relacionadas à CT&I do País, promovendo posteriormente a padronização de cada unidade, a construção e a divulgação de serviços de informação e, mais recentemente, o intercâmbio internacional para a comunidade científica do Brasil. (BALANCIERI, 2004, p. 69).

A plataforma é classificada como um sistema de gestão de CT&I, visto que atende os requisitos de captura e armazenamento de dados e fornece uma interface para intercâmbio de informações relacionadas ao domínio.

De acordo com Balancieri (2004), suas informações podem ser utilizadas para auxiliar na gestão de Ciência, Tecnologia e Inovação (CT&I), tais como fortalecer o apoio a linhas de pesquisa prioritárias para o desenvolvimento socioeconômico e cultural, a execução mais eficiente das pesquisas e a conversão mais rápida dos resultados obtidos para aplicação na sociedade.

Dentro do escopo de um governo eletrônico, a plataforma representa um insumo importante para a tomada de decisão. Por exemplo, a realização da analise bibliométrica permite que o uso, a disseminação e outros aspectos quantitativos das produções possam ser mensurados. Tais estudos podem ser realizados dentro da área da

cienciometria, que envolve estudos quantitativos das atividades científicas. Os resultados desses estudos podem ser aplicados na gestão de CT&I para o direcionamento dos esforços às políticas científicas vigentes. (MACIAS-CHAPULA, 1998).

Conceitualmente, a Plataforma Lattes é composta por uma estrutura em pirâmide, com quatro níveis de sustentação: unidades de informação, sistema e fontes de informação, portais e serviços web e sistemas de conhecimento.

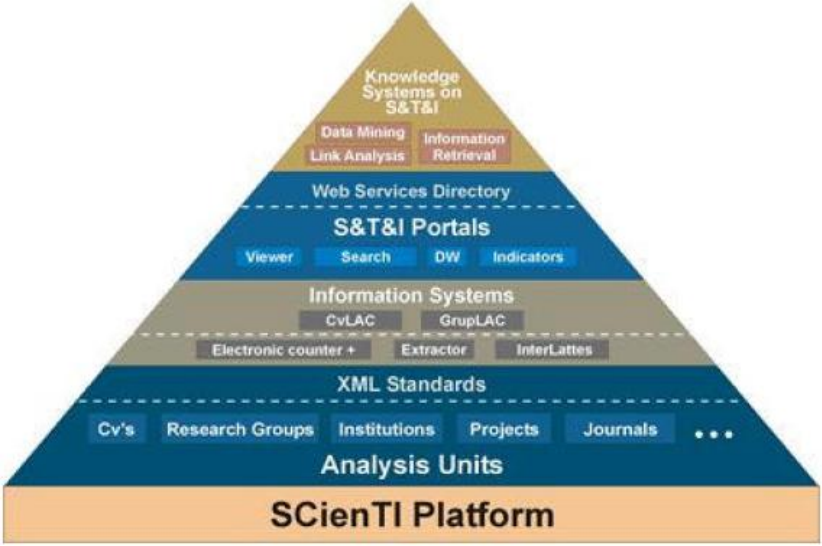
Balancieri (2004, p. 70) atesta a respeito das quatro camadas presentes no modelo conceitual:

> As quatro camadas conceituais da arquitetura da Plataforma Lattes formam um conjunto de elementos tecnológicos e metodológicos: unidades de análise e normalização da informação de CT&I, sistemas e ferramentas de infra-estrutura de informação, diretórios de informação sobre CT&I, portais relacionados com CT&I e sistemas de extração de conhecimento de CT&I.

Juntas, as quatro camadas formam um ecossistema que permite que tanto a alimentação quanto a recuperação de informações e indicadores possam ser realizadas de forma integrada. Considerando a diversidade das informações captadas e o nível de integração fornecido, a plataforma por si só consegue atender as necessidades estratégicas de informação para a implantação e direcionamento da CT&I no país.

A Figura 4 apresenta a arquitetura conceitual da Plataforma Lattes.

Figura 4 – Arquitetura conceitual da Plataforma Lattes.



Fonte: STELA, 2002 apud BALANCIERI, 2004.

O nível de base (tratado na figura anterior como *Analysis Units*) abrange as unidades de análise da plataforma: currículo, grupos de pesquisa, projetos de pesquisa e instituições. Cada uma das unidades foi identificada e especificada através da observação dos subdomínios da área de CT&I e pode ser conceituada da seguinte forma (PACHECO, 2003):

a) currículo: unidade de análise composta pelo conjunto de informações que descrevem a atividade profissional dos usuários cadastrados na plataforma. Esses usuários são todos os indivíduos que possuem alguma relação com a área de CT&I (pesquisadores, estudantes, docentes, gestores de C&T, técnicos de governo, administradores e representantes da sociedade);
b) grupos de pesquisa: unidade de análise composta pelas informações referentes aos grupos de pesquisa. Cada grupo de pesquisa pode ser definido como um conjunto de indivíduos organizados hierarquicamente, baseando-se em critérios de experiência e liderança no terreno científico ou tecnológico no qual trabalham;
c) projetos de pesquisa: unidade de análise que abrange as atividades de pesquisa, desenvolvimento ou extensão que são

executadas por um pesquisador ou um grupo de pesquisadores. Essas atividades têm objetivos, metodologia e duração definidas, e possuem como foco um tema ou objeto específico;

d) instituições: unidade de análise que contempla as informações sobre organizações, institutos, empresas, universidades e outros organismos que estejam ligados de alguma forma a CT&I.

O estudo de caso realizado nesta abordagem será realizado sobre a unidade de análise denominada currículo. Ela a disponibilizada na plataforma através do módulo Currículo Lattes, o qual será descrito em mais detalhes na próxima subseção.

O segundo nível conceitual envolve os sistemas que atuam sobre as unidades de informação *(Information Systems* na figura anterior). Esses sistemas são concebidos com o intuito de auxiliar o processo de gestão de informação, viabilizando a sua coleta para a construção das bases e, ao mesmo tempo, fornecendo instrumentos de coleta aos atores encarregados de direcionar a CT&I em suas funções tanto em organizações ou institutos de pesquisa quanto a nível nacional (PACHECO, 2003; BALANCIERI, 2004).

Os sistemas de envio de currículos, recebimento e certificação de arquivos, de carga e transferência de informações para bases de dados, de visualização e extratores de informação – assim como outros sistemas que contribuem com informações ou extraiam conhecimento a partir dos registros existentes – constituem elementos que integram o segundo nível da plataforma.

O terceiro nível conceitual é composto pelos portais e serviços *web* que utilizam a base de informações da Plataforma Lattes como elemento central para sua operação (*S&T&I Portals*). Desde serviços orientados à divulgação de indicadores nacionais de CT&I a instrumentos de gestão de comunidades virtuais compõem este nível (BALANCIERI, 2004).

Os Portais Lattes são os principais elementos constituintes desse nível. Eles são compostos por um conjunto de cinco portais, que possuem foco em diferentes subdomínios de interesse da CT&I. Esses portais estão descritos em mais detalhes no Quadro 6.

Quadro 6 – Componentes da camada de portais da Plataforma Lattes.

Componente	Descrição
Site da Plataforma	Site principal da Plataforma, que contém links para cada unidade de análise (com buscas e publicação dos respectivos

Lattes[5]	sistemas de informação), para os itens de abertura da Plataforma (i.e., sistemas de integração com outras agências, links para as comunidades virtuais e para instituições que replicam o site Lattes), para os sites Lattes temáticos (descritos a seguir) e para a Rede SCienTI.
Diretório de Grupos[6]	Site do projeto "Diretório de Grupos de Pesquisa no Brasil" com links para cada censo realizado sobre a pesquisa brasileira na unidade de grupos de pesquisa.
Lattes Egressos	Site temático que publicava as informações de egressos de cursos de graduação, especialização, MBAs, mestrado e doutorado das instituições brasileiras.
Demografia Curricular	Site temático que registrava a produção e publicação de diversos indicadores curriculares segundo instituição de afiliação indicada nos currículos.
Investimentos em C&T[7]	Site temático que disponibiliza indicadores sobre investimentos do CNPq em atividades científicas e tecnológicas.

Fonte: PACHECO, 2003, p. 24.

Parte dos sites listados no quadro anterior foi atualizada e/ou aglutinada nas novas versões dos sites da Plataforma Lattes e do Diretório de Grupos. Dessa forma, os indicadores demográficos, que antes eram acessíveis através de módulos isolados, atualmente podem ser obtidos através desses sites.

Por fim, o quarto nível da arquitetura conceitual aborda os sistemas, algoritmos, técnicas e demais instrumentos de gestão e extração de conhecimento *(Knowledge Systems on S&T&I*). Esses instrumentos contemplam algoritmos e sistemas desenvolvidos no âmbito do projeto da plataforma assim como estudos propostos pela comunidade científica que revelam novos conhecimentos sobre a CT&I no país. Esses conhecimentos são extraídos através da utilização da base de dados da plataforma como entrada para obtenção dos dados ou para validação de novas abordagens e modelos dentro da área de CT&I (PACHECO, 2003; BALANCIERI, 2004).

O Quadro 7 apresenta alguns dos componentes da camada de sistemas de conhecimento já produzidos na Plataforma Lattes.

5 Disponível em: <http://lattes.cnpq.br>
6 Disponível em: <http://lattes.cnpq.br/web/dgp>
7 Disponível em: <http://fomentolattes.cnpq.br/fomentoLattes>

Quadro 7 – Componentes da camada de sistemas de conhecimento da Plataforma Lattes

Componente	Descrição
Estratificação (Grupos)	Algoritmo de base estatística utilizado para classificação de grupos de pesquisa (Guimarães et al., 1999) que considera as avaliações da pós-graduação nacional (CAPES), dos bolsistas de produtividade e pesquisa (CNPq) e a produção dos pesquisadores dos grupos.
RedesGP (Grupos)	Projeto que visa à construção de um sistema de análise da atividade científica e tecnológica organizada na forma de redes de pesquisa. Utiliza como base as informações do Diretório de Grupos de Pesquisa e tem fundamentos na área de cientometria e em algoritmos de extração de conhecimento (e.g., *link analysis*).
Lattes Egressos (Currículos)	Projeto e site temático que, por meio de *link analysis*, permite a apresentação de análises sobre o perfil de egressos de cursos de uma instituição. As informações são apresentadas de forma gráfica no site "Lattes Egressos".
Demografia Curricular (Currículos)	Projeto e site temático que, por meio de sistemas com base na técnica OLAP permite uma ampla variedade de análises sobre os currículos de uma determinada instituição.
Investimentos em C&T (Projetos)	Projeto e site temático que, por meio de sistemas com base na técnica OLAP permite a produção dinâmica de indicadores sobre os investimentos realizados pelo CNPq ou por outras agências.

Fonte: PACHECO, 2003, p. 25.

A camada de sistemas de conhecimento é o que permite a plataforma ultrapassar o estado de instrumento agregador informacional, conferindo a ela a capacidade de extrair ativos de conhecimento a partir de seus registros.

Os componentes elencados nesse nível estruturam-se sobre os portais e serviços *web* presentes no nível anterior, que por sua vez dependem das informações registradas nas bases de dados. É possível traçar um paralelo entre esses componentes e os *sites* da Plataforma Lattes, assim como entre os *sites* e as fontes de informação. Essas fontes são preenchidas pelos instrumentos de carga apresentados no segundo nível, o que justifica a abordagem conceitual em pirâmide proposta por Pacheco e Balancieri.

O conhecimento extraído a partir do topo dessa arquitetura pode ser crucial para a tomada de decisão correta dentro do âmbito da gestão de CT&I. A plataforma pode ser considerada um instrumento de

governo eletrônico inteligente, já que além do caráter agregador, é capaz de produzir novas informações relevantes para a gestão.

A próxima subseção descreve em mais detalhes o Currículo Lattes, instrumento o qual é utilizado para extração das informações dentro do estudo de caso proposto no Capítulo 5 .

2.6.1 Currículo Lattes

O Currículo Lattes (Curriculum Lattes/CV Lattes) é o repositório de currículos dos pesquisadores da Plataforma Lattes. Ele armazena os dados relativos a produções bibliográficas, técnicas, artísticas e culturais.

Através do sistema de currículos, são levantadas informações estratégicas para o MCT, CNPq, FINEP e CAPES/MEC, como parâmetros para avaliar a competência de candidatos para a obtenção de bolsas, seleção de consultores, membros de comitês e grupos assessores e a própria pesquisa e pós-graduação brasileiras (AMORIN, 2003).

Os dados dos currículos estão estruturados de forma hierárquica e concentrados nos módulos Dados Gerais e Produção.

O módulo de Dados Gerais concentra informações pessoais e informações profissionais. As seções de cadastro do currículo quanto aos Dados Gerais são as seguintes:

a) Identificação;
b) Endereço;
c) Formação acadêmica/titulação;
d) Atuação profissional;
e) Área de atuação;
f) Idiomas;
g) Prêmios e títulos;
h) Linhas de pesquisa;
i) Outras informações relevantes;

Já o modulo Produção armazena toda a produção bibliográfica, técnica e artística, além de orientações e outras participações intelectuais. O currículo armazena as informações quanto à produção do autor na seguinte estrutura:

a) Produção bibliográfica:
    - Trabalho em anais;
    - Artigo em periódico;
    - Livro e capítulo de livro;
    - Texto em jornal ou revista;

- Tradução;
- Outra;

b) Produção técnica:
   - Softwares;
   - Produtos tecnológicos;
   - Processos ou técnicas;
   - Trabalhos técnicos;
   - Outros:
     i. cartas, mapas ou similares;
     ii. curso de curta duração;
     iii. desenvolvimento de material didático ou instrucional;
     iv. editoração;
     v. manutenção de obra artística;
     vi. maquetes;
     vii. organização de evento;
     viii. programa de rádio ou televisão;
     ix. relatório de pesquisa;
     x. apresentações de trabalhos;
     xi. outra produção técnica;

c) Produção artística/cultural:
   - Apresentações de obras artísticas;
   - Arranjos musicais;
   - Composições musicais;
   - Programas de rádio ou TV;
   - Obras de artes visuais;
   - Outra produção artística e cultural;

d) Orientações concluídas;
e) Demais trabalhos;
f) Toda produção;
g) Propriedade intelectual;
h) Trabalhos mais relevantes;

As informações cadastradas no Currículo Lattes podem ser visualizadas por outros usuários através do sistema de busca e visualização de currículos.

O currículo também pode ser exportado em formato XML, de forma que outros sistemas possam utilizar as suas informações. O formato de exportação está definido em um arquivo DTD, que contém as declarações de marcação utilizadas. Ele foi definido pelo grupo de

trabalho Conscientias, responsável por definir a Linguagem de Marcação da Plataforma Lattes.

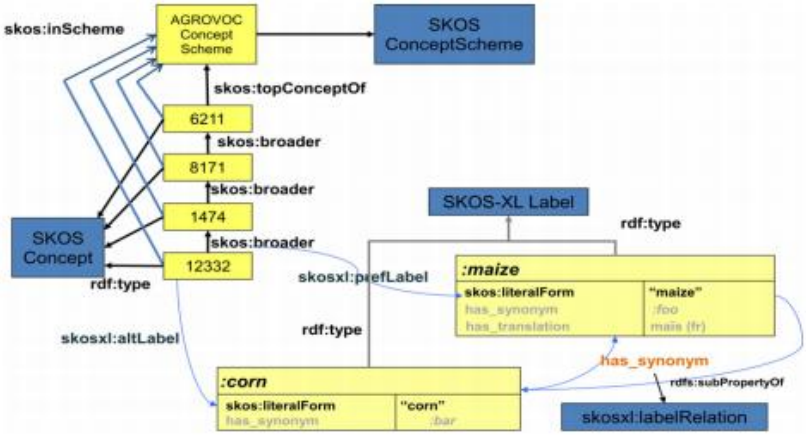
2.7 AGROVOC

O tesauro AGROVOC (junção das palavras agricultura e vocabulário) é um vocabulário controlado para a área de ciência e tecnologia da agricultura – os termos abrangem também os domínios da silvicultura, criação animal e nutrição humana. Seu propósito inicial era o de auxiliar a catalogar as publicações científicas na área, tendo sua primeira versão publicada ainda na década de 1980. O formato adotado para a sua apresentação era o papel, o qual continuou sendo meio até os anos 2000. (CARACCIOLO et al., 2013)

A partir dos anos 2000, o AGROVOC realizou uma transição: trocar a mídia física pela mídia digital. O gerenciamento de dados do tesauro passou a ser realizado através de bases de dados relacionais, com o objetivo de facilitar sua manutenção. No entanto, restrições do acesso a esse tipo de informação dificultaram o desenvolvimento do vocabulário, tais como a limitação de sua distribuição em *dumps* e *web services* e a dificuldade da comunidade (que é distribuída) em atualizar a base. Uma nova migração foi realizada à medida que as tecnologias da *web* semântica evoluíram e possibilitaram essa migração. Atualmente, o AGROVOC é mantido e distribuído em SKOS com a extensão SKOS-XL, assimilando as vantagens dessas novas tecnologias. (CARACCIOLO et al., 2012).

A Figura 5 apresenta a estrutura interna de um fragmento do AGROVOC. O fragmento em questão apresenta a estrutura semântica utilizada para gerenciar as informações sobre o conceito do termo “milho” dentro do tesauro.

Figura 5 – Fragmento do AGROVOC expresso em SKOS-XL



Fonte: CARACCIOLO et al., 2012, p. 68.

Na figura acima, é possível notar a utilização das propriedades da SKOS para a estruturação dos conceitos. A extensão SKOS-XL é utilizada para separar a semântica da hierarquia dos diferentes nomes que o conceito possui nas linguagens atendidas pelo tesauro.

3 PROCEDIMENTOS METODOLÓGICOS

O conhecimento é um dos pilares para a evolução da sociedade. Essa concepção ficou ainda mais em evidência com a cultura da sociedade do conhecimento.

As organizações perceberam o valor que o conhecimento possui e investem em sua produção. Atualmente, ele pode ser considerado sinal de desenvolvimento econômico, ao aliá-lo com as inovações e avanços tecnológicos resultantes de seu uso. (LIMA; MIOTO, 2007).

Do ponto de vista da ciência, o conhecimento científico é o único tipo de conhecimento válido. Sua motivação é a necessidade do homem de compreender os fenômenos da realidade da forma mais transparente possível, além de uma compreensão apresentada pelo senso comum e de uma percepção sensorial. Sua produção, baseada em princípios explicativos e que, ao serem seguidos, produzem o mesmo resultado, é o que caracteriza esse tipo de conhecimento. (GOHN, 2005; FONSECA, 2009; KÖCHE, 1997).

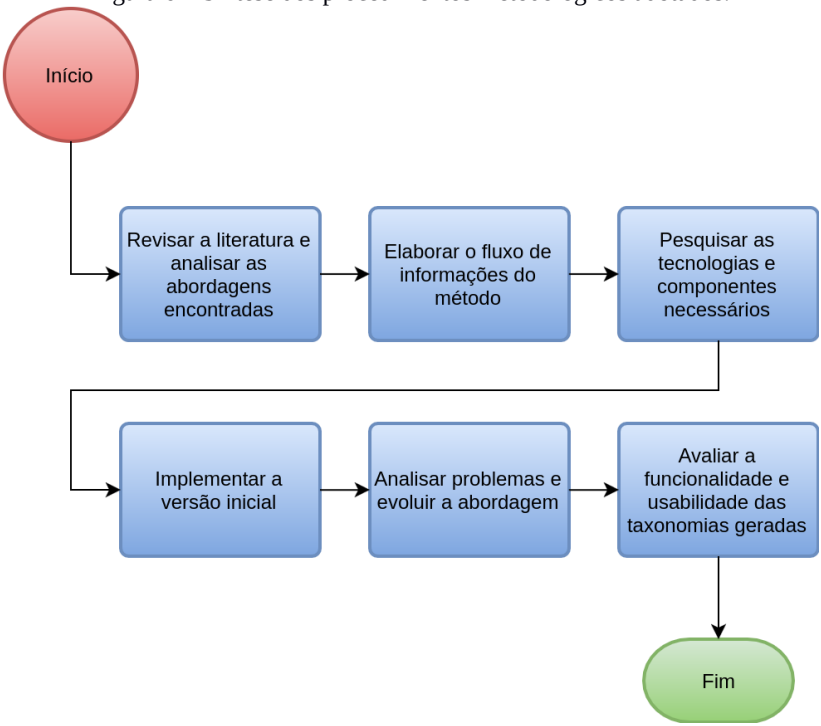
O conhecimento científico, com esses requisitos, precede de um processo de pesquisa. De acordo com Lima e Mioto (2007, p. 39) o processo de pesquisa é "uma atividade científica básica que, através da indagação e (re)construção da realidade, alimenta a atividade de ensino e a atualiza frente à realidade".

Com esse trabalho, pretende-se enriquecer o conhecimento obtido no cenário de construção de taxonomias. Seguindo os preceitos científicos para a produção de conhecimento, foi adotado um processo metodológico para que o desenvolvimento desse trabalho pudesse ser realizado. Esse processo pode ser divido em seis macroetapas:

- Revisar a literatura e analisar as abordagens encontradas;
- Elaborar o fluxo de informação do método;
- Pesquisar as tecnologias e componentes necessários;
- Implementar a versão inicial;
- Analisar problemas e evoluir a abordagem;
- Avaliar a funcionalidade e a usabilidade das taxonomias geradas.

A Figura 6 apresenta o diagrama com a sequência de passos executados no desenvolvimento desse trabalho. As próximas subseções relatam em mais detalhes esses passos.

Figura 6 – Síntese dos procedimentos metodológicos adotados.



Fonte: Elaboração do autor, 2017.

3.1 REVISAR A LITERATURA E ANALISAR AS ABORDAGENS ENCONTRADAS

Antes de construir quaisquer conhecimentos novos, foi realizada uma etapa de revisão bibliográfica e de busca sistemática para integrar os conhecimentos já existentes.

A revisão bibliográfica iniciou a pesquisa apresentada neste trabalho. A revisão buscou consolidar os conhecimentos dentro do Programa de Pós-Graduação de Engenharia e Gestão do Conhecimento da UFSC. Inicialmente, definiram-se as diferenças entre os conceitos de dado, informação e conhecimento, tendo por sequência as delimitações e responsabilidades das áreas de Engenharia, Gestão e Mídias do Conhecimento.

Após a definição dos conceitos inerentes ao programa, buscou-se apresentar e consolidar a definição do conceito de taxonomia segundo os autores da área, envolvendo as suas classificações e aplicações.

Posteriormente, foi realizada uma busca sistemática para averiguar o estado atual da área de construção de taxonomias utilizando procedimentos automatizados, tema o qual é o foco dessa pesquisa.

A busca sistemática é uma das etapas da revisão sistemática. De acordo com Sampaio e Mancini (2007, p. 84), uma revisão sistemática proporciona aos seus utilizadores a coleta de um resumo de evidências sobre uma determinada temática, mediante "a aplicação de métodos explícitos e sistematizados de busca, apreciação crítica e síntese da informação selecionada". A revisão sistemática permite obter os parâmetros pelos quais o estudo foi realizado, sendo possível realizá-lo da mesma forma posteriormente ao se utilizar os mesmos parâmetros.

As consultas foram realizadas em três bases de conhecimento: ScienceDirect, Scopus e Web of Science. A cláusula de consulta utilizada para realizar as consultas foi a seguinte:

automatic AND *taxonomy* AND (*method* OR *methodology* OR *creation*)

A utilização dos operadores AND e OR deve-se a uma otimização de múltiplas de consultas. Dessa forma, a busca pode ser decomposta entre os seguintes critérios:

- *automatic taxonomy method*;
- *automatic taxonomy methodology*;
- *automatic taxonomy creation*;

Por se tratar de uma base com artigos em inglês, apenas a consulta com os termos na língua inglesa foi realizada.

Considerando o universo de análise e o interesse em se verificar o estado atual das técnicas de construção, os resultados foram filtrados de forma que somente as publicações realizadas entre os anos de 2013 e 2015 (ano no qual a busca sistemática foi realizada) sejam incluídas na análise. O Quadro 8 apresenta a quantidade de artigos encontrados por ano em cada base de dados.

Quadro 8 - Número de publicações encontradas em cada base de dados

Base de dados	**Quantidade de artigos encontrados**		
	2013	**2014**	**2015**
ScienceDirect	1.589	1.700	1.980
Scopus	40	48	36
Web of Science	25	34	33

Dessa forma, optou-se por se utilizar unicamente a base de dados ScienceDirect, visto que a quantidade de artigos que foram recuperados a partir da cláusula de consulta foi consideravelmente maior.

Do conjunto de artigos encontrados, foram selecionados os 100 primeiros para análise, de acordo com o critério de relevância disponibilizado pela própria base. Os estudos coletados foram analisados por título, resumo e conteúdo. Os artigos com títulos ou resumos que indiciavam pistas de não haver relação com o tema de análise foram excluídos da análise. Por fim, os artigos que em seu conteúdo não abordavam a construção de taxonomias também foram excluídos do relatório final. As técnicas utilizadas foram coletadas a partir da interpretação dos textos e sumarizadas para indicar o estado da arte na construção de taxonomias. Após a coleta, definiu-se a abordagem a ser utilizada para a construção de taxonomias.

No contexto dessa pesquisa, a abordagem elaborada foi baseada na proposta de Chernyak e Mirkin (2014), que utilizaram a Wikipedia em russo para auxiliar na montagem da árvore taxonômica a partir do terceiro grau de profundidade. O motivo de utilização dessa abordagem foi a simplicidade e acessibilidade da implementação dessa abordagem, comparada a complexidade existente em abordagens estatísticas. No entanto, neste trabalho optou-se por utilizar a DBpedia para a construção de taxonomias desde o nível raiz da taxonomia, ao contrário da abordagem de Chernyak e Mirkin (2014), que alimenta a proposta de taxonomia manualmente com os termos que devem ocupar os dois níveis iniciais.

Com a escolha da abordagem, iniciou-se para a construção do fluxo de informações do método, descrito na próxima seção.

3.2 IDENTIFICAR O FLUXO DE INFORMAÇÕES DO MÉTODO

Após a realização da revisão bibliográfica e analisar as abordagens encontradas para a construção de taxonomias, iniciou-se de fato a construção do método apresentado nesta pesquisa.

Esta etapa levou em conta a necessidade de se identificar o fluxo de informações entre os componentes de forma que seja possível gerar uma taxonomia no final do fluxo. Os componentes identificados necessários para o funcionamento do método são:

- Um repositório de documentos;
- Um componente de reconhecimento de entidades;
- Uma base de conhecimento;

Tendo em vista estes componentes, foi traçado o fluxo de informações – o qual é apresentado na seção 4.2 Após sua construção, a etapa seguinte da pesquisa foi buscar e escolher as tecnologias e componentes necessários para que as informações possam ser encaminhadas através dos componentes. Esta etapa está descrita na próxima seção.

3.3 PESQUISAR TECNOLOGIAS E COMPONENTES NECESSÁRIOS

Utilizando como base o fluxo desenvolvido na etapa anterior, iniciou-se uma pesquisa para a busca das tecnologias e componentes necessários para atender as necessidades para geração de taxonomias.

O custo de desenvolvimento de uma solução que atende os requisitos propostos pelo método a partir de um rascunho em branco demanda o investimento de uma grande quantidade de tempo, além de condicionar a existência de *bugs* e pontos de falha. Considerando essa premissa, foi realizada uma pesquisa para reutilizar tecnologias já existentes para a codificação em *software* do método proposto.

O aparato tecnológico selecionado e as adaptações que foram necessárias para a sua aplicação na implementação do método estão descritos na seção 4.1 As tecnologias e bibliotecas selecionadas podem ser interligadas através da linguagem de programação Java, que foi escolhida para a escrita do código referente às etapas do método.

Com linguagem e tecnologias selecionadas, foi iniciada a implementação em formato de *software* do método, descrita na próxima seção.

3.4 IMPLEMENTAR A VERSÃO INICIAL

Nessa etapa, iniciou-se a construção do código-fonte responsável por transformar o método em uma aplicação prática. Para isso, foi desenvolvido um *software* que contivesse a lógica de passos necessária para o tratamento da informação e utilização dos componentes identificados durante a busca de componentes e tecnologias necessários (seção 3.3). Tentou-se construir a implementação mais simples possível nessa fase de desenvolvimento, visando atender a demanda por resultados iniciais.

Os detalhes de implementação que puderam ser abstraídos na fase de construção do fluxo de informações do método foram analisados e complementados à medida que foram necessários para a escrita do

código-fonte. Dentre esses detalhes, está a utilização do resumo inicial apresentado nos currículos da Plataforma Lattes como campo de análise para o reconhecimento de entidades e a utilização de SPARQL para a realização de consultas na DBpedia.

O desenvolvimento do código foi realizado com uma base pequena de currículos, com aproximadamente duzentos registros de pesquisadores pertencentes a uma mesma instituição.

Esta fase de desenvolvimento foi relativamente curta, pois visava somente o encaixe das tecnologias encontradas e a execução do fluxo de informações do método. Questões como desempenho e aprimoramento dos resultados obtidos foram tratadas após os primeiros testes, que são abordadas na próxima seção.

3.5 ANALISAR PROBLEMAS E EVOLUIR A ABORDAGEM

Nesta fase, foram tratados os principais problemas encontrados durante o desenvolvimento do método, visando melhorar os resultados obtidos ao se requisitar a geração de taxonomias através do *software* variando os parâmetros de entrada informados.

Podem-se classificar os problemas encontrados em três grupos: problemas com o reconhecimento de entidades, problemas de desempenho e *bugs* e erros de código identificados durante o desenvolvimento. Os tratamentos realizados na implementação do método estão descritos na seção 4.1

Também nesta fase da pesquisa foi adicionado o suporte às estatísticas de geração da taxonomia e de acerto no reconhecimento de entidades, necessárias para a avaliação das taxonomias geradas pela implementação do método.

3.6 AVALIAR A FUNCIONALIDADE E A USABILIDADE DAS TAXONOMIAS GERADAS

Com a implementação em *software* do método devidamente ajustada, as taxonomias geradas foram avaliadas com base em uma série de estatísticas envolvendo seu formato e aproveitamento do *corpus*.

As estatísticas foram identificadas com base nos metadados encontrados sobre grafos, árvores e taxonomias. Também foram contempladas estatísticas de acerto do modelo de reconhecimento de entidades e de utilização dos termos dentro da estrutura taxonômica. Essas medidas estão detalhadas no Capítulo 5

A verificação do funcionamento do método ocorreu através de sua aplicação em um conjunto de currículos dentro de um domínio específico. As taxonomias geradas foram comparadas com uma taxonomia de referência da área de conhecimento, buscando analisar o índice de acerto dos termos encontrados nas taxonomias geradas dentro da taxonomia de referência.

No contexto desse trabalho, foi selecionado um conjunto de currículos da área de agricultura para a avaliação das taxonomias geradas. Essas taxonomias foram comparadas com o tesauro AGROVOC. Ao final, os resultados da análise foram coletados e registrados no formato de tabelas, que podem ser visualizadas nos apêndices A, B, C e D.

4 MÉTODO PARA CONSTRUÇÃO DE TAXONOMIAS

A busca de informações e a execução dos procedimentos metodológicos descritos no Capítulo 3 resultaram na concepção do método de construção de taxonomias proposto neste capítulo. O texto das próximas subseções descreve-o em mais detalhes, abordando as variações de fluxo e os parâmetros que podem ser configurados para a sua execução.

O código-fonte correspondente a implementação do método (utilizado para a realização dos experimentos descritos no capítulo 5) pode ser visualizado no endereço <https://github.com/mtslohn/taxonomy-extractor>.

4.1 TECNOLOGIAS E COMPONENTES UTILIZADOS

A implementação do método em formato de *software* foi realizada com o auxílio de algumas tecnologias e componentes específicos. Esta seção descreve os artefatos utilizados, além de indicar os ajustes que foram necessários para a melhoria dos resultados obtidos pelo método.

Para o reconhecimento de entidades foi adotada a biblioteca Apache OpenNLP. A Apache OpenNLP é uma biblioteca escrita na linguagem de programação Java que oferece funções de processamento de linguagem natural em texto. Dentre o rol de funções, estão o suporte a tokenização, segmentação de sentenças, analise sintática, resolução de correferências e reconhecimento de entidades nomeadas. (FONSECA et al, 2015). O componente NameFinder da Apache OpenNLP foi utilizado para extrair as entidades a partir do resumo dos currículos. Esse componente é capaz de criar e utilizar modelos para o reconhecimento de entidades em qualquer idioma a partir de um *corpus* de treinamento.

O modelo de reconhecimento de entidades foi gerado a partir do *corpus* Amazônia. Esse *corpus* é disponibilizado dentro da coleção de artefatos do projeto Floresta Sintá(c)tica[8] – um *treebank*[9] que cobre diferentes variantes de *corpora*, envolvendo jornais, portais colaborativos e entrevistas transcritas. Todo o material é anotado automaticamente através do analisador automático PALAVRAS, sendo partes desse conteúdo também analisada e validada manualmente por revisores. (FREITAS; AFONSO, 2008).

8 Disponível em: <http://www.linguateca.pt/Floresta/>
9 Conjunto de árvores sintaticamente anotadas

Os *corpora* desse projeto são estruturados no formato Árvores Deitadas (ad), criado especificamente para o *treebank*. Para cada frase catalogada, é registrada uma ou mais análises (em caso de ambiguidades) denotando a construção frasal. O Quadro 9 apresenta um exemplo de frase anotada dentro dos *corpora*.

Quadro 9 – Exemplo de frase anotada no formato Árvores Deitadas (ad)

```
CF2-5 Manchete estréia novo jornalístico

A1
STA:fcl
=SUBJ:prop('Manchete' F S) Manchete
=P:v-fin('estrear' PR 3S IND) estréia
=ACC:Np
==>N:adj('novo' M S) novo
==H:n('jornalístico' M S) jornalístico
```

Fonte: FREITAS; AFONSO, 2008.

Cada análise é registrada através do código A1, A2, A3, e assim sucessivamente, até A(n). O código *fcl* indica que a frase é uma oração finita (*finite clause*), pois possui um verbo na forma finita. O sujeito da oração é anotado através do marcador SUBJ. No contexto da frase, ele identifica a palavra "Manchete" como substantivo feminino e singular. O marcador P é uma etiqueta de função, que caracteriza o termo "estrear" como verbo na forma da terceira pessoa do presente do indicativo. Já os valores Np, n e adj complementam a anotação da frase com as informações do sintagma nominal.

Existe uma série de diferentes marcadores que são utilizados para diferentes orações e frases registradas no *corpora*. No entanto, não é necessário compreender em sua plenitude o funcionamento interno para sua utilização dentro do contexto desse trabalho.

A utilização de arquivos no formato Árvores Deitadas dentro do componente NameFinder não é possível nativamente, mas a biblioteca Apache OpenNLP oferece o ferramental necessário para converter o *corpus* para o formato de treinamento utilizado no reconhecimento de entidades. Esse procedimento foi adotado no trabalho para que a geração do modelo de reconhecimento pudesse ser realizada a partir do *corpus* escolhido.

As entidades reconhecidas inicialmente através do OpenNLP apresentaram alguns problemas durante a coleta inicial. Muitos dos termos coletados eram terminados com uma preposição – o que indica

que o processamento realizado pelo algoritmo de reconhecimento trouxe lixo em sua coleta. Após uma análise dos termos, optou-se por realizar a filtragem dos termos coletados: as preposições encontradas no final dos termos compostos reconhecidos foram removidas.

Outro tratamento realizado foi a identificação e a filtragem das entidades coletadas de acordo com o tipo: entidades identificadas do tipo *numeric* e *time* são removidas da análise, por não representarem termos comumente encontrados em uma taxonomia.

A base de conhecimento elegida para a construção do método foi a DBpedia. O motivo de utilização dessa base em relação a outras opções disponíveis deve-se ao volume de informação existente e a qualidade de conteúdo que a mesma possui. O sistema de categorias presente na DBpedia é o principal ponto de interesse desse método. As classificações hierárquicas presentes nesse sistema, que é utilizado para organizar os artigos cadastrados na Wikipedia, pode ser aplicado para a identificação de relacionamentos de generalização e especialização entre termos.

Além da escolha do componente para o reconhecimento de entidades, foi necessária a pesquisa por uma biblioteca de navegação em bases de conhecimento capaz de endereçar e consultar a quantidade massiva de dados existentes na DBpedia. A estrutura da DBpedia é um grafo estruturado no padrão RDF (*Resource Description Framework*), que permite a utilização de consultas no formato SPARQL. O *framework* Apache Jena foi adotado para a realização de consultas nos arquivos que compõem o grafo da DBpedia, visto o conjunto de funcionalidades embutidas que atende plenamente as necessidades no contexto de utilização do método e a popularidade na comunidade de desenvolvimento de *software*.

Várias abordagens para consideradas para a realização de consultas na DBpedia: uma única consulta SPARQL utilizando parâmetros de profundidade para recuperação de termos, várias consultas SPARQL para montagem de uma árvore com os resultados, consultas montadas especificamente para o módulo ARQ[10] (o qual é o interpretador SPARQL do Apache Jena) e consultas montadas programaticamente através dos objetos nativos da API do Jena. De ambas as abordagens, a opção de utilizar a API nativa do Jena foi a escolhida. O motivo principal da escolha da API nativa foi o tempo de processamento para interpretar uma consulta montada em SPARQL/ARQ. Por causa do alto número de consultas realizadas, o uso

[10] O nome do módulo é constituído em letras maiúsculas – não representa uma sigla.

de linguagens de consulta causa grande impacto no tempo de geração da taxonomia, devido à necessidade de se embutir uma camada de tradução da linguagem para o formato de objetos da API do Jena. Ao utilizar diretamente a API nativa, elimina-se essa camada. Taxonomias que levavam cerca de uma hora para serem processadas e exibidas passaram a ser computadas em minutos.

Foram necessárias também otimizações em nível de código para lidar com o volume de informação processado, principalmente ao identificar as coocorrências de termos entre os currículos. Para o gerenciamento de coleções de objetos (tais como listas e mapas), foi utilizada a biblioteca Trove[11] em substituição as coleções nativas do Java. A biblioteca utilizada possui recursos que reduzem a utilização de memória e aceleram o processamento de operações de consulta e alteração de dados. A implementação também foi alterada em determinados pontos para utilizar operações baseadas em *hashs* em vez de pesquisas iterativas dentro de listas, diminuindo o tempo necessário para localizar os objetos necessários para montar a taxonomia.

Outra evolução realizada foi a transformação dos repositórios de dados semânticos utilizados pela abordagem de arquivos RDF para bases em formato TDB. Essas bases são criadas e gerenciadas pelo componente TDB da biblioteca Apache Jena. A transformação desses arquivos permitiu com que não fosse mais necessário manter todas as bases de conhecimento na memória principal do sistema. Em resultados práticos, isso reduziu o tempo de levantamento da base de conhecimento e o consumo de recursos para a execução de consultas.

É importante ressaltar que a escolha dos componentes citados anteriormente pode ser alterada em aplicações futuras do método. É possível utilizar outras ferramentas com funcionalidades equivalentes para a execução dos recursos desejados. Inclusive, isso possibilita a coleta de resultados diferentes, visto as diferenças tecnológicas entre as ferramentas.

4.2 DESCRIÇÃO DO MÉTODO

O método é composto pelo encaminhamento e tratamento de um fluxo de informações através de diversos componentes. Estes componentes agem como intermediadores de processamento, realizando o tratamento necessário da informação para que ela possa ser utilizada

11 Disponível em: <http://trove.starlight-systems.com/>

pelo componente seguinte. Neste método, são utilizados como componentes:

- Um conjunto de currículos do Currículo Lattes;
- Uma biblioteca de *software* para reconhecimento de entidades;
- A base de conhecimento DBpedia, em conjunto com uma biblioteca de *software* para realização de consultas sobre suas informações.

A Plataforma Lattes pode ser alterada por quaisquer outros repositórios de documentos textuais. No entanto, com o intuito de demonstrar com maior proximidade a implementação que foi realizada para a avaliação do método, os passos serão descritos tendo como base este ambiente como repositório de conteúdo.

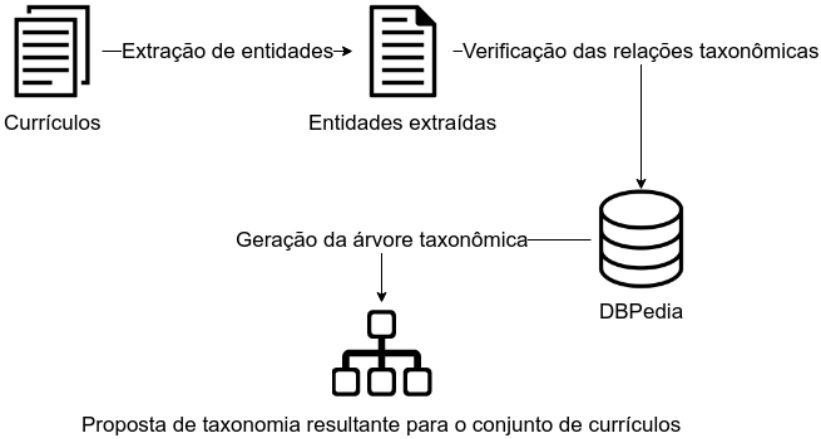
Da mesma forma que o repositório de conteúdo textual, a base de conhecimento pode ser alterada para quaisquer outras que possuam uma representação de hierarquia de termos. Como já justificado anteriormente, foi utilizada na implementação do método a base de conhecimento DBpedia – sendo esta a base na qual está estruturado o texto descritivo do método.

O processamento do método parte da base de currículos disponíveis na plataforma. Cada currículo possui um conjunto de informações estruturadas sobre um pesquisador. Inicialmente, o resumo textual dos currículos de cada pesquisador foi definido como conteúdo para a extração de termos a serem utilizados pela taxonomia. A partir desses currículos é realizado um processo de reconhecimento de entidades, que extrai os principais termos mencionados.

O próximo passo é a verificação das relações entre os termos extraídos anteriormente. Cada par de termos validado (de acordo com os critérios adotados pela variação do método utilizada) é encaminhado para a DBpedia. Caso exista uma relação hierárquica válida entre as entidades (de acordo com os parâmetros definidos ao utilizar o método), ela é adicionada à árvore taxonômica. Caso contrário, a relação é descartada. Por fim, essas relações são organizadas em formato de árvore, respeitando a hierarquia entre os termos.

Para melhor compreensão do método desenvolvido, a Figura 7 apresenta o fluxo de informações do método.

Figura 7 – Fluxo de informações do método



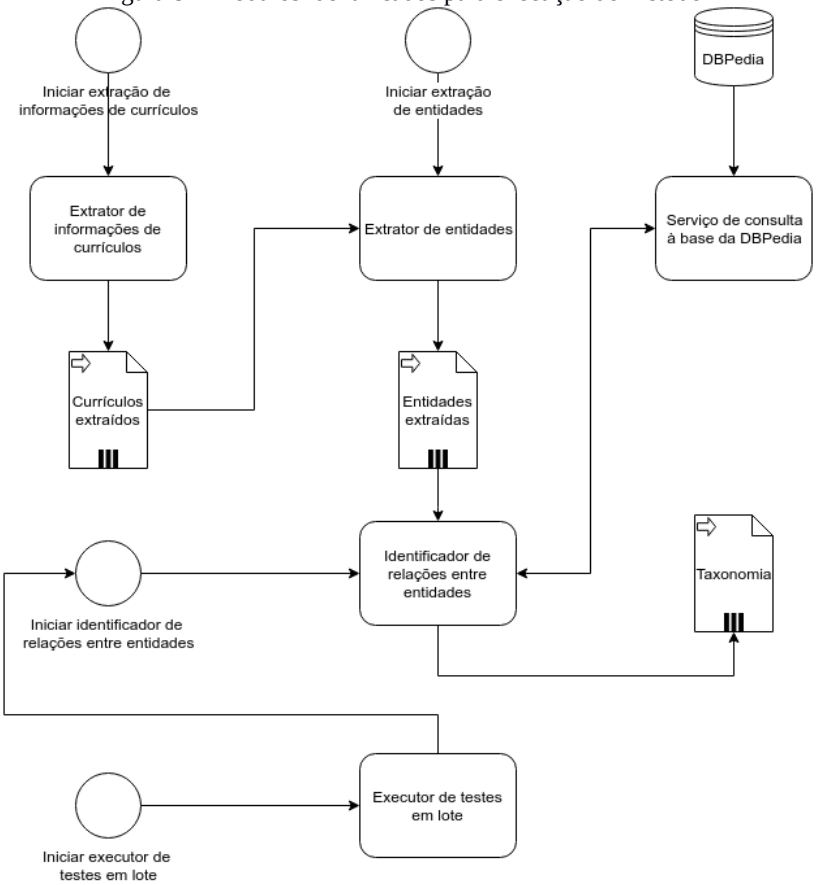
Fonte: Elaboração do autor, 2017.

Visando desempenho, o método foi modularizado em etapas que podem ser inicializadas ignorando a execução sequencial de suas etapas anteriores, desde que já existam as informações necessárias para sua execução armazenadas previamente. Dessa forma, a realização de baterias de testes pode ser efetuada a partir de determinado ponto do método, diminuindo o tempo de espera para a geração das taxonomias. Foram identificados e implementados cinco módulos:

- Extrator de informações de currículos;
- Serviço de consultas à base da DBpedia;
- Reconhecedor de entidades;
- Identificador de relações entre entidades;
- Executor de testes em lote.

A Figura 8 apresenta a interligação lógica entre os módulos e seus possíveis pontos de inicialização dentro do método.

Figura 8 – Módulos identificados para execução do método



Fonte: Elaboração do autor, 2017.

Os módulos trabalham de forma interdependente, pois o resultado ou o processamento de um módulo é a entrada para o módulo seguinte. Os pontos de partida (marcados com círculos) indicam os módulos que podem ser inicializados independentemente para critérios de teste e execução do método.

O método possui duas variações de algoritmo-base, que tratam da forma de se testar a validade de relações taxonômicas:

- Na primeira variação, a ocorrência de um par de termos dentro do *corpus* basta para que se teste a existência de uma relação hierárquica entre elas na DBpedia;

- Na segunda variação, a coocorrência dos termos dentro de cada currículo é utilizada como critério de validação para identificar possíveis relações hierárquicas;

Ambas as variações são abordadas em maiores detalhes nas subseções seguintes.

4.2.1 Método baseado em coocorrências dentro do *corpus*

Essa versão do método adota para identificação e validação das relações hierárquicas entre dois termos os seguintes critérios:

- A ocorrência de ambos os termos na lista de entidades reconhecidas dentro dos currículos;
- A permanência dos termos nessa listagem após a filtragem dos termos de acordo com a sua ocorrência;
- A existência de uma relação hierárquica direta ou indireta (dependendo da configuração do parâmetro de consideração de relações indiretas) entre ambos os termos na base de conhecimento da DBpedia.

Esses critérios são executados ao longo da sequência de passos do método. A vantagem nesse cenário é a simplicidade de processamento, visto que não é necessária nenhuma estrutura adicional além da lista de entidades para se realizar a verificação das relações hierárquicas entre os termos.

O algoritmo com a sequência de passos executados por essa versão do método é o seguinte:

1) Carregar modelo de reconhecimento de entidades;
2) Extrair entidades identificadas nos resumos dos currículos;
3) Adicionar palavras-chaves registradas por currículo;
4) Criar lista filtrada das entidades encontradas nos currículos (de acordo com o critério para filtragem das contagem das entidades informado como parâmetro);
5) Para cada combinação de pares de entidades ainda não testada:
    5.1) Consultar na DBpedia a existência de relações entre as entidades;
    5.2) Verificar se existe relação hierárquica direta ou indireta (de acordo com o critério de verificação de relações indiretas entre as entidades, informado como parâmetro);
        5.2.1) Se sim: Adicionar a relação entre os pares e os termos intermediários identificados;

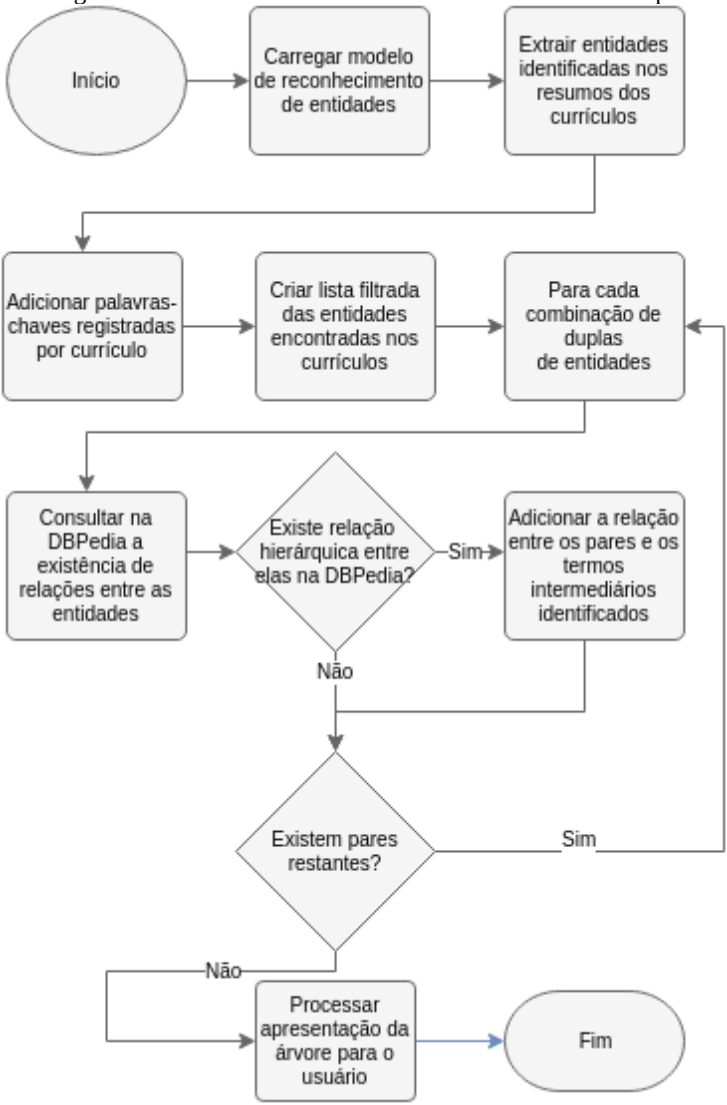
5.3) Verificar se existem combinações de palavras ainda não testadas;

5.3.1) Se sim: Voltar ao passo 5;

6) Terminar a montagem da estrutura da árvore taxonômica e consolidar as estatísticas de geração.

Com o intuito de facilitar a compreensão da sequência de passos do algoritmo, a Figura 9 apresenta um fluxograma com a sequência de passos da execução dessa variação do método.

Figura 9 – Método baseado em coocorrências dentro do corpus



Fonte: Elaboração do autor, 2017.

4.2.2 Método baseado em coocorrências dentro do currículo

Em vez de considerar somente a existência de relações hierárquicas entre os termos identificados no corpus como critério-base de validação (exceto os parâmetros), essa versão do método filtra também as possíveis relações considerando a existência de coocorrência das palavras testadas dentro de pelo menos um dos currículos em análise.

Dessa forma, os critérios adotados por essa versão do método para identificação e validação das relações hierárquicas são os seguintes:

- A ocorrência de ambos os termos na lista de entidades reconhecidas dentro dos currículos;
- A permanência dos termos nessa listagem após a filtragem dos termos de acordo com a sua ocorrência;
- A coocorrência do par de termos em pelo menos n currículos das entidades testadas (sendo n um valor informado como parâmetro para o método);
- A existência de uma relação hierárquica direta ou indireta (dependendo da configuração do parâmetro de consideração de relações indiretas) entre ambos os termos na base de conhecimento da DBpedia.

O algoritmo executado nessa versão é semelhante ao executado para a versão baseada em coocorrências dentro do *corpus*, com a adição da verificação de coocorrências para cada uma dos pares de entidades testados:

1) Carregar modelo de reconhecimento de entidades;
2) Extrair entidades identificadas nos resumos dos currículos;
3) Adicionar palavras-chaves registradas por currículo;
4) Criar lista filtrada das entidades encontradas nos currículos (de acordo com o critério para filtragem das contagem das entidades informado como parâmetro);
5) Identificar os pares de entidades que coocorrem em cada currículo;
6) Remover os pares de entidades que não atingiram o valor mínimo de coocorrências (informado como parâmetro);
7) Para cada combinação de pares de entidades identificados no passo 5:
    7.1) Consultar na DBpedia a existência de relações entre as entidades;
    7.2) Verificar se existe relação hierárquica direta ou indireta (de acordo com o critério de verificação de relações indiretas entre as entidades, que é informado como parâmetro);

7.2.1) Se sim: Adicionar a relação entre os pares e os termos intermediários identificados;

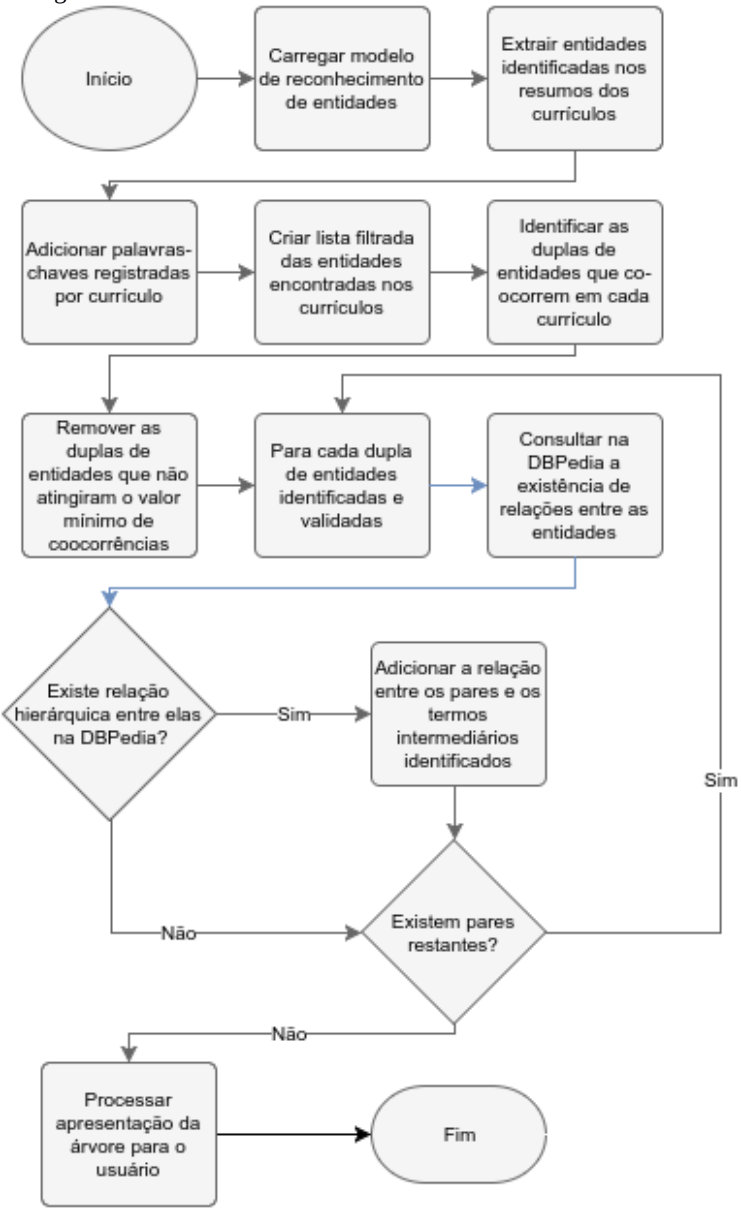
7.3) Verificar se existem combinações de palavras ainda não testadas;

7.3.1) Se sim: Voltar ao passo 7;

8) Terminar a montagem da estrutura da árvore taxonômica e consolidar as estatísticas de geração.

Da mesma forma como apresentada na seção anterior, a Figura 10 apresenta um fluxograma com a sequência de passos realizados durante a execução do algoritmo.

Figura 10 – Método baseado em coocorrências dentro do currículo



Fonte: Elaboração do autor, 2017.

4.3 PARÂMETROS

Além das variações listadas na seção anterior, o método fornece a possibilidade de parametrizar o funcionamento de determinadas funcionalidades, visando refinar os resultados obtidos com a execução do processo. O Quadro 10 apresenta esses parâmetros, que serão detalhados e justificados no decorrer do texto dessa seção.

Quadro 10 - Parâmetros de entrada para as abordagens

Número mínimo de ocorrências de uma entidade	Valida as entidades reconhecidas como termos candidatos para a busca de relações hierárquicas. Varia de um a infinito.
Número máximo de graus de distância entre dois termos na DBpedia	Valida as relações hierárquicas de acordo com o número de graus de distância entre os dois termos em análise de hierarquia na DBpedia. Varia de um a infinito.
Número mínimo de coocorrências de cada par de termos para validar a relação hierárquica	Valida as relações hierárquicas sob um critério de coocorrência mínima. Varia de um a infinito.

O primeiro dos parâmetros que podem ser informados é o número mínimo de ocorrências de uma entidade. As mudanças realizadas nesse parâmetro possibilitam que entidades que tenham sido reconhecidas erroneamente possam ser excluídas da taxonomia ao se verificar sua baixa ocorrência. Além disso, o parâmetro interfere no total de entidades da taxonomia, tornando-a mais enxuta e simplificando seu entendimento. O contraponto é que um valor alto para esse parâmetro pode acabar por remover termos que são importantes para o domínio de análise.

Outro parâmetro que pode ser informado é relacionado ao nível de interação na base da DBpedia: o número de graus de distância entre dois termos dentro da hierarquia da DBpedia para que a relação entre eles seja considerada válida para a construção da taxonomia. Isso permite que, mesmo quando houver apenas uma relação indireta entre dois termos dentro da DBpedia, essa relação hierárquica seja considerada válida e, dessa forma, possível de ser adicionada na

taxonomia. O valor um indica que é necessária uma relação direta entre os dois termos. Valores inteiros maiores que um indica que pode haver até n - 1 termos intermediários (sendo n o valor do parâmetro) entre os termos analisados na DBpedia para que a relação possa ser considerada válida.

Por fim, é disponibilizado também um terceiro parâmetro, envolvendo o número mínimo de coocorrências de cada par de termos dentro dos currículos analisados para que a mesma possa ser considerada válida para o teste na DBpedia. Ele é exclusivo para uso na variação do método baseado em coocorrências dentro do currículo. O parâmetro permite que pares de entidades que ocorreram com pouca frequência não interfiram nas relações apresentadas na árvore taxonômica. Dessa forma, apenas as relações mais relevantes dentro do domínio são consideradas na construção da taxonomia.

Durante o desenvolvimento, também foram realizados testes com duas funções modificadoras que alteram a estrutura das taxonomias. Ambas as funções atuam fazendo uma espécie de faxina na árvore gerada, trazendo resultados diferenciados para o usuário:

- Reposicionar termos: Essa função analisa os termos raízes que foram identificados durante o processo de montagem da taxonomia e verifica se é possível realocá-los em algum nodo filho fora da raiz. O acionamento dessa função permite que os termos estejam posicionados por completo dentro de um termo pai, assim como diminui o número de termos raízes na taxonomia. O contraponto do acionamento dessa função é que a mesma pode induzir a união incorreta de termos de semânticas distintas.
- Remover termos não utilizados nos currículos: Essa função tem aplicação ao montar árvores utilizando a pesquisa hierarquizada de termos. Dessa forma, os termos existentes na taxonomia que não foram encontrados através do processo de reconhecimento de entidades são removidos.

Os resultados dessas estratégias melhoram parcialmente as árvores geradas, porém mascaram um pouco o esquema de construção da árvore. Os resultados gerados por esses modificadores não foram abordados no conjunto de análise apresentados nos apêndices, visto que ambos necessitavam de codificação mais especializada para que seus resultados pudessem ter maior confiabilidade.

5 EXPERIMENTOS E RESULTADOS

Essa seção descreve os experimentos e resultados obtidos com a execução da implementação do método para diversos conjuntos de parâmetros.

São descritas inicialmente as estatísticas coletadas após a construção das taxonomias. Em sequência, algumas taxonomias geradas pelo método são apresentadas ao leitor, considerando as possibilidades de utilização de parâmetros e variações do método. Por fim, é apresentada uma análise sobre a bateria de testes realizada para verificação do funcionamento do método, comparando os resultados obtidos em relação ao tesauro AGROVOC, que é referência na área de agricultura, pecuária e abastecimento – foco deste estudo de caso.

5.1 ESTATÍSTICAS OBTIDAS

A análise dos resultados obtidos é realizada a partir de estatísticas coletadas sobre as taxonomias de acordo com a variação do método empregada, os parâmetros informados e o *corpus* utilizado. Além da taxonomia, são avaliados dados sobre o processo de reconhecimento de entidades (mapeando assim a quantidade de termos que de fato entraram para a avaliação da existência de relacionamentos hierárquicos) e do pertencimento desses termos à taxonomia de referência utilizada para comparação.

Dessa forma, a estatísticas coletadas foram agrupadas em três grupos. Em ambos os cenários, cada *token* indica uma palavra existente no texto analisado.

O primeiro grupo trata das estatísticas envolvendo o processo de reconhecimento de entidades. O Quadro 11 apresenta essas estatísticas.

Quadro 11 - Estatísticas coletadas sobre o processo de reconhecimento de entidades

Número de *tokens*	Número de *tokens* identificados durante o processo de reconhecimento de entidades
Tokens reconhecidos	Total de *tokens* encontrados no conjunto de termos que foram reconhecidos durante o processo de reconhecimento de entidades
Fator de reconhecimento	Coeficiente do total de *tokens* distintos reconhecidos sobre o total de *tokens* encontrados no *corpus*. Varia de zero a um.

O segundo grupo de estatísticas coletadas refere-se aos metadados coletados sobre a taxonomia construída pelo método. Ao contrário do quadro anterior, as estatísticas coletadas na seção são uma reflexão direta da taxonomia. O Quadro 12 apresenta essas estatísticas.

Quadro 12 - Estatísticas coletadas sobre a taxonomia

Tokens utilizados	Total de *tokens* provenientes do processo de reconhecimento de entidades que foram efetivamente utilizados na taxonomia.
Fator de utilização de tokens	Coeficiente do total de *tokens* utilizados na taxonomia provenientes do processo de reconhecimento de entidades sobre o total de *tokens* encontrados no *corpus*. Varia de zero a um.
Contagem de termos	Contagem do total de termos encontrados na taxonomia.
Nível máximo	Nível de profundidade máximo da taxonomia, alcançado por pelo menos um termo.
Número de expansões	Contagem do número de termos que possuem filhos na árvore taxonômica. O nome do parâmetro deve-se o fato de que cada termo pai expande filhos para um nível seguinte.
Soma do número de termos expandidos	Soma de todos os termos que surgiram devido a uma expansão na árvore taxonômica. Por conceito, somente os termos do nível raiz não entram nessa contagem.
Média de termos por expansão	Número médio de filhos que um termo pai tem durante uma expansão.
Nível médio do termo	Média do nível de localização de um termo dentro da taxonomia.
Número de termos cíclicos	Número de termos que se repetem dentro da taxonomia.
Fator de termos cíclicos	Porcentagem dos termos repetidos dentro da taxonomia.
Horizontalidade	Razão do nível máximo ocupado por pelo menos um termo da taxonomia sobre o total de termos existentes.
Verticalidade	Razão do total de termos existentes sobre o nível máximo ocupado por pelo menos um termo da

	taxonomia – é a operação inversa à realizada para identificar a horizontalidade.

Para critérios de avaliação do método, foram elaboradas estatísticas em específico relacionadas à comparação com o tesauro AGROVOC. Seguindo o padrão dos quadros anteriores, o Quadro 13 apresenta a descrição dessas estatísticas.

Quadro 13 - Estatísticas coletadas na comparação das taxonomias geradas com o tesauro AGROVOC

Número de termos encontrados no AGROVOC	Número de termos da taxonomia que possuem equivalentes no tesauro AGROVOC.
Porcentagem de termos encontrados no AGROVOC	Porcentagens dos termos que foram identificados também no tesauro AGROVOC.
Número de relacionamento hierárquicos encontrados no AGROVOC	Número de relações hierárquicas semelhantes (sendo diretas ou indiretas) encontradas também no tesauro AGROVOC.

5.2 RESULTADOS OBTIDOS

Essa seção apresenta algumas taxonomias geradas através da implementação do método proposto nesse trabalho. Com fins de facilitar e diferenciar as taxonomias apresentadas nessa seção, as taxonomias serão rotuladas com letras.

A justificativa para a adoção dos parâmetros em cada uma das execuções do método é permitir a geração de taxonomias com tamanhos e estruturas que possam ser apresentadas nesse documento. Taxonomias com centenas de termos não seriam possíveis de serem representadas em papel. Por outro lado, taxonomias com pouquíssimos termos não teriam expressividade e utilidade. Dessa forma, procurou-se ponderar nos valores informados para os parâmetros com o objetivo de encontrar um meio-termo entre esses números, de forma que os resultados obtidos possam ser facilmente compreendidos no texto desse trabalho.

A implementação do método operou sobre um corpus de 20.000 currículos, extraídos a partir de um subconjunto de currículos com atuação na área da agricultura, pecuária e abastecimento. Eles foram

obtidos a partir da filtragem dos currículos pela área de conhecimento das Ciências Agrárias. Desta amostra, foram utilizados os campos contendo o resumo do currículo do pesquisador e as palavras-chaves registradas para cada produção científica. Sobre o campo de resumo do currículo, ainda foi aplicado um processo de reconhecimento de entidades, objetivando extrair apenas os termos mais importantes do texto. Este subconjunto de currículos foi selecionado tendo em vista a comparação dos resultados obtidos com o tesauro AGROVOC, que por sua vez constitui parte do processo avaliativo empregado neste trabalho.

Já o tamanho da amostra foi selecionado devido a limitações da implementação desenvolvida para demonstrar a viabilidade do método e possibilitar a sua avaliação. O *corpus* de currículos da Plataforma Lattes pertencentes às áreas selecionadas na amostra possui 283.838 registros. As estruturas utilizadas durante a implementação tiveram como objetivo principal demonstrar a viabilidade do método, ignorando compensações a serem realizadas para processar grandes quantidades de dados. Com vistas nesse cenário, as estruturas de dados elaboradas para representar as informações não são robustas o bastante para armazenar a imensa quantidade de informação. Dessa forma, foi selecionado um subconjunto a partir do recorte de currículos que seja capaz de demonstrar a viabilidade do método e, ao mesmo tempo, seja restrito o bastante para que as estruturas de dados elaboradas comportem a quantidade de informação gerada e manipulada.

A taxonomia inicial apresentada nessa seção é a taxonomia A. Os parâmetros informados para a geração dessa taxonomia são apresentados no Quadro 14.

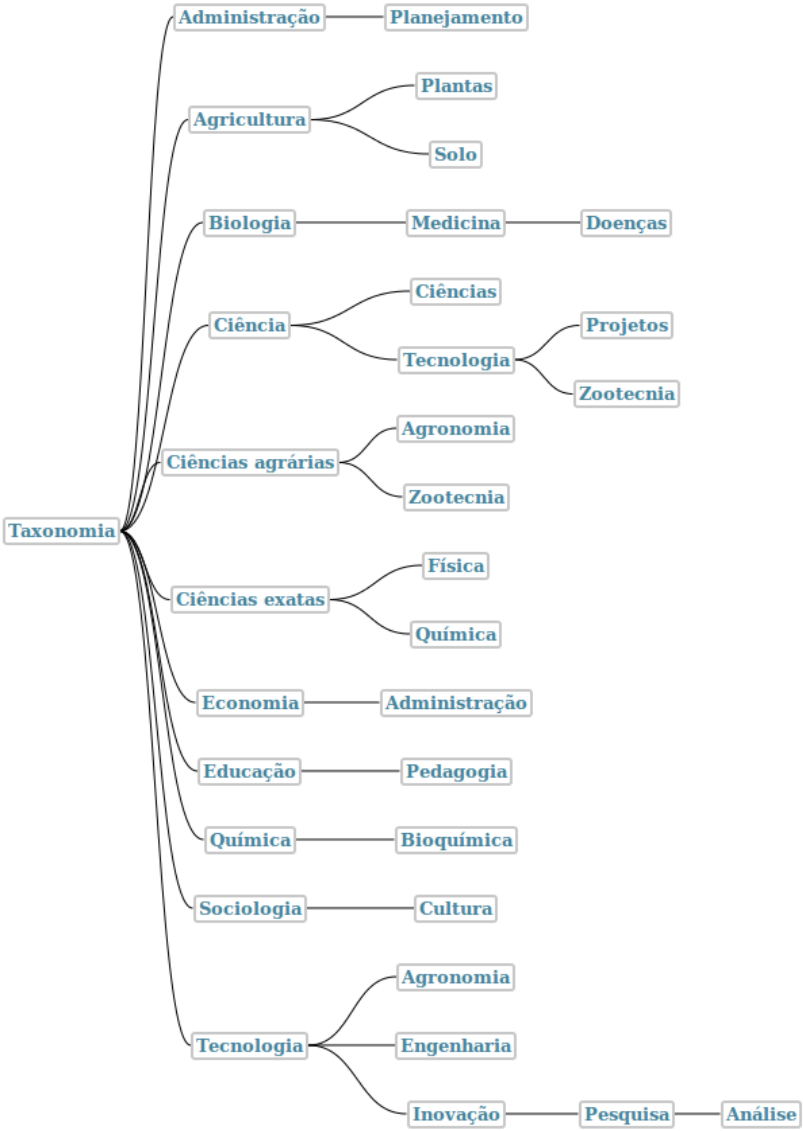
Quadro 14 - Parâmetros utilizados para a geração da taxonomia A

Variação do método utilizada	Abordagem de coocorrência dentro do *corpus*
Número de currículos analisados	20.000
Número mínimo de ocorrências de cada entidade	500
Número de níveis considerados na hierarquia da DBpedia	1

O método foi executado para os parâmetros informados, sendo entregue uma taxonomia no final do processo. Essa taxonomia foi

traduzida para o formato de representação gráfica, que pode ser visualizada na Figura 11.

Figura 11 – Representação gráfica da taxonomia A



Fonte: Elaboração do autor, 2017.

Ao gerar essa taxonomia, foi possível coletar uma série de estatísticas – as mesmas mencionadas na seção 5.1 O Quadro 15 apresenta as estatísticas coletadas.

Quadro 15 – Estatísticas obtidas sobre a taxonomia A

Número de *tokens*	9.325.838
Tokens reconhecidos	3.052.738
Fator de reconhecimento	0,3273420
Tokens utilizados	70.742
Fator de utilização de *tokens*	0,0075
Contagem de termos	33
Nível máximo	4
Número de expansões	15
Soma do número de termos expandidos	22
Média de termos por expansão	1,467
Nível médio do termo	1,848
Número de termos cíclicos	5
Fator de termos cíclicos	0,152
Horizontalidade	0,121
Verticalidade	8,250
Número de termos encontrados no AGROVOC	23
Porcentagem de termos encontrados no AGROVOC	69,697
Número de relacionamento hierárquicos encontrados no AGROVOC	2

Uma consideração marcante sobre a implementação do método, que também se repete para as outras taxonomias geradas, são as faixas de termos reconhecidos a partir dos currículos e os termos efetivamente utilizados na taxonomia.

O número de termos reconhecidos é considerável: o fator de reconhecimento é de 0,327342, o que indica que mais de 32% dos termos encontrados nos currículos foram reconhecidos. Esse fator de reconhecimento é obtido exclusivamente pela biblioteca de reconhecimento de entidades – no contexto desse trabalho, a Apache OpenNLP.

Já o número de *tokens* efetivamente utilizados na taxonomia pode ser considerado uma partícula do universo de termos identificados durante o reconhecimento de entidades, visto que a razão de aproveitamento apresentada é de 0,75%. Isso se deve a filtragem realizada através do parâmetro de número mínimo de ocorrências de

cada termo para que ele possa ser considerado válido para inclusão na taxonomia. A ausência desse parâmetro permitiria com que muitos termos classificados incorretamente como válidos pela biblioteca de reconhecimento entrassem na taxonomia, reduzindo assim a sua qualidade. Além disso, pesa o fator de que diversos termos identificados não possuem relações hierárquicas com outros termos provenientes do processo de reconhecimento de entidades.

Sobre a taxonomia gerada, é possível verificar a existência de termos majoritariamente nos primeiro e segundo níveis da taxonomia (o que explica a estatística de nível médio do termo calculada em 1,848). Além disso, a taxonomia é verticalizada, com mais termos em níveis base do que em níveis mais profundos da hierarquia.

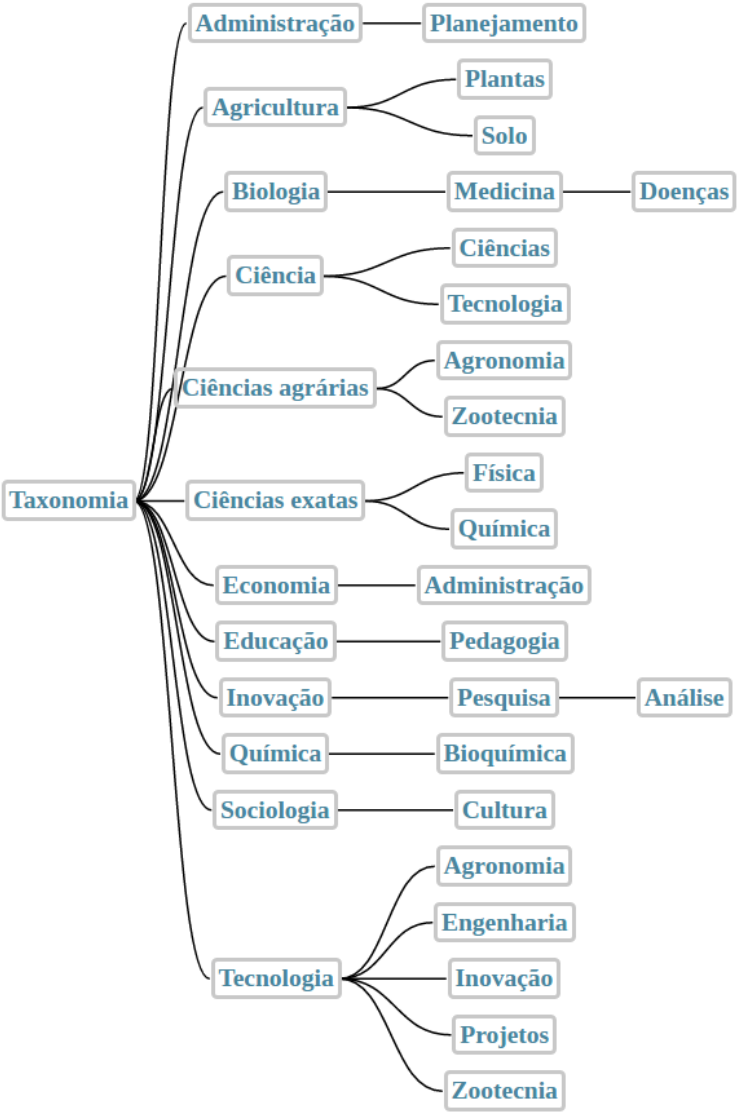
A taxonomia B foi gerada com parâmetros semelhantes aos da taxonomia A, exceto pelo fato de utilizar a variação do método de identificação de relações hierárquicas através da coocorrência de pares de termos dentro do currículo. O Quadro 16 apresenta o conjunto de parâmetros informados para a geração do método.

Quadro 16 - Parâmetros utilizados para a geração da taxonomia B

Variação do método utilizada	Abordagem de coocorrência dentro do currículo
Número de currículos analisados	20.000
Número mínimo de ocorrências de cada entidade	500
Número de níveis considerados na hierarquia da DBpedia	1
Número mínimo de coocorrências de cada par de termos para validar a relação hierárquica	1

A utilização de uma estratégia de identificação de relações diferente, mesmo com a adoção dos outros parâmetros iguais aos informados para a geração da taxonomia A, causa diferenças no resultado final. A Figura 12 apresenta em forma de imagem a taxonomia gerada pelo método.

Figura 12 – Representação gráfica da taxonomia B



Fonte: Elaboração do autor, 2017.

Essa taxonomia centraliza mais termos no segundo nível da árvore que a taxonomia A. O conceito Tecnologia também possui sua capacidade de agregação ampliada, ao agrupar cinco termos filhos na

árvore, em vez dos três representados na taxonomia anterior. O nível máximo dessa segunda taxonomia também é menor.

O panorama das estatísticas coletadas para essa taxonomia pode ser conferido no Quadro 17.

Quadro 17 - Estatísticas obtidas sobre a taxonomia B

Número de *tokens*	9.325.838
Tokens reconhecidos	3.052.738
Fator de reconhecimento	0,3273420
Tokens utilizados	62.782
Fator de utilização de *tokens*	0,007
Contagem de termos	32
Nível máximo	3
Número de expansões	13
Soma do número de termos expandidos	21
Média de termos por expansão	1,615
Nível médio do termo	1,719
Número de termos cíclicos	4
Fator de termos cíclicos	0,125
Horizontalidade	0,094
Verticalidade	10,667
Número de termos encontrados no AGROVOC	22
Porcentagem de termos encontrados no AGROVOC	68,750
Número de relacionamento hierárquicos encontrados no AGROVOC	3

Comparando os valores obtidos no Quadro 15 e no Quadro 17, é possível perceber outras nuances sobre as taxonomias. O número de *tokens* utilizados continua o mesmo, porém a taxonomia “B” possui um termo a mais em relação à taxonomia “A” – mudança que é refletida também no aumento do número de termos cíclicos. A diferença entre as taxonomias é causada pelo método de montagem da árvore: na taxonomia “A”, cada combinação de entidades identificada no método foi considerada uma relação válida; já na taxonomia “B”, as combinações de entidades foram validadas de acordo com sua ocorrência em um mesmo currículo, sendo adicionadas em uma lista para verificação posterior na DBpedia. A árvore taxonômica da última taxonomia é mais verticalizada em relação à primeira, visto a diferença de coeficientes (11,333 contra 8,250).

É possível conferir também que os índices de acerto do AGROVOC se mantiveram próximos. A taxonomia “B” apresenta uma leve vantagem na estatística de número de termos encontrados no

tesauro. No entanto, o termo adicional na verdade é um termo cíclico. Dessa forma, é possível considerar que os resultados da execução das abordagens semelhantes para esse critério.

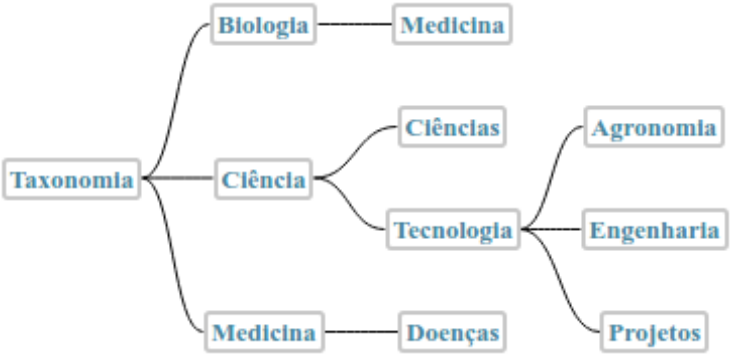
Por fim, a taxonomia “C” objetiva realizar uma análise ainda mais restritiva, procurando utilizar somente termos de maior frequência na composição da árvore. Além de filtrar as entidades de acordo com o número mínimo de ocorrências delas dentro do corpus, apenas são escolhidos para verificação na Wikipedia os pares de entidades que coocorrem em pelo menos 1000 dos currículos selecionados para análise. O Quadro 18 apresenta os parâmetros utilizados para a geração da taxonomia.

Quadro 18 - Parâmetros utilizados para a geração da taxonomia C

Variação do método utilizada	Abordagem de coocorrência dentro do currículo
Número de currículos analisados	20.000
Número mínimo de ocorrências de cada entidade	1.000
Número de níveis considerados na hierarquia da DBpedia	1
Número mínimo de coocorrências de cada par de termos para validar a relação hierárquica	1.000

Devido ao caráter mais restritivo dos parâmetros informados, a taxonomia gerada possui menos termos e relações hierárquicas. A Figura 13 mostra a representação gráfica da taxonomia gerada.

Figura 13 – Representação gráfica da taxonomia C



Fonte: Elaboração do autor, 2017.

O número de termos elegidos para essa taxonomia é de apenas dez. Um desses termos ainda é cíclico – no caso, o termo Medicina ocorre em ambas as extremidades da taxonomia.

As estatísticas coletadas sobre a taxonomia C podem ser visualizadas no Quadro 19.

Quadro 19 - Estatísticas obtidas sobre a taxonomia C

Número de *tokens*	9.325.838
Tokens reconhecidos	3.052.738
Fator de reconhecimento	0,3273420
Tokens utilizados	28.163
Fator de utilização de *tokens*	0,00
Contagem de termos	10
Nível máximo	3
Número de expansões	4
Soma do número de termos expandidos	7
Média de termos por expansão	1,750
Nível médio do termo	2,000
Número de termos cíclicos	1
Fator de termos cíclicos	0,100
Horizontalidade	0,300
Verticalidade	3,333
Número de termos encontrados no AGROVOC	7
Porcentagem de termos encontrados no AGROVOC	70,000
Número de relacionamento hierárquicos encontrados no AGROVOC	0

Conforme as informações apresentadas no quadro, os números menores de termos e relações hierárquicas afetam as estatísticas absolutas da taxonomia, mas não as baseadas em porcentagens ou razões. O nível médio da taxonomia é próximo aos apresentados nas taxonomias A e B, assim como a porcentagem de termos encontrados no AGROVOC.

5.3 ANÁLISE DOS RESULTADOS

Nesta seção é possível verificar o comportamento de execução do método elaborado nesta pesquisa sobre os currículos da Plataforma Lattes. Os resultados foram coletados a partir da execução de baterias de testes envolvendo os parâmetros informados nos quadros.

Os parâmetros utilizados levaram em consideração as limitações da implementação em *software* do método. Foi utilizada uma amostra de 20.000 currículos da área das Ciências Agrárias em todos os cenários apresentados nesta seção. A variação dos valores empregados nos testes deste experimento tiveram como premissa os seguintes objetivos:

- A variação dos valores empregados no parâmetro de número mínimo de ocorrências mostra a influência que a redução do conjunto de termos relevantes dentro do domínio possui nas taxonomias resultantes.
- Já a variação do número de níveis considerados na DBpedia exemplifica como a forma de utilização da base de conhecimento age sobre os resultados obtidos.
- Por fim, a variação do número mínimo de coocorrências de cada par de termos (nos casos em que este parâmetro se emprega) demonstra o impacto que a utilização de estratégias de verificação de frequências conjuntas mais restritivas exerce na geração das taxonomias.

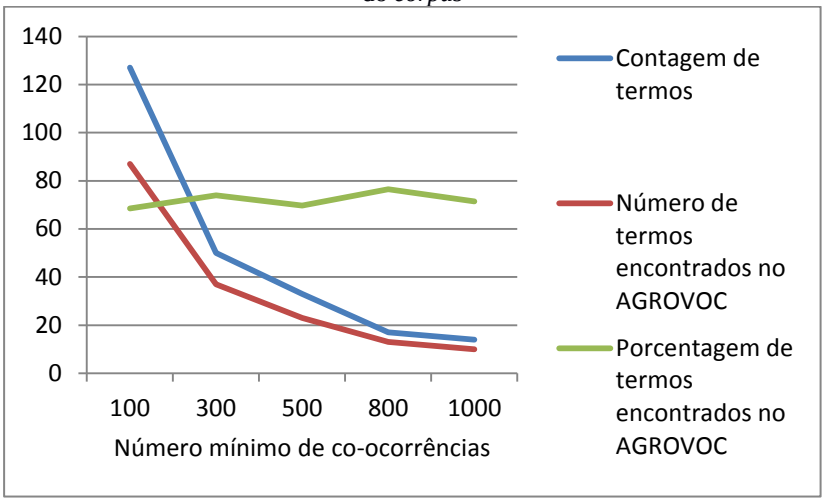
A primeira análise realizada nessa seção refere-se à porcentagem de temos encontrados no AGROVOC ao utilizar uma determinada série de parâmetros nas diferentes abordagens. O Quadro 20 apresenta os parâmetros utilizados nos testes para a geração do gráfico de análise.

Quadro 20 - Parâmetros utilizados para a geração das taxonomias do gráfico A

Variação do método utilizada	Abordagem de coocorrência dentro do *corpus*
Número de currículos analisados	20.000
Número mínimo de ocorrências de cada entidade	100-1.000
Número de níveis considerados na hierarquia da DBpedia	1

A Figura 14 apresenta a contagem de termos obtida nas taxonomias à medida que o parâmetro de filtragem mínima de ocorrência de entidades é alterado entre 100 e 1000 vezes e a porcentagem de termos encontrados que pertencem ao AGROVOC.

Figura 14 – Gráfico A: Contagem de termos X Porcentagem de termos encontrados no AGROVOC para a abordagem baseada em coocorrências dentro do *corpus*



Fonte: Elaboração do autor, 2017.

É possível observar que, mesmo ao aumentar o valor do parâmetro de filtragem da ocorrência mínima de entidades, a porcentagem de termos encontrados no AGROVOC tende a se manter em um mesmo intervalo no gráfico, dentro da faixa de 60% a 80%. A oscilação da porcentagem ocorre por conta da variação dos termos contidos no conjunto de entidades válidas. O tamanho desse conjunto varia em função do domínio (neste caso, da qualidade do texto informado pelos pesquisadores no Currículo Lattes) e do parâmetro de mínimo de ocorrências de cada entidade. Com a variação desse conjunto, o número de termos selecionados que também existem na AGROVOC pode variar.

Para a análise B, foram adotados os mesmos parâmetros, porém houve a alteração da variação do método utilizada – dessa vez, foi adotada a abordagem de verificação de coocorrências dentro do currículo, com o número mínimo de coocorrências para cada par de termos definido em 1. O Quadro 21 apresenta os parâmetros utilizados neste teste.

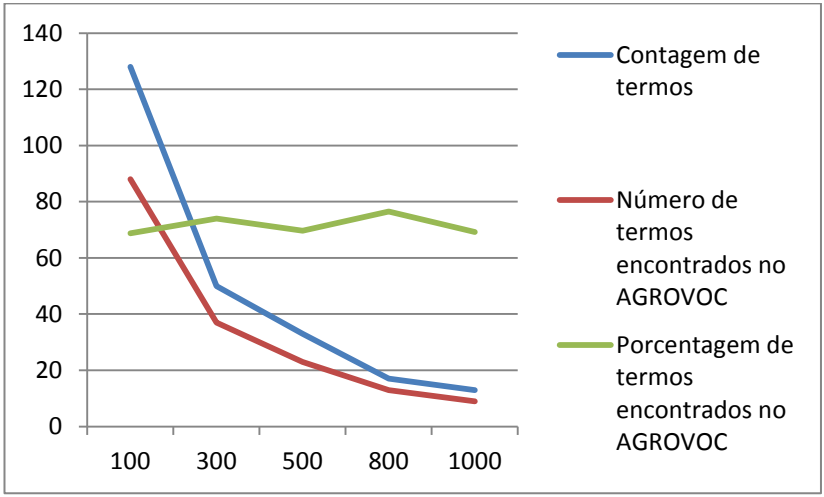
Quadro 21 - Parâmetros utilizados para a geração das taxonomias do gráfico B

Variação do método utilizada	Abordagem de coocorrência

	dentro do currículo
Número de currículos analisados	20.000
Número mínimo de ocorrências de cada entidade	100-1000
Número de níveis considerados na hierarquia da DBpedia	1
Número mínimo de coocorrências de cada par de termos para validar a relação hierárquica	1

A Figura 15 apresenta a contagem de termos obtidos nas taxonomias (na medida em que o parâmetro de filtragem mínima de ocorrência de entidades é alterado entre 100 e 1000 vezes), o número de termos encontrados no AGROVOC e a porcentagem correspondente de termos encontrados para essa variação do método.

Figura 15 – Gráfico B: Contagem de termos X Porcentagem de termos encontrados no AGROVOC para a abordagem baseada em coocorrências dentro do currículo



Fonte: Elaboração do autor, 2017.

Os resultados obtidos são bastante semelhantes aos encontrados para as taxonomias geradas a partir da abordagem baseada em coocorrências dentro do *corpus* como um todo. Dessa forma, a verificação da existência ou não de coocorrências dentro de um mesmo

currículo não impacta de maneira significativa no número de termos obtidos nas taxonomias geradas através do método.

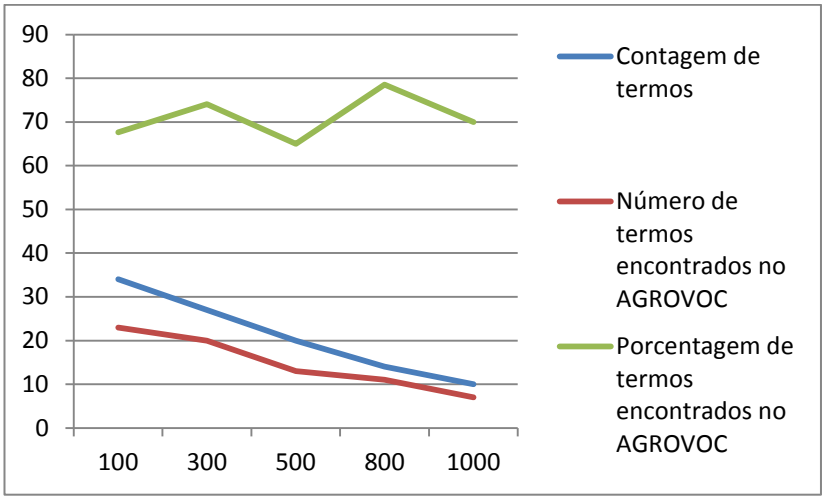
No entanto, a filtragem das relações hierárquicas elegíveis de verificação na DBpedia de acordo com a quantidade de coocorrências pode influenciar bastante nas taxonomias geradas. O Quadro 22 apresenta os parâmetros utilizados para a realização da análise apresentada no Gráfico C.

Quadro 22 - Parâmetros utilizados para a geração das taxonomias do gráfico C

Variação do método utilizada	Abordagem de coocorrência dentro do currículo
Número de currículos analisados	20.000
Número mínimo de ocorrências de cada entidade	100-1000
Número de níveis considerados na hierarquia da DBpedia	1
Número mínimo de coocorrências de cada par de termos para validar a relação hierárquica	1000

A Figura 16 apresenta o Gráfico C, que comprova a afirmação ao demonstrar a quantidade de termos obtidos e a porcentagem de termos encontrados no AGROVOC, filtrando as relações hierárquicas da árvore de forma que somente sejam consideradas as que ocorreram pelo menos 1000 vezes nos currículos.

Figura 16 – Gráfico C: Contagem de termos X Porcentagem de termos encontrados no AGROVOC para a abordagem baseada em coocorrências dentro do currículo, com filtragem da quantidade mínima das coocorrências nos currículos fixada em 1000 vezes



Fonte: Elaboração do autor, 2017.

A contagem de termos obtida diminui substancialmente com a adição do parâmetro de filtragem de ocorrência mínima das relações. A porcentagem de termos encontrados no AGROVOC oscila próximo dos valores encontrados nas outras abordagens, mas devido ao número menor de termos encontrados, as curvas existentes no gráfico tornam-se mais acentuadas.

O número de termos cíclicos encontrados nas árvores aumenta de acordo com o número de termos existentes dentro da árvore. Uma das formas de se provocar esse aumento de termos na árvore é alterar o parâmetro de níveis considerados na hierarquia da DBpedia. O Quadro 23 apresenta os parâmetros utilizados para a geração do gráfico que comprova essa afirmação.

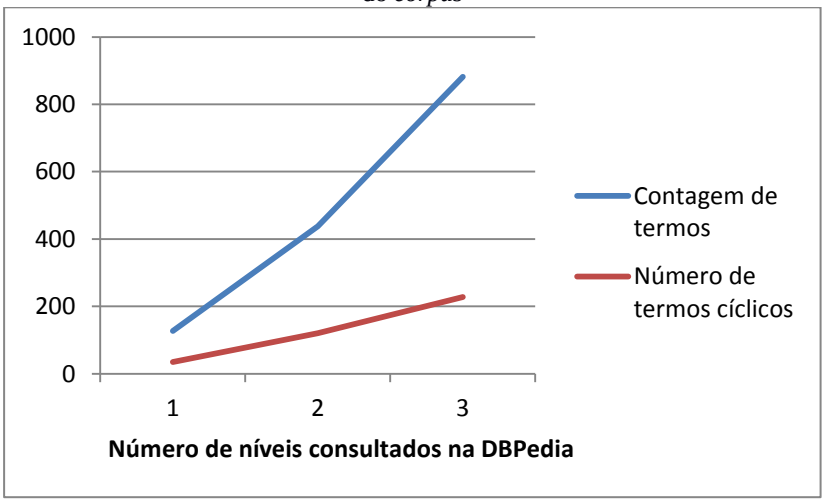
Quadro 23 - Parâmetros utilizados para a geração das taxonomias do gráfico D

Variação do método utilizada	Abordagem de coocorrência dentro do *corpus*
Número de currículos analisados	20.000
Número mínimo de ocorrências de cada entidade	100

Número de níveis considerados na hierarquia da DBpedia	1-3

Através desses parâmetros, foi possível gerar o gráfico apresentado na Figura 17, que compara o total de termos existentes na árvore com a quantidade de termos cíclicos encontrados.

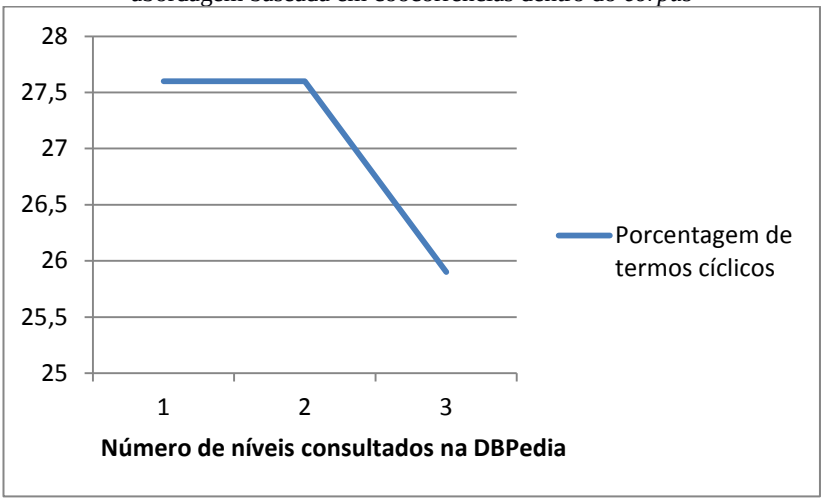
Figura 17 – Gráfico D-1: Contagem de termos X Número de termos cíclicos encontrados no AGROVOC para a abordagem baseada em coocorrências dentro do *corpus*



Fonte: Elaboração do autor, 2017.

Embora a quantidade de termos cíclicos aumente, o porcentual de termos cíclicos se mantém em relação ao total de termos presentes na taxonomia. A Figura 18 mostra essa proporção para cada um dos níveis considerados na hierarquia da DBpedia nesta análise.

Figura 18 – Gráfico D-2: Porcentagem de termos cíclicos encontrados para a abordagem baseada em coocorrências dentro do *corpus*



Fonte: Elaboração do autor, 2017.

Conforme apresenta o gráfico, a diferença entre os percentuais obtidos para cada execução da implementação do método é pequena. Ao considerar todas as baterias de testes, o percentual de termos cíclicos oscilou entre 7,7% e 32%, visto que o fluxo gerado pelo método e os parâmetros informados influenciam nos resultados obtidos.

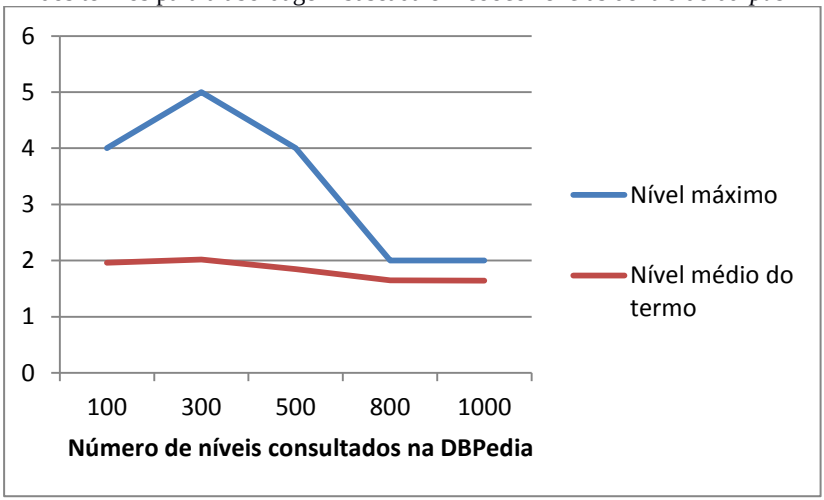
A próxima análise visa demonstrar a variação do nível máximo de profundidade alcançado por pelo menos um dos termos e a média do nível dos termos dentro das árvores produzidas durante a variação dos parâmetros informados para a execução do método. O Quadro 24 apresenta os parâmetros utilizados para a produção das taxonomias utilizadas nessa análise.

Quadro 24 - Parâmetros utilizados para a geração das taxonomias do gráfico E

Variação do método utilizada	Abordagem de coocorrência dentro do *corpus*
Número de currículos analisados	20.000
Número mínimo de ocorrências de cada entidade	100-1000
Número de níveis considerados na hierarquia da DBpedia	1

A Figura 19 apresenta as estatísticas descritas no parágrafo anterior para as variações dos parâmetros usadas na análise.

Figura 19 – Gráfico E: Nível máximo de profundidade na árvore X nível médio dos termos para a abordagem baseada em coocorrências dentro do *corpus*



Fonte: Elaboração do autor, 2017.

O nível máximo da árvore variou de forma assíncrona com a quantidade de termos existentes na árvore. Essa medida possui esse comportamento por estar mais vinculada aos relacionamentos existentes entre os termos na DBpedia. Já o nível médio do termo manteve-se aproximadamente o mesmo para as execuções do método, tendendo a uma leve queda à medida que a contagem de termos na árvore se reduz.

6 CONSIDERAÇÕES FINAIS E TRABALHOS FUTUROS

Esta seção apresenta as conclusões e considerações finais realizadas a partir da análise dos resultados obtidos com a execução do experimento.

O conteúdo desse capítulo segmenta-se em três seções. A seção inicial trata do alcance dos objetivos propostos no Capítulo 1. Por sua vez, a segunda seção aborda as contribuições realizadas por essa pesquisa ao cenário acadêmico e organizacional quanto à construção de taxonomias. Por último, a terceira seção apresenta sugestões de trabalhos futuros que podem ser originados a partir dessa pesquisa.

6.1 PÓS-PESQUISA E ALCANCE DOS OBJETIVOS

Taxonomias são artefatos constituintes de soluções da Engenharia do Conhecimento, tanto em sua disposição principal (nas quais estas cumprem o papel de organizar e classificar hierarquicamente) quanto como elementos estruturais para a construção de outros artefatos para a organização e representação do conhecimento.

A contribuição realizada por esse trabalho tem por objetivo auxiliar na construção destes artefatos. A partir dos experimentos realizados, foi possível verificar que o método desenvolvido permite automatizar parcialmente ou plenamente a construção de taxonomias. Conforme pode ser visto nas seções iniciais, a construção desse método foi fixada como objetivo principal desse trabalho. Dessa forma, pode-se afirmar que esse objetivo foi alcançado ao se verificar os resultados obtidos e as taxonomias geradas.

Além de um objetivo principal, essa pesquisa também demarcou objetivos específicos. O primeiro destes objetivos trata da apresentação do cenário atual da construção de taxonomias de forma automatizada, que pode ser visto na seção 2.3.2 dentro do referencial teórico. A apresentação deste cenário permitiu o levantamento das publicações na área e a seleção de componentes e abordagens a serem utilizadas no método. Os objetivos seguintes abordaram a proposição do método a partir do referencial teórico levantado, a sua derivação em *software* e a sua avaliação no formato de um experimento. Todas estes objetivos foram atingidos.

O método proposto auxilia na eliminação do gargalo inicial para a construção de taxonomias. Ao fornecer propostas de taxonomias iniciais, o artefato desenvolvido permite que as etapas iniciais para sua construção sejam condensadas e ocupem menos tempo. Dessa forma,

parte dos recursos que estariam destinados para a sua construção pode ser movidos para outras áreas prioritárias.

A construção deste método utilizou como base outras abordagens já publicadas anteriormente, embora o formato e cenário de aplicação deste método não tenham sido identificados em nenhuma dessas pesquisas. A abordagem utilizando a DBpedia para a verificação de relações taxonômicas é explorada também por Chernyak e Mirkin (2014) e Xavier e Lima (2009). Esta escolha evidencia a importância que o *crowdworking* pode exercer na construção de soluções de Engenharia do Conhecimento.

6.2 CONTRIBUIÇÕES REALIZADAS

Esta pesquisa entrega como artefato um método flexível o bastante ao ponto de permitir diversas variações de tecnologias para a obtenção de taxonomias. Seu acionamento pode ser realizado de forma modular: é possível fornecer uma lista de termos para que se construam as taxonomias ou adotar a estratégia de extrair as entidades relevantes de um repositório de documentos. As taxonomias resultantes podem ser aprimoradas posteriormente por especialistas, tornando elas ainda mais apropriadas aos domínios aos quais se destinam.

A flexibilidade adotada nos módulos apresentados permite que diversos domínios de informação possam ser utilizados como recurso de entrada para a geração de taxonomias. No contexto desse trabalho foi utilizada a Plataforma Lattes, mas outros repositórios organizacionais de informação (inclusive privados) também podem servir como entrada para processamento pelo método. Manuais, relatórios e outros documentos são exemplos de artefatos que podem ser utilizados. Os resultados podem variar em relação à qualidade das taxonomias geradas, mas a alteração dos parâmetros informados na entrada do método pode auxiliar a obter taxonomias mais adequadas aos seus cenários de aplicação.

É possível também alterar a estrutura utilizada para a recuperação de informações hierárquicas entre dois termos. No método foi proposta a utilização da DBpedia. Outras bases de conhecimento que tratem informações e conceitos de forma hierárquica também podem ser utilizadas. A integração com a tecnologia SPARQL abstrai os detalhes sobre formato de representação do conhecimento, o que atribui flexibilidade a essa escolha.

Além disso, o método também permite a geração de taxonomias em outros idiomas, bastando para isso reunir os modelos de

reconhecimento de entidades e as árvores categóricas da DBpedia específicos desses idiomas.

6.3 TRABALHOS FUTUROS

As sugestões de trabalhos futuros apresentadas nessa seção são resultantes de oportunidades que surgiram durante o desenvolvimento do trabalho. Por questões de delimitação de escopo e de falta de tempo hábil para pesquisa essas oportunidades não foram exploradas.

Quanto à aplicação do método, a primeira proposta é a utilização da implementação sobre uma base de informações diferente do subconjunto de currículos da Plataforma Lattes utilizado nesse trabalho. A análise realizada contemplou um subconjunto relacionado a currículos com especialidades em agricultura, pecuária e abastecimento. É possível aplicar o método e coletar os resultados sobre os domínios das ciências exatas, biológicas ou da saúde, por exemplo, e verificar se seu funcionamento também é adequado nessas ocasiões.

Seguindo a mesma linha, é possível também aplicar o artefato em outros domínios de informação, conforme discutido na seção anterior. Repositórios de documentos organizacionais, publicações e perfis disponibilizados em mídias sociais também podem ser utilizados para a montagem das taxonomias, enriquecendo a análise e apuração do método para esses cenários.

É possível também utilizar o método aplicando outras bases de conhecimento. A DBpedia foi escolhida devido ao volume de dados contidos nela. A utilização de outras bases auxilia na avaliação do método e pode até aprimorar os resultados obtidos em alguns cenários.

Outra possibilidade, em especial em combinação com um conjunto de documentos em um idioma em específico, é a realização de testes do método com outros idiomas nos quais a DBpedia é disponibilizada. A qualidade das bases de conhecimento influencia nos resultados obtidos. As Wikipedias em inglês, sueco e cebuano são as derivações que possuem mais artigos. Testes obtidos com as bases de conhecimento desses idiomas podem afirmar se o método também pode ser utilizado nessas circunstâncias.

Quanto à implementação do método, destaca-se a possibilidade de alterar o motor de reconhecimento de entidades utilizado. Pelo fato dela ser construída através de componentes modularizados, é possível alterar o módulo de reconhecimento de entidades sem interferir nas funcionalidades dos demais componentes. O motor de reconhecimento escolhido nessa pesquisa foi o Apache OpenNLP. Existem outros

motores para o reconhecimento, inclusive alguns construídos de forma direcionada para o português.

Em relação à estratégia de identificação de relações entre termos, propõe-se o estudo e aplicação de outros modelos de correlação. Trabalhos como o de Bovo (2011) e o de Sérgio, Silva e Gonçalves (2016) enumeram uma série de modelos matemáticos presentes na literatura para identificar com maior precisão a existência de correlações entre pares de termos. Estes modelos podem ser aplicados para restringir de forma mais refinada o conjunto de relações identificadas pelo método e, consequentemente, possibilitando que este último possa construir taxonomias mais relevantes para o domínio.

Também é sugerida a melhoria da implementação do método, de forma que a mesma possa lidar com quantidades maiores de informação. O *software* codificado apresentou instabilidade ao lidar com quantidades maiores de 20.000 currículos. Isso ocorre devido à falta de estruturas de armazenamento apropriadas para computar as variáveis utilizadas durante o processamento. Aprimoramentos nesses pontos podem trazer resultados mais refinados e apropriados para os domínios em estudo.

REFERÊNCIAS

ABÁNADES, Miguel Á. et al. An algebraic taxonomy for locus computation in dynamic geometry. **Computer-aided Design**, v. 56, p.22-33, nov. 2014. Elsevier BV. http://dx.doi.org/10.1016/j.cad.2014.06.008.

ABOLFAZLI, Saeid et al. Rich Mobile Applications: Genesis, taxonomy, and open issues. **Journal Of Network And Computer Applications**, v. 40, p.345-362, abr. 2014. Elsevier BV. http://dx.doi.org/10.1016/j.jnca.2013.09.009.

AGANETTE, Elisangela; ALVARENGA, Lídia; SOUZA, Renato Rocha. Elementos constitutivos do conceito de Taxonomia. **Informação & Sociedade: Estudos**, João Pessoa, v. 3, n. 20, p.77-93, set./dez. 2010.

AHMED, Ejaz et al. Seamless application execution in mobile cloud computing: Motivation, taxonomy, and open challenges. **Journal Of Network And Computer Applications**, v. 52, n. 1, p.154-172, jun. 2015. Elsevier BV. http://dx.doi.org/10.1016/j.jnca.2015.03.001.

_____. Application optimization in mobile cloud computing: Motivation, taxonomies, and open challenges. **Journal Of Network And Computer Applications**, v. 52, p.52-68, jun. 2015. Elsevier BV. http://dx.doi.org/10.1016/j.jnca.2015.02.003.

ALALWAN, Jaffar Ahmad. A taxonomy for decision support capabilities of enterprise content management systems. **The Journal Of High Technology Management Research**, v. 24, n. 1, p.10-17, 2013. Elsevier BV. http://dx.doi.org/10.1016/j.hitech.2013.02.001.

ALVARENGA, Lídia. Representação do conhecimento na perspectiva da ciência da informação em tempo e espaço digitais. **Encontros Bibli**: revista eletrônica de biblioteconomia e ciência da informação, v. 8, n. 15, p.18-40, 1º sem. 2003. Universidade Federal de Santa Catarina (UFSC). http://dx.doi.org/10.5007/1518-2924.2003v8n15p18. Disponível em: <https://periodicos.ufsc.br/index.php/eb/article/view/1518-2924.2003v8n15p18/5233>. Acesso em: 20 fev. 2017.

AMARAL, Daniela Oliveira Ferreira do. **O Reconhecimento de Entidades Nomeadas por meio de Conditional Random Fields para a Língua Portuguesa**. 2013. 100 f. Dissertação (Mestrado) - Curso de Ciência da Computação, Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, 2013. Disponível em: <http://tede.pucrs.br/tde_arquivos/4/TDE-2014-04-24T051906Z-4975/Publico/457280.pdf>. Acesso em: 17 fev. 2017.

AMORIN, C. V. Organização do currículo – plataforma Lattes Curriculum vitae organization – the Lattes software platform. **Pesquisa Odontológica Brasileira**, v. 17, n. Supl 1, p. 18–22, 2003.

ANDRIANI, Mateus Lohn et al. Um experimento envolvendo a geração de mapas de tópicos automatizada a partir dos dados abertos do Sistema de Convênios (SICONV). In: LINKED OPEN DATA BRASIL, 1., 2014, Florianópolis. **Anais do LOD Brasil**. Florianópolis: EGC/UFSC, 2014. v. 1, p. 71 - 84.

BALANCIERI, Renato. **Análise de redes de pesquisa em uma plataforma de gestão em ciência e tecnologia**: uma aplicação à Plataforma Lattes. 2004. Dissertação (Mestrado em Engenharia de Produção) – Universidade Federal de Santa Catarina, 2004.

BEYDOUN, Ghassan et al. Providing metrics and automatic enhancement for hierarchical taxonomies. **Information Processing & Management**, v. 49, n. 1, p.67-82, jan. 2013. Elsevier BV. http://dx.doi.org/10.1016/j.ipm.2012.01.006.

BIZER, C. et al. **DBpedia - A crystallization point for the Web of Data.** Journal of Web Semantics, v. 7, n. 3, p. 154–165, 2009.

BLACKBURN, B. **Taxonomy design types**. AIIM E-doc Magazine, Maryland, USA. v.20, n.3, p.14-16, maio/jun. 2006.

BOCCATO, Vera Regina Casari. Os sistemas de organização do conhecimento nas perspectivas atuais das normas internacionais de construção. **InCID**: Revista de Ciência da Informação e Documentação, Ribeirão Preto, v. 2, n. 1, p. 165-192, jan./jun. 2011. Disponível em: <http://revistas.ffclrp.usp.br/incid/article/view/44/pdf>. Acesso em: 21 fev. 2017.

BOLLACKER, K.; COOK, R.; TUFTS, P. **Freebase**: A shared database of structured general human knowledge. Proceedings of the national conference on Artificial Intelligence, v. 22, n. 2, p. 1962, 2007.

BOLLACKER, K. et al. **Freebase**: A Collaboratively Created Graph Database For Structuring Human Knowledge. p. 1247–1249, 2008.

BOVO, Alessandro Botelho. **Um modelo de descoberta de conhecimento inerente à evolução temporal dos relacionamentos entre elementos textuais.** 2011. 155 f. Tese (Doutorado) - Curso de Engenharia e Gestão do Conhecimento, Centro Tecnológico, Universidade Federal de Santa Catarina, Florianópolis, 2010. Disponível em: <http://btd.egc.ufsc.br/wp-content/uploads/2011/05/Alessandro_Botelho_Bovo.pdf>. Acesso em: 21 mar. 2017.

BRÄSCHER, Marisa; CAFÉ, Lígia. **Organização da Informação ou Organização do Conhecimento?**. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 9. 2008, São Paulo, Anais... São Paulo: ANCIB, 2008. Disponível em: < http://cmapspublic.ihmc.us/rid=1KR7TM7S9-S3HDKP-5STP/BRASCHER%20CAF%C3%89(2008)-1835.pdf>. Acesso em: 20 fev. 2017.

BREWSTER, Christopher; WILKS, Yorick. Onologies, Taxonomies, Thesauri: Learning from Texts. In: DEEGAN, Marilyn (Org.). **The Keyword Project**: Unlocking Content through Computational Linguistics. London: Kings College, 2004. (Proceedings of the The Use of Computational Linguistics in the Extraction of Keyword Information from Digital Library Content: Workshop 5-6 February). Disponível em: <http://www.cbrewster.com/papers/KeyWord_FMO.pdf>. Acesso em: 25 mar. 2017.

BUCCELLA, Agustina et al. Marine ecology service reuse through taxonomy-oriented SPL development. **Computers & Geosciences**, v. 73, p.108-121, dez. 2014. Elsevier BV. http://dx.doi.org/10.1016/j.cageo.2014.09.004.

CAFARELLA, M. et al. **KnowItNow**: fast, scalable information extraction from the web. Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, n. October, p. 563–570, 2005.

CARACCIOLO, Caterina et al. Thesaurus maintenance, alignment and publication as linked data: the AGROVOC use case. **International Journal Of Metadata, Semantics And Ontologies**, [s.l.], v. 7, n. 1, p.65-75, 2012. Inderscience Publishers. http://dx.doi.org/10.1504/ijmso.2012.048511. Disponível em: <http://eprints.rclis.org/17735/1/IJMSO 7_1_Paper 6_PublishedVersion[1].pdf>. Acesso em: 11 set. 2016.

_____. The AGROVOC Linked Dataset. **Semantic Web**, [s.l.], v. 4, n. 3, p.341-348, 2013. IOS Press. http://dx.doi.org/10.3233/SW-130106. Disponível em: <http:// eprints.rclis.org /20648/11/SW106.pdf>. Acesso em: 11 set. 2016.

CARDENAS, Y. G**. Modelo de Ontologia para Representação de Jogos Digitais de Disseminação do Conhecimento**. 2014. 149 f. Dissertação (Mestrado) - Curso do Programa de Pós-graduação em Engenharia e Gestão do Conhecimento, Departamento de Centro Tecnológico, Universidade Federal de Santa Catarina, Florianópolis, 2014. Disponível em: <http://btd.egc.ufsc.br/wp-content/uploads/2014/05/Yuri-Gomes-Cardenas.pdf>. Acesso em: 30 out. 2014.

CARLAN, Eliana. **Sistemas de Organização do Conhecimento**: uma reflexão no contexto da Ciência da Informação. 2010. 195 f. Dissertação (Mestrado) - Curso de Ciência da Informação, Departamento de Ciência da Informação e Documentação, Universidade de Brasília, Brasília, 2010. Disponível em: <http://eprints.rclis.org/15298/1/Carlan-Eliana-Dissertacao.pdf>. Acesso em: 21 fev. 2017

CARLSON, Andrew et al. Toward an Architecture for Never-Ending Language Learning. In: CONFERENCE ON ARTIFICIAL INTELLIGENCE, 24., 2010, Atlanta. **Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI-10)**. Association for the Advancement of Artificial Intelligence, 2010. p. 1306 - 1313.

CARREIRA, Paulo; RESENDES, Sílvia; SANTOS, André C.. Towards automatic conflict detection in home and building automation systems. **Pervasive And Mobile Computing**, v. 12, p.37-57, jun. 2014. Elsevier BV. http://dx.doi.org/10.1016/j.pmcj.2013.06.001.

CECI, Flávio. **Um modelo semi-automático para a construção e manutenção de ontologias a partir de bases de documentos não estruturados**. 2010. 131 f. Dissertação (Mestrado) - Curso do Programa de Pós-graduação em Engenharia e Gestão do Conhecimento, Departamento de Centro Tecnológico, Universidade Federal de Santa Catarina, Florianópolis, 2010. Disponível em: <https://repositorio.ufsc.br/handle/123456789/94643>. Acesso em: 07 out. 2014.

CHERNYAK, Ekaterina; MIRKIN, Boris. A Method for Refining a Taxonomy by Using Annotated Suffix Trees and Wikipedia Resources. **Procedia Computer Science**, v. 31, n. 1, p.193-200, dez. 2014. Elsevier BV. http://dx.doi.org/10.1016/j.procs.2014.05.260.

CHILTON, Lydia B. et al. Cascade: crowdsourcing taxonomy creation. In: SIGCHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS, 13., 2013, Paris. **Proceedings of the SIGCHI Conference on Human Factors in Computing Systems**. New York: ACM, 2013. p. 1 - 10.

CONWAY, Susan; SLIGAR, Char. Building taxonomies. In.: _____. **Unlocking knowledge assets**. Redmont: Microsoft Press, 2002. Cap. 6. Disponível em: <https://www.microsoft.com/mspress/books/sampchap/5516.aspx#SampleChapter>. Acesso em: 28 nov. 2015.

CORONA, Igino; GIACINTO, Giorgio; ROLI, Fabio. Adversarial attacks against intrusion detection systems: Taxonomy, solutions and open issues. **Information Sciences**, [s.l.], v. 239, p.201-225, ago. 2013. Elsevier BV. http://dx.doi.org/10.1016/j.ins.2013.03.022.

CUPANI, A. La peculiaridad del conocimiento tecnológico. **Scientia Studia**, v. 4, n. 3, p. 353-71, 2006.

DERRAC, Joaquín; GARCÍA, Salvador; HERRERA, Francisco. Fuzzy nearest neighbor algorithms: Taxonomy, experimental analysis and prospects. **Information Sciences**, v. 260, p.98-119, mar. 2014. Elsevier BV. http://dx.doi.org/10.1016/j.ins.2013.10.038.

DUKARIC, Robert; JURIC, Matjaz B.. Towards a unified taxonomy and architecture of cloud frameworks. **Future Generation Computer**

Systems, v. 29, n. 5, p.1196-1210, jul. 2013. Elsevier BV. http://dx.doi.org/10.1016/j.future.2012.09.006.

ETZIONI, O. et al. Unsupervised named-entity extraction from the Web: An experimental study. **Artificial Intelligence**, v. 165, n. 1, p. 91–134, 2005.

FACHIN, G. R. B. **Ontologia de referência para periódico científico digital**. 2011. 401 f. Tese (Doutorado) - Curso do Programa de Pós-graduação em Engenharia e Gestão do Conhecimento, Departamento de Centro Tecnológico, Universidade Federal de Santa Catarina, Florianópolis, 2011. Disponível em: <http://btd.egc.ufsc.br/wp-content/uploads/2011/10/Gleisy-Fachin.pdf> Acesso em: 30 out. 2014.

FATEMA, Kaniz et al. A survey of Cloud monitoring tools: Taxonomy, capabilities and objectives. **Journal Of Parallel And Distributed Computing**, v. 74, n. 10, p.2918-2933, out. 2014. Elsevier BV. http://dx.doi.org/10.1016/j.jpdc.2014.06.007.

FERNÁNDEZ-MARTÍNEZ, F.; GARCÍA, A. Hernández; MARÍA, F. Díaz de. Succeeding metadata based annotation scheme and visual tips for the automatic assessment of video aesthetic quality in car commercials. **Expert Systems With Applications**, v. 42, n. 1, p.293-305, jan. 2015. Elsevier BV. http://dx.doi.org/10.1016/j.eswa.2014.07.033.

FONSECA, Evandro B. et al. Reconhecimento de Entidades Nomeadas para o Portugues Usando o OpenNLP. In: ENCONTRO NACIONAL DE INTELIGÊNCIA ARTIFICIAL E COMPUTACIONAL (ENIAC), 12., 2015, Natal. **XII Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)**, Natal, Online, 2015. p. 1 - 5. Disponível em: <http://www. lbd.dcc.ufmg.br/colecoes/eniac/ 2015/011.pdf>. Acesso em: 14 ago. 2016.

FONSECA, R. C. V.. **Metodologia Do Trabalho Científico**. 1ª ed. Curitiba: IESDE Brasil S.A., 2009.

FREITAS, Cláudia; AFONSO, Susana (Eds.). **Árvores deitadas**: Descrição do formato e descrição das opções de análise na Floresta Sintá(c)tica. 2008. 151 p. O projeto da Floresta Sintáctica é (parcialmente) financiado pela Fundação para a Ciência e Tecnologia,

co-financiada pelo POSI, através do projecto POSI/PLP/43931/2001 (Linguateca). Disponível em: <http://www.linguateca.pt/Floresta/BibliaFlorestal/>. Acesso em: 17 ago. 2016.

FUJITA, M. N. S. L. Organização e representação do conhecimento no Brasil: análise de aspectos conceituais e da produção científica do ENANCIB no período de 2005 a 2007. **Tendências da Pesquisa Brasileira em Ciência da Informação**, v. 1, n. 1, p. 1-32, 2008. Disponível em: <http://basessibi.c3sl.ufpr.br/brapci/v/a/7781>. Acesso em: 19 Fev. 2017

GALAR, Mikel et al. A survey of fingerprint classification Part I: Taxonomies on feature extraction methods and learning models. **Knowledge-based Systems**, v. 81, p.76-97, jun. 2015. Elsevier BV. http://dx.doi.org/10.1016/j.knosys.2015.02.008.

GARCÍA, José Luís Garcia. **Ciencias de Joseleg**: Los reinos se multiplican. 2013. Disponível em: <http://cienciasdejoseleg.blogspot.com.br/2013/08/los-reinos-se-multiplican.html>. Acesso em: 21 nov. 2016.
GIL, Antônio Carlos. Como elaborar projetos de pesquisa. 4. ed. São Paulo: Atlas, 2002.

GHANDI, Somayé; MASEHIAN, Ellips. Review and taxonomies of assembly and disassembly path planning problems and approaches. **Computer-aided Design**, v. 67-68, p.58-86, out. 2015. Elsevier BV. http://dx.doi.org/10.1016/j.cad.2015.05.001.

GOHN, Maria da Glória Marcondes. A pesquisa na produção do conhecimento: questões metodológicas. **ECCOS-Revista Científica**, v. 7, n. 2, p. 253-274, 2005.

GOOD, Benjamin M. et al. Fast, Cheap and Out of Control: A Zero Curation Model for Ontology Development. In: PACIFIC SYMPOSIUM ON BIOCOMPUTING, 11., 2006, Grand Wailea. **Proceedings...** On-line: PSB, 2006. p. 128 - 139. Disponível em: <http://psb.stanford.edu/psb-online/proceedings/psb06/good.pdf>. Acesso em: 17 mar. 2017.´

GRUBER, Thomas R. A translation approach to portable ontology specifications. **Knowledge acquisition**, v. 5, n. 2, p. 199-220, 1993.

GUPTA, Ravi Kumar; GURUMOORTHY, Balan. Classification, representation, and automatic extraction of deformation features in sheet metal parts. **Computer-aided Design**, v. 45, n. 11, p.1469-1484, nov. 2013. Elsevier BV. http://dx.doi.org/10.1016/j.cad.2013.06.010.

HODGE, Gail. Knowledge Organization Systems: An Overview. In: HODGE, Gail. **Systems of Knowledge Organization for Digital Libraries**: Beyond Traditional Authority Files. Washington Dc: The Digital Library Federation, 2000. Cap. 1. p. 3-9. Disponível em: <https://www.clir.org/pubs/reports/reports/pub91/pub91.pdf>. Acesso em: 21 fev. 2017.

HOQUE, N. et al. Network attacks: Taxonomy, tools and systems. **Journal Of Network And Computer Applications**, v. 40, p.307-324, abr. 2014. Elsevier BV. http://dx.doi.org/10.1016/j.jnca.2013.08.001.

HOSSEINI, Mahmood et al. Crowdsourcing: A taxonomy and systematic mapping study. **Computer Science Review**, v. 17, p.43-69, ago. 2015. Elsevier BV. http://dx.doi.org/10.1016/j.cosrev.2015.05.001.

IJNTEMA, W. et al. A lexico-semantic pattern language for learning ontology instances from text. **Web Semantics: Science, Services and Agents on the World Wide Web**, v. 15, p. 37–50, set. 2012. Disponível em: <http://linkinghub.elsevier.com/retrieve/pii/S1570826812000121>. Acesso em: 15 ago. 2014.

LIN, J.; MENDELZON, **A. Knowledge base merging by majority. Dynamic Worlds**: From the Frame Problem to Knowledge Management, p. 195–217, 1999.

KASNECI, G. et al. The YAGO-NAGA approach to knowledge discovery. **SIGMOD Rec.**, v. 37, n. 4, p. 41–47, 2008. Disponível em: <http://portal.acm.org/citation.cfm?id=1519103.1519110&coll=Portal&dl=GUIDE&CFID=52635999&CFTOKEN=19805102\nhttp://portal.acm.org/ft_gateway.cfm?id=1519110&type=pdf&coll=Portal&dl=GUIDE&CFID=52635999&CFTOKEN=19805102>.

KENDAL, Simon L.; CREEN, Malcolm. **An introduction to knowledge engineering**. 1. ed. London: Springer, 2007.

KIM, Tae Hun; CENFETELLI, Ronald T.; BENBASAT, Izak. Organizational Performance With Environmental Knowledge Intensity: Resource Vs. Knowledge-Based Performance. In: INTERNATIONAL CONFERENCE ON INFORMATION SYSTEMS, 33., 2012, Orlando. **Proceedings of the international conference on information systems**, Orlando: On-line, 2012. p. 1 - 11. Disponível em: <https://pdfs.semanticscholar.org/d73a/4f48b334f6bfad9a0a3008c7ed3b1d1a52b8.pdf>. Acesso em: 22 out. 2016.

KNIJFF, J.; FRASINCAR, F.; HOGENBOOM, F. Domain taxonomy learning from text: The subsumption method versus hierarchical clustering. **Data and Knowledge Engineering**, v. 83, p. 54–69, jan. 2013. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0169023X12000973>. Acesso em: 20 mar. 2015.

KNOX, Rita E.; LOGAN, Debra. **What Taxonomies Do for the Enterprise**. 2003. Elaborada por Gartner Research. Disponível em: <https://www.gartner.com/doc/409155/taxonomies-enterprise>. Acesso em: 20 out. 2016.

KÖCHE, J. C. **Fundamentos de metodologia científica: teoria da ciência e prática da pesquisa**. Petrópolis: Vozes, 1997.

LASTRES, Helena Maria Martins; FERRAZ, João Carlos. Economia da Informação, do Conhecimento e do Aprendizado. In: LASTRES, Helena Maria Martins; ALBAGLI, Sarita (Org.). **Informação e globalização na era do conhecimento**. Rio de Janeiro: Campus, 1999. p. 27-57.

LIMA, Telma CS; MIOTO, Regina Célia Tamaso. Procedimentos metodológicos na construção do conhecimento científico: a pesquisa bibliográfica. **Revista Katálysis**, v. 10, n. 1, p. 37-45, 2007.

LIU, Jieyao et al. Application partitioning algorithms in mobile cloud computing: Taxonomy, review and future directions. **Journal Of Network And Computer Applications**, v. 48, p.99-117, fev. 2015. Elsevier BV. http://dx.doi.org/10.1016/j.jnca.2014.09.009.

MACIAS-CHAPULA, C. A. O papel da informetria e da cienciometria e sua perspectiva nacional e internacional. **Ciência da Informação**, v. 27, n. 2, p. 134–140, 1998.

MEIJER, K.; FRASINCAR, F.; HOGENBOOM, F. A semantic approach for extracting domain taxonomies from text. **Decision Support Systems**, v. 62, p. 78–93, jun. 2014. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0167923614001031>. Acesso em: 24 mai. 2015.

MELGAR-SASIETA, H. C. A. S.; BEPPLER, F. D.; PACHECO, R. C. S. A memória organizacional no contexto da engenharia do conhecimento. **DataGramaZero**, v. 12, n. 4, 2011. Disponível em: <http://basessibi.c3sl.ufpr.br/brapci/v/a/10951>. Acesso em: 25 Mar. 2017.

MENA-CHALCO, J. P.; CESAR JUNIOR, R. M. scriptLattes: an open-source knowledge extraction system from the Lattes platform. **Journal of the Brazilian Computer Society**, v. 15, n. 4, p. 31–39, 2009. Disponível em: <http://link.springer.com/10.1007/BF03194511>. Acesso em: 26 nov. 2015.

MEIJER, Kevin; FRASINCAR, Flavius; HOGENBOOM, Frederik. A semantic approach for extracting domain taxonomies from text. **Decision Support Systems**, v. 62, p.78-93, jun. 2014. Elsevier BV. http://dx.doi.org/10.1016/j.dss.2014.03.006.

MENDES, Pablo N. et al. DBpedia spotlight: shedding light on the web of documents. In: INTERNATIONAL CONFERENCE ON SEMANTIC SYSTEMS, 7., 2011, Graz. **Proceedings of the 7th International Conference on Semantic Systems**. New York: ACM, 2011. p. 1 - 8.

NADEAU, David; SEKINE, Satoshi. A survey of named entity recognition and classification. **Lingvisticæ Investigationes: International Journal of Linguistics and Language Resources**, v. 30, n. 1, p.3-26, 10 ago. 2007. John Benjamins Publishing Company. http://dx.doi.org/10.1075/li.30.1. Disponível em: <https://nlp.cs.nyu.edu/sekine/papers/li07.pdf>. Acesso em: 17 fev. 2017.

NATIONAL INFORMATION STANDARDS ORGANIZATION (2005). **ANSI/NISO Z39.19-200**3: guidelines for the construction,

format, and management of monolingual thesauri. 2005. Disponível em: <http://www.niso.org/kst/reports/standards/kfile_download?id%3Austring%3Aiso-8859-1=Z39-19-2005.pdf&pt=RkGKiXzW643YeUaYUqZ1BFwDhIG4-24RJbcZBWg8uE4vWdpZsJDs4RjLz0t90_d5_ymGsj_IKVaGZww13HuDlYn5U74YdfA-3TffjxYQ25QrtR8PONuJLqxvo-l0NIr5>. Acesso em: 12 out. 2015.

OCHOA, J. L. et al. A semantic role labelling-based framework for learning ontologies from Spanish documents. **Expert Systems with Applications**, v. 40, n. 6, p. 2058–2068, maio 2013. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0957417412011311>. Acesso em: 20 ago. 2014.

PACHECO, R. C. S. **Uma metodologia de desenvolvimento de plataformas de governo para geração e divulgação de informações e de conhecimento**. Artigo apresentado em cumprimento a requisito parcial de concurso para professor no INE/UFSC. 35 p. Florianópolis, 14 de janeiro de 2003.

PACHECO, R. C. S.; KERN, V. M. Uma ontologia comum para a integração de bases de informações e conhecimento sobre ciência e tecnologia. **Ciência da Informação**, v. 30, n. 3, p. 56–63, 2001.

PALOMARES, Iván et al. Consensus under a fuzzy context: Taxonomy, analysis framework AFRYCA and experimental case of study. **Information Fusion**, v. 20, p.252-271, nov. 2014. Elsevier BV. http://dx.doi.org/10.1016/j.inffus.2014.03.002.

PELLUCCI, Paulo Roberto Simões et al. Utilização de técnicas de aprendizado de máquina no reconhecimento de entidades nomeadas no Português. **E-xacta**, Belo Horizonte, v. 4, n. 1, p.73-81, 24 jul. 2011. Revista Exacta. http://dx.doi.org/10.18674/exacta.v4i1.305. Disponível em: <http://revistas.unibh.br/index.php/dcet/article/download/305/164>. Acesso em: 18 fev. 2017.

PLOSKER, G. Taxonomies: facts and opportunities for information professionals. Online. **ABI/Inform Global**, v.1, n.29, jan./fev. 2005. p. 58-60.

PONCHIROLLI, O.; FIALHO, F. Gestão estratégica do conhecimento como parte da estratégia empresarial. **Revista FAE**, Curitiba, p. 127–

138, 2005. Disponível em: <http://www.fae.edu/publicacoes/pdf/revista_da_fae/rev_fae_v8_n1/rev_fae_v8_n1_11.pdf>.

PREECE, A. et al. Better knowledge management through knowledge engineering. IEEE Intelligent Systems, v. 16, n. 1, p.36-43, jan. 2001. Institute of Electrical and Electronics Engineers (IEEE). http://dx.doi.org/10.1109/5254.912383.

PUNERA, K. V. K. **Enhanced Classification through Exploitation of Hierarchical Structures.** 2007. 234 f. Dissertação (Doutorado em Filosofia) – Faculty of the Gratuate School, University of Texas, Austin. Disponível em: <https://repositories2.lib.utexas.edu/bitstream/handle/2152/3265/punerak87642.pdf?sequence=2>. Acesso em: 13 out. 2015.

R, Sujatha; RAO, Bandaru Rama Krishna. Taxonomy Construction Techniques: Issues and Challenges. **Indian Journal Of Computer Science And Engineering**, v. 2, n. 5, p.661-671, out./nov. 2011. Disponível em: <http://www.ijcse.com/docs/INDJCSE11-02-05-006.pdf >. Acesso em: 17 out. 2016.

SAMPAIO, R. F.; MANCINI, M. C.. Estudos de revisão sistemática: um guia para síntese criteriosa da evidência científica. **Revista Brasileira de Fisioterapia,** São Carlos, v. 11, n. 1, p.83-89, fev. 2007. Bimestral. FapUNIFESP (SciELO). http://dx.doi.org/10.1590/s1413-35552007000100013. Disponível em: <http://www.scielo.br/scielo.php?pid=S1413-35552007000100013&script=sci_abstract&tlng=pt>. Acesso em: 09 jul. 2016.

SANG, Erik F. Tjong Kim; MEULDER, Fien de. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In: CONFERENCE ON NATURAL LANGUAGE LEARNING AT HLT-NAACL, 7., 2003, Edmonton. **Proceedings of the seventh conference on natural language learning**. 2003. v. 4, p. 142 - 147. Disponível em: <http://delivery.acm.org/10.1145/1120000/1119195/p142-tjong_kim_sang.pdf>. Acesso em: 17 fev. 2017.

SANTOS, M. T. ; CORRÊA, R. F. . Estudo da Construção da Taxonomia do Programa de Pós-Graduação em Letras. In: XIV ENCONTRO REGIONAL DE ESTUDANTES DE BIBLIOTECONOMIA, DOCUMENTAÇÃO, CIÊNCIA DA

INFORMAÇÃO E GESTÃO DA INFORMAÇÃO - EREBD, 2011, São Luís. **Anais do XIV Encontro Regional de Estudantes de Biblioteconomia**, Documentação, Ciência da Informação e Gestão da Informação, 2011.

SANTOS, P. M.; ZANCANARO, A.; NAKAYAMA, M. Pesquisas qualitativas em engenharia e gestão do conhecimento: uma revisão sistemática. **Informação & Informação**, v. 20, n. 1, p. 209, 2015. Disponível em: <http://www.uel.br/revistas/uel/index.php/informacao/article/view/15118>. Acesso em: 11 jan. 2016.

SCHREIBER, A. et al. **Knowledge Engineering and Management**: The CommonKADS Methodology. 1ª ed. Cambridge: MIT, 1999.

SELLAMI, Z.; CAMPS, V.; AUSSENAC-GILLES, N. DYNAMO-MAS: a Multi-Agent System for Ontology Evolution from Text. **Journal on Data Semantics**, v. 2, n. 2-3, p. 145–161, 28 maio 2013. Disponível em: <http://link.springer.com/10.1007/s13740-013-0025-1>. Acesso em: 22 set. 2014.

SÉRGIO, Marina Carradore; SILVA, Thales do Nascimento da; GONÇALVES, Alexandre Leopoldo. Descoberta de Conhecimento a partir de informações não estruturadas por meio de técnicas de correlação e associação. **Em Questão**, v. 22, n. 2, p.87-113, 11 ago. 2016. Programa de Pós-graduação em Comunicação e Informação da Universidade Federal do Rio Grande do Sul. http://dx.doi.org/10.19132/1808-5245222.87-113.

ŜEVA, Jurica; SCHATTEN, Markus; GRD, Petra. Open Directory Project based universal taxonomy for Personalization of Online (Re)sources. **Expert Systems With Applications**, v. 42, n. 17-18, p.6306-6314, out. 2015. Elsevier BV. http://dx.doi.org/10.1016/j.eswa.2015.04.033.

SILVA, Edna Lúcia da; MENEZES, Estera Muszkat. **Metodologia da pesquisa e elaboração de dissertação**. 4. ed. rev. atual. Florianópolis: UFSC, 2005.

SIQUEIRA, M. C. **Gestão Estratégica da Informação.** 1. ed. Rio de Janeiro: Brasport, 2005.

STUDER, R.; BENJAMINS, V.; FENSEL, D. Knowledge engineering: principles and methods. **Data knowledge Engineering**, Amsterdam, v. 25, n. 1-2, p. 161–197, 1998.

SUÁREZ-FIGUEROA, Mari Carmen; GÓMEZ-PÉREZ, Asunción; VILLAZÓN-TERRAZAS, Boris. How to Write and Use the Ontology Requirements Specification Document. In: ON THE MOVE TO MEANINGFUL INTERNET SYSTEMS: OTM, 2009, 00., 2009, Vilamoura. **Proceedings....** Berlin: Springer-verlag, 2009. p. 966 - 982. Disponível em: <http://delicias.dia.fi.upm.es/wiki/images/b/b0/ORequirementsSpecification.pdf>. Acesso em: 25 mar. 2017.

SUCHANEK, Fabian M.; KASNECI, Gjergji; WEIKUM, Gerhard. Yago: a core of semantic knowledge. In: INTERNATIONAL CONFERENCE ON WORLD WIDE WEB, 16., 2007, Banff. **Proceedings of the 16th international conference on World Wide Web**. New York: Acm, 2007. p. 697 - 706.

SUDA, Martin e WEIDENBACH, Christoph e WISCHNEWSKI, Patrick. On the saturation of YAGO. GIESL, J.; HÄHNLE, R. (Org.). . **Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010. v. 6173 LNAI. p. 441–456. Disponível em: <http://dx.doi.org/10.1007/978-3-642-14203-1_38>.

TERRA, J. C.; BAX, M. P. Portais corporativos: instrumento de gestão de informação e de conhecimento. In: Isis Paim. (Org.). **A Gestão da Informação e do Conhecimento**. 1 ed. Belo Horizonte, 2003, p. 33-53.

TRISTÃO, Ana Maria Delazari; FACHIN, Gleisy Regina Bóries; ALARCON, Orestes Estevam. **Sistema de classificação facetada e tesauros: instrumentos para organização do conhecimento**. Ciência da Informação, 2004, v. 33, n. 2, p. 161-171. http://dx.doi.org/10.1590/S0100-19652004000200017

TU, D.; CHEN, L.; CHEN, G. Automatic multi-way domain concept hierarchy construction from customer reviews. **Neurocomputing**, v. 147, p. 472–484, jan. 2015. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0925231214008108>. Acesso em: 24 maio 2015.

UZUNOV, Anton V.; FERNANDEZ, Eduardo B.. An extensible pattern-based library and taxonomy of security threats for distributed systems. **Computer Standards & Interfaces**, v. 36, n. 4, p.734-747, jun. 2014. Elsevier BV. http://dx.doi.org/10.1016/j.csi.2013.12.008.

VATSALAN, Dinusha; CHRISTEN, Peter; VERYKIOS, Vassilios S.. A taxonomy of privacy-preserving record linkage techniques. **Information Systems**, v. 38, n. 6, p.946-969, set. 2013. Elsevier BV. http://dx.doi.org/10.1016/j.is.2012.11.005.

VITAL, Luciane Paula. **Recomendações para construção de taxonomia em portais corporativos**. 2007. 113 f. Dissertação (Mestrado) – Programa de Pós-Graduação em Ciência da Informação, Universidade Federal de Santa Catarina, Florianópolis, 2007.

XAVIER, Clarissa Castellã; LIMA, Vera Lúcia Strube de. Construção de uma Estrutura Ontológica de Domínio a partir da Wikipédia. In: SIMPÓSIO BRASILEIRO DE TECNOLOGIAS DA INFORMAÇÃO E DA LÍNGUA, 7., 2009, São Carlos. **Anais...** São Carlos: Universidade de São Paulo, 2009. p. 1 - 10. Disponível em: <http://www.inf.pucrs.br/linatural/Docs/publicacoes/STIL_Clarissa.pdf>. Acesso em: 30 mar. 2017.

ZAMIL, Mohammed G.h. Al; AL-RADAIDEH, Qasem. Automatic extraction of ontological relations from Arabic text. **Journal Of King Saud University - Computer And Information Sciences**, v. 26, n. 4, p.462-472, dez. 2014. Elsevier BV. http://dx.doi.org/10.1016/j.jksuci.2014.06.007.

ZESCH, T.; MÜLLER, C.; GUREVYCH, I. Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), p. 1646–1652, 2008.

ZOUAQ, A.; NKAMBOU, R. Evaluating the Generation of Domain Ontologies in the Knowledge Puzzle Project. **IEEE Transactions on Knowledge and Data Engineering**, v. 21, n. 11, p. 1559–1572, nov. 2009. Disponível em: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4752828>. Acesso em: 16 nov. 2013.

GLOSSÁRIO

CORPUS: Conjunto de documentos sobre um determinado tema que servem como base de análise.

CORPORA: Conjunto de *corpus*.

FRAMEWORK: Biblioteca de *softwares* que fornecem diversas funcionalidades para a construção de novos *softwares*.

HASH: Identificador computacional de tamanho fixo resultante da aplicação de um algoritmo de dispersão sobre um conjunto de dados de tamanho fixo ou variável.

METADADOS: São dados sobre os dados principais utilizados em um contexto. São importantes no contexto computacional, pois definem e tipificam corretamente o dado para a compreensão por máquinas.

TREEBANK: *Corpus* de texto analisado e anotado com informações sintáticas e/ou semânticas.

TOKEN: Subdivisão dos elementos que compõem um texto. Em geral, refere-se a cada palavra existente em um trecho de análise, mas dependendo do processo de tokenização realizado, pode referenciar conjuntos de palavras ou até mesmo um texto por inteiro.

TOKENIZAÇÃO: Processo de análise e divisão do texto em *tokens*.

APÊNDICE A - RESULTADOS DAS BATERIAS DE TESTES DO RECONHECIMENTO DE ENTIDADES

Quadro 25 - Estatísticas sobre o reconhecimento de entidades obtidas para todas as variações

Número de currículos analisados	20.000
Número de *tokens*	9.325.838
Tokens reconhecidos	3.052.738
Fator de reconhecimento	0,3273420

APÊNDICE B - RESULTADOS DAS BATERIAS DE TESTES PARA A ABORDAGEM DE COOCORRÊNCIA DENTRO DO CORPUS

Quadro 26 - Estatísticas obtidas para o parâmetro mínimo de 100 ocorrências nos currículos para cada termo utilizado na taxonomia

Máximo de graus de distância da consulta na DBpedia	1	2	3
Tokens utilizados	93.013	129.571	162.911
Fator de utilização de tokens	0,010	0,014	0,017
Contagem de termos	127	438	882
Nível máximo	4	7	15
Número de expansões	44	189	402
Soma do número de termos expandidos	97	381	818
Média de termos por expansão	2,205	2,016	2,035
Nível médio do termo	1,961	2,852	4,193
Número de termos cíclicos	35	121	228
Fator de termos cíclicos	0,276	0,276	0,259
Horizontalidade	0,031	0,016	0,017
Verticalidade	31,750	62,571	58,800
Número de termos encontrados no AGROVOC	87	211	333
Porcentagem de termos encontrados no AGROVOC	68,504	48,174	37,755
Número de relacionamento hierárquicos encontrados no AGROVOC	11	23	61

Quadro 27 - Estatísticas obtidas para o parâmetro mínimo de 300 ocorrências nos currículos para cada termo utilizado na taxonomia

Máximo de graus de distância da consulta na DBpedia	1	2	3
Tokens utilizados	76.799	109.421	128.283
Fator de utilização de tokens	0,008	0,012	0,014
Contagem de termos	50	187	339
Nível máximo	5	7	13
Número de expansões	20	87	163
Soma do número de termos expandidos	36	163	313
Média de termos por expansão	1,800	1,874	1,920
Nível médio do termo	2,020	2,920	4,354
Número de termos cíclicos	9	52	92
Fator de termos cíclicos	0,180	0,278	0,271
Horizontalidade	0,100	0,037	0,038
Verticalidade	10,000	26,714	26,077

Número de termos encontrados no AGROVOC	37	96	139
Porcentagem de termos encontrados no AGROVOC	74,000	51,337	41,003
Número de relacionamento hierárquicos encontrados no AGROVOC	3	15	32

Quadro 28 - Estatísticas obtidas para o parâmetro mínimo de 500 ocorrências nos currículos para cada termo utilizado na taxonomia

Máximo de graus de distância da consulta na DBpedia	1	2	3
Tokens utilizados	70.742	95.519	106.812
Fator de utilização de tokens	0,008	0,010	0,011
Contagem de termos	33	103	168
Nível máximo	4	8	8
Número de expansões	15	51	86
Soma do número de termos expandidos	22	89	151
Média de termos por expansão	1,467	1,745	1,756
Nível médio do termo	1,848	3,146	3,631
Número de termos cíclicos	5	28	50
Fator de termos cíclicos	0,152	0,272	0,298
Horizontalidade	0,121	0,078	0,048
Verticalidade	8,250	12,875	21,000
Número de termos encontrados no AGROVOC	23	59	75
Porcentagem de termos encontrados no AGROVOC	69,697	57,282	44,643
Número de relacionamento hierárquicos encontrados no AGROVOC	2	25	22

Quadro 29 - Estatísticas obtidas para o parâmetro mínimo de 800 ocorrências nos currículos para cada termo utilizado na taxonomia

Máximo de graus de distância da consulta na DBpedia	1	2	3
Tokens utilizados	36.628	63.232	73.310
Fator de utilização de tokens	0,004	0,007	0,008
Contagem de termos	17	60	97
Nível máximo	2	5	7
Número de expansões	6	31	53
Soma do número de termos expandidos	11	50	86
Média de termos por expansão	1,833	1,613	1,623
Nível médio do termo	1,647	2,400	3,206
Número de termos cíclicos	2	16	25
Fator de termos cíclicos	0,118	0,267	0,258
Horizontalidade	0,118	0,083	0,072

Verticalidade	8,500	12,000	13,857
Número de termos encontrados no AGROVOC	13	35	48
Porcentagem de termos encontrados no AGROVOC	76,471	58,333	49,485
Número de relacionamento hierárquicos encontrados no AGROVOC	0	4	13

Quadro 30 - Estatísticas obtidas para o parâmetro mínimo de 1000 ocorrências nos currículos para cada termo utilizado na taxonomia

Máximo de graus de distância da consulta na DBpedia	1	2	3
Tokens utilizados	33.081	59.685	61.734
Fator de utilização de tokens	0,004	0,006	0,007
Contagem de termos	14	54	78
Nível máximo	2	7	9
Número de expansões	5	29	42
Soma do número de termos expandidos	9	46	70
Média de termos por expansão	1,800	1,586	1,667
Nível médio do termo	1,643	2,685	3,564
Número de termos cíclicos	2	13	21
Fator de termos cíclicos	0,143	0,241	0,269
Horizontalidade	0,143	0,130	0,115
Verticalidade	7,000	7,714	8,667
Número de termos encontrados no AGROVOC	10	30	40
Porcentagem de termos encontrados no AGROVOC	71,429	55,556	51,282
Número de relacionamento hierárquicos encontrados no AGROVOC	0	4	15

APÊNDICE C - RESULTADOS DAS BATERIAS DE TESTES PARA A ABORDAGEM DE COOCORRÊNCIA DENTRO DO CURRÍCULO (COM PELO MENOS UMA COOCORRÊNCIA NO *CORPUS* PARA CADA RELAÇÃO HIERÁRQUICA REPRESENTADA)

Quadro 31 - Estatísticas obtidas para o parâmetro mínimo de 100 ocorrências nos currículos para cada termo utilizado na taxonomia

Máximo de graus de distância da consulta na DBpedia	1	2	3
Tokens utilizados	85.053	127.350	160.118
Fator de utilização de tokens	0,009	0,014	0,017
Contagem de termos	126	429	963
Nível máximo	4	8	16
Número de expansões	47	188	442
Soma do número de termos expandidos	95	390	917
Média de termos por expansão	2,021	2,074	2,075
Nível médio do termo	1,921	3,371	5,769
Número de termos cíclicos	35	114	219
Fator de termos cíclicos	0,278	0,266	0,227
Horizontalidade	0,032	0,019	0,017
Verticalidade	31,500	53,625	60,188
Número de termos encontrados no AGROVOC	87	219	403
Porcentagem de termos encontrados no AGROVOC	69,048	51,049	41,848
Número de relacionamento hierárquicos encontrados no AGROVOC	10	37	218

Quadro 32 - Estatísticas obtidas para o parâmetro mínimo de 300 ocorrências nos currículos para cada termo utilizado na taxonomia

Máximo de graus de distância da consulta na DBpedia	1	2	3
Tokens utilizados	68.839	104.610	126.668
Fator de utilização de tokens	0,007	0,011	0,014
Contagem de termos	49	191	330
Nível máximo	3	9	11
Número de expansões	20	89	160
Soma do número de termos expandidos	35	171	307
Média de termos por expansão	1,750	1,921	1,919
Nível médio do termo	1,837	3,068	3,979
Número de termos cíclicos	8	53	96
Fator de termos cíclicos	0,163	0,277	0,291
Horizontalidade	0,061	0,047	0,033

Verticalidade	16,333	21,222	30,000
Número de termos encontrados no AGROVOC	36	103	136
Porcentagem de termos encontrados no AGROVOC	73,469	53,927	41,212
Número de relacionamento hierárquicos encontrados no AGROVOC	2	24	27

Quadro 33 - Estatísticas obtidas para o parâmetro mínimo de 500 ocorrências nos currículos para cada termo utilizado na taxonomia

Máximo de graus de distância da consulta na DBpedia	1	2	3
Tokens utilizados	62.782	90.708	105.771
Fator de utilização de tokens	0,007	0,010	0,011
Contagem de termos	32	91	154
Nível máximo	3	7	9
Número de expansões	13	43	79
Soma do número de termos expandidos	21	76	139
Média de termos por expansão	1,615	1,767	1,759
Nível médio do termo	1,719	2,945	3,649
Número de termos cíclicos	4	29	47
Fator de termos cíclicos	0,125	0,319	0,305
Horizontalidade	0,094	0,077	0,058
Verticalidade	10,667	13,000	17,111
Número de termos encontrados no AGROVOC	22	53	70
Porcentagem de termos encontrados no AGROVOC	68,750	58,242	45,455
Número de relacionamento hierárquicos encontrados no AGROVOC	3	17	19

Quadro 34 - Estatísticas obtidas para o parâmetro mínimo de 800 ocorrências nos currículos para cada termo utilizado na taxonomia

Máximo de graus de distância da consulta na DBpedia	1	2	3
Tokens utilizados	36.628	63.232	73.310
Fator de utilização de tokens	0,004	0,007	0,008
Contagem de termos	17	56	94
Nível máximo	2	6	9
Número de expansões	6	27	46
Soma do número de termos expandidos	11	46	86
Média de termos por expansão	1,833	1,704	1,870
Nível médio do termo	1,647	2,696	4,277
Número de termos cíclicos	2	15	26
Fator de termos cíclicos	0,118	0,268	0,277

Horizontalidade	0,118	0,107	0,096
Verticalidade	8,500	9,333	10,444
Número de termos encontrados no AGROVOC	13	33	47
Porcentagem de termos encontrados no AGROVOC	76,471	58,929	50,000
Número de relacionamento hierárquicos encontrados no AGROVOC	0	3	11

Quadro 35 - Estatísticas obtidas para o parâmetro mínimo de 1000 ocorrências nos currículos para cada termo utilizado na taxonomia

Máximo de graus de distância da consulta na DBpedia	1	2	3
Tokens utilizados	33.081	59.685	61.734
Fator de utilização de tokens	0,004	0,006	0,007
Contagem de termos	13	50	79
Nível máximo	3	6	8
Número de expansões	5	26	42
Soma do número de termos expandidos	9	42	73
Média de termos por expansão	1,800	1,615	1,738
Nível médio do termo	1,923	2,780	3,797
Número de termos cíclicos	1	12	23
Fator de termos cíclicos	0,077	0,240	0,291
Horizontalidade	0,231	0,120	0,101
Verticalidade	4,333	8,333	9,875
Número de termos encontrados no AGROVOC	9	29	42
Porcentagem de termos encontrados no AGROVOC	69,231	58,000	53,165
Número de relacionamento hierárquicos encontrados no AGROVOC	0	3	10

APÊNDICE D - RESULTADOS DAS BATERIAS DE TESTES PARA A ABORDAGEM DE COOCORRÊNCIA DENTRO DO CURRÍCULO (COM PELO MENOS 1000 COOCORRÊNCIAS NO *CORPUS* PARA CADA RELAÇÃO HIERÁRQUICA REPRESENTADA A)

Quadro 36 - Estatísticas obtidas para o parâmetro mínimo de 100 ocorrências nos currículos para cada termo utilizado na taxonomia

Máximo de graus de distância da consulta na DBpedia	1	2	3
Tokens utilizados	56.454	90.394	109.837
Fator de utilização de tokens	0,006	0,010	0,012
Contagem de termos	34	112	230
Nível máximo	3	6	9
Número de expansões	12	45	107
Soma do número de termos expandidos	24	99	216
Média de termos por expansão	2,000	2,200	2,019
Nível médio do termo	1,765	2,812	3,900
Número de termos cíclicos	5	28	62
Fator de termos cíclicos	0,147	0,250	0,270
Horizontalidade	0,088	0,054	0,039
Verticalidade	11,333	18,667	25,556
Número de termos encontrados no AGROVOC	23	61	107
Porcentagem de termos encontrados no AGROVOC	67,647	54,464	46,522
Número de relacionamento hierárquicos encontrados no AGROVOC	2	11	36

Quadro 37 - Estatísticas obtidas para o parâmetro mínimo de 300 ocorrências nos currículos para cada termo utilizado na taxonomia

Máximo de graus de distância da consulta na DBpedia	1	2	3
Tokens utilizados	53.920	84.747	101.931
Fator de utilização de tokens	0,006	0,009	0,011
Contagem de termos	25	72	139
Nível máximo	2	5	9
Número de expansões	9	31	71
Soma do número de termos expandidos	16	59	126
Média de termos por expansão	1,778	1,903	1,775
Nível médio do termo	1,640	2,333	3,374
Número de termos cíclicos	3	18	32
Fator de termos cíclicos	0,120	0,250	0,230
Horizontalidade	0,080	0,069	0,065

Verticalidade	12,500	14,400	15,444
Número de termos encontrados no AGROVOC	18	49	73
Porcentagem de termos encontrados no AGROVOC	72,000	68,056	52,518
Número de relacionamento hierárquicos encontrados no AGROVOC	1	8	20

Quadro 38 - Estatísticas obtidas para o parâmetro mínimo de 500 ocorrências nos currículos para cada termo utilizado na taxonomia

Máximo de graus de distância da consulta na DBpedia	1	2	3
Tokens utilizados	47.940	75.433	90.789
Fator de utilização de tokens	0,005	0,008	0,010
Contagem de termos	19	57	100
Nível máximo	3	5	7
Número de expansões	7	26	53
Soma do número de termos expandidos	13	47	88
Média de termos por expansão	1,857	1,808	1,660
Nível médio do termo	1,737	2,596	3,060
Número de termos cíclicos	2	13	26
Fator de termos cíclicos	0,105	0,228	0,260
Horizontalidade	0,158	0,088	0,070
Verticalidade	6,333	11,400	14,286
Número de termos encontrados no AGROVOC	12	36	48
Porcentagem de termos encontrados no AGROVOC	63,158	63,158	48,000
Número de relacionamento hierárquicos encontrados no AGROVOC	1	8	17

Quadro 39 - Estatísticas obtidas para o parâmetro mínimo de 800 ocorrências nos currículos para cada termo utilizado na taxonomia

Máximo de graus de distância da consulta na DBpedia	1	2	3
Tokens utilizados	31.710	54.312	68.202
Fator de utilização de tokens	0,003	0,006	0,007
Contagem de termos	14	38	73
Nível máximo	2	6	8
Número de expansões	5	19	35
Soma do número de termos expandidos	9	32	66
Média de termos por expansão	1,800	1,684	1,886
Nível médio do termo	1,643	2,789	3,973
Número de termos cíclicos	2	8	16
Fator de termos cíclicos	0,143	0,211	0,219

Horizontalidade	0,143	0,158	0,110
Verticalidade	7,000	6,333	9,125
Número de termos encontrados no AGROVOC	11	25	38
Porcentagem de termos encontrados no AGROVOC	78,571	65,789	52,055
Número de relacionamento hierárquicos encontrados no AGROVOC	0	3	11

Quadro 40 - Estatísticas obtidas para o parâmetro mínimo de 1000 ocorrências nos currículos para cada termo utilizado na taxonomia

Máximo de graus de distância da consulta na DBpedia	1	2	3
Tokens utilizados	28.163	50.765	57.495
Fator de utilização de tokens	0,003	0,005	0,006
Contagem de termos	10	35	64
Nível máximo	3	6	8
Número de expansões	4	18	32
Soma do número de termos expandidos	7	31	60
Média de termos por expansão	1,750	1,722	1,875
Nível médio do termo	2,000	2,886	3,859
Número de termos cíclicos	1	8	17
Fator de termos cíclicos	0,100	0,229	0,266
Horizontalidade	0,300	0,171	0,125
Verticalidade	3,333	5,833	8,000
Número de termos encontrados no AGROVOC	7	22	35
Porcentagem de termos encontrados no AGROVOC	70,000	62,857	54,688
Número de relacionamento hierárquicos encontrados no AGROVOC	0	3	10