ANDRÉ IPPOLITO

**Uma Abordagem de Múltiplos Aspectos para Alinhamento de Ontologias
Baseado em Cluster Ensembles Bayesianos**

São Paulo

2017

ANDRÉ IPPOLITO

**Uma Abordagem de Múltiplos Aspectos para Alinhamento de Ontologias
Baseado em Cluster Ensembles Bayesianos**

Tese apresentada à Escola Politécnica da
Universidade de São Paulo para obtenção
do Título de Doutor em Ciências

São Paulo
2017

ANDRÉ IPPOLITO

**Uma Abordagem de Múltiplos Aspectos para Alinhamento de Ontologias
Baseado em Cluster Ensembles Bayesianos**

Tese apresentada à Escola Politécnica da
Universidade de São Paulo para obtenção
do Título de Doutor em Ciências

Área de Concentração: Engenharia de
Computação

Orientador: Prof. Dr. Jorge Rady de
Almeida Jr.

São Paulo
2017

Este exemplar foi revisado e corrigido em relação à versão original, sob responsabilidade única do autor e com a anuência de seu orientador.

São Paulo, _____ de _____ de _____

Assinatura do autor: _____

Assinatura do orientador: _____

Catálogo-na-publicação

Ippolito, André

Uma Abordagem de Múltiplos Aspectos para Alinhamento de Ontologias Baseado em Cluster Ensembles Bayesianos / A. Ippolito – versão corr. -- São Paulo, 2017. 232 p.

Tese (Doutorado) - Escola Politécnica da Universidade de São Paulo. Departamento de Engenharia de Computação e Sistemas Digitais.

1.Ontologias 2.Inferência bayesiana e redes de crença 3.Teoria dos Grafos I.Universidade de São Paulo. Escola Politécnica. Departamento de Engenharia de Computação e Sistemas Digitais II.t.

Dedico esta Tese à minha família, em especial aos meus pais, pela paciência e pelo incentivo em alcançar mais este relevante objetivo e cujo apoio e compreensão foram fundamentais para vencer este árduo desafio da minha carreira.

AGRADECIMENTOS

Ao meu orientador, Prof. Dr. Jorge Rady de Almeida Júnior, pela orientação, bom senso e experiência, que ajudaram na condução e conclusão deste trabalho com sucesso.

Aos demais professores e aos funcionários da Escola Politécnica da Universidade de São Paulo com quem tive oportunidade de conviver e aprender, seja na graduação ou na pós-graduação, e que contribuíram na minha formação acadêmica, profissional e pessoal.

RESUMO

Ontologias são especificações formais e explícitas usadas para descrever entidades de um domínio e seus relacionamentos. Estatísticas recentes do projeto *Linked Open Data* (LOD) indicam a existência de milhares de ontologias heterogêneas publicadas na nuvem do LOD, impondo um desafio para a integração de ontologias. Um passo fundamental na integração é o emparelhamento, processo que obtém elementos correspondentes entre ontologias heterogêneas. Visando superar o desafio de efetuar o emparelhamento em larga escala, desenvolveu-se uma estratégia baseada em clusterização das ontologias, a qual particiona as ontologias em subontologias, clusteriza as subontologias e restringe o processo de emparelhamento aos elementos de um mesmo *cluster*. Porém, observa-se que as soluções do estado da arte necessitam explorar mais os múltiplos aspectos que as subontologias possuem. As clusterizações de cada aspecto podem ser combinadas, por meio de um consenso. *Cluster Ensembles* é uma técnica que permite obter esse consenso. Além disso, estudos comparativos indicaram que o uso de *Cluster Ensembles* Bayesianos (CEB) resulta em uma clusterização de maior acurácia do que a obtida por outras técnicas de *Cluster Ensembles*. Um dos principais objetivos deste trabalho foi desenvolver uma nova metodologia de emparelhamento de ontologias baseada em clusterização consensual de múltiplos aspectos de comunidades, de forma a estruturar um arcabouço metodológico, por meio do qual diferentes técnicas e aspectos podem ser incorporados e testados. De acordo com a metodologia desenvolvida neste trabalho, inicialmente aplicaram-se técnicas de Detecção de Comunidades para particionar as ontologias. Em seguida, consideraram-se os seguintes aspectos das comunidades obtidas: terminológico, estrutural e extensional. Fez-se, separadamente, a clusterização das comunidades segundo cada aspecto e aplicaram-se diferentes técnicas de clusterização consensual para obter um consenso entre as clusterizações de cada aspecto: CEB, técnicas baseadas em similaridades e técnicas baseadas em métodos diretos. Para os diferentes consensos, o processo de emparelhamento foi feito apenas entre elementos das ontologias que pertencessem a um mesmo *cluster* consensual. As soluções consensuais destacaram-se nos estudos de caso efetuados quanto à precisão e cobertura dos alinhamentos, enquanto a solução baseada no aspecto terminológico destacou-se quanto ao valor de *F-measure*. A principal contribuição deste trabalho relaciona-se à metodologia desenvolvida, que constitui um arcabouço metodológico, por meio do qual diferentes aspectos e técnicas podem ser incorporados e testados quanto ao seu desempenho de clusterização e de alinhamento de ontologias.

Palavras-Chave: Ontologia. Emparelhamento. Clusterização. *Cluster Ensembles* Bayesianos. Detecção de Comunidades. Arcabouço.

ABSTRACT

Ontologies are formal and explicit specifications used to describe entities of a domain and its relationships. Recent statistics of the Linked Open Data (LOD) project indicate the existence of thousands of heterogeneous ontologies in the LOD cloud, posing a challenge to ontology integration. A fundamental step in integration is matching, a process that finds correspondent elements between heterogeneous ontologies. Aiming to overcome the challenge of large-scale ontology matching, researchers developed a strategy based on clustering, which divides ontologies into subontologies, clusters subontologies and restricts the matching process to elements of the same cluster. However, state-of-the-art solutions need to explore more the multiple aspects that subontologies have. Clustering solutions of each aspect can be combined, by means of a consensus. Cluster Ensembles is a technique that allows obtaining this consensus. Besides, comparative studies indicated that Bayesian Cluster Ensembles has higher clustering accuracy than other Cluster Ensembles techniques. One of the main goals of this work was to develop a new methodology for ontology matching based on consensus clustering of multiple aspects of communities, structuring a methodological framework that enables the use and tests of different techniques and aspects. According to the methodology adopted in this work, initially, Community Detection techniques were applied to partition the ontologies. In the sequence, the following aspects of the communities were considered: terminological, structural and extensional. Clustering according to each aspect was performed separately and different consensus clustering techniques were applied to obtain a consensus among clustering solutions of each aspect: Bayesian Cluster Ensembles, techniques based on similarities and techniques based on direct methods. For the different consensuses, matching was done only between elements of the two ontologies that belonged to the same consensual cluster. For the case studies applied in this work, the consensual solutions were a standout in precision and recall, while the terminological-based solution was a standout in F-measure. The main contribution of this work is related to the developed methodology, which constitutes a methodological framework, through which different aspects and techniques can be incorporated and tested concerning their ontology clustering and alignment performance.

Keywords: Ontology. Matching. Clustering. Bayesian Cluster Ensembles. Community Detection. Framework.

LISTA DE FIGURAS

Figura 1.1 – Emparelhamento de ontologias baseado em clusterização.....	17
Figura 2.1 - Exemplo de triplas representadas em um grafo.....	21
Figura 2.2 – Alinhamento entre duas ontologias.....	25
Figura 2.3 – Classificação de técnicas de emparelhamento.	13
Figura 3.1 – Exemplo de um dendograma obtido com clusterização hierárquica.	34
Figura 4.1 – Modelo do Espaço Vetorial representando os documentos d_1 , d_2 , d_3 e a consulta q , considerando apenas dois termos distintos (“gossip” e “jealous”) nesses documentos e na consulta.	44
Figura 4.2 – Diagrama da Decomposição de Valor Singular Truncada.	46
Figura 4.3 – Gráfico do custo computacional do ICA, com o eixo vertical representando o tempo de CPU em segundos e o eixo horizontal representando a quantidade de palavras.....	50
Figura 5.1 – Grafo representando comunidades em uma empresa: círculos tracejados delimitam os vértices e arcos do grafo que compõem cada comunidade.	52
Figura 5.2 – Exemplo de grafo representando a relação entre as classes “Part” e “PageRange”.	57
Figura 5.3 – Exemplo de grafo representando a relação entre as classes “Book” , “Collection”, “Proceedings” e “Monograph”.	57
Figura 6.1 – Processo generativo para Cluster Ensembles Bayesianos.....	68
Figura 6.2 – Modelo gráfico probabilístico para Cluster Ensembles Bayesianos.	69
Figura 6.3 – Gráfico do tempo de execução dos CEB em função da quantidade de clusterizações de base.	70
Figura 7.1 – Processo para o emparelhamento de ontologias baseado em clusterização usando MapReduce.....	73
Figura 8.1 – Fluxograma da metodologia proposta.	80
Figura 8.3 – Modelagem da ontologia 1 por meio de grafo.	83
Figura 8.4 – Modelagem da ontologia 2 por meio de grafo.	84
Figura 8.5 – Partição do grafo da ontologia 1.....	85
Figura 8.6 – Partição do grafo da ontologia 2.....	85

Figura 8.7 – Fluxograma da clusterização segundo o aspecto terminológico. .87	87
Figura 8.8 – Fluxograma da clusterização segundo o aspecto extensional.91	91
Figura 8.9 – Cluster Ensembles dos aspectos terminológico, extensional e topológico.....93	93
Figura 8.10 – Sequência de comandos para instanciar o servidor Rserve com o console do R Studio.....98	98
Figura 8.11 – Execução da classe “Main”, ao abrir o projeto do sistema com a ferramenta Eclipse.99	99
Figura 8.12 – Entrada de dados no console do Eclipse, com a especificação pelo usuário dos caminhos para os arquivos e nomes dos arquivos que contêm as ontologias a serem emparelhadas.....100	100
Figura 8.13 – Exemplo de estrutura de armazenamento em diretórios para armazenar as ontologias a serem emparelhadas.....100	100
Figura 8.14 – Script do R com seta indicando a parametrização que precisa ser feita quanto à quantidade de clusters kc a ser gerada pelas técnicas de clusterização.....101	101
Figura 8.17 – Exemplo de gráfico de largura de silhueta para um resultado de clusterização das comunidades das ontologias conforme o aspecto terminológico.....103	103
Figura 8.18 – Script do R para a clusterização do aspecto extensional baseado em ICA.....105	105
Figura 8.19 – Exemplo de coleção de arquivos texto com os conjuntos de termos de cada comunidade das ontologias.....106	106
Figura 8.20 – Parametrização da quantidade de clusters kc para o script de clusterização do aspecto topológico. No exemplo da figura, foi utilizado o valor de kc igual a 2.....107	107
Figura 8.21 – Script do Matlab para efetuar a clusterização consensual dos três aspectos.....108	108
Figura 8.22 – Saída do Matlab, com a seta na figura indicando os rótulos da clusterização consensual de maior informação mútua normalizada média, dentre as obtidas com CSPA, HGPA e MCLA.....109	109
Figura 8.23 – Script do CEB no Matlab, indicando os quatro primeiros parâmetros a serem especificados pelo usuário: “base_labels”, “number_baseclusterers”, “true_labels” e “Palpha”.....110	110

Figura 8.24 – Script do CEB no Matlab, indicando a parametrização do vetor de matrizes Pbeta	112
Figura 8.25 – Saída do Matlab após a execução do script de CEB, indicando-se por uma seta o vetor de rótulos da clusterização consensual obtida.....	112
Figura 8.26 – Console do Eclipse com destaque para os dados de entrada que devem ser especificados pelo usuário, relativos ao caminho e nome do arquivo com a clusterização a ser usada como base para o emparelhamento.....	113
Figura 8.27 – Resultados de precisão cobertura e F-measure para o alinhamento, após o usuário fornecer o caminho e o nome do arquivo com o alinhamento de referência.	114
Figura 9.1 – Grafo de parte da ontologia 101 da OAEI.	118
Figura 9.2 – Grafo de parte da ontologia 304 da OAEI.	118
Figura 9.3 – Grafo da ontologia 101.....	119
Figura 9.4 – Grafo da ontologia 304.....	119
Figura 9.5 – Partição da ontologia 101 em comunidades obtida com a técnica de Brandes.....	123
Figura 9.6 – Partição da ontologia 304 em comunidades obtida com a técnica de Brandes.....	125
Figura 9.7 – Gráfico das larguras de silhueta de cada cluster, para a clusterização de maior largura de silhueta média, ao usar ICA.....	127
Figura 9.8 – Gráfico das larguras de silhueta de cada cluster para a clusterização de maior largura de silhueta média, ao usar LSA.....	128
Figura 9.9 – Gráfico das larguras de silhueta de cada cluster para a clusterização de maior largura de silhueta média, ao usar a técnica ICA.....	130
Figura 9.10 – Gráfico das larguras de silhueta de cada cluster para a clusterização de maior largura de silhueta média, ao usar a técnica LSA.....	132
Figura 9.11 – Gráfico das larguras de silhueta de cada cluster para a clusterização de maior largura de silhueta média, ao usar o K-means.....	133
Figura 9.12 – Gráfico indicando a variação das métricas de alinhamento conforme a quantidade de clusters.	141
Figura 9.13 – Gráfico indicando a variação da largura de silhueta conforme a quantidade de clusters.	142
Figura A1 – Grafo da ontologia 101.	163
Figura A2 – Grafo da ontologia 304.	164

LISTA DE TABELAS

Tabela 1.1 – Ontologias da Linked Open Data Cloud.....	15
Tabela 2.1 – Exemplos de modelagens em OWL 2.....	23
Tabela 7.1 – Visão geral das técnicas utilizadas nos trabalhos relacionados...	79
Tabela 8.1 – Mapeamento entre identificadores dos nós e os nomes das classes das ontologias do exemplo da metodologia.....	86
Tabela 8.2 – Textos contidos em cada comunidade para o exemplo da metodologia.....	88
Tabela 8.3 – Resultado da clusterização do aspecto terminológico com a maior largura de silhueta.	88
Tabela 8.4 – Restrições e probabilidades dos conceitos para o exemplo da metodologia.....	90
Tabela 8.5 - Resultado da clusterização do aspecto extensional com a maior largura de silhueta.	91
Tabela 8.6 – Valores dos atributos topológicos para o exemplo da metodologia.	92
Tabela 8.7 – Resultado da clusterização do aspecto topológico com a maior largura de silhueta.	92
Tabela 8.8 – Resultado da clusterização consensual com CSPA.	93
Tabela 8.9 – Resultado da clusterização consensual com CEB.	93
Tabela 8.10 – Alinhamento entre classes para o exemplo da metodologia.	94
Tabela 8.11 – Ferramentas utilizadas na metodologia.....	95
Tabela 9.1 – Tabela com as quantidades de classes, instâncias e propriedades das ontologias do teste benchmarks usadas nos estudos de caso.....	117
Tabela 9.3 – Valores de modularidade resultantes da aplicação das técnicas de Detecção de Comunidades às ontologias 101 e 304.....	121
Tabela 9.4 – Identificadores de nós, nomes das classes correspondentes e rótulos das comunidades as quais as classes pertencem, para representar os agrupamentos resultantes da aplicação da técnica de Brandes à ontologia 101.	122
Tabela 9.5 – Identificadores de nós, respectivos nomes das classes e comunidades obtidos com a técnica de Brandes para a ontologia 304.....	124

Tabela 9.6 – Resultados de largura de silhueta média das clusterizações para ICA com variação de k.	126
Tabela 9.7 – Resultado de clusterização das comunidades para a máxima largura de silhueta média com ICA.....	127
Tabela 9.8 – Resultados de largura de silhueta média das clusterizações para LSA com variação de k.	128
Tabela 9.9 – Resultado de clusterização das comunidades para a máxima largura de silhueta média com LSA.....	128
Tabela 9.10 – Resultados de largura de silhueta média das clusterizações para ICA com variação de k.	130
Tabela 9.11 – Resultado de clusterização das comunidades para a máxima largura de silhueta média.	131
Tabela 9.12 – Resultados de largura de silhueta média das clusterizações para LSA com variação de k.	131
Tabela 9.13 – Resultado de clusterização das comunidades para a máxima largura de silhueta média.	132
Tabela 9.14 – Resultados de largura de silhueta média das clusterizações com variação de k.....	133
Tabela 9.15 – Resultado de clusterização das comunidades para a máxima largura de silhueta média.	134
Tabela 9.16 – Resultados de precisão, cobertura e F-measure com base nas técnicas de clusterização.	135
Tabela 9.17 – Resultados de modularidade para as técnicas de Detecção de Comunidades aplicadas os estudos de caso.	136
Tabela 9.18 – Resultados de precisão, cobertura e F-measure para cada estudo de caso, conforme a técnica consensual ou o aspecto usados na solução.....	79
Tabela 9.19 – Quantidade de vezes em que a aplicação de uma dada técnica ou aspecto conseguiu atingir a máxima métrica de alinhamento em um estudo de caso.	140
Tabela 9.20 – Valores médios de precisão, cobertura e F-measure conforme a quantidade de clusters.	140
Tabela 9.21 – Valores médios de largura de silhueta dos estudos de caso conforme a quantidade de clusters.	141

Tabela 9.22 – Quantidade de vezes em que cada técnica de clusterização obteve a máxima largura de silhueta de um estudo de caso.....	142
Tabela 9.23 – Quantidade de vezes em que cada técnica de clusterização obteve a máxima largura de silhueta de um estudo de caso.....	143
Tabela D1 - Alinhamento e similaridades com base nas técnicas CSPA, HGPA e MCLA.	213
Tabela D2 - Alinhamento e similaridades com base em CEB com três clusters.	214
Tabela D3 - Alinhamento e similaridades com base no aspecto terminológico.	216
Tabela D4 - Alinhamento e similaridades com base no aspecto topológico. ..	218
Tabela D5 - Alinhamento e similaridades com base no aspecto extensional..	218
Tabela D6 - Alinhamento de referência da OAEI.	220
Tabela E1 – Resultados de clusterização e alinhamento para as demais ontologias do teste de benchmarks.	222

LISTA DE ABREVIATURAS E SIGLAS

ACM	<i>Association for Computing Machinery</i>
CEB	<i>Cluster Ensembles Bayesianos</i>
CSPA	<i>Cluster-based Similarity Partitioning Algorithm</i>
EON	<i>Evaluation of Ontologies and Ontology-based Tools</i>
HGPA	<i>HyperGraph Partitioning Algorithm</i>
HTML	<i>HyperText Markup Language</i>
ICA	<i>Independent Component Analysis</i>
IEEE	<i>Institute of Electrical and Electronics Engineers</i>
INRIA	<i>Institut National de Recherche en Informatique et en Automatique</i>
LOD	<i>Linked Open Data</i>
LSA	<i>Latent Semantic Analysis</i>
MCLA	<i>Meta-CLustering Algorithm</i>
MIT	<i>Massachusetts Institute of Technology</i>
OAEI	<i>Ontology Alignment Evaluation Initiative</i>
OWL	<i>Web Ontology Language</i>
PAM	<i>Partitioning Around Medoids</i>
RDF	<i>Resource Description Framework</i>
TF-IDF	<i>Term Frequency-Inverse Document Frequency</i>
UPGMA	<i>Unweighted Pair Group Method with Arithmetic Mean</i>
UML	<i>Unified Modeling Language</i>

URI	<i>Uniform Resource Identifier</i>
WOMO	<i>Workshop on Modular Ontologies</i>
W3C	<i>World Wide Web Consortium</i>

SUMÁRIO

1 INTRODUÇÃO	14
1.1 Escopo do Trabalho.....	14
1.2 Motivação e Justificativas	14
1.3 Objetivos.....	18
1.4 Organização do Texto.....	19
2 ONTOLOGIAS	20
2.1 RDF	20
2.2 OWL.....	21
2.3 Lógica de Descrição	22
2.4 Emparelhamento de Ontologias	24
2.4.1 Técnicas de Emparelhamento	25
2.4.2 Avaliação do Emparelhamento	29
2.5 Distâncias entre Ontologias	29
3 CLUSTERIZAÇÃO	32
3.1 Clusterização Hierárquica	33
3.2 Clusterização Particional	35
3.3 Avaliação de Clusterização.....	37
4 CLUSTERIZAÇÃO DE DOCUMENTOS.....	41
4.1 Modelo do Espaço Vetorial	42
4.2 Técnicas de Redução de Dimensionalidade	44
4.2.1 Análise Semântica Latente	45
4.2.2 Análise de Componentes Independentes	46
5 CLUSTERIZAÇÃO DE GRAFOS	51
5.1 Comunidades.....	51
5.2 Modularidade de Comunidades	52
5.3 Topologia de Comunidades	53
5.3.1 Excentricidade de vértices	53
5.3.2 Densidade de arcos	54
5.3.3 Diâmetro do grafo	54
5.3.4 Comprimento do caminho médio	55
5.4 Detecção de Comunidades.....	55
5.5 Detecção de Comunidades em Ontologias.....	56
5.5.1 Passeios Aleatórios	58

5.5.2 Otimização Gulosa de Modularidade	60
5.5.3 Método de Potts.....	61
5.5.4 Técnica de Brandes	62
6 CLUSTER ENSEMBLES	64
6.1 Cluster Ensembles Baseados em Similaridades	65
6.2 Cluster Ensembles Baseados em Métodos Diretos	66
6.3 Cluster Ensembles Bayesianos	67
7 REVISÃO DA LITERATURA	71
7.1 A Clustering-Based Approach for Large-scale Ontology Matching (Algergawy; Massmann; Rahm, 2011).....	71
7.2 VDoc+: a Virtual Document Based Approach for Matching Large Ontologies using MapReduce (Zhang; Hu; Qu, 2012)	72
7.3 Improving Clustering-Based Schema Matching Using Latent Semantic Indexing (ALGERGAWY et al., 2015).....	73
7.4 Efficient Multidimensional Blocking for Link Discovery without Losing Recall (Isele; Jentzsch; Bizer, 2011).....	74
7.5 Alignment-Based Partitioning of Large-scale Ontologies (Hamdi et al., 2010).75	
7.6 Dimensional Clustering of Linked Data: Techniques and Applications (Ferrara et al., 2015).....	76
7.7 Análise	77
8 METODOLOGIA PROPOSTA	80
8.1 Modelagem das Ontologias como Grafos.....	81
8.2 Obtenção das Partições dos Grafos das Ontologias	84
8.3 Clusterização das Comunidades Baseada no Aspecto Terminológico	86
8.4 Clusterização das Comunidades Baseada no Aspecto Extensional	89
8.5 Clusterização das Comunidades Baseada no Aspecto Topológico	91
8.6 Clusterização Consensual dos Diferentes Aspectos.....	92
8.7 Emparelhamento.....	93
8.8 Ferramentas	94
8.9 Diagramas de classes e de sequência	94
8.10 Manual de utilização do sistema.....	97
8.10.1 Instanciação de um servidor para integração entre o Eclipse e o R Studio 98	
8.10.2 Parametrização das ontologias a serem emparelhadas	98

8.10.3 Parametrização e execução das técnicas de clusterização dos aspectos terminológico, extensional e topológico	100
8.10.4 Parametrização e execução das técnicas de clusterização consensual... ..	106
8.10.5 Parametrização e execução do emparelhamento.....	113
9 ESTUDOS DE CASO.....	115
9.1 Resultados de Detecção de Comunidades nas Ontologias (ontologia 101 x ontologia 304)	117
9.2 Resultados de Clusterização das Comunidades Baseada no Aspecto Terminológico (ontologia 101 x ontologia 304)	125
9.2.1 Resultados ao usar ICA	126
9.2.2 Resultados ao usar LSA	127
9.3 Resultados de Clusterização das Comunidades Baseada no Aspecto Extensional (ontologia 101 x ontologia 304)	129
9.3.1 Resultados ao usar ICA	129
9.3.2 Resultados ao usar LSA	131
9.4 Resultados de Clusterização das Comunidades Baseada no Aspecto Topológico (ontologia 101 x ontologia 304)	132
9.5 Resultados de Clusterização Consensual dos Aspectos (ontologia 101 x ontologia 304)	134
9.6 Resultados de Alinhamento (ontologia 101 x ontologia 304)	134
9.7 Principais Resultados de Detecção de Comunidades, de Alinhamento e de Clusterização dos Estudos de Caso	135
9.7.1 Resultados de Detecção de Comunidades.....	136
9.7.2 Resultados de Alinhamento	137
9.7.3 Resultados de Clusterização	141
10 DISCUSSÃO	144
10.1 Detecção de Comunidades nas Ontologias 101 e 304	144
10.2 Clusterizações das Comunidades Baseadas nos Aspectos Terminológico, Extensional e Topológico (ontologias 101 e 304)	144
10.2.1 Clusterização das Comunidades baseada no Aspecto Terminológico (ontologias 101 e 304)	144
10.2.2 Clusterização das Comunidades baseada no Aspecto Extensional (ontologias 101 e 304)	145

10.2.3 Clusterização das Comunidades baseada no Aspecto Topológico (ontologias 101 e 304)	146
10.3 Alinhamento das ontologias 101 e 304	146
10.4 Discussão dos Principais Resultados de Detecção de Comunidades, de Alinhamento e de Clusterização dos Estudos de Caso	147
10.4.1 Detecção de Comunidades	147
10.4.2 Alinhamento	147
10.4.3 Clusterização	148
10.5 Trabalhos Futuros	149
11 CONCLUSÕES	151
12 REFERÊNCIAS	154
ANEXO A – Grafos das ontologias 101 e 304	163
ANEXO B – Ontologias 101 e 304 em OWL	165
ANEXO C – Diagramas UML do sistema	209
ANEXO D - Alinhamentos das ontologias 101 e 304	213
ANEXO E - Resultados de clusterização e de alinhamento das demais ontologias do teste benchmarks	222

1 INTRODUÇÃO

Neste capítulo, são definidos o escopo, a motivação, as justificativas e os objetivos deste trabalho. O escopo delimita quais elementos das ontologias foram emparelhados ao aplicar a metodologia de emparelhamento proposta. A motivação e as justificativas contextualizam o cenário e as dificuldades que impulsionam e fundamentam o desenvolvimento de uma nova metodologia de emparelhamento de ontologias. Nos objetivos, define-se o que o trabalho visou alcançar. Ao final, é apresentada a estrutura da organização do texto. Este trabalho representa um desenvolvimento e uma evolução em relação ao trabalho de Ippolito e Almeida Júnior (2016), estendendo-o com uma avaliação experimental composta de uma série de estudos de caso efetuados para validar a metodologia ora criada.

1.1 Escopo do Trabalho

Este trabalho tem como escopo efetuar o emparelhamento entre classes e propriedades de duas ontologias com base em técnicas de clusterização. O emparelhamento é o processo de descoberta de semelhanças entre elementos de ontologias heterogêneas, que define pares de elementos correspondentes entre si. O conjunto de pares correspondentes, ou conjunto de correspondências do emparelhamento, é denominado alinhamento e é importante na integração de ontologias heterogêneas.

1.2 Motivação e Justificativas

Conforme Cardoso (2010), a informação existente na *web*, veiculada por meio de páginas em *HyperText Markup Language* (HTML), permite a leitura e a compreensão de seu conteúdo por humanos. Porém, seu significado intrínseco não é suficientemente explicitado de forma a permitir sua interpretação por computadores. A comunidade do *World Wide Web Consortium* (W3C), voltada para a especificação de padrões para a *web*, tem empenhado esforços para desenvolver tecnologias que superem essa limitação. Nesse sentido, criaram-se grupos específicos no W3C, engajados na criação de tecnologias relacionadas ao conceito de *Web Semântica* (BERNERS-LEE; HENDLER; LASSILA, 2001).

Berners-Lee, Hendler e Lassila (2001) definem que a *Web Semântica* é uma extensão da *web* atual, na qual o significado (ou a semântica) da informação é bem definido, de forma a tornar seu conteúdo interpretável por computadores, viabilizando que o homem e a máquina interajam cooperativamente. Uma forma de modelar semanticamente o significado da informação de um determinado domínio é usar ontologias, que são modelos de dados que permitem especificar formalmente as relações entre termos, possibilitando revelar significados e desambiguar termos, além de permitir inferir regras implícitas de um domínio. Assim sendo, as ontologias exercem papel fundamental para viabilizar a implantação da *Web Semântica*.

Em função da importância das ontologias, Hendler (2001) previu a tendência de organizações desenvolverem suas próprias ontologias, para representar entidades e seus relacionamentos, permitindo inferir regras sobre seus domínios. Na prática, tem-se observado a publicação de ontologias na *web*, referentes a diversos domínios. A adoção dos princípios e recomendações do projeto *Linked Open Data* (LOD) (BIZER; HEATH; BERNERS-LEE, 2009), por exemplo, permite que organizações publiquem seus dados na *web* na forma de ontologias, possibilitando compartilhar dados relacionados a um amplo espectro de domínios, associados a diferentes tópicos, tais como dados governamentais, acadêmicos, científicos, geográficos, dentre outros. Um quadro geral, de acordo com estatísticas feitas em agosto de 2014 (SCHMACHTENBERG; BIZER; PAULHEIM, 2014), é mostrado na tabela 1.1, na qual são discriminados quantidades e percentuais de ontologias da *LOD Cloud*, por domínio, totalizando 1.014 ontologias.

Tabela 1.1 – Ontologias da *Linked Open Data Cloud*.

Fonte: Schmachtenberg, Bizer e Paulheim (2014)

Domínio	Número de ontologias	%
Governo	183	18,05
Publicações	96	9,47
Ciências	83	8,19
Conteúdo gerado por usuários	48	4,73
Vários domínios	41	4,04
Mídia	22	2,17
Geográfico	21	2,07
Redes sociais	520	51,28
Totais	1.014	100,00

Além disso, observa-se um notório crescimento da quantidade de ontologias na LOD *Cloud* ao longo do tempo. Em setembro de 2011, havia 295 ontologias (JENTZSCH; CYGANIAK; BIZER, 2011), progredindo para 870 ontologias em meados de 2013 (HU; RODRIGUES; VIEL, 2014) e aumentando para 9.960 ontologias em outubro de 2016 (AUER et al., 2016).

Com o aumento progressivo da quantidade de ontologias, desenvolvidas por diferentes equipes, que adotam diferentes vocabulários e padrões para as ontologias, surge o desafio de integrar ontologias cada vez mais numerosas e heterogêneas. A integração de ontologias adquire importância em ambientes corporativos, por exemplo, no caso de fusões entre duas ou mais empresas, pois, integrando as suas respectivas ontologias, podem-se inferir regras sobre o domínio de negócio resultante da fusão, tendo como base uma ontologia única, que é obtida da mesclagem das ontologias originais.

Para Euzenat e Shvaiko (2013), um pré-requisito essencial e necessário para o processo de integração é efetuar o emparelhamento dos elementos das ontologias. O emparelhamento é o processo de obtenção dos pares de elementos das ontologias que são semanticamente equivalentes entre si. Com base no resultado de emparelhamento, faz-se a mesclagem das ontologias, gerando uma ontologia única, a qual fornece uma visão integrada das ontologias iniciais.

Com o crescimento da quantidade e da heterogeneidade das ontologias, tornando o emparelhamento mais complexo, Euzenat e Shvaiko (2013) consideram a solução baseada em clusterização uma estratégia relevante para o emparelhamento. Esta estratégia pode ser visualizada esquematicamente na figura 1.1, na qual há quatro passos. Dadas duas ontologias para as quais se deseja obter o emparelhamento, no passo 1 fragmentam-se as ontologias em ontologias menores, denominadas subontologias ou blocos. Cada subontologia é um agrupamento de elementos da ontologia conectados entre si por relacionamentos e com características semelhantes entre si. No passo 2, agrupam-se subontologias das duas ontologias que sejam semelhantes entre si em *clusters* (fase de clusterização). Podem fazer parte do mesmo *cluster* duas ou mais subontologias. No passo 3, o emparelhamento é feito apenas entre elementos de subontologias semelhantes, ou seja, entre elementos de um mesmo *cluster*. Ao final, no passo 4, unem-se todos os resultados de emparelhamento, gerando o emparelhamento final. O processo para gerar subontologias do passo 1 é também conhecido como Modularização

Figura 1.1 – Emparelhamento de ontologias baseado em clusterização.

Fonte: adaptado de Algergawy, Massmann e Rahm (2011)

(STUCKENSCHMIDT; SPACCAPIETRA, 2009), havendo inclusive uma linha de pesquisas nessa área, com realizações de *workshops*, como é o caso do *Workshop on Modular Ontologies* (WOMO). Podem-se enumerar alguns trabalhos de destaque no WOMO, tais como os de Coskun et al. (2011), Vescovo et al. (2012) e Gatens, Konev e Wolter (2013).

Segundo Euzenat e Shvaiko (2013), essa estratégia objetiva evitar a comparação entre todos os elementos das ontologias (produto cartesiano de comparações), durante o processo de emparelhamento. Em alguns casos, ao adotar essa abordagem baseada em clusterização, podem-se ter subontologias de uma ontologia descartadas nas comparações de emparelhamento, caso não haja, como contrapartida, subontologias semelhantes na outra ontologia em comparação. Evitam-se, assim, comparações desnecessárias, o que melhora o desempenho do processo de emparelhamento. Para Rahm (2011), essa estratégia de clusterização é considerada promissora para o emparelhamento de ontologias em larga escala. Segundo Shvaiko e Euzenat (2013), soluções de emparelhamento baseadas em clusterização ainda carecem de pesquisas, necessitando-se explorar novas metodologias, visando facilitar o processo de emparelhamento.

As análises de alguns dos principais trabalhos de emparelhamento de ontologias baseados nessa estratégia apontam para algumas lacunas. Na fase de clusterização por exemplo, os trabalhos de Algergawy, Massmann e Rahm (2011), Algergawy et al. (2015) e Zhang, Hu e Qu (2012) consideram apenas o vocabulário das ontologias (termos contidos em rótulos de classes e em anotações). Outros aspectos das ontologias, tais como o estrutural ou o extensional (relacionado às instâncias das classes das ontologias), podem ser explorados, visando melhorar a clusterização e, por consequência, o resultado final de emparelhamento. Ferrara et al. (2015) analisam diferentes aspectos, associados a diferentes tópicos de uma mesma ontologia, aplicando *Cluster Ensembles* para combinar clusterizações relacionadas aos tópicos, encontrando um consenso entre as clusterizações. O consenso baseia-se na interseção de *clusters*, ou seja, no agrupamento de mesmos elementos em diferentes clusterizações.

Outras abordagens de *Cluster Ensembles* podem propiciar melhores clusterizações consensuais. Wang, Shan e Banerjee (2011), por exemplo, desenvolveram a técnica de *Cluster Ensembles* Bayesianos (CEB) e concluíram em seus experimentos que a técnica gerou uma clusterização mais precisa do que a obtida por outras técnicas de *Cluster Ensembles*. Pode-se explorar uma sinergia entre os diferentes aspectos das ontologias, usando CEB, verificando se há maior precisão ou cobertura no alinhamento resultante, comparativamente a soluções que se baseiam exclusivamente nos aspectos em separado.

A precisão é uma das métricas usadas para avaliar o alinhamento, medindo o percentual de correspondências corretas em relação ao total de correspondências retornadas por uma técnica ou metodologia de emparelhamento. Já a cobertura mede o percentual de correspondências que são retornadas em relação ao total de correspondências consideradas corretas, ou seja, mede-se o quanto das correspondências corretas a metodologia ou técnica consegue cobrir.

1.3 Objetivos

O principal objetivo deste trabalho é desenvolver uma nova metodologia para o emparelhamento de ontologias, tendo como base a clusterização consensual de múltiplos aspectos das ontologias. O trabalho visa também criar um arcabouço metodológico, no qual possam ser inseridas e testadas diferentes técnicas e aspectos

nas diversas fases da metodologia proposta, permitindo avaliar os resultados de emparelhamento provenientes de diferentes abordagens.

1.4 Organização do Texto

O presente texto organiza-se conforme a seguinte estrutura: no capítulo 2, são explicados alguns dos principais conceitos relativos a ontologias; no capítulo 3, explanam-se conceitos gerais de clusterização, suas principais técnicas e critérios de avaliação da qualidade da clusterização; nos capítulos 4, 5 e 6 explicam-se técnicas específicas de clusterização: Clusterização de Documentos, Clusterização de Grafos e *Cluster Ensembles*, respectivamente; no capítulo 7, faz-se a revisão da literatura de emparelhamento de ontologias baseado em clusterização; no capítulo 8, é explanada a metodologia adotada neste trabalho e é apresentado um manual de utilização do sistema que implementa a metodologia; no capítulo 9, descrevem-se os estudos de caso e apresentam-se os resultados; no capítulo 10, os resultados são discutidos e são propostos trabalhos futuros; no capítulo 11, são feitas as conclusões; o capítulo 12 é reservado às referências bibliográficas. Os capítulos 2 a 6 compreendem os conceitos e técnicas que foram aplicados na metodologia (capítulo 8). Sua apresentação, explicação e detalhamento fizeram-se necessários para facilitar o entendimento da metodologia. A metodologia, por sua vez, foi desenvolvida para cobrir lacunas e deficiências encontradas com base na revisão da literatura (capítulo 7).

Os anexos são os seguintes: anexo A, com o detalhamento das ontologias 101 e 304 de um dos estudos de caso, por meio de sua representação na forma de grafos; anexo B, contendo as ontologias 101 e 304 expressas em *Web Ontology Language* (OWL); anexo C, com os diagramas em *Unified Modeling Language* (UML), referentes ao sistema desenvolvido para implementar a metodologia; anexo D, com os alinhamentos das ontologias 101 e 304 obtidos ao aplicar a metodologia aos estudos de caso; anexo E, contendo os resultados de clusterização e de alinhamento das demais ontologias emparelhadas nos estudos de caso.

2 ONTOLOGIAS

Neste capítulo, são explicadas as principais definições acerca de ontologias e que estão relacionadas aos conceitos aplicados na metodologia proposta neste trabalho. Também são explanadas as linguagens de modelagem e de representação usadas na metodologia ora proposta. São elas: RDF, OWL e Lógica de Descrição. Além disso, são explicados as técnicas e os critérios de avaliação de emparelhamento, bem como as principais distâncias entre ontologias, visando embasar o entendimento da metodologia desenvolvida neste trabalho.

Para Mc Guinness e Harmelen (2004), ontologias descrevem entidades do mundo real e a forma como essas entidades se relacionam. Gruber (1993) define que ontologias são especificações formais e explícitas, as quais representam o conhecimento relativo a um domínio, possuindo descrições de objetos e de seus relacionamentos.

Para Hitzler et al. (2012), um agrupamento de objetos de uma ontologia com mesma característica é representado por uma classe. De acordo com Mc Guinness e Harmelen (2004), ontologias devem prover descrições de classes, seus relacionamentos e suas propriedades, para um dado domínio de interesse. Para Krötzsch, Simancík e Horrocks (2012) classes são conceitos, objetos de uma classe são indivíduos associados a um conceito e relações entre conceitos são papéis.

De acordo com Breitman, Casanova e Truszkowski (2007), ontologias podem ser expressas por meio de linguagens, tais como o *Resource Description Framework* (RDF) e a *Web Ontology Language* (OWL). A seguir, alguns dos principais conceitos relativos a RDF, OWL e Lógica de Descrição são explanados. Na sequência, técnicas de emparelhamento de ontologias e métricas para avaliação de emparelhamento são apresentadas.

2.1 RDF

Breitman, Casanova e Truszkowski (2007) definem o RDF como uma linguagem para representar informações sobre recursos. Para Manola, Miller e Mc Bride (2014), recursos podem ser objetos ou conceitos abstratos de uma ontologia, ou seja, são entidades do mundo real.

O RDF permite definir relações entre recursos por meio de sentenças. Conforme Manola, Miller e Mc Bride (2014), cada sentença em RDF é uma tripla, composta de sujeito, predicado e objeto. O sujeito e o objeto representam dois recursos, enquanto o predicado representa o relacionamento entre eles. As triplas em RDF podem ser representadas em um grafo. Sujeitos e objetos correspondem aos nós do grafo e os predicados correspondem aos arcos, conforme exemplo da figura 2.1.

Figura 2.1 - Exemplo de triplas representadas em um grafo.

Fonte: adaptado de Manola, Miller e Mc Bride (2014)

2.2 OWL

Hitzler et al. (2012) definem que a OWL é uma linguagem de modelagem do conhecimento humano, a qual permite modelar ontologias. Por meio da OWL, pode-se representar conhecimento complexo sobre objetos, grupos de objetos e seus relacionamentos, usando Lógica de Descrição (BAADER; NUTT, 2003). A OWL é considerada uma extensão do RDF, pois adiciona mais expressividade ao RDF sob o ponto de vista semântico (SEMANTIC WEB STANDARDS, 2011). A OWL aproveita a sintaxe do RDF para interoperabilidade, ao usar o RDF como uma linguagem comum para a troca de informações entre ontologias (W3C OWL WORKING GROUP, 2012).

A versão mais recente da OWL é a OWL 2 (HITZLER et al., 2012), que possibilita expressar e compartilhar o conhecimento de um dado domínio de interesse, além de permitir efetuar inferências. Na OWL 2, objetos são indivíduos, categorias de objetos são classes e relações são propriedades. Propriedades são subdivididas em propriedades de objetos e propriedades de tipos de dados. Propriedades de objetos relacionam objetos entre si e propriedades de tipos de dados atribuem valores a objetos. Há também propriedades denominadas anotações, as quais permitem associar metadados a elementos da ontologia.

De acordo com Hitzler et al. (2012), a OWL 2 permite modelar objetos, propriedades de objetos, classes, hierarquias entre classes, disjunções entre classes, hierarquias entre propriedades, igualdade e desigualdade entre objetos, tipos de dados, classes complexas (classes resultantes da união, interseção ou complemento entre classes), enumerações de indivíduos, restrições para cardinalidades de propriedades etc. Na tabela 2.1, são reproduzidos alguns dos exemplos de modelagens expressas com a sintaxe funcional da OWL 2, dados por Hitzler et al. (2012).

2.3 Lógica de Descrição

Conforme Krötzsch, Simancík e Horrocks (2012), Lógica de Descrição é uma família de linguagens de representação do conhecimento, baseada em uma semântica formal, a qual usa fragmentos de Lógica de Primeira Ordem (GENESERETH; NILSSON, 1987). A Lógica de Descrição é amplamente usada na modelagem de ontologias, tal como ocorre nas ontologias modeladas com OWL.

De acordo com Krötzsch, Simancík e Horrocks (2012), uma ontologia baseada em Lógica de Descrição não descreve completamente uma situação específica ou estado do mundo. As afirmações contidas na ontologia, denominadas axiomas, referem-se a um conhecimento parcial, sendo possíveis outros estados do mundo, consistentes com a ontologia. A semântica usada em Lógica de Descrição considera todas as possíveis situações, ou estados do mundo, em que os axiomas são válidos. Esta característica é comumente denominada *Open World Assumption*. Com uso da Lógica de Descrição, podem-se inferir informações adicionais, a partir de fatos explícitos contidos em uma ontologia. Quanto mais expressiva é a linguagem de representação, mais complexa é a tarefa de inferência.

Tabela 2.1 – Exemplos de modelagens em OWL 2.

Fonte: Hitzler et al. (2012)

Exemplo	Sintaxe	Descrição
Descrição de objetos	<i>ClassAssertion</i> (:Pessoa :Maria)	“Maria” é um indivíduo (ou instância) pertencente à classe “Pessoa”.
Hierarquia entre classes	<i>SubClassOf</i> (:Mulher :Pessoa)	Mulher é uma pessoa (a classe “Pessoa” é uma generalização da classe “Mulher”).
Equivalência entre classes	<i>EquivalentClasses</i> (:Pessoa :Humano)	Pessoa equivale a humano.
Propriedades de objetos	<i>ObjectPropertyAssertion</i> (:temEsposa :João :Maria)	João é casado com Maria.
Hierarquia entre propriedades	<i>SubObjectPropertyOf</i> (:temEsposa :temCompanheiro)	A propriedade “temCompanheiro” é uma generalização da propriedade “temEsposa”.
Classes complexas	<i>EquivalentClasses</i> (:Mãe <i>ObjectIntersectionOf</i> (:Mulher :Genitor))	A classe “Mãe” equivale à interseção entre a classe “Mulher” e “Genitor”

Conforme Krötzsch, Simancík e Horrocks (2012), a Lógica de Descrição possibilita modelar as relações entre entidades de um domínio. Há três tipos de entidades: conceitos, papéis e nomes individuais. Conceitos (ou classes) representam grupos de indivíduos similares, papéis representam relações binárias entre conceitos e nomes individuais são indivíduos (ou objetos) do domínio.

Para Krötzsch, Simancík e Horrocks (2012), as afirmações contidas na ontologia, representadas por axiomas, permitem classificar indivíduos, estabelecer relações entre conceitos ou indivíduos e definir propriedades de papéis. Há três tipos de axiomas: axiomas relativos a asserções (**ABox**), axiomas terminológicos (**TBox**) e axiomas relacionais (**RBox**).

Os axiomas do tipo **ABox** capturam o conhecimento referentes a indivíduos, permitindo estabelecer a qual conceito um indivíduo pertence e qual a relação entre indivíduos. Pode ser feita uma asserção de conceito, tal como **Mãe(Julia)**, que afirma que “Júlia é uma mãe”. Asserções de papel descrevem relações entre indivíduos,

como, por exemplo, **genitorDe(Julia, João)**, que afirma que “Júlia é a genitora de João”.

Conforme Krötzsch, Simancík e Horrocks (2012), os axiomas do tipo **TBox** descrevem relações entre conceitos, sendo que a inclusão de conceitos é um tipo de **TBox**. Seja, por exemplo, o axioma **Mãe $\sqsubseteq$ Genitor**, o qual afirma que o conceito “Mãe” herda características do conceito “Genitor”, ou “Genitor” é uma generalização de “Mãe”. A equivalência entre conceitos também é um axioma do tipo **TBox**, o qual é usado para estabelecer que dois conceitos têm as mesmas instâncias, como é o caso do axioma **Pessoa $\equiv$ Humano**.

Um **RBox** refere-se a axiomas para as propriedades de papéis. Há axiomas do tipo **RBox** para inclusão e equivalência. Um exemplo de inclusão é o axioma **genitorDe $\sqsubseteq$ ancestralDe**, o qual define que **genitorDe** é um subpapal de **ancestralDe**, estabelecendo que todos indivíduos ligados pela relação **genitorDe** são também ligados pela relação **ancestralDe**, ou seja, todo indivíduo que é genitor de um indivíduo também é seu ancestral.

2.4 Emparelhamento de Ontologias

De acordo com Euzenat e Shvaiko (2013), emparelhamento é o processo de encontrar entidades correspondentes em ontologias diferentes. O conjunto de correspondências obtido denomina-se alinhamento, sendo este o resultado do processo de emparelhamento. O emparelhamento é importante na integração de ontologias, pois é o passo que precede a mesclagem de ontologias. O alinhamento obtido com o emparelhamento é o dado de entrada para que se possa fazer a mesclagem. A partir da mesclagem é possível uma visão integrada de diferentes ontologias, representada por uma ontologia única, sobre a qual podem-se realizar inferências relativas ao domínio que resulta da integração. Uma ilustração de possíveis correspondências entre duas ontologias é feita na figura 2.2.

Dada a importância do emparelhamento, há iniciativas voltadas para o desenvolvimento acadêmico e corporativo de pesquisas relacionadas ao tema, como é o caso da *Ontology Alignment Evaluation Initiative* (OAEI) (ONTOLOGY ALIGNMENT EVALUATION INITIATIVE, 2017). Por meio da OAEI, são promovidos

Figura 2.2 – Alinhamento entre duas ontologias.

Fonte: Shvaiko (2006b)

eventos anuais, em que pesquisadores são motivados a apresentar suas soluções de emparelhamento, fomentando, assim, o desenvolvimento dessa área de pesquisa em ontologias.

2.4.1 Técnicas de Emparelhamento

Diversas técnicas de emparelhamento visam resolver heterogeneidades entre elementos de ontologias, descobrindo semelhanças e estabelecendo correspondências entre esses elementos. Euzenat e Shvaiko (2013) propõem uma taxonomia para classificar as técnicas de emparelhamento em níveis hierárquicos, conforme a figura 2.3.

No primeiro nível hierárquico da figura 2.3, Euzenat e Shvaiko (2013) subdividem as técnicas de emparelhamento em elementares ou estruturais. As técnicas elementares são usadas para encontrar similaridades entre elementos heterogêneos de ontologias, analisados isoladamente, comparando-se nomes, tipos de dados e restrições dos elementos. As técnicas estruturais visam descobrir elementos correspondentes com base na análise das estruturas das ontologias como um todo.

Em um segundo nível hierárquico da figura 2.3, Euzenat e Shvaiko (2013) subdividem tanto as técnicas elementares quanto as estruturais em sintáticas e semânticas.

Figura 2.3 – Classificação de técnicas de emparelhamento.

Fonte: Euzenat e Shvaiko (2013)

- **Técnicas Elementares Sintáticas**

Para Euzenat e Shvaiko (2013), as técnicas elementares sintáticas encontram correspondências do emparelhamento com base em recursos informais, similaridade de cadeias de caracteres, linguagem ou restrições.

Recursos informais podem ser documentos anotados (ou referenciados) por classes de uma ontologia. Assim, classes que anotam um mesmo termo em um documento são consideradas correspondentes.

Correspondências podem ser estabelecidas também com base na similaridade de caracteres dos nomes de classes ou propriedades de duas ontologias diferentes. A premissa é que, quanto maior for a similaridade entre as cadeias de caracteres, mais provável é que as classes ou propriedades associadas às cadeias sejam equivalentes. Usam-se funções que calculam distâncias entre termos, sendo que, quanto menor a distância, maior a similaridade. Neste grupo, Euzenat e Shvaiko (2013) destacam as seguintes técnicas: *prefix*, *suffix*, distância de Levenshtein, *n*-gram. Quanto à complexidade computacional, podem-se descrever, por exemplo, a distância de Levenshtein (UKKONEN, 1983) e o *n*-gram (NADESAN; ALAGARSAMY, 2013). Para uma comparação das cadeias de caracteres de duas palavras, p_1 e p_2 , a distância de Levenshtein tem complexidade no tempo igual a $O(\text{len}_{p_1} \times \text{len}_{p_2})$, onde len_{p_1} é a quantidade de caracteres da palavra p_1 e len_{p_2} é a quantidade de caracteres da palavra p_2 . Já o *n*-gram tem complexidade $O(n \log n)$, onde, dadas duas palavras p_1 e p_2 a serem comparadas, n é o tamanho da cadeia de caracteres usada para comparar essas palavras.

Técnicas baseadas em linguagem analisam características morfológicas de termos e subdividem-se em técnicas de *tokenization*, de lematização e baseadas em dicionários. Técnicas de *tokenization* desmembram um texto em unidades (*tokens*). Cada *token* é igual a uma sequência de caracteres que forma um termo. Na lematização, os termos são reduzidos aos seus radicais (verbos no gerúndio, por exemplo, são reduzidos para o infinitivo). Técnicas que se baseiam em dicionários fazem buscas em léxicos, tais como o WordNet (MILLER, 1995), para encontrar palavras com grafias diferentes, porém com mesmo significado. No caso específico do WordNet, usualmente conhecido como um *thesaurus*, há não só sinônimos, mas também hiperônimos, que são termos com significado mais abrangente

(generalizações) em relação a uma dada palavra. Um exemplo é a palavra “biografia”, que é hiperônimo de “autobiografia”.

Técnicas que usam restrições fazem emparelhamento analisando tipos de dados, cardinalidades ou chaves de elementos das ontologias.

- **Técnicas Elementares Semânticas**

De acordo com Euzenat e Shvaiko (2013), as técnicas elementares semânticas baseiam-se em recursos formais, utilizando ontologias de domínio para auxiliar o emparelhamento. Essas técnicas podem inferir correspondências entre duas ontologias com base em uma terceira ontologia, que cubra o domínio das outras duas, ou podem-se reusar correspondências previamente armazenadas, que são reaproveitadas para estabelecer correspondências entre duas ontologias em comparação.

- **Técnicas Estruturais Sintáticas**

Para Euzenat e Shvaiko (2013), as técnicas estruturais sintáticas baseiam-se na análise estrutural de grafos e taxonomias associados às ontologias. Parte-se da premissa de que elementos dos grafos ou das taxonomias que são vizinhos a elementos correspondentes entre si também são correspondentes. Incluem-se também no grupo das estruturais sintáticas as técnicas baseadas em instâncias, as quais comparam objetos pertencentes a diferentes classes das ontologias para verificar se as classes são equivalentes. Essas técnicas também são denominadas extensionais.

- **Técnicas Estruturais Semânticas**

Uma vertente das técnicas estruturais semânticas baseia-se em Lógica Proposicional (POSPESEL, 1979). No trabalho de Shvaiko (2006a), inicialmente, os conceitos de duas ontologias em comparação são expressos por meio de axiomas (subseção 2.3). Faz-se a união dos axiomas das duas ontologias, gerando uma base de conhecimento. Para obter o emparelhamento, testa-se a validade de correspondências entre elementos das ontologias em relação à base de conhecimento. Testa-se se uma determinada correspondência é válida, verificando-

se se a base de conhecimento tem como consequência lógica o axioma que define a correspondência entre um par de elementos.

2.4.2 Avaliação do Emparelhamento

Para avaliar resultados de emparelhamento, são usadas as medidas de precisão, cobertura (ou *recall*) e o *F-measure*. Trata-se de métricas oriundas da Recuperação de Informação e criadas por van Rijsbergen (1979).

Conforme Euzenat e Shvaiko (2013), a precisão é igual ao número de correspondências corretas encontradas, dividido pelo número total de correspondências obtidas. A cobertura é igual ao número de correspondências corretas obtidas dividido pelo número total de correspondências de um resultado de alinhamento de referência, usado como gabarito, também denominado de *gold standard* ou *benchmark*. O *F-measure* usualmente é calculado fazendo-se a média harmônica entre a precisão e a cobertura.

Assim, sendo, a precisão, a cobertura e o *F-measure* são expressos por meio das seguintes fórmulas:

$$\text{Precisão} = \frac{\text{Número de correspondências corretas}}{\text{Número total de correspondências}} \quad (1)$$

$$\text{Cobertura} = \frac{\text{Número de correspondências corretas}}{\text{Número de correspondências de referência}} \quad (2)$$

$$\text{F-measure} = \frac{2 \times \text{precisão} \times \text{cobertura}}{\text{precisão} + \text{cobertura}} \quad (3)$$

2.5 Distâncias entre Ontologias

Conforme Euzenat et al. (2009), distâncias entre ontologias são calculadas com o intuito de compará-las antes de ser feito o emparelhamento. A distância obtida serve de base para decidir se uma técnica de emparelhamento deve ser aplicada ou não para alinhar as ontologias. Comparam-se as ontologias como um todo. Euzenat et al. (2009) definem algumas das principais técnicas e métricas usadas para obter

distâncias entre ontologias. Dentre elas, têm-se as distâncias baseadas no Modelo do Espaço Vetorial e as distâncias semânticas.

Segundo Euzenat et al. (2009), na distância baseada no Modelo do Espaço Vetorial, cada ontologia é representada por um vetor, no qual termos contidos em anotações e nomes de elementos das ontologias são as dimensões de um espaço vetorial. São atribuídos pesos aos termos conforme sua relevância. As distâncias entre ontologias representadas nesse espaço correspondem a distâncias angulares entre os vetores representativos das ontologias.

Euzenat et al. (2009) definem também uma distância semântica entre ontologias, que se baseia em suas consequências lógicas. A distância δ_{is} é dada por:

$$\delta_{is}(o, o') = 1 - \frac{|CL(o) \cap CL(o')|}{|CL(o) \cup CL(o')|} \quad (4)$$

Onde:

o, o' = duas ontologias em comparação;

CL = função que retorna as consequências lógicas de uma ontologia.

Segundo Euzenat et al. (2009), o conjunto de consequências lógicas de uma ontologia é usualmente infinito, podendo-se tentar recorrer a reduções do tamanho das ontologias. Porém, essa redução não é única, o que dificulta a obtenção da distância semântica entre ontologias.

Hu et al. (2006) definem também diversas métricas aplicáveis a ontologias. Uma delas usa conceitos alicerçados em probabilidades. Hu et al. (2006) definem a probabilidade $p(C_k)$ de um conceito C_k com base em sua distância $\Delta(C_k)$ em relação a um conceito ideal C_o , para o qual não há restrição. A restrição pode ser calculada como sendo igual ao inverso da quantidade de objetos instanciados pelo conceito, de forma que C_o tem restrição igual a zero, ou seja, C_o instancia o máximo de objetos possíveis. Assim, a probabilidade $p(C_k)$ é igual a:

$$p(C_k) = \frac{1 - \Delta(C_k)}{\sum_{j=0}^n (1 - \Delta(C_j))} \quad (5)$$

Onde:

$p(C_k)$ = probabilidade associada ao conceito C_k ;

$\Delta(C_k)$ = distância do conceito C_k em relação ao conceito ideal;

$\Delta(C_j)$ = distância do conceito C_j em relação ao conceito ideal;

n = quantidade de conceitos na ontologia.

3 CLUSTERIZAÇÃO

Neste capítulo, são explicados os conceitos fundamentais sobre clusterização, tais como *clusters*, instâncias e rótulos, e que são relevantes para a compreensão da metodologia ora desenvolvida. São definidas também as principais medidas de distância usadas na metodologia, medidas essas importantes para mensurar a similaridade entre objetos, a qual é fundamental para descobrir padrões nos dados. Esses padrões, por sua vez, permitem definir agrupamentos com características similares, ou seja, os *clusters*. Além disso, são explicadas as técnicas de clusterização e os critérios de avaliação de clusterização aplicados na metodologia proposta neste trabalho.

Conforme Jain, Murty e Flynn (1999), a clusterização agrupa uma coleção de objetos em categorias, sendo cada categoria um agrupamento de objetos com características similares. As categorias são denominadas *clusters*, de forma que objetos em um mesmo *cluster* têm maior similaridade entre si do que objetos pertencentes a *clusters* diferentes. Na clusterização, os objetos são usualmente denominados instâncias e é atribuído um rótulo a cada *cluster*, com o intuito de categorizá-lo.

- **Medidas de Distância**

Segundo Jain, Murty e Flynn (1999), na clusterização é preciso definir uma medida de proximidade entre as instâncias. A proximidade pode ser aferida por meio de distâncias, sendo que, quanto menor a distância, maior a similaridade entre os elementos comparados. Dentre as medidas, há duas métricas de distância usualmente aplicadas: a distância Euclidiana e a distância de Manhattan.

Usando as definições e notações de Jain, Murty e Flynn (1999), um conjunto de n instâncias é representado por $X = \{x_1, x_2, \dots, x_n\}$, sendo a i -ésima instância representada por $x_i = (x_{i1}, x_{i2}, \dots, x_{id})$, em que d é a quantidade de atributos de uma instância ou sua dimensão. Assim sendo, conforme Jain, Murty e Flynn (1999), para duas instâncias $x_i = (x_{i1}, x_{i2}, \dots, x_{id})$ e $x_j = (x_{j1}, x_{j2}, \dots, x_{jd})$, de dimensão d , a distância Euclidiana DE entre elas é calculada por meio da seguinte fórmula:

$$DE(x_i, x_j) = (\sum_{k=1}^d |x_{ik} - x_{jk}|^2)^{1/2} \quad (6)$$

De acordo com Xu e Wunsch (2005), a distância de Manhattan DM entre essas instâncias é dada por:

$$DM(x_i, x_j) = \sum_{k=1}^d |x_{ik} - x_{jk}| \quad (7)$$

Segundo afirmam Xu e Wunsch (2005), a distância Euclidiana tende a gerar *clusters* com formato esférico e a distância de Manhattan tende a gerar *clusters* de formato semelhante ao de um paralelepípedo.

A seguir são descritas as técnicas de clusterização hierárquica e particional. Ao final deste capítulo, métricas para avaliar a clusterização são apresentadas.

3.1 Clusterização Hierárquica

Conforme Xu e Wunsch (2005), para um conjunto de dados com N instâncias, define-se uma matriz NxN, denominada matriz de proximidade, cuja entrada na linha i e na coluna j representa a medida de similaridade entre a i-ésima instância e a j-ésima instância. Nesse sentido, Xu e Wunsch (2005) definem que algoritmos de clusterização hierárquica organizam dados em uma hierarquia, de acordo com a matriz de proximidade. Os resultados da aplicação desses algoritmos são representados em uma árvore binária ou dendograma, conforme ilustra a figura 3.1. De acordo com Jain, Murty e Flynn (1999), o dendograma é uma hierarquia que representa o agrupamento aninhado de instâncias, sendo que a composição desses agrupamentos varia conforme a similaridade entre as instâncias.

Conforme Xu e Wunsch (2005), o nó da raiz da hierarquia de um dendograma (figura 3.1) corresponde ao conjunto total de dados. Ao percorrer a estrutura do dendograma descendentemente, a hierarquia formada define *clusters* que agrupam uma quantidade de instâncias cada vez menor, de forma que, ao chegar aos nós folha da estrutura, tem-se um *cluster* para cada instância do conjunto de dados. O eixo vertical do dendograma representa a similaridade entre as instâncias. Conforme a figura 3.1, o eixo da similaridade representa valores positivos e crescentes conforme progride-se descendentemente pelo eixo. Assim sendo, de acordo com Jain, Murty e Flynn (1999), pode-se impor um nível de similaridade para as instâncias dos *clusters*,

Figura 3.1 – Exemplo de um dendograma obtido com clusterização hierárquica.

Fonte: adaptado de Jain, Murty e Flynn (1999)

por meio de um valor de corte, traçando-se uma reta horizontal no valor de similaridade desejado. A reta intercepta, na hierarquia do dendograma, a estrutura de *clusters* correspondente à similaridade imposta.

Conforme Jain, Murty e Flynn (1999), a clusterização hierárquica pode ser divisiva (*top-down*) ou aglomerativa (*bottom-up*). Na divisiva, todas as instâncias estão inicialmente em um único *cluster* e é feita a subdivisão em *clusters* até satisfazer-se um determinado critério de parada. Xu e Wunsch (2005) enumeram alguns algoritmos divisivos, tais como o DIANA (KAUFMAN; ROUSSEEUW, 1990) e o MONA (KAUFMAN; ROUSSEEUW, 1990).

Na clusterização hierárquica aglomerativa, parte-se de uma estrutura em que cada *cluster* é formado por uma única instância (cada instância é um *cluster* distinto) e procede-se à mesclagem dos *clusters* até que haja uma convergência. Xu e Wunsch (2005) destacam alguns algoritmos aglomerativos, tais como o *single-link* e o *complete-link*. Segundo Jain, Murty e Flynn (1999), no *single-link*, a distância entre *clusters* é igual à mínima distância entre os padrões (atributos dos dados) existentes nos *clusters*. Já no *complete-link*, essa distância é igual à máxima distância entre os

padrões existentes nos *clusters*. Jain, Murty e Flynn (1999) afirmam que o *complete-link* gera *clusters* compactos (com elementos bem próximos uns dos outros), enquanto, no *single-link*, os *clusters* são alongados, ou seja, seus elementos estão mais dispersos ou espalhados. Dentre os algoritmos aglomerativos, têm-se o *Unweighted Pair Group Method with Arithmetic Mean* (UPGMA) (SOKAL; MICHENER, 1958) e o método de Ward (WARD, 1963).

De acordo com Xu e Wunsch (2005), devido ao alto custo computacional da clusterização hierárquica divisiva, seu uso é menos comum do que a clusterização hierárquica aglomerativa.

Jain, Murty e Flynn (1999) descrevem os passos da clusterização hierárquica aglomerativa:

Passo 1: obter a matriz de proximidade, contendo a distância entre os pares de instâncias;

Passo 2: encontrar os pares de *clusters* mais similares, usando a matriz de proximidade. Mesclar os *clusters* e atualizar a matriz de proximidade;

Passo 3: se todas as instâncias estiverem em um *cluster*, parar. Do contrário, voltar ao passo 2.

Segundo Jain, Murty e Flynn (1999), algoritmos hierárquicos têm mais versatilidade do que os particionais, tais como o *K-means*, pois conseguem detectar *clusters* de formato alongado (com instâncias mais dispersas ou mais espalhadas espacialmente). Esses tipos de *clusters* também são denominados de *clusters* não isotrópicos. Já os particionais costumam apresentar melhores resultados com dados que apresentem *clusters* de formato mais compacto (*clusters* isotrópicos).

3.2 Clusterização Particional

Conforme definido por Jain, Murty e Flynn (1999), algoritmos particionais obtêm uma única partição dos dados, em vez de obter uma estrutura de *clusters*, como ocorre no caso de algoritmos hierárquicos. Jain, Murty e Flynn (1999) destacam o algoritmo *K-means*. Há outros algoritmos particionais, tais como *Partitioning Around Medoids* (PAM) (KAUFMAN; ROUSSEEUW, 1990) e CLARA (KAUFMAN; ROUSSEEUW, 1990). A seguir, explica-se o *K-means*.

- **K-means**

Segundo Jain (2010), o *K-means* determina *K clusters* em um conjunto de dados. O valor de *K* não é obtido automaticamente pelo algoritmo, sendo necessário fornecer valores aleatórios de *K* como semente e avaliar a qual valor de *K* corresponde o melhor resultado de clusterização. O resultado é avaliado comparando-se a um *gold standard*, quando disponível, ou verificando-se, por meio de métricas apropriadas (como a largura de silhueta, por exemplo), se os *clusters* isoladamente são bem compactos e se diferentes *clusters* são bem separados entre si.

No *K-means*, dado um valor de *K* aleatório, o conjunto de *clusters* é obtido minimizando-se o erro quadrático entre o elemento médio de cada *cluster* e os demais elementos do *cluster*. Jain, Murty e Flynn (1999) formalizam este critério por meio da seguinte função:

$$e^2(X, L) = \sum_{j=1}^K \sum_{i=1}^{n_j} |x_i^{(j)} - c_j|^2 \quad (8)$$

Onde:

$e^2(X, L)$ = erro quadrático para uma clusterização *L* de um conjunto de dados *X*;

K = número de *clusters* fornecido como semente ao *K-means*;

n_j = número de instâncias do *j*-ésimo *cluster*;

$x_i^{(j)}$ = *i*-ésima instância do *j*-ésimo *cluster*;

c_j = centroide do *j*-ésimo *cluster*;

De acordo com Jain (2010), os passos do *K-means* são os seguintes:

- 1) Selecionar aleatoriamente uma partição inicial com *K clusters*;
- 2) Atribuir elementos a um *cluster* cujo centroide seja mais próximo desses elementos, gerando, com isso, uma nova partição;
- 3) Recalcular os centroides dos *clusters*;
- 4) Repetir 2 e 3 até estabilizar a composição dos *clusters*. Isto ocorre se as instâncias não são mais redistribuídas entre os *clusters* ou se o erro quadrático passa a não ter um decréscimo significativo.

Segundo Jain, Murty e Flynn (1999), para a clusterização de conjuntos grandes de dados, o tempo de execução de algoritmos particionais costuma ser menor do que o dos algoritmos hierárquicos aglomerativos, tendo em vista o alto custo computacional destes últimos por causa da construção do dendograma. A complexidade de tempo do *K-means* é $O(n)$, sendo n o número de instâncias. Para os hierárquicos aglomerativos, a complexidade de tempo é $O(n^2 \log n)$. Um dos problemas apontados por Jain, Murty e Flynn (1999) em relação ao *K-means* é a sua sensibilidade em relação às escolhas da partição inicial e do valor de K . Se estes não forem adequadamente escolhidos, a função critério do *K-means* pode convergir para um mínimo local.

3.3 Avaliação de Clusterização

De acordo com Manning, Raghavan e Schütze (2009), a avaliação do resultado de uma clusterização pode ser feita por meio de critérios internos ou externos. Conforme Handl, Knowles e Kell (2005), critérios internos avaliam a qualidade da clusterização com base nas características intrínsecas dos dados. São usados quando não há um *benchmark* ou *gold standard* de clusterização disponível como referência para avaliar a clusterização. Para Manning, Raghavan e Schütze (2009), critérios externos baseiam-se em um *benchmark* de clusterização, formado por classes pré-definidas, que fornecem a clusterização considerada correta, a qual é a referência para avaliar a clusterização obtida ao aplicar um determinado algoritmo.

Handl, Knowles e Kell (2005) definem algumas características que uma clusterização de qualidade deve ter e que embasam a aplicação de critérios internos. São elas: compactação; conectividade e separação. A compactação mede a homogeneidade de um *cluster* com base na distância média entre elementos de um *cluster* e seu centroide. A conectividade é proporcional à quantidade de elementos vizinhos a um dado elemento que é atribuída ao mesmo *cluster* desse elemento. A separação baseia-se na distância entre centroides de diferentes *clusters*.

Conforme Handl, Knowles e Kell (2005), na avaliação segundo critérios internos, aplicam-se métricas para avaliar a presença dessas características em um resultado de clusterização. Métricas como a largura de silhueta (ROUSSEEUW, 1987) e o índice de Dunn (DUNN, 1974) usam formulações que permitem avaliar a

compactação e a separação de forma combinada. Para a conectividade, há também tem uma expressão matemática própria.

A largura de silhueta baseia-se na distância média entre elementos de um mesmo *cluster*, mensurando assim a compactação de um *cluster*, e baseia-se também na distância média entre elementos de um *cluster* e elementos do *cluster* mais próximo, para aferir a separação entre *clusters*. A fórmula da largura de silhueta é dada por:

$$S(i) = \frac{b_i - a_i}{\max(b_i, a_i)} \quad (9)$$

Onde:

a_i = distância média entre o elemento i e todos elementos do mesmo *cluster* de i ;

b_i = distância média entre o elemento i e os elementos do *cluster* mais próximo;

Conforme Rousseeuw (1987), a largura de silhueta varia entre -1 e 1, sendo que, em uma dada clusterização, se um objeto tem largura de silhueta próxima ou igual a 1 ($a_i \ll b_i$), isto significa que o objeto foi bem agrupado pela técnica de clusterização, ou seja, o objeto é bem compacto em relação aos elementos de seu *cluster* e está bem separado dos elementos dos *clusters* mais próximos. Por outro lado, se ao aplicar uma dada técnica de clusterização, resultar largura de silhueta próxima de -1 ($a_i > b_i$) para um objeto, significa que a técnica de clusterização não agrupou o objeto devidamente. Uma largura de silhueta próxima de zero ($a_i = b_i$) para um dado objeto significa que não se conseguiu definir se esse objeto deve ficar em um agrupamento ou em outro, ou seja, sua atribuição a um *cluster* ou a outro é indiferente e ele está em uma posição intermediária.

O índice de Dunn é baseado no quociente entre a mínima distância entre diferentes *clusters* e a máxima distância entre elementos de um mesmo *cluster*. Esse índice é expresso por:

$$D = \min_{C_k \in C} \left(\min_{C_l \in C} \frac{\text{dist}(C_k, C_l)}{\max_{C_m \in C} \text{diâm}(C_m)} \right) \quad (10)$$

Onde:

C = conjunto de *clusters*;

$\text{dist}(C_k, C_l)$ = mínima distância entre elementos pertencentes aos *clusters*

C_k e C_l ;

$\text{diâm}(C_m)$ = máxima distância entre elementos do *cluster* C_m .

De acordo com Handl, Knowles e Kell (2005), a conectividade é mensurada por meio da seguinte fórmula:

$$\text{CON} = \sum_{i=1}^N \sum_{j=1}^L x_i, \text{nn}_{i(j)} \quad (11)$$

Onde:

N = número total de elementos de um *cluster*;

L = número de vizinhos que contribuem para a conectividade;

$\text{nn}_{i(j)}$ = j -ésimo vizinho do elemento i ;

$$x_i, \text{nn}_{i(j)} = \begin{cases} \frac{1}{j} & \text{se } \nexists C_k \text{ tal que } i \text{ e } \text{nn}_{i(j)} \in C_k \\ 0, & \text{caso contrário} \end{cases}$$

Segundo Handl, Knowles e Kell (2005), como os algoritmos de clusterização baseiam-se na otimização de um critério de clusterização, os *clusters* resultantes de cada algoritmo costumam ter um viés em relação a determinada característica (compactação, conectividade ou separação). Conforme Handl, Knowles e Kell (2005), o *K-means* e o UPGMA, por exemplo, tendem a ter maior compactação e o *single-link*, melhor conectividade. Como consequência, determinados algoritmos apresentam melhores resultados quanto a determinadas métricas. Assim sendo, Handl, Knowles e Kell (2005) consideram uma boa solução de clusterização aquela que apresenta bons resultados para métricas que avaliam diferentes características simultaneamente. Nesse sentido, Handl, Knowles e Kell (2005) recomendam o uso de métricas que combinam a análise de diferentes características da clusterização. Essa combinação pode ser obtida tanto ao aplicar a largura de silhueta quanto ao aplicar o índice de Dunn. Handl, Knowles e Kell (2005) acrescentam que o uso da largura de

silhueta é mais vantajoso do que o índice de Dunn, pois este último é mais sensível ao ruído dos dados.

Para a avaliação de clusterização com uso de critérios externos, Manning, Raghavan e Schütze (2009) destacam as seguintes métricas: informação mútua normalizada (COVER; THOMAS, 1991); pureza (ZHAO; KARYPIS, 2001); índice de Rand (RAND, 1971); *F-measure* (VAN RIJSBERGEN, 1979). A informação mútua normalizada baseia-se na probabilidade de um elemento estar associado simultaneamente a um *cluster* e a uma classe do *benchmark* de clusterização. A pureza baseia-se na classe predominante para um determinado *cluster*. O índice de Rand considera a atribuição de elementos a um mesmo *cluster* como um processo de decisão, medindo a porcentagem de decisões corretas com base no *benchmark* de clusterização. O *F-measure* faz uma ponderação de erros e acertos nessas decisões.

4 CLUSTERIZAÇÃO DE DOCUMENTOS

Os conceitos de modelagem, as técnicas e o processo associados à Clusterização de Documentos, explanados neste capítulo, são diretamente aplicados em uma das fases da metodologia criada neste trabalho. Portanto, dedicou-se este capítulo à explicação do Modelo do Espaço Vetorial, das técnicas de pré-processamento de textos (*tokenization*, remoção de *stop words* etc.) e de redução de dimensionalidade (Análise Semântica Latente e Análise de Componentes Independentes). A maioria delas tem sido aplicadas em trabalhos do estado da arte de emparelhamento de ontologias baseado em clusterização. No entanto, técnicas como a Análise de Componentes Independentes, por exemplo, foram pouco exploradas nos trabalhos relacionados e mereceram uma atenção diferenciada ao serem também incluídas neste capítulo.

Manning, Raghavan e Schütze (2009) decompõem o processo de Clusterização de Documentos nas seguintes fases: pré-processamento; modelagem da representação dos documentos; redução de dimensionalidade; aplicação de técnicas de clusterização propriamente ditas.

Conforme Manning, Raghavan e Schütze (2009), previamente ao processo de Clusterização de Documentos, é necessário fazer um pré-processamento do texto contido nos documentos. Dentre as técnicas usadas nessa fase, Manning, Raghavan e Schütze (2009) destacam as técnicas de *tokenization*, remoção de *stop words*, *stemming* e lematização. Técnicas de *tokenization* quebram um texto em unidades, denominadas *tokens*, sendo cada *token* correspondente à sequência de caracteres que forma um termo. A remoção de *stop words* é feita para retirar palavras que não acrescentam significado, tais como artigos, preposições, conjunções etc. No *stemming*, são retirados afixos dos termos. Na lematização, os termos são reduzidos aos seus radicais, fazendo com que verbos no gerúndio, por exemplo, sejam reduzidos à sua forma no infinitivo.

Para efetuar a Clusterização de Documentos, Manning, Raghavan e Schütze (2009) apontam como fundamental a utilização de um modelo para representação dos documentos. Um desses modelos é o Modelo do Espaço Vetorial, que representa um conjunto de documentos como vetores, em um espaço cuja dimensão inicial é igual ao número de palavras distintas da coleção. A modelagem vetorial permite aferir similaridades entre documentos por meio de distâncias angulares entre os vetores

representativos de cada documento, possibilitando agrupar documentos com base nessas similaridades, usando-se técnicas de clusterização.

Associada ao Modelo do Espaço Vetorial de uma coleção de documentos, há a matriz termo-documento, em que cada linha corresponde a um documento e cada coluna corresponde a um termo diferente da coleção, sendo que as entradas na matriz possuem um valor correspondente à relevância do termo na coleção. Manning, Raghavan e Schütze (2009) observam que até mesmo pequenas coleções de documentos podem gerar uma matriz termo-documento, associada ao Modelo do Espaço Vetorial, com dezenas de milhares de linhas ou colunas. Em função disso, técnicas de redução de dimensionalidade foram desenvolvidas para facilitar a clusterização. Dentre as principais técnicas de redução de dimensionalidade, têm-se: Análise Semântica Latente (LANDAUER; FOLTZ; LAHAM, 1998) e a Análise de Componentes Independentes (HYVÄRINEN, 2000).

Após a redução de dimensionalidade, aplicam-se técnicas de clusterização, com base nas distâncias entre documentos, que podem ser calculadas usando a distância Euclidiana (capítulo 3).

A seguir, são explanados o Modelo do Espaço Vetorial e as principais técnicas de redução de dimensionalidade.

4.1 Modelo do Espaço Vetorial

No Modelo do Espaço Vetorial, conforme Manning, Raghavan e Shütze (2009), dada uma coleção de N documentos, cada documento é modelado como um vetor de n dimensões, em que n é igual à quantidade de palavras distintas da coleção de documentos. Cada dimensão está associada a uma palavra diferente e tem um valor que é proporcional à importância do termo na coleção de documentos.

Manning, Raghavan e Shütze (2009) calculam o valor de cada dimensão por meio da métrica *Term Frequency-Inverse Document Frequency* (TF-IDF). Para Manning, Raghavan e Shütze (2009), o TF de um termo t em um documento D , dado por $TF_{t,D}$, é a frequência de um termo t em um documento D e visa valorizar palavras frequentes. O IDF é usado para ponderar a importância de palavras raras do conjunto de documentos. Manning, Raghavan e Shütze (2009) definem o IDF de um termo t , dado por IDF_t , por meio da seguinte fórmula:

$$IDF_t = \log \frac{N}{DF_t} \quad (12)$$

Onde:

N = quantidade de documentos da coleção;

DF_t = quantidade de documentos da coleção que contêm o termo t ;

Para calcular o TF-IDF de um termo t em um documento D , Manning, Raghavan e Shütze (2009) efetuam o produto $TF_{t,D} \times IDF_t$. Desta forma, com o TF-IDF pondera-se simultaneamente a frequência de uma palavra no documento e a sua raridade na coleção.

Uma ilustração do Modelo do Espaço Vetorial é feita na figura 4.1, na qual três documentos, d_1 , d_2 e d_3 , são representados pelos vetores $\vec{v}(d_1)$, $\vec{v}(d_2)$ e $\vec{v}(d_3)$, respectivamente. Uma consulta por documentos é representada pelo vetor $\vec{v}(q)$. Os eixos horizontal e vertical representam termos presentes na coleção de documentos e o ângulo Θ representa a similaridade entre d_1 e d_2 . Observa-se que, quanto mais frequente for um determinado termo em um documento, mais o vetor que representa esse documento aproxima-se do eixo que representa o termo. No caso da figura 4.1, o documento d_1 tem maior ocorrência da palavra “*gossip*”, por isso é mais próximo do eixo que representa esse termo, enquanto o documento d_3 tem maior ocorrência da palavra “*jealous*”, ficando próximo ao eixo dessa palavra. O documento d_2 e a consulta q tem uma frequência mais balanceada desses dois termos, ficando em uma posição intermediária. Quanto mais similares os documentos entre si ou a consulta em relação aos documentos, menor é o ângulo Θ entre os vetores que representam esses elementos.

Então, o cálculo da similaridade entre d_1 e d_2 pode ser feito por meio da similaridade do cosseno (MANNING; RAGHAVAN; SHÜTZE, 2009). Calcula-se o produto interno dos vetores que representam os documentos e normaliza-se esse produto, dividindo-o pelo produto dos comprimentos euclidianos dos vetores. Com base na similaridade entre documentos, pode-se calcular também a distância, que é igual ao complemento da similaridade em relação à unidade, de forma que, quanto maior a similaridade entre os documentos, menor é a distância entre eles. Manning, Raghavan e Shütze (2009) definem a similaridade do cosseno para d_1 e d_2 , dada por $\text{sim}(d_1, d_2)$, por meio da fórmula 13 a seguir.

Figura 4.1 – Modelo do Espaço Vetorial representando os documentos d_1 , d_2 , d_3 e a consulta q , considerando apenas dois termos distintos (“*gossip*” e “*jealous*”) nesses documentos e na consulta.

Fonte: Manning, Raghavan e Shütze (2009)

$$\text{sim}(d_1, d_2) = \frac{\vec{v}(d_1) \cdot \vec{v}(d_2)}{|\vec{v}(d_1)| |\vec{v}(d_2)|} \quad (13)$$

Com base nas similaridades entre os documentos de uma coleção, monta-se uma matriz M de similaridades, na qual cada entrada $M(i,j)$ da matriz é igual à similaridade entre o documento i e o documento j . Em seguida, aplicam-se algoritmos de clusterização à matriz M , para gerar agrupamentos de documentos similares.

4.2 Técnicas de Redução de Dimensionalidade

Steinbach, Ertöz e Kumar (2003) ressaltam problemas resultantes da alta dimensionalidade dos dados, afirmando que dados se tornam mais esparsos conforme a dimensionalidade aumenta. Segundo Steinbach, Ertöz e Kumar (2003), no caso da clusterização de dados com alta dimensionalidade de atributos, a diferença

relativa entre a distância mínima de duas instâncias e a distância máxima de duas instâncias tende a zero, dificultando a identificação de *clusters*, pois a distância entre instâncias tende a ficar uniforme.

Conforme Steinbach, Ertöz e Kumar (2003), para contornar esse problema, faz-se a redução da dimensionalidade, visando manter apenas os atributos mais relevantes. Removem-se, então, aqueles atributos que têm pouca variação ou que sejam altamente correlacionados com outros atributos.

Na sequência, descrevem-se as técnicas de Análise Semântica Latente e Análise de Componentes Independentes.

4.2.1 Análise Semântica Latente

A Análise Semântica Latente (MARTIN; BERRY, 2006), ou *Latent Semantic Analysis* (LSA) em inglês, é aplicada a uma representação inicial da coleção de documentos, correspondente ao Modelo do Espaço Vetorial. Monta-se a matriz termo-documento A , com cada elemento $A(i,j)$ inicialmente correspondendo, por exemplo, ao TF-IDF relacionado ao termo i e ao documento j . Na sequência, são aplicadas decomposições ortogonais à matriz A . Uma dada matriz Q é ortogonal se o produto de sua matriz transposta por ela mesma resulta na matriz identidade, ou seja, se $Q^T \times Q = I$.

Aplica-se a Decomposição de Valor Singular para fazer as decomposições ortogonais. A Decomposição de Valor Singular da matriz A , para uma dada coleção com m termos e n documentos, é representada por meio do seguinte produto matricial:

$$A = U \times \Sigma \times V^T \quad (14)$$

Onde:

U = matriz ortogonal, tal que $U^T \times U = I_m$;

Σ = matriz diagonal;

V = matriz ortogonal, tal que $V^T \times V = I_n$;

Na Decomposição de Valor Singular, a dimensão da matriz A é reduzida à dimensão k . Essa redução é denominada Decomposição de Valor Singular Truncada. De acordo com Martin e Berry (2006), a redução é feita por meio de uma

matriz A_k , tal que $A_k = U_k \times \Sigma_k \times V_k^T$. Para tanto, consideram-se apenas as k primeiras colunas de U , os k primeiros valores da diagonal de Σ e as k primeiras colunas de V . Um diagrama da Decomposição de Valor Singular Truncada é dado na figura 4.2. Conforme Hofmann (2001), essa decomposição baseia-se no princípio de que há uma distribuição probabilística gaussiana associada aos dados. Segundo Martin e Berry (2006), ao reduzir a dimensionalidade da matriz A para k , obtém-se a estrutura semântica latente mais relevante para os termos e documentos da coleção inicialmente considerada. Conforme Martin e Berry (2006), no espaço vetorial de dimensão k resultante da Decomposição de Valor Singular Truncada, palavras e documentos similares ficam próximos. Wu, Potdar e Chang (2008) afirmam que a complexidade da Análise Semântica Latente é $O(N^2 k^3)$, sendo N a soma da quantidade de termos com a quantidade de documentos e k a dimensão resultante da Decomposição de Valor Singular.

Figura 4.2 – Diagrama da Decomposição de Valor Singular Truncada.

Fonte: Martin e Berry (2006)

4.2.2 Análise de Componentes Independentes

Hyvärinen e Oja (2000) ilustram a aplicação de Análise de Componentes Independentes, ou *Independent Component Analysis* (ICA) em inglês, com um problema de separação de sinais, denominado “*cocktail-party problem*”. Sejam duas pessoas, conversando em uma sala, com dois microfones que gravam as suas conversas. Modelam-se os sinais gravados por cada um dos microfones como sendo

combinações lineares dos sinais emitidos separadamente por cada pessoa. Assim, para um dado instante t , tem-se:

$$\begin{cases} x_1(t) = a_{11}s_1 + a_{12}s_2 & (15) \\ x_2(t) = a_{21}s_1 + a_{22}s_2 & (16) \end{cases}$$

Onde:

$x_1(t)$ = sinal gravado pelo microfone 1 no instante t ;

$x_2(t)$ = sinal gravado pelo microfone 2 no instante t ;

a_{ij} = parâmetros relativos às distâncias entre as pessoas e os microfones;

s_1 = sinal emitido pela pessoa 1;

s_2 = sinal emitido pela pessoa 2.

Hyvärinen e Oja (2000) definem que a Análise de Componentes Independentes corresponde a um modelo generativo, no qual dados observados são gerados por uma mistura de componentes, que correspondem a variáveis latentes, a serem estimadas. Assim sendo, o objetivo é estimar s_1 e s_2 com base na observação de $x_1(t)$ e $x_2(t)$. Assume-se que s_1 e s_2 são componentes estatisticamente independentes, o que permite separá-los das misturas $x_1(t)$ e $x_2(t)$.

Hyvärinen (2013) define a notação e os passos necessários para efetuar a Análise de Componentes Independentes. As variáveis observadas são denotadas por $x_i(t)$, para $i=1,\dots,n$ e $t=1,\dots,T$, sendo i um índice para as variáveis observadas e t é um índice temporal. Então, as variáveis observadas $x_i(t)$ são combinações lineares das variáveis latentes $s_j(t)$, $j=1,\dots,m$, com coeficientes a_{ij} , conforme as expressões (15) e (16). Em (15) e (16), $s_j(t)$ são os componentes independentes e a_{ij} são coeficientes da mistura. Assim, tem-se:

$$x_i(t) = \sum_{j=1}^m a_{ij} s_j(t) \quad (17)$$

Hyvärinen (2013) usa também uma notação matricial, pela qual tem-se:

$$\mathbf{X} = \mathbf{AS} \quad (18)$$

Onde:

X = matriz $n \times T$ para os dados observados $x_i(t)$;

A = matriz $n \times n$ de coeficientes da mistura;

S = matriz $m \times T$ para os componentes independentes $s_i(t)$.

Hyvärinen (2013) divide a Análise de Componentes Independentes em duas etapas: uma transformação linear (*whitening*) dos dados observados e a estimativa dos componentes independentes. Opera-se uma transformação linear sobre os dados **X**, com uso de uma matriz **V**, tal que **Z=VX**, sendo que para **Z** deve-se ter:

$$\frac{1}{T} \mathbf{Z} \mathbf{Z}^T = \mathbf{I} \quad (19)$$

Onde:

I= matriz de identidade.

Ao substituir (18) em **Z=VX**, tem-se:

$$\mathbf{Z} = \mathbf{VX} = \mathbf{VAS} = \tilde{\mathbf{A}}\mathbf{S} \quad (20)$$

Conforme Hyvärinen (2013), a matriz **V** é usada para a ortogonalização de **A**, gerando-se a matriz $\tilde{\mathbf{A}}$, por meio da multiplicação **VA**. Com a ortogonalização de **A**, reduzem-se os graus de liberdade do modelo, o que permite otimizações, visando a obtenção dos componentes independentes.

De acordo com Hyvärinen (2013), a estimativa de $\tilde{\mathbf{A}}$ é feita otimizando-se uma função objetiva. Há três abordagens, as quais conduzem à mesma função objetiva: estimativa da máxima verossimilhança (PHAM; GARRAT, 1997); minimização da informação mútua (COMON, 1994); uso de Redes Neurais (BELL; SEJNOWSKI, 1995).

Hyvärinen (2013) fornece a seguinte formulação para a função objetiva L:

$$L = \sum_{i=1}^n \sum_{t=1}^T \log f_i(\mathbf{w}_i^T \mathbf{z}(t)) \quad (21)$$

Onde:

f_i = função de densidade de probabilidade de s_i ;

$\mathbf{w}_i^T$ = linhas da matriz inversa de $\tilde{\mathbf{A}}$;

$\mathbf{z}(t)$ = dados observados após a transformação linear.

Conforme Hyvärinen e Oja (2000), a Análise de Componentes Independentes é aplicável a diversos problemas, tais como extração de características em imagens, identificação de sinais cerebrais em eletroencefalogramas, descoberta de fatores intrínsecos em dados financeiros, redução de dimensionalidade em textos, dentre outros.

Tang et al. (2005) e Honkela, Hyvärinen e Väyrynen (2010) compararam a Análise de Componentes Independentes com a Análise Semântica Latente, em relação à redução de dimensionalidade em documentos. Tang et al. (2005) concluíram que a aplicação de Análise de Componentes Independentes forneceu resultados de maior acurácia. O estudo de Honkela, Hyvärinen e Väyrynen (2010) indicou que a Análise de Componentes Independentes permitiu descobrir atributos mais relevantes em documentos do que a Análise Semântica Latente. O estudo aponta que o uso de ICA gera melhores resultados se a distribuição probabilística dos dados é não gaussiana. Nesse estudo, também foi feita uma análise do custo computacional da Análise de Componentes Independentes, em termos de tempo de execução, conforme aumentou-se o número de palavras dos textos e usando-se diferentes tamanhos de matrizes termo-documento. Concluiu-se que o tempo de execução do ICA foi mais sensível ao aumento do número de colunas da matriz termo-documento. Uma ilustração dessa constatação é dada na figura 4.3, em que cada tipo de linha unindo pontos do gráfico está associado a um diferente tamanho de matriz termo-documento $n \times m$ (n linhas e m colunas), conforme a legenda. Nota-se a sensibilidade maior do tempo de execução ao aumentar-se o número de colunas da matriz.

Figura 4.3 – Gráfico do custo computacional do ICA, com o eixo vertical representando o tempo de CPU em segundos e o eixo horizontal representando a quantidade de palavras.

Fonte: Honkela, Hyvärinen e Vährynen (2010)

5 CLUSTERIZAÇÃO DE GRAFOS

Neste capítulo, explicam-se alguns dos principais conceitos relacionados à Teoria dos Grafos, tais como comunidades, modularidade, topologia e detecção de comunidades. São explanadas algumas das técnicas de detecção de comunidades aplicáveis aos grafos associados às estruturas de ontologias. Trata-se de conceitos e técnicas relacionados a uma das fases da metodologia desenvolvida neste trabalho, na qual considerou-se o aspecto topológico das ontologias. Sua explanação visa fornecer alicerces para o entendimento da metodologia.

Para Schaeffer (2007), um grafo G é um par de conjuntos $G = (V, E)$, sendo V o conjunto de vértices e E o conjunto de arcos ou arestas conectando os vértices. Um subgrafo G^S de G é um par (S, E_S) , tal que $S \subseteq V$ e $E_S \subseteq E$. Segundo Schaeffer (2007), clusterização de grafos é a tarefa de agrupar vértices e arcos de um grafo em *clusters*, considerando a estrutura do grafo, de forma que haja muitos arcos em um mesmo *cluster* e relativamente poucos arcos entre *clusters*. Conforme Schaeffer (2007), o objetivo da clusterização de grafos é minimizar o número de arcos entre os *clusters* de um grafo. De acordo com Schaeffer (2007), um *cluster* em um grafo é considerado também como uma comunidade.

5.1 Comunidades

Para Porter, Onnela e Mucha (2009), uma comunidade em um grafo é um grupo de nós densamente conectados entre si, porém conectados de maneira esparsa a outros grupos densos do grafo. Para Fortunato (2010), comunidades de um grafo são agrupamentos de vértices com propriedades em comum, formando subgrafos, sendo denominadas também *clusters* ou módulos. Fortunato (2010) cita como exemplo de grafo de comunidades uma rede social, na qual se formam grupos de colegas de trabalho, de familiares ou de amigos com características comuns. Na figura 5.1, Schaeffer (2007) ilustra um exemplo de comunidades em uma empresa, que correspondem a seus departamentos, delimitados com base na estrutura de relações de trabalho entre empregados.

Fortunato (2010) introduz alguns conceitos relevantes para comunidades em grafos. O grau interno de um vértice é o número de arestas que conectam o vértice aos vértices de sua comunidade. O grau externo de um vértice é o número de arestas

Figura 5.1 – Grafo representando comunidades em uma empresa: círculos tracejados delimitam os vértices e arcos do grafo que compõem cada comunidade.
Fonte: Schaeffer (2007)

que ligam o vértice de uma comunidade com vértices externos à comunidade. Se o grau externo de uma comunidade é próximo de zero, para Fortunato (2010) a comunidade forma um bom *cluster*. Assim, uma comunidade deve ser tal que o seu grau interno seja superior ao seu grau externo. O grau externo da comunidade define o tamanho de corte da comunidade.

5.2 Modularidade de Comunidades

De acordo com Fortunato (2010), para avaliar a qualidade de uma comunidade usa-se a função de modularidade. Essa função compara a densidade de arestas em uma comunidade com a densidade de arestas de um grafo aleatório (sem estrutura de comunidade). Neste último, para qualquer par de vértices, há a mesma probabilidade de haver uma aresta conectando-os, ou seja, não há uma estrutura de comunidades, pois não existe concorrência de forças para gerar grupos. Para Fortunato (2010), usando-se o conceito de modularidade, um subgrafo forma uma comunidade se o número de suas arestas internas supera o número de arestas internas do subgrafo aleatório correspondente. Formalmente, a modularização é dada por (PORTER; ONNELA; MUCHA, 2009):

$$Q = \sum_i (e_{ii} - b_i^2) \quad (22)$$

Onde:

e_{ii} = fração das extremidades dos arcos que pertencem à comunidade i ;

$b_i = \sum_j e_{ij}$ = fração das extremidades dos arcos que pertencem à comunidade i , para as quais a outra extremidade do arco pertence à comunidade j .

Conforme Clauset, Newman e Moore (2004), valores de modularidade acima de 0,3 indicam a existência de uma estrutura característica de uma comunidade.

5.3 Topologia de Comunidades

Os trabalhos de Harary (1994), Wasserman e Faust (1994) e West (1996) descrevem alguns dos principais atributos topológicos para grafos, tais como: a excentricidade de vértices, a densidade de arcos, o diâmetro do grafo e o comprimento do caminho médio do grafo.

5.3.1 Excentricidade de vértices

De acordo com Harary (1994), a distância entre dois vértices de um grafo é igual ao menor número de arcos que precisam ser percorridos para partir de um dos vértices e chegar ao outro. Harary (1994) define a excentricidade E de um vértice v de um grafo, dada por $E(v)$, como sendo igual à máxima distância entre o vértice v e todos os outros vértices do grafo:

$$E(v) = \text{máxima } (d(u_i, v)) \quad (23)$$

Onde:

$d(u_i, v)$ = distância entre o vértice u_i e o vértice v , para $i=1, \dots, n$, sendo n o número total de vértices do grafo.

Para um subgrafo (comunidade de um grafo), pode-se definir a excentricidade média do subgrafo, dada por E_m , como sendo igual à soma das excentricidades dos vértices do subgrafo, dividida pelo número de vértices do subgrafo:

$$E_m = \frac{\sum_{i=1}^n E(v_i)}{n} \quad (24)$$

Onde:

$E(v_i)$ = excentricidade do i -ésimo vértice do subgrafo;

n = número de vértices do subgrafo.

5.3.2 Densidade de arcos

Wasserman e Faust (1994) conceituam a densidade de arcos como sendo o quociente entre a quantidade de arcos existente em um grafo e a máxima quantidade de arcos possíveis interligando os vértices do grafo. Para um subgrafo (comunidade), a densidade de arcos DA pode ser definida como o quociente entre o número de arcos do subgrafo e o número de arcos do grafo:

$$DA = \frac{NE_s}{NE_g} \quad (25)$$

Onde:

NE_s = número de arcos do subgrafo;

NE_g = número de arcos do grafo.

5.3.3 Diâmetro do grafo

Conforme Harary (1994), o diâmetro D de um grafo é a maior distância d entre os pares de vértices existentes no grafo, ou seja, D é a máxima excentricidade de um grafo. Assim, dado um grafo G , com n vértices, para um par de vértices i e j , com $i \neq j$, $i=1, \dots, n$ e $j=1, \dots, n$, tem-se:

$$D = \text{máxima } (d(i,j)) \quad (26)$$

Onde:

$d(i,j)$ = distância entre o vértice i e o vértice j .

5.3.4 Comprimento do caminho médio

De acordo com West (1996), o comprimento do caminho médio L_m de um grafo é dado pela média das distâncias entre os vértices de um grafo.

$$L_m = \frac{\sum_{i=1}^p d_i}{p} \quad (27)$$

Onde:

d_i = distância entre os vértices do i -ésimo par de vértices do grafo;

p = número de pares de vértices do grafo.

5.4 Detecção de Comunidades

Segundo Fortunato (2010), o objetivo da Detecção de Comunidades é identificar módulos e estruturas hierárquicas em grafos, com base na análise da topologia dos grafos. Várias técnicas foram criadas para detectar comunidades. Fortunato (2010) destaca o trabalho seminal de Newman e Girvan (2004), que se baseia na medida de intermediação, a qual considera os caminhos mais curtos entre os vértices. Arcos entre comunidades são detectados de acordo com essa métrica. Outras técnicas foram desenvolvidas, sendo baseadas, por exemplo, na otimização da modularidade (NEWMAN, 2004). Há uma miríade de técnicas de Detecção de Comunidades. Fortunato (2010) faz a descrição e a análise de um amplo espectro dessas técnicas.

A Detecção de Comunidades é importante para auxiliar em diversas aplicações. No caso de sistemas de recomendação (REDDY et al., 2002), por exemplo, sua aplicação permite identificar grupos de consumidores de acordo com seus principais interesses de compras, tendo como base a rede de relacionamentos entre clientes e fornecedores. A Detecção de Comunidades é aplicável também a ontologias. Coskun et al. (2011) aplicam algoritmos de Detecção de Comunidades ao

grafo associado à estrutura de uma dada ontologia para identificar agrupamentos de conceitos. Adota-se a premissa de que subdomínios em uma ontologia correspondem a subgrafos mais densamente conectados na estrutura da ontologia, ou seja, formam comunidades. De acordo com Coskun et al. (2011), com a partição obtida ao agrupar conceitos de uma ontologia, identificam-se seus subdomínios, o que facilita o entendimento do seu conteúdo.

5.5 Detecção de Comunidades em Ontologias

Coskun et al. (2011) efetuaram um estudo voltado para detecção de comunidades em grafos associados a ontologias. Inicialmente, para obter o grafo de uma ontologia, Coskun et al. (2011) usaram três abordagens diferentes:

- 1) Os sujeitos e os objetos de uma ontologia em RDF (subseção 2.1) são convertidos em nós e os predicados correspondem aos arcos;
- 2) Os predicados são considerados nós, criando-se um arco entre o sujeito e o predicado, e um arco entre o predicado e o objeto;
- 3) Uma propriedade de uma classe define um arco, o qual conecta a classe pertencente ao domínio da propriedade com a classe pertencente ao escopo da propriedade. Este tipo de propriedade é denominado propriedade de objeto. Na figura 5.2, ilustra-se um exemplo, extraído de ontologias da OAEI, no qual uma propriedade de objeto, denominada “pages”, tem como domínio a classe “Part” e como escopo a classe “PageRange”.

Tran, Wan e Haase (2009) também utilizaram a abordagem do item 3, mas definiram uma modelagem adicional, na qual propriedades de herança correspondem a arcos do grafo, ligando o nó associado a uma dada classe com o nó correspondente à subclasse dessa classe. Na figura 5.3, ilustra-se a modelagem da relação de herança, com base em uma ontologia da OAEI. Um nó associado à classe “Book” conecta-se aos nós correspondentes às suas subclasses “Collection”, “Proceedings” e “Monograph”. A conexão dá-se por meio da relação de herança “is-a”, representada um por arco.

Coskun et al. (2011) aplicaram aos grafos das ontologias cinco algoritmos de Detecção de Comunidades, comparando os resultados obtidos com cada técnica. As técnicas usadas foram: partição espectral (NEWMAN, 2006); centralidade de intermediação de arestas (NEWMAN; GIRVAN, 2004), passeios aleatórios (PONS;

Figura 5.2 – Exemplo de grafo representando a relação entre as classes “Part” e “PageRange”.

Figura 5.3 – Exemplo de grafo representando a relação entre as classes “Book”, “Collection”, “Proceedings” e “Monograph”.

LATAPY, 2005); otimização gulosa de modularidade (CLAUSET; NEWMAN; MOORE, 2004); método de Potts (REICHARDT; BORNHOLDT, 2006).

Coskun et al. (2011) avaliaram as comunidades obtidas, comparando as comunidades com uma referência (*gold standard*) definida por especialistas em ontologias ou disponível na *web*. Não houve diferenças significativas nos resultados obtidos com as diferentes abordagens de conversão das ontologias em grafos. As técnicas que geraram comunidades mais similares às de referência foram: passeios aleatórios; otimização gulosa de modularidade; método de Potts.

A seguir, descrevem-se as técnicas aplicadas no estudo de Coskun et al. (2011) que obtiveram os melhores resultados na detecção de comunidades em ontologias. Adicionalmente, descreve-se a técnica de Brandes et al. (2008).

5.5.1 Passeios Aleatórios

O conceito associado à técnica de Pons e Latapy (2005) é que, se um grafo tem uma forte estrutura de comunidades, um passeio aleatório em uma comunidade tem longa duração, por causa da alta densidade de arestas internas, que aumenta o número de caminhos. A técnica de Pons e Latapy (2005) usa a distância entre vértices baseada em passeios aleatórios. A distância é obtida a partir de probabilidades de movimentos aleatórios entre vértices. Aplica-se o algoritmo hierárquico aglomerativo (subseção 3.1) de Ward (WARD, 1963) às distâncias para gerar os agrupamentos.

Seja uma matriz de adjacência A , para um grafo G , tendo-se $A_{ij} = 1$ se os vértices i e j estão conectados e $A_{ij} = 0$, caso contrário. Então, segundo Pons e Latapy (2005), a probabilidade P_{ij} de transição do vértice i para o vértice j , sendo o grau de i dado por $d(i)$, é:

$$P_{ij} = \frac{A_{ij}}{d(i)} \quad (28)$$

Além disso, para um caminho de comprimento t , Pons e Latapy (2005) definem que a probabilidade de transitar de um vértice i para um vértice j é igual a $(P_{ij})^t$. Pons e Latapy (2005) formulam também a relação entre os graus dos vértices e as probabilidades de transição entre vértices, em um passeio aleatório de comprimento t , como sendo igual a:

$$d(i) (P_{ij})^t = d(j) (P_{ji})^t \quad (29)$$

Adicionalmente, Pons e Latapy (2005) definem que a distância r_{ij} entre os vértices i e j é igual a:

$$r_{ij} = \sqrt{\sum_{k=1}^n \frac{(P_{ik}^t - P_{jk}^t)^2}{d(k)}} \quad (30)$$

Para Pons e Latapy (2005), a probabilidade $(P_{Cj})^t$ de caminhar de uma comunidade C para um vértice j em t passos é:

$$(P_{Cj})^t = \frac{1}{|C|} \sum_{i \in C} (P_{ij})^t \quad (31)$$

Com base nas equações (29) e (30), de acordo com Pons e Latapy (2005), a distância r_{C1C2} entre as comunidades C1 e C2 é dada por:

$$r_{C1C2} = \sqrt{\sum_{k=1}^n \frac{(P_{C1k}^t - P_{C2k}^t)^2}{d(k)}} \quad (32)$$

Pons e Latapy (2005) definem também a média σ_k do quadrado das distâncias entre cada vértice e a sua comunidade por meio da fórmula:

$$\sigma_k = \frac{1}{n} \sum_{C \in P_k} \sum_{i \in C} r_{iC}^2 \quad (33)$$

Onde:

C = comunidade;

P_k = partição de um grafo em comunidades;

r_{iC}^2 = quadrado da distância entre o vértice i e sua comunidade C.

Na solução de Pons e Latapy (2005), parte-se de uma partição inicial P_1 do grafo, no qual cada vértice corresponde a uma comunidade. Calculam-se as distâncias entre vértices adjacentes e efetuam-se os seguintes passos:

- 1) Escolhem-se duas comunidades de forma a minimizar σ_k ;
- 2) Agrupam-se as duas comunidades escolhidas em uma mesma comunidade, gerando uma nova partição;
- 3) Calculam-se as distâncias entre as comunidades adjacentes da nova partição e retorna-se ao passo 1.

Assim sendo, a solução de Pons e Latapy (2005) é iterativa, repetindo-se iterativamente os passos 1 a 3, sendo cada repetição desses três passos uma iteração. Para k iterações, obtêm-se k partições. Pons e Latapy (2005) escolhem como solução final de clusterização a partição que tem a máxima modularidade, dada pela eq. (25). A complexidade no tempo dos passeios aleatórios é dada por $O(n^2 d)$, sendo n o número de nós do grafo e d a profundidade do dendograma associado às partições obtidas ao aplicar os passeios aleatórios.

5.5.2 Otimização Gulosa de Modularidade

Segundo Fortunato (2010), na otimização da modularidade (NEWMAN, 2004), as comunidades de um grafo são obtidas maximizando-se a modularidade. Essa técnica baseia-se na clusterização hierárquica aglomerativa (subseção 3.1) dos vértices.

Inicia-se com cada vértice do grafo correspondendo a uma comunidade, fazendo-se a mesclagem de vértices que maximize a modularidade. Há mudanças na modularidade, provenientes de novas mesclagens, e o objetivo é descobrir a mesclagem que dê o maior acréscimo de modularidade. O processo é iterativo e repete-se até que se obtenha uma única comunidade. Newman (2004) baseia-se na matriz de adjacência do grafo, efetuando mesclagens repetidamente.

Clauset, Newman e Moore (2004) consideram essa solução computacionalmente custosa e propõem uma otimização gulosa da modularidade. A premissa usada é que cada comunidade corresponda a um vértice, formando-se um multigrafo, no qual um grupo de arcos conecta um vértice ao outro. Clauset, Newman e Moore (2004) utilizam uma matriz ΔQ , na qual cada entrada ΔQ_{ij} é o acréscimo de modularidade obtido na mesclagem das comunidades i e j . Essa matriz considera apenas os pares de comunidades conectados por uma ou mais arestas.

O algoritmo de Clauset, Newman e Moore (2004) tem os seguintes passos:

- 1) Cálculo dos valores iniciais de ΔQ_{ij} e armazenamento do máximo valor de cada linha de ΔQ em uma árvore hierárquica H ;
- 2) Seleção do maior valor de ΔQ_{ij} em H ;
- 3) Mesclagem das comunidades i e j correspondentes ao maior valor de ΔQ_{ij} ;
- 4) Atualização de ΔQ e H ;
- 5) Retorno ao passo 2 até obter uma comunidade.

A complexidade no tempo para a otimização gulosa de modularidade é dada por $O(m d \log n)$, sendo m a quantidade de arcos, d é a profundidade do dendograma das sucessivas partições obtidas e n é a quantidade de nós do grafo (FORTUNATO, 2010).

5.5.3 Método de Potts

Este método, criado por Reichardt e Bornholdt (2006), baseia-se no modelo de Potts, que originalmente é usado para o estudo do comportamento de materiais magnéticos. No modelo, partículas com momentos magnéticos são denominadas *spins*, as quais podem assumir vários estados.

No modelo de Potts, a energia H entre dois *spins* i e j é modelada de acordo com a equação:

$$H = - \sum_{i < j} J_{ij} \delta(s_i, s_j) \quad (34)$$

Onde:

J_{ij} = força de interação ou ângulo entre os *spins* i e j ;

s_i = estado do *spin* i ;

s_j = estado do *spin* j ;

$$\delta(s_i, s_j) = \text{delta de Kronecker} = \begin{cases} 1, & \text{se } s_i = s_j \\ 0, & \text{se } s_i \neq s_j \end{cases}$$

Para aplicar o modelo de Potts a redes de comunidades, faz-se com que cada nó da rede corresponda a um *spin* e cada comunidade corresponda a um estado. Dois nós i e j pertencem a uma mesma comunidade se os *spins* a que correspondem têm o mesmo estado. Isto implica que eles tenham entre si a mesma energia de interação H , dada por $-J_{ij}$, pois s_i é igual a s_j (mesmo estado). Se os *spins* i e j têm estados diferentes, não têm interação entre si e os seus respectivos vértices pertencem a comunidades diferentes. Segundo Porter, Onnela e Mucha (2009), o método de Potts baseia-se na premissa de que a otimização da modularidade de uma rede de comunidades equivale a minimização da energia de um modelo de Potts. Essa minimização corresponde ao *ground state* do modelo de Potts, ou seja, quando se atinge a temperatura igual a zero. Segundo Fortunato (2010), em grafos esparsos, o modelo de Potts tem complexidade no tempo dada por $O(n^{2+\Theta})$, sendo n o número de nós do grafo e $\Theta \sim 1,2$.

5.5.4 Técnica de Brandes

A técnica de Brandes et al. (2008) encontra comunidades em grafos maximizando a modularidade por meio de Programação Linear Inteira (PLI) (PAPADIMITRIOU; STEIGLITZ, 1998). São definidas variáveis de decisão $X_{uv} \in \{0,1\}$, para todos os pares de nós (u,v) de um dado grafo. As variáveis são interpretadas como uma relação de equivalência entre os nós do grafo, representando, assim, uma clusterização. Para essas variáveis, impõem-se as propriedades de reflexividade, simetria e transitividade, expressas respectivamente por (35), (36) e (37):

$$\forall u: X_{uu} = 1 \quad (35)$$

$$\forall u,v: X_{uv} = X_{vu} \quad (36)$$

$$\forall u,v,w: \begin{cases} X_{uv} + X_{vw} - 2 X_{uw} \leq 1 \\ X_{uw} + X_{uv} - 2 X_{vw} \leq 1 \\ X_{vw} + X_{uw} - 2 X_{uv} \leq 1 \end{cases} \quad (37)$$

Nessa técnica, a modularidade Q é calculada por:

$$Q = \frac{1}{2m} \sum_{(u,v)} \left(\epsilon_{uv} - \frac{\deg(u)\deg(v)}{2m} \right) X_{uv} \quad (38)$$

Onde:

m = quantidade de arcos do grafo;

(u,v) = par de nós (u,v) do grafo;

$$\epsilon_{uv} = \begin{cases} 1, & \text{se } (u,v) \in \text{conjunto de arcos do grafo} \\ 0, & \text{caso contrário} \end{cases}$$

$\text{deg}(i)$ = grau do nó i .

Conforme Brandes et al. (2008), algumas técnicas que otimizam a modularidade (otimização gulosa, por exemplo), de início clusterizam, em um mesmo grupo, nós de diferentes graus (um dos nós com baixo grau e o outro com alto grau), objetivando, com isso, equilibrar a soma dos quadrados dos graus. Ao final da geração de agrupamentos, esses nós de diferentes graus ficam em agrupamentos diferentes. Isto implica na formação de pequenos *clusters* com distribuições de grau não simétricas, impossibilitando obter o resultado que otimiza a modularidade. O estudo de Brandes et al. (2008) constatou que a técnica de Brandes conseguiu superar essa deficiência. Brandes et al. (2008) comprovou também em seu estudo que a otimização de modularidade de sua solução é um problema NP-completo, ou seja, não há como resolver o problema em tempo polinomial. Problemas NP-completos são resolvidos em tempo exponencial (LEVIN, 2016).

6 CLUSTER ENSEMBLES

Neste capítulo, são explicados os principais conceitos e técnicas de *Cluster Ensembles*, uma área do conhecimento voltada para encontrar combinações de diferentes soluções de clusterização, as quais correspondem a um consenso entre essas clusterizações, consenso esse que reúne, em uma única solução, características diversas, que podem estar relacionadas às diferentes técnicas aplicadas ou a diferentes aspectos dos dados que foram considerados. Alguns dos conceitos explicados são: clusterizações de base, vetor de rótulos dos *clusters*, clusterização consensual e matriz de co-associação. As técnicas explanadas são: baseadas em similaridades, baseadas em métodos diretos e probabilística. Este capítulo abarcou as definições e técnicas que foram usadas em um dos passos da metodologia ora desenvolvida, visando justamente fornecer um introito que propicie embasamento teórico previamente à apresentação da metodologia.

Ghosh e Acharya (2014) definem que *Cluster Ensembles* é uma técnica usada para combinar múltiplas soluções de clusterização individuais (clusterizações de base), gerando uma solução única e consolidada, denominada solução consensual. Conforme Ghosh e Acharya (2014), essa solução consensual potencialmente resulta em uma clusterização melhor do que a obtida pela aplicação isolada de algoritmos de clusterização. Ghosh e Acharya (2014) afirmam que *Cluster Ensembles* podem ser usados na solução de diversos problemas de clusterização e destacam algumas aplicações práticas em Segmentação de Imagens, Bioinformática e Recuperação de Informação.

Ghosh e Acharya (2014) descrevem algumas contribuições que o uso de *Cluster Ensembles* pode trazer. Dentre elas, está a qualidade da solução de clusterização, uma vez que *Cluster Ensembles* pode aumentar a acurácia dos resultados, comparativamente às soluções de base, tomadas individualmente. *Cluster Ensembles*, segundo Ghosh e Acharya (2014), podem prover também soluções mais robustas, pois, ao combinar algoritmos que se baseiam em diferentes modelos, que assumem diferentes premissas, tende-se a gerar bons resultados de clusterização para conjuntos de dados com características diversas. De acordo com Ghosh e Acharya (2014), *Cluster Ensembles* podem ser aplicados também com o intuito de combinar múltiplos aspectos ou visões dos dados, considerando-se diferentes atributos em cada aspecto ou visão e combinando-os com *Cluster Ensembles*.

Ghosh e Acharya (2014) apresentam algumas das notações usuais para soluções de *Cluster Ensembles*. Considerem-se r clusterizações de base, para um conjunto de dados $X = \{x_i\}$, $i=1, \dots, n$. A q -ésima clusterização de base contém $k^{(q)}$ *clusters*, com $q=1, \dots, r$. O conjunto de *clusters* da q -ésima clusterização é representado por um vetor de *clusters*, dado por $\lambda^{(q)} = \{C_L \mid L=1, 2, \dots, k^{(q)}\}$, também denominado vetor de rótulos dos *clusters*. Convenciona-se também uma função de consenso Γ , que mapeia um conjunto de clusterizações de base a uma clusterização consolidada e consensual, de forma que $\Gamma: \lambda^{(q)} \rightarrow \hat{\lambda}$. O conjunto de clusterizações de base também é usualmente denotado por Λ . Convenciona-se também uma matriz de co-associação para as clusterizações de base. Seja a q -ésima clusterização de base e os elementos i e j do conjunto de dados, a matriz de co-associação $S^{(q)}$ é dada por:

$$S_{ij}^{(q)} = \begin{cases} 1, & \text{se } (i,j) \in C_L \text{ (} \lambda^{(q)} \text{)} \\ 0, & \text{caso contrário} \end{cases} \quad (39)$$

Ghosh e Acharya (2014) subdividem os tipos de algoritmos de *Cluster Ensembles* conforme três abordagens: baseada em similaridades; baseada em métodos diretos; probabilística.

6.1 *Cluster Ensembles* Baseados em Similaridades

Segundo Ghosh e Acharya (2014), nesta abordagem, calcula-se a média ponderada das matrizes de co-associação das r clusterizações de base, com $q=1, \dots, r$. Essa média resulta na matriz consensual S , em que w_q é o peso atribuído à q -ésima clusterização de base:

$$S = \frac{1}{r} \sum_{q=1}^r w_q S^{(q)} \quad (40)$$

A matriz S corresponde à fração de vezes em que um par de elementos do conjunto de dados fica agrupado em um mesmo *cluster*, considerando as r clusterizações de base.

Um exemplo de algoritmo que aplica esta abordagem é o *Cluster-based Similarity Partitioning Algorithm* (CSPA), desenvolvido por Strehl e Ghosh (2003). Usa-se o conceito de matriz de co-associação entre elementos a serem clusterizados. Dados r algoritmos de clusterização e n elementos, obtêm-se r matrizes de co-associação, $n \times n$. Os valores das entradas de cada matriz são iguais a 1 caso o elemento correspondente à linha seja similar ao elemento correspondente à coluna e igual a zero, caso contrário. Strehl e Ghosh (2003) calculam a média das matrizes e, com base na matriz média, reclusterizam os objetos usando o algoritmo de METIS (KARYPIS; KUMAR, 1998). Conforme Strehl e Ghosh (2003), apesar do CSPA ser simples, ele tem complexidade computacional quadrática em n , sendo expressa por $O(n^2kr)$.

6.2 *Cluster Ensembles* Baseados em Métodos Diretos

Nesta abordagem, de acordo com Ghosh e Acharya (2014), consideram-se apenas os rótulos resultantes das clusterizações de base, otimizando-se uma função objetiva, sem considerar uma matriz de co-associação. Strehl e Ghosh (2003) desenvolveram dois algoritmos que se enquadram na abordagem por métodos diretos: o *HyperGraph Partitioning Algorithm* (HGPA) e o *Meta-CLustering Algorithm* (MCLA).

O HGPA baseia-se no conceito de hipergrafos. Um hipergrafo é um grafo no qual os seus arcos, denominados hiperarcos, podem conectar qualquer conjunto de vértices. Dado um conjunto de clusterizações, mapeia-se cada *cluster* a um hiperarco e o conjunto de *clusters* é mapeado a um hipergrafo.

No HGPA, particiona-se o hipergrafo de forma a cortar o mínimo de hiperarcos. O objetivo é particionar o hipergrafo em k subgrafos de aproximadamente mesmo tamanho. Para Strehl e Ghosh (2003), esta condição torna o algoritmo inadequado para casos em que os *clusters* não são naturalmente balanceados. Strehl e Ghosh (2003) usam o algoritmo HMETIS (KARYPIS et al., 1997) para particionar o hipergrafo. A complexidade computacional do HGPA é dada por $O(nkr)$.

No MCLA, fazem-se cortes em hiperarcos do hipergrafo, definindo-se *clusters*. Atribui-se um dado elemento do hipergrafo a um *cluster* conforme a força de seu relacionamento com o hiperarco que foi cortado. Usa-se a métrica de Jaccard (JACCARD, 1901) para medir similaridades e a clusterização é feita com o algoritmo

METIS (KARYPIS; KUMAR, 1998). A complexidade computacional do MCLA é dada por $O(nk^2r^2)$.

6.3 *Cluster Ensembles Bayesianos*

Dentre os algoritmos que usam a abordagem probabilística, destacam-se os CEB. Conforme Wang, Shan e Banerjee (2011), dado um conjunto de dados $X = \{x_1, \dots, x_n\}$ que se almeja clusterizar, CEB assumem a existência de k *clusters* consensuais, sendo que cada ponto de dados x_i ($i=1, \dots, n$) tem um grau de pertinência em relação aos k *clusters*, ou seja, cada ponto tem um grau de participação diferente em múltiplos *clusters* consensuais (*soft clustering*). Assim, cada ponto de dados participa simultaneamente em todos os *clusters* consensuais, em diferentes proporções, dadas por probabilidades. Pressupõe-se um processo generativo probabilístico, no qual as clusterizações de base são geradas a partir dos k *clusters* consensuais.

O processo generativo é ilustrado por meio do exemplo da figura 6.1. Há duas matrizes: a matriz B, na parte superior da figura 6.1, representando as clusterizações de base e a matriz C, na parte inferior, referente ao *clusters* consensuais 1 e 2. Na matriz B, há sete linhas, correspondentes a sete pontos de dados x_i ($i=1, \dots, 7$). Há três resultados de clusterização de base λ_i ($i=1, \dots, 3$), representados pelas colunas de B, sendo que as entradas da matriz B são os rótulos das clusterizações de base. No processo generativo, os rótulos das clusterizações de base (matriz B) são obtidos a partir de amostragens probabilísticas, obtidas das clusterizações consensuais 1 e 2 (matriz C). Os rótulos das clusterizações de base seguem distribuições probabilísticas discretas (matriz C). No exemplo da figura 6.1, para determinar qual *cluster* consensual gerou o rótulo para o ponto de dado x_1 da clusterização de base λ_1 (linha 1 e coluna 1 da matriz B), devem-se analisar as distribuições probabilísticas discretas da matriz C associadas à clusterização de base λ_1 (coluna 1 da matriz C). Dessa análise, verifica-se que a maior probabilidade é de 0,7, proveniente da distribuição discreta do *cluster* consensual 2 (linha 2 de C) e associada ao rótulo 3 da clusterização de base. Assim, no processo generativo probabilístico dos CEB do exemplo, o rótulo 3 de x_1 para λ_1 foi gerado pelo *cluster* consensual 2.

Segundo Wang, Shan e Banerjee (2011), o objetivo dos CEB é combinar as clusterizações de base dos elementos do conjunto de dados, gerando uma

clusterização de consenso. Para tanto, faz-se o processo generativo de forma inversa, isto é, dadas as clusterizações de base observadas, inferem-se os graus de pertinência dos pontos de dados. Assim, no processo de inferência, objetiva-se inferir as variáveis latentes, denotadas por θ e z (figura 6.1). A variável θ é representada pelos histogramas à esquerda da figura 6.1, correspondendo ao grau de pertinência do ponto x_i em relação aos *clusters* consensuais. A variável z corresponde ao rótulo consensual atribuído ao ponto x_i após a inferência. α e β são parâmetros probabilísticos do modelo. A inferência, feita diretamente, é computacionalmente intratável, sendo que, na prática, recorre-se a aproximações, tais como Inferência Variacional (JORDAN et al., 1999) ou Amostragem de Gibbs (CASELLA; GEORGE, 1992).

Figura 6.1 – Processo generativo para *Cluster Ensembles* Bayesianos.
Fonte: adaptado de Banerjee (2012)

O modelo gráfico probabilístico (AIROLDI, 2007) para CEB é ilustrado na figura 6.2. Em um Modelo Gráfico Probabilístico, as variáveis observadas são representadas por nós sombreados e as variáveis não observadas (latentes), as quais se deseja inferir, são representadas por nós não sombreados. Os arcos indicam dependências condicionais entre as variáveis e os retângulos representam quantidades relacionadas às clusterizações e aos dados, indicadas nos cantos inferiores dos retângulos. No caso de CEB, θ e z são variáveis latentes, x representa as clusterizações de base observadas para os pontos de dados x_i , sendo α e β os parâmetros do modelo. M , N e k são, respectivamente, a quantidade de algoritmos que gerou as clusterizações de base, a quantidade de pontos do conjunto de dados e a quantidade de *clusters* consensuais.

Figura 6.2 – Modelo gráfico probabilístico para *Cluster Ensembles* Bayesianos.

Fonte: Wang, Shan e Banerjee (2011)

Wang, Shan e Banerjee (2011) efetuaram um experimento, no qual compararam a clusterização obtida por meio de CEB com a clusterização obtida com MCLA, CSPA e HGPA. Compararam-se também os resultados de CEB com resultados obtidos por algoritmos de clusterização aplicados isoladamente (*K-means*, *fuzzy c-means*, METIS e propagação de afinidade). Comparou-se a precisão da clusterização obtida, com base em uma clusterização de referência, medindo-se a quantidade de pontos de dados que foram corretamente atribuídos a um *cluster*. Wang, Shan e Banerjee (2011) concluíram que CEB superou os demais algoritmos na maioria dos casos. Além disso, no citado experimento, foi feita a análise do custo computacional dos CEB: ao aumentar o número de clusterizações de base, verificou-se o tempo necessário para executar os CEB. O gráfico obtido foi o indicado na figura 6.3, na qual se observa um crescimento exponencial.

Figura 6.3 – Gráfico do tempo de execução dos CEB em função da quantidade de clusterizações de base.

Fonte: Adaptado de Wang, Shan e Banerjee (2011)

7 REVISÃO DA LITERATURA

Neste capítulo, são descritos e analisados os principais trabalhos relacionados ao emparelhamento de ontologias baseado em clusterização. A pesquisa e a revisão bibliográfica iniciaram-se em 2013 e foram selecionados os trabalhos mais recentes. As principais palavras-chave usadas na pesquisa foram: “*matching*”, “*alignment*”, “*clustering*”, “*cluster*”, “*ontology*”, “*partition*”, “*partitioning*” etc. As principais fontes de pesquisa utilizadas foram as seguintes bases: *Association for Computing Machinery* (ACM), *Institute of Electrical and Electronics Engineers* (IEEE), Google Acadêmico, Google, Scopus, *Microsoft Academic Search*, ResearcherID e *Web of Science*.

Verificou-se quais as técnicas utilizadas em cada fase das metodologias adotadas nos principais trabalhos, com o intuito de identificar deficiências e lacunas no estado da arte. Em resumo, observou-se que os trabalhos relacionados se restringiram ao aspecto terminológico para fazer a clusterização, não tratando de outros aspectos das ontologias, além de não aplicarem técnicas de Detecção de Comunidades. Também exploraram pouco as técnicas de *Cluster Ensembles*.

7.1 A Clustering-Based Approach for Large-scale Ontology Matching (ALGERGAWY; MASSMANN; RAHM, 2011)

Algergawy, Massmann e Rahm (2011) desenvolveram uma solução para emparelhamento de duas ontologias baseada em clusterização. A metodologia tem três fases: obtenção de subgrafos dos grafos correspondentes às estruturas das ontologias; clusterização dos subgrafos; emparelhamento entre elementos de subgrafos semelhantes.

Inicialmente, calcula-se, para cada ontologia, a similaridade estrutural entre os conceitos de uma mesma ontologia. Algergawy, Massmann e Rahm (2011) obtêm a similaridade com base na quantidade de conceitos ancestrais, descendentes e equivalentes que os conceitos têm em comum. A partir dessas similaridades, aplica-se, para cada ontologia, um algoritmo de clusterização hierárquica aglomerativa (subseção 3.1), denominado AHSCAN (YURUK et al., 2009). Usa-se uma abordagem *bottom-up*, considerando-se, inicialmente, que cada classe é um *cluster*. As classes são mescladas de forma iterativa, gerando uma estrutura de *clusters*, a qual permite obter a clusterização conforme uma determinada similaridade selecionada. A

clusterização assim obtida resulta, para cada ontologia, em uma partição do grafo associado à ontologia.

Na sequência, Algergawy, Massmann e Rahm (2011) usam o Modelo do Espaço Vetorial (subseção 4.1) para clusterizar os subgrafos. Extraem-se os nomes de classes e propriedades e os textos de anotações de cada subgrafo, considerando cada subgrafo como um documento. Aplica-se o TF-IDF e as similaridades entre subgrafos são calculadas por meio da similaridade de cossenos (subseção 4.1). Com base nos resultados, constrói-se uma matriz de similaridades dos pares de subgrafos das duas ontologias em comparação. O passo seguinte é fazer emparelhamento dos elementos dos pares de subgrafos mais similares. Foi usada a ferramenta COMA++ (DO; RAHM, 2007) na fase de emparelhamento.

7.2 VDoc+: a Virtual Document Based Approach for Matching Large Ontologies using MapReduce (ZHANG; HU; QU, 2012)

No trabalho de Zhang, Hu e Qu (2012) desenvolveu-se uma solução de emparelhamento de ontologias usando o conceito de documentos virtuais (QU; HU; CHENG, 2006) e o Modelo do Espaço Vetorial (subseção 4.1).

Um documento virtual obtém-se extraindo os termos das descrições de classes de uma ontologia. A descrição de uma classe é obtida por Zhang, Hu e Qu (2012) a partir da união do nome local de seu *Uniform Resource Identifier* (URI), seus comentários e seu nome. O documento virtual de Zhang, Hu e Qu (2012) é composto de pesos associados à relevância das palavras no documento, sendo os pesos obtidos por meio do TF-IDF (subseção 4.1).

Zhang, Hu e Qu (2012) subdividem o emparelhamento em quatro etapas: obter descrições de classes; trocar informações entre classes vizinhas para gerar documentos virtuais; usar o Modelo do Espaço Vetorial com o TF-IDF (subseção 4.1) para agrupar documentos virtuais; fazer emparelhamento entre elementos pertencentes a um mesmo agrupamento. A ferramenta MapReduce (DEAN; GHEMAWAT, 2008) é usada para executar as etapas da solução.

Na figura 7.1, tem-se uma visão do processo usado na solução. Inicialmente, são atribuídos identificadores únicos para cada documento virtual. A função Map do MapReduce cria pares de chave-valor: a chave corresponde a uma palavra e o valor corresponde ao documento virtual em que a palavra ocorre. Em seguida, agrupam-

Figura 7.1 – Processo para o emparelhamento de ontologias baseado em clusterização usando MapReduce.

Fonte: adaptado de Zhang, Hu e Qu (2012)

se os documentos virtuais conforme a chave, ordenando-se as chaves de acordo com a relevância da palavra a que correspondem. Essa relevância baseia-se no TF-IDF. Para cada um dos agrupamentos, a função Reduce do MapReduce calcula, de forma paralela, as similaridades entre documentos virtuais de um mesmo agrupamento com base na similaridade de cossenos (subseção 4.1). Como cada documento virtual está associado a uma classe da ontologia, as similaridades entre os documentos virtuais são consideradas as similaridades entre as classes correspondentes. Assim sendo, com base nas similaridades entre documentos, obtidas pela similaridade do cosseno, obtém-se o alinhamento entre as classes.

7.3 Improving Clustering-Based Schema Matching Using Latent Semantic Indexing (ALGERGAWY et al., 2015)

Algergawy et al. (2015) desenvolveram uma solução para emparelhamento de esquemas baseada em clusterização, aplicável também a ontologias. Gera-se inicialmente uma partição dos grafos associados aos esquemas, subdividindo-os em subgrafos. Em seguida, é feita a clusterização dos subgrafos com base na terminologia usada para os nomes de elementos dos subgrafos. A clusterização faz-se com auxílio de Análise Semântica Latente (subseção 4.2.2).

Na partição dos esquemas, modelam-se os esquemas como grafos. Para esta fase, usam-se as ideias de Algergawy, Massmann e Rahm (2011), cujo trabalho é descrito em 7.1. Para cada esquema, gera-se uma partição, formada por um conjunto de subgrafos.

Para a clusterização, Algergawy et al. (2015) consideram cada subgrafo como um documento. Para o conjunto de subgrafos de um dos esquemas, inicialmente faz-se um pré-processamento dos termos associados a elementos dos esquemas, aplicando-se *tokenization* e remoção de *stop words* (capítulo 4). Em seguida, constrói-se uma matriz termo-documento, na qual cada entrada da matriz é igual ao número de vezes que um termo ocorre em cada subgrafo. Aplica-se uma função de entropia (MITCHELL, 1997) à matriz e, na sequência, faz-se a Análise Semântica Latente (subseção 4.2.2) da matriz resultante. Repetem-se esses procedimentos para o conjunto de subgrafos do outro esquema. Na sequência, é feita a comparação entre subgrafos dos dois esquemas, aplicando-se a similaridade do cosseno (subseção 4.1). Os pares de subgrafos dos dois esquemas com similaridade do cosseno entre si maior do que um dado limite são agrupados em um mesmo *cluster*. Para fazer o emparelhamento, aplicam-se técnicas de similaridade sintática (subseção 2.4.1) aos elementos contidos no mesmo *cluster*.

7.4 Efficient Multidimensional Blocking for Link Discovery without Losing Recall (ISELE; JENTZSCH; BIZER, 2011)

Isele, Jentzsch e Bizer (2011) propõem uma solução para emparelhamento de ontologias do LOD baseada na clusterização de entidades das ontologias. Isele, Jentzsch e Bizer (2011) geram blocos de entidades semelhantes das ontologias, visando a redução do número de comparações entre entidades, ao comparar apenas os elementos de blocos semelhantes. O algoritmo desenvolvido por Isele, Jentzsch e Bizer (2011), denominado MultiBlock, agrupa entidades em blocos considerando métricas de similaridade de caracteres, de dados numéricos e de dados geográficos. O MultiBlock desdobra-se em três fases: geração de índices; agregação de índices; geração de pares de entidades a serem comparados.

Para a geração de índices, inicialmente calcula-se cada uma das similaridades consideradas pelo MultiBlock, para todos os pares de entidades das ontologias do LOD. Cada similaridade corresponde a um índice. Para cada índice, há uma função para geração de blocos, que atribui entidades a um mesmo bloco com base na similaridade entre as entidades. Na fase seguinte, gera-se um índice composto, que pode ser obtido por meio da média ponderada das similaridades obtidas na geração

de índices. Alternativamente, pode ser obtido o índice composto calculando-se os valores mínimo ou máximo das similaridades.

Na sequência, pares que pertencem a um mesmo bloco são selecionados para serem comparados. É feito o cálculo de múltiplas similaridades entre os pares. Os pares são considerados correspondentes se o resultado combinado das similaridades for igual ou maior do que um limite de similaridade pré-determinado.

7.5 Alignment-Based Partitioning of Large-scale Ontologies (HAMDI et al., 2010)

Hamdi et al. (2010) propõem uma solução para emparelhamento de duas ontologias baseado em clusterização. Inicialmente, identificam-se pares de conceitos com rótulos idênticos, em duas ontologias diferentes. Essa identificação é feita com uso de um algoritmo de emparelhamento para similaridade exata de caracteres. Os pares alinhados são denominados âncoras. Encontram-se subconjuntos de elementos das ontologias (subontologias ou blocos) semelhantes entre si. O alinhamento é feito somente entre os subconjuntos similares.

O trabalho de Hamdi et al. (2010) inspira-se no trabalho de Hu, Qu e Cheng (2008). Calculam-se pesos para relacionamentos entre conceitos de ontologias, obtidos com base na similaridade de Wu e Palmer (1994). Essa similaridade considera o número de relacionamentos existentes entre o conceito raiz da ontologia e determinado conceito em análise. Hamdi et al. (2010) baseiam-se também em conceitos como coesão e acoplamento. A coesão de Hamdi et al. (2010) é proporcional aos pesos dos relacionamentos de uma mesma subontologia e o acoplamento é proporcional aos pesos dos relacionamentos que conectam conceitos de diferentes subontologias. Inicialmente, o algoritmo considera um conceito como sendo um bloco e impõe-se a quantidade de blocos a ser gerada ou a quantidade de conceitos máxima de cada bloco. O algoritmo é iterativo, sendo que, em cada iteração, escolhem-se os blocos de maior coesão e de maior acoplamento, que são mesclados. Na iteração seguinte, recalculam-se a coesão e o acoplamento para os blocos mesclados e o processo repete-se até que seja satisfeita uma das seguintes condições: atingir-se a quantidade de blocos inicialmente estipulada ou a quantidade máxima de conceitos por bloco ou não haver mais blocos com coesão maior do que zero.

Concluída obtenção de subontologias das duas ontologias em comparação, inicia-se a obtenção de pares de subontologias semelhantes, que serão submetidos ao processo de alinhamento. A obtenção de pares de blocos similares baseia-se na análise das âncoras que unem os blocos. Quanto maior é a quantidade de âncoras em comum entre um par de blocos, maior é a sua similaridade. O emparelhamento é feito considerando-se pares correspondentes aqueles que tenham a similaridade maior do que um dado valor mínimo. Para o emparelhamento, são aplicadas técnicas sintáticas e semânticas (subseção 2.4.1).

7.6 Dimensional Clustering of Linked Data: Techniques and Applications (FERRARA et al., 2015)

Ferrara et al. (2015) fazem o emparelhamento entre elementos de ontologias do LOD que se refiram a um mesmo tema. Cada tema é considerado uma dimensão. Com base no resultado de emparelhamento, clusterizam-se elementos das ontologias segundo cada uma das dimensões temáticas. Ao final, aplicam-se técnicas de *Cluster Ensembles* (capítulo 6) para combinar as clusterizações de cada dimensão. A metodologia de Ferrara et al. (2015) subdivide-se nas seguintes etapas: especificação de dimensões temáticas; emparelhamento; clusterização.

Ferrara et al. (2015) usam, em seu experimento, uma ontologia da DBpedia sobre Ciência da Computação, existente no LOD. Os diferentes temas, que correspondem às dimensões temáticas, são previamente definidos. No sistema correspondente à solução de Ferrara et al. (2015), o usuário define os temas ou utilizam-se temas pré-definidos. A especificação das dimensões temáticas consiste, segundo Ferrara et al. (2015), na escolha de diferentes propriedades, dentre os elementos das ontologias, relacionadas a diferentes dimensões. No caso do experimento de Ferrara et al. (2015), as propriedades são as especialidades de cientistas, suas filiações e dados de suas nacionalidades.

Após a especificação das dimensões, é feito o emparelhamento apenas entre propriedades pertencentes a uma mesma dimensão temática. No emparelhamento, aplicam-se técnicas de similaridade de caracteres (subseção 2.4.1).

Com base nas similaridades obtidas na fase de emparelhamento, Ferrara et al. (2015) obtêm, para cada dimensão temática, uma matriz de similaridades M . Cada entrada $M(i,j)$ da matriz é igual à similaridade entre os elementos i e j da ontologia,

conforme uma dada dimensão temática. Ferrara et al. (2015) aplicam às matrizes de cada dimensão um algoritmo de clusterização hierárquica (subseção 3.1), obtendo-se, assim, uma clusterização para cada dimensão temática. No caso do experimento de Ferrara et al. (2015), há duas dimensões temáticas, gerando-se, portanto, duas clusterizações: uma associada ao tema profissional e outra relacionada ao tema geográfico. Ou seja, obtiveram-se dois conjuntos de *clusters*: um conjunto de *clusters* de cientistas conforme suas filiações e premiações acadêmicas e um conjunto de *clusters* conforme seus locais de nascimento.

Em seguida, aplicam-se técnicas de *Clusters Ensembles* com o objetivo de combinar os *clusters* dos diferentes temas. Ferrara et al. (2015) fazem a combinação impondo, por exemplo, que apenas elementos comuns entre dois *clusters* temáticos diferentes façam parte de um *cluster* final, ou seja, consideram a interseção ao combinar *clusters* temáticos. Ferrara et al. (2015) usam também critérios de combinação baseados em informação mútua (COVER; THOMAS, 1991) entre *clusters*, fazendo com que sejam combinados os *clusters* que compartilhem mesmas informações.

7.7 Análise

Nos trabalhos de Algergawy, Massmann e Rahm (2011), Algergawy et al. (2015), Hamdi et al. (2010), Isele, Jentzsch e Bizer (2011) e Zhang, Hu e Qu (2012), algumas lacunas foram observadas. Não são exploradas técnicas de Detecção de Comunidades (subseção 5.4) para particionar o grafo associado às estruturas das ontologias. Essas técnicas descobrem agrupamentos em grafos com base na densidade de arcos, delimitando comunidades, e caracterizam-se pela sua escalabilidade e qualidade de clusterização ao serem aplicados a redes de grafos grandes e complexas (FORTUNATO, 2010). Esses trabalhos também não combinam resultados de diferentes algoritmos de clusterização. Isto pode ser feito com uso de *Cluster Ensembles* (capítulo 6), tornando a solução mais robusta, com potencial de fornecer resultados mais precisos do que os obtidos com a aplicação de uma técnica de clusterização em isolado (GHOSH; ACHARYA, 2014).

Ferrara et al. (2015) usam técnicas de *Cluster Ensembles* para combinar *clusters* de diferentes tópicos das ontologias. Combinam-se *clusters* impondo-se que apenas elementos comuns entre dois *clusters* temáticos façam parte de um *cluster*

final. Técnicas mais elaboradas podem ser exploradas, fornecendo melhores clusterizações. Wang, Shan e Banerjee (2011), por exemplo, verificaram que a aplicação de CEB (subseção 6.3) resultou em uma clusterização de maior acurácia do que a obtida por outras técnicas de *Cluster Ensembles*. Além disso, a solução de Ferrara et al. (2015) tem outras limitações. Depende-se do prévio conhecimento dos temas de uma ontologia, sem fazer uso de técnicas de Clusterização de Documentos (capítulo 4) para a descoberta automática desses temas. Também não são tratados os aspectos estrutural e extensional das ontologias.

Algergawy, Massmann e Rahm (2011), Algergawy et al. (2015) e Zhang, Hu e Qu (2012) aplicam técnicas de clusterização que consideram apenas o aspecto terminológico das ontologias, sem ponderar outros aspectos, tais como o estrutural ou o extensional. Esses aspectos podem ser considerados em diferentes clusterizações, que podem ser combinadas, gerando uma clusterização consensual potencialmente mais precisa do que as individuais.

Algergawy et al. (2015) aplicam Análise Semântica Latente (subseção 4.2.2) para efetuar a redução de dimensionalidade. Outras técnicas podem ser utilizadas, com potencial para gerar melhores resultados. No trabalho de Honkela, Hyvärinen e Väyrynen (2010), por exemplo, constatou-se que o uso da Análise de Componentes Independentes (subseção 4.2.3) para redução de dimensionalidade permitiu a descoberta de termos mais relevantes em coleções de documentos do que a Análise Semântica Latente. Este resultado pode ser estendido à terminologia usada em ontologias, considerando diferentes subontologias pertencentes a uma ontologia como diferentes documentos, com a possibilidade da Análise de Componentes Independentes revelar termos relevantes nas subontologias, inclusive melhorando a qualidade da clusterização.

Isele, Jentzsch e Bizer (2011) aplicam uma função de geração de *clusters* de elementos de ontologias, que se baseia em um emparelhamento prévio entre todos os pares de elementos. Assim, necessita-se que seja feita a comparação do produto cartesiano dos elementos, antes de gerar os *clusters*. Esse emparelhamento prévio à clusterização precisa ser feito também nas soluções de Hamdi et al. (2010) e Ferrara et al. (2015) e reduz a eficiência do processo de emparelhamento, pois implica em um maior número de comparações entre os elementos das ontologias.

Na tabela 7.1 apresenta-se uma visão geral das técnicas utilizadas nos trabalhos relacionados e indicação, por meio de números, das deficiências do estado da arte e respectivas técnicas e aspectos com potencial para suplantar essas lacunas.

Tabela 7.1 – Visão geral das técnicas utilizadas nos trabalhos relacionados.

Trabalho	Técnicas de clusterização e de redução de dimensionalidade					Técnicas de emparelhamento		
	Clusterização hierárquica aglomerativa	Modelo do Espaço Vetorial	<i>Cluster Ensembles</i> baseados em interseção	Indexação Semântica Latente	Geração de <i>clusters</i> baseada em emparelhamento prévio	Elementares		Estruturais
						Sintáticas	Semânticas	Sintáticas
Algergawy, Massmann e Rahm (2011)	1 X	2 X	3	4		X	X	X
Zhang, Hu e Qu (2012)		X				X		
Algergawy et al. (2015)	X	X		X		X		X
Isele, Jentzsch e Bizer (2011)					X	X		
Hamdi et al. (2010)	X				X	X	X	X
Ferrara et al. (2015)	X		X		X	X		

Técnicas e aspectos que podem ser utilizados para melhorar o estado da arte:

- 1 Detecção de Comunidades
- 2 Outros aspectos (estrutural e extensional)
- 3 CEB
- 4 ICA

8 METODOLOGIA PROPOSTA

Neste capítulo, é explicada a metodologia proposta neste trabalho. A metodologia desenvolvida utiliza os conceitos e aplica as técnicas explanadas nos capítulos 2 a 6, com o objetivo de preencher as lacunas e suprir as deficiências observadas na análise do estado da arte, feita no capítulo 7. Assim sendo, a metodologia subdivide-se em cinco etapas principais: modelagem das ontologias como grafos; detecção de comunidades nos grafos das ontologias; clusterização das comunidades segundo diferentes aspectos; obtenção de um consenso entre as clusterizações dos diferentes aspectos; emparelhamento entre elementos de ontologias de um mesmo *cluster* consensual. O fluxograma da metodologia está representado na figura 8.1. Visando tornar mais didática a explanação da metodologia, usaram-se, como exemplo, ontologias mais simples e menores, adaptadas de ontologias da OAEI, indicando-se os principais resultados de cada etapa, juntamente com ilustrações.

Figura 8.1 – Fluxograma da metodologia proposta.

Considerando duas ontologias para as quais se deseja obter o alinhamento, inicialmente modela-se cada ontologia como um grafo, fazendo com que classes das ontologias correspondam a nós dos grafos e representando propriedades que conectam classes como arcos dos grafos. Em seguida, particionam-se os grafos

assim obtidos, aplicando-se técnicas de Detecção de Comunidades (subseção 5.5). Na sequência, consideram-se três aspectos das comunidades: terminológico, topológico e extensional.

A escolha desses aspectos foi feita com base em uma analogia em relação às técnicas de emparelhamento (subseção 2.4.1). Essas técnicas consideram, dentre outras, características associadas aos termos, às estruturas e às instâncias dos elementos (classes ou propriedades) existentes nas ontologias em comparação. Da mesma forma que técnicas de emparelhamento estruturais e baseadas em instâncias descobrem semelhanças entre elementos que as técnicas baseadas em cadeias de caracteres não conseguem descobrir, adotaram-se o aspecto estrutural e o das instâncias como base para modelagens adicionais das comunidades, visando auxiliar a agrupar comunidades que tenham semelhanças que o aspecto terminológico por si só não consegue descobrir.

Desta forma, após modelarem-se as comunidades com base nesses três aspectos em separado, são obtidas três clusterizações, uma para cada aspecto. A seguir, obtém-se um consenso entre as três clusterizações, combinando-se as clusterizações de cada aspecto, de forma a gerar uma clusterização consensual. Ao final, faz-se o emparelhamento entre elementos de comunidades que pertençam a um mesmo *cluster* consensual, desde que sejam comunidades de diferentes ontologias. A representação gráfica da metodologia é apresentada na figura 8.2.

8.1 Modelagem das Ontologias como Grafos

Dadas duas ontologias para as quais se objetiva encontrar o alinhamento, inicialmente, é feita a modelagem de cada ontologia como se fosse um grafo. Adota-se a modelagem de Tran, Wan e Haase (2009) (subseção 5.5): classes das ontologias são mapeadas a nós do grafo e propriedades são mapeadas a arcos. Nessa modelagem, uma dada propriedade de objeto (propriedade com domínio e escopo) corresponde a um arco do grafo. O arco assim definido conecta dois nós: um deles corresponde a classe do domínio da propriedade e o outro corresponde à classe do escopo da propriedade. Na modelagem adotada, também se representam propriedades de herança das ontologias. Essas propriedades são mapeadas a arcos do grafo, sendo que, nesses casos, o arco conecta o nó que corresponde à classe mãe com o nó que corresponde à sua classe filha.

Figura 8.2 – Representação gráfica da metodologia proposta.

Para ilustrar a modelagem descrita, considerem-se duas ontologias, adaptadas de ontologias da OAEI, representadas por meio dos grafos das figuras 8.3 e 8.4, respectivamente. Essas ontologias serão usadas como exemplo, ao longo deste capítulo, para facilitar a compreensão da metodologia e serão denominadas ontologia 1 e ontologia 2.

Nas figuras 8.3 e 8.4, os nós dos grafos são representados por retângulos, rotulados com os nomes das classes. Os arcos dos grafos correspondem a relações de herança entre as classes e a propriedades de objeto. As propriedades de herança são representadas por arcos com linhas contínuas e as propriedades de objeto são representadas por arcos com linhas tracejadas. Diferentes cores dos arcos das figuras 8.3 e 8.4 representam diferentes relações, conforme explicam as legendas colocadas

abaixo das respectivas figuras. Como exemplo de relação de herança, tem-se, na figura 8.3, a classe “Report”, que herda características da classe “Reference”, sendo a relação de herança entre essas classes representada por um arco contínuo de cor azul. Para propriedades de objeto, tem-se, por exemplo, na figura 8.3, a propriedade de objeto “pages”, representada por um arco descontínuo verde que conecta a classe de domínio “Part” com a classe de escopo “PageRange”.

Figura 8.3 – Modelagem da ontologia 1 por meio de grafo.

Legenda:

☒	— has subclass
☒	— parts (Domain>Range)
☒	— chapters(Subclass all)
☒	— pages (Domain>Range)
☒	— book (Domain>Range)
☒	— communications (Domain>Range)
☒	— parts(Subclass all)
☒	— chapters(Subclass all)
☒	— collection (Domain>Range)
☒	— proceedings (Domain>Range)

Figura 8.4 – Modelagem da ontologia 2 por meio de grafo.

Legenda:

8.2 Obtenção das Partições dos Grafos das Ontologias

Após modelarem-se as ontologias como grafos, aplicam-se técnicas de Detecção de Comunidades, para cada um dos grafos em separado. São aplicadas as técnicas de Detecção de Comunidades que obtiveram os melhores resultados no estudo de Coskun et al. (2011): passeios aleatórios (PONS; LATAPY, 2005); otimização gulosa de modularidade (CLAUSET; NEWMAN; MOORE, 2004); método de Potts (REICHARDT; BORNHOLDT, 2006). Adicionalmente, é aplicada a técnica de Brandes et al. (2008). Com o intuito de comparar o desempenho das técnicas, calcula-se a modularidade (subseção 5.2) da partição resultante da aplicação de cada técnica, adotando-se a partição que tiver a maior modularidade.

Ao fazer isso para a ontologia 1 do exemplo da metodologia, resultou que a técnica de Brandes e o método de Potts forneceram a maior modularidade, sendo igual a 0,31125. Para a ontologia 2, a maior modularidade obtida foi igual a 0,4923, gerada pela técnica de Brandes, método de Potts e otimização gulosa de modularidade. Nas figuras 8.5 e 8.6, são mostradas as partições assim geradas. Nós com a mesma cor pertencem à mesma comunidade. Em vermelho, representam-se os arcos que ligam nós de diferentes comunidades (arcos externos). Os demais arcos ligam nós de uma mesma comunidade (arcos internos às comunidades). Na tabela 8.1, faz-se um mapeamento entre o número identificador dos nós (ID) e as classes correspondentes nas ontologias.

Figura 8.5 – Partição do grafo da ontologia 1.

Figura 8.6 – Partição do grafo da ontologia 2.

Tabela 8.1 – Mapeamento entre identificadores dos nós e os nomes das classes das ontologias do exemplo da metodologia.

Ontologia 1		Ontologia 2	
ID do nó	Nome da classe	ID do nó	Nome da classe
V1	DeliverableReport	V1	Entry
V2	Report	V2	Informal
V3	Reference	V3	Book
V4	Misc	V4	Chapter
V5	Book	V5	Booklet
V6	Part	V6	Manual
V7	Article	V7	Unpublished
V8	InBook	V8	TechReport
V9	Monograph	V9	Misc
V10	PageRange	V10	LectureNotes
V11	InCollection	V11	Proceedings
V12	BookPart	V12	Book
V13	Collection	V13	InCollection
V14	InProceedings	V14	Monography
V15	Proceedings	V15	Collection

8.3 Clusterização das Comunidades Baseada no Aspecto Terminológico

Para cada uma das comunidades, inicialmente extrai-se todo o vocabulário existente em seus elementos: nomes de classes e de propriedades; anotações (metadados com descrições textuais explicando o significado dos nomes). Cada comunidade é tratada como se fosse um documento. Na sequência, faz-se o pré-processamento (capítulo 4) do conjunto de palavras obtido: removem-se *stop words*, efetua-se o *stemming* e a lematização. Em seguida, aplica-se o Modelo do Espaço Vetorial (subseção 4.1) aos textos pré-processados. Após isso, faz-se a redução de dimensionalidade (subseção 4.2). São feitas duas reduções, uma aplicando-se Análise de Componentes Independentes e outra com Análise Semântica Latente.

Para cada redução, um diferente subespaço do espaço vetorial inicial é obtido. Para cada subespaço, considera-se, entre as comunidades, a distância do cosseno e aplicam-se os seguintes algoritmos de clusterização hierárquica e particional (subseções 3.1 e 3.2): *single-link*, *complete-link*, UPGMA, WARD, DIANA, CLARA, PAM e *K-means*. Para cada subespaço, avalia-se a clusterização obtida por cada algoritmo, calculando-se a largura de silhueta (subseção 3.3) de cada clusterização. Seleciona-se a clusterização para a qual corresponder a maior largura. O fluxograma relacionado a esta etapa é dado pela figura 8.7.

Figura 8.7 – Fluxograma da clusterização segundo o aspecto terminológico.

Tendo como base as comunidades das ontologias 1 e 2 do exemplo deste capítulo (figuras 8.5 e 8.6) e com auxílio da ferramenta R (RSTUDIO, 2016), foram feitos os passos relacionados à clusterização do aspecto terminológico. Os textos iniciais associados a cada comunidade são discriminados na tabela 8.2. As comunidades de 1 a 4 são da ontologia 1 e as de 5 a 7 são da ontologia 2. A título ilustrativo, efetuou-se a redução de dimensionalidade com Análise de Componentes Independentes, aplicando-se, em seguida, as técnicas de clusterização acima mencionadas. A maior largura de silhueta foi igual a 0,81, obtida por todas as técnicas de clusterização, exceto CLARA. A clusterização obtida com essa largura de silhueta

é representada na tabela 8.3, em que rótulos das comunidades estão na primeira linha e rótulos dos *clusters* estão na segunda linha. A tabela indica que as comunidades 1 e 6 estão no *cluster* 1, as comunidades 2 e 7 estão no *cluster* 2, a comunidade 3 ficou isolada e as comunidades 4 e 5 estão no *cluster* 4.

Tabela 8.2 – Textos contidos em cada comunidade para o exemplo da metodologia.

Comunidade	Texto contido nos nomes de classes e nas anotações relativas às classes
1	DeliverableReport: a report delivered for accomplishing a contract. Report: a report published by an institution with some explicit policy. Reference: Base class for all entries. Misc: Use this type when nothing else fits.
2	Book: a book that may be a monograph or a collection of written texts. Monograph: a book that is a single entity, as opposed to a collection. BookPart: A chapter (or section or whatever) of a book having its own title. Collection: A book that is collection of texts or articles.
3	Part: a part of something (either Book or Proceedings). Article: An article from a journal or magazine. InBook: a subpart of a book given by a range of pages. Collection: a book that is collection of texts or articles. PageRange: a range of pages. InCollection: a part of a book having its own title.
4	InProceedings: An article in a conference proceedings. Proceedings: The proceedings of a conference. A proceedings may be implicitly defined with an inproceedings.
5	Entry: Base class for all entries. Book: a single book that may be a monography or a collection of written texts. Proceedings: the proceeding of a conference. A proceedings may be implicitly defined with an inproceedings.
6	Informal: a document that that was informally published or not published. Booklet: a work that is printed and bound, but without a named publisher or sponsoring institution. Manual: technical documentation. Unpublished: a document having an author and title, but not formally published. TechReport: a report published by a school or other institution, usually numbered within a series. Misc: use this type when nothing else fits. LectureNotes: lecture notes.
7	Chapter: the chapter (or section or whatever) number in which the entry is found. Book: a work that may be composed of several parts or chapter. May be a monography, a collection or a proceedings. InCollection: a part of a book having its own title. Monography: a book that is a single entity, as opposed to a collection. Collection: a book that is collection of texts or articles.

Tabela 8.3 – Resultado da clusterização do aspecto terminológico com a maior largura de silhueta.

comunidades	1	2	3	4	5	6	7
<i>clusters</i>	1	2	3	4	4	1	2

8.4 Clusterização das Comunidades Baseada no Aspecto Extensional

Para o aspecto extensional, consideram-se características relacionadas às instâncias (objetos ou indivíduos) contidas nas comunidades das ontologias. Para cada comunidade, calculam-se as restrições e probabilidades associadas a seus conceitos, de acordo com Hu et al. (2006) (fórmula 5 da subseção 2.5). As restrições e probabilidades obtidas para o exemplo usado neste capítulo são discriminados na tabela 8.4. Na última coluna da tabela, a soma refere-se à somatória da parcela $1-\Delta C_k$ para os conceitos de uma mesma comunidade. Adotaram-se quantidades de instâncias aleatórias.

Na sequência, gera-se uma matriz, na qual as linhas correspondem a comunidades e as colunas correspondem aos diferentes conceitos existentes nas comunidades, sendo que as entradas da matriz são os valores das probabilidades dos conceitos. Faz-se a redução de dimensionalidade (subseção 4.2). São feitas duas reduções, uma aplicando-se Análise de Componentes Independentes e outra com Análise Semântica Latente.

Em seguida, para cada espaço dimensional reduzido (subespaço), calculam-se as distâncias entre os pares de comunidades de ontologias diferentes, usando-se a distância Euclidiana. Para cada subespaço, usam-se as respectivas distâncias para aplicar os seguintes algoritmos de clusterização hierárquica e particional (subseções 3.1 e 3.2): *single-link*, *complete-link*, UPGMA, WARD, DIANA, PAM e *K-means*. Para cada subespaço, avalia-se a clusterização de cada algoritmo, calculando-se a largura de silhueta (subseção 3.3) e selecionando-se a clusterização para a qual corresponder a maior largura. O fluxograma para a obtenção das clusterizações das comunidades das ontologias segundo o aspecto extensional é dado pela figura 8.8.

Para as ontologias usadas como exemplo neste capítulo, aplicou-se a ferramenta R para a redução da dimensionalidade, cálculo de distâncias e clusterização. Usando Análise de Componentes Independentes e comparando os algoritmos de clusterização supramencionados, resultou que o *K-means* gerou a maior largura de silhueta, igual a 0,57, agrupando as comunidades conforme indica a tabela 8.5.

Tabela 8.4 – Restrições e probabilidades dos conceitos para o exemplo da metodologia.

Ontologia	Comunidade	Conceito	n.º de instâncias	Restrição = ΔC_k	$1 - \Delta C_k$	Probabilidade = $(1 - \Delta C_k) / \text{Soma}$
1	C1	DeliverableReport	50	0,0200	0,9800	0,3448
1	C1	Report	1	1,0000	0,0000	0,0000
1	C1	Reference	8	0,1250	0,8750	0,3078
1	C1	Misc	80	0,0125	0,9875	0,3474
1	C2	Book	60	0,0167	0,9833	0,2494
1	C2	Monograph	75	0,0133	0,9867	0,2502
1	C2	BookPart	65	0,0154	0,9846	0,2497
1	C2	Collection	90	0,0111	0,9889	0,2508
1	C3	Part	65	0,0154	0,9846	0,2828
1	C3	Article	19	0,0526	0,9474	0,2721
1	C3	InBook	4	0,2500	0,7500	0,2154
1	C3	PageRange	5	0,2000	0,8000	0,2298
1	C3	InCollection	1	1,0000	0,0000	0,0000
1	C4	InProceedings	7	0,1429	0,8571	0,5000
1	C4	Proceedings	7	0,1429	0,8571	0,5000
2	C5	Entry	70	0,0143	0,9857	0,3372
2	C5	Book	80	0,0125	0,9875	0,3378
2	C5	Proceedings	20	0,0500	0,9500	0,3250
2	C6	Informal	60	0,0167	0,9833	0,1442
2	C6	Booklet	75	0,0133	0,9867	0,1447
2	C6	Manual	65	0,0154	0,9846	0,1444
2	C6	Unpublished	90	0,0111	0,9889	0,1451
2	C6	TechReport	50	0,0200	0,9800	0,1437
2	C6	Misc	44	0,0227	0,9773	0,1433
2	C6	LectureNotes	12	0,0833	0,9167	0,1345
2	C7	Chapter	1	1,0000	0,0000	0,0000
2	C7	Book	103	0,0097	0,9903	0,2559
2	C7	InCollection	27	0,0370	0,9630	0,2488
2	C7	Monography	87	0,0115	0,9885	0,2554
2	C7	Collection	14	0,0714	0,9286	0,2399

Figura 8.8 – Fluxograma da clusterização segundo o aspecto extensional.

Tabela 8.5 - Resultado da clusterização do aspecto extensional com a maior largura de silhueta.

comunidades	1	2	3	4	5	6	7
clusters	1	2	2	3	4	2	2

8.5 Clusterização das Comunidades Baseada no Aspecto Topológico

Para a clusterização baseada no aspecto topológico, inicialmente calculam-se, para cada comunidade, as seguintes métricas de topologia (subseção 5.3): excentricidade média, densidade de arcos, diâmetro e comprimento do caminho médio. As métricas são consideradas os atributos de cada comunidade. Como são apenas quatro atributos, não é feita a redução de dimensionalidade. Com base nos atributos, calculam-se as distâncias entre os pares de comunidades das duas ontologias em comparação, usando-se a distância Euclidiana (capítulo 3). Aplicam-se os algoritmos *single-link*, *complete-link*, UPGMA, WARD, DIANA, PAM e *K-means* (subseção 3.2). Avalia-se a clusterização estrutural obtida por cada algoritmo com largura de silhueta (subseção 3.3), selecionando-se a clusterização para a qual resultar a maior largura.

Ao aplicar esses passos ao exemplo usado neste capítulo, resultam, primeiramente, os atributos topológicos conforme tabela 8.6, em que Em é a excentricidade média de cada comunidade, DA é a densidade de arcos das comunidades, D é o diâmetro de cada comunidade e Lm é o comprimento do caminho médio das comunidades.

Tabela 8.6 – Valores dos atributos topológicos para o exemplo da metodologia.

Comunidades	Em	DA	D	Lm
1	2,5000	0,1500	3,0000	1,6667
2	2,0000	0,2000	2,0000	1,3333
3	1,8000	0,2000	2,0000	1,6000
4	1,0000	0,0500	1,0000	1,0000
5	1,6667	0,1429	2,0000	1,3333
6	1,8571	0,4286	2,0000	1,7143
7	3,2000	0,2857	4,0000	2,0000

Com base nesses valores, com auxílio da ferramenta R, fez-se o cálculo das distâncias e aplicaram-se as técnicas de clusterização acima citadas. Obteve-se com o *K-means* a máxima largura de silhueta, igual a 0,76, com a clusterização dada pela tabela 8.7.

Tabela 8.7 – Resultado da clusterização do aspecto topológico com a maior largura de silhueta.

comunidades	1	2	3	4	5	6	7
clusters	1	2	2	2	2	2	1

8.6 Clusterização Consensual dos Diferentes Aspectos

Nesta etapa, faz-se a combinação dos melhores resultados de clusterização (de maior largura de silhueta) de cada um dos três aspectos. Aplicam-se *Cluster Ensembles* (capítulo 6) para obter a combinação, que representa uma clusterização consensual dos aspectos. São aplicadas as seguintes técnicas: CEB, explanada na subseção 6.3, CSPA, HGPA e MCLA, explanadas nas subseções 6.1 e 6.2. Assim, têm-se, a princípio, quatro resultados de clusterização consensual possíveis. Uma ilustração do fluxograma desta etapa é dada pela figura 8.9.

Para o exemplo usado neste capítulo, ao serem aplicadas as técnicas de clusterização consensual CSPA, HGPA e MCLA às clusterizações dos aspectos terminológico, extensional e topológico, a clusterização consensual de maior informação mútua (subseção 3.3) foi obtida com CSPA. Esse resultado de clusterização é dado pela tabela 8.8. A clusterização obtida com CEB está representada na tabela 8.9. Para obter esses resultados, utilizaram-se as

implementações das técnicas consensuais da ferramenta Matlab (MATLAB, 2016), discriminadas na subseção 8.8.

Figura 8.9 – *Cluster Ensembles* dos aspectos terminológico, extensional e topológico.

Tabela 8.8 – Resultado da clusterização consensual com CSPA.

comunidades	1	2	3	4	5	6	7
clusters	1	2	4	3	3	4	2

Tabela 8.9 – Resultado da clusterização consensual com CEB.

comunidades	1	2	3	4	5	6	7
clusters	1	3	2	2	2	3	3

8.7 Emparelhamento

Obtida a clusterização consensual entre os três aspectos, é feito o emparelhamento entre elementos das ontologias que pertençam a um mesmo *cluster* consensual. Os elementos considerados são as classes e as propriedades, comparando-se os nomes desses elementos.

Para o emparelhamento, são aplicadas aos nomes dos elementos as técnicas Levenshtein e n-gram, as quais se baseiam em similaridade de caracteres (subseção 2.4.1), e uma técnica baseada em consulta ao dicionário semântico do WordNet (subseção 2.4.1). Para cada par de elementos, calcula-se a média das similaridades

obtidas pelas técnicas de Levenshtein e n-gram. Verifica-se também se os nomes dos elementos são sinônimos pelo WordNet. Considera-se que um dado par de elementos é similar, fazendo parte do alinhamento, se a média das similaridades com Levenshtein e n-gram for maior ou igual a 0,7 ou se os elementos forem sinônimos segundo o WordNet, caso em que se atribui similaridade igual a 1. Assim, usa-se um *threshold* (valor mínimo) de 0,7 para o alinhamento.

O alinhamento assim obtido é comparado com um *gold standard* de referência para o alinhamento, que é o gabarito do alinhamento fornecido pela OAEI. A comparação do alinhamento obtido pela metodologia com o alinhamento de referência é feita com base nas métricas de cobertura, precisão e *F-measure* (subseção 2.4.2).

No caso do exemplo da metodologia, resultou o mesmo alinhamento para as duas soluções consensuais (CSPA e CEB). Na tabela 8.10, é discriminado esse alinhamento.

Tabela 8.10 – Alinhamento entre classes para o exemplo da metodologia.

Classe da ontologia 1	Classe da ontologia 2
Book	Book
Monograph	Monography
Collection	Collection
Proceedings	Proceedings

8.8 Ferramentas

As ferramentas utilizadas neste trabalho são: Eclipse, RStudio, Matlab, Protegé e ObjectAid. Na tabela 8.11, discriminam-se as finalidades das ferramentas, pacotes e bibliotecas usadas.

8.9 Diagramas de classes e de sequência

No desenvolvimento do sistema que implementa a metodologia proposta, foi escolhida a linguagem Java de programação, que é uma linguagem orientada a objetos. Previamente à efetiva implementação do sistema, foram elaborados os diagramas de classes e de sequência. Estes diagramas fornecem uma visão de alto nível do sistema, a qual permite uma melhor compreensão da arquitetura da solução, visando facilitar o entendimento do funcionamento do sistema. A inclusão do código

Tabela 8.11 – Ferramentas utilizadas na metodologia.

Finalidade	Ferramenta, biblioteca ou pacote
Ambiente para desenvolvimento em Java	Eclipse Java Mars (ECLIPSE MARS 2 PACKAGES, 2016)
Ambiente para desenvolvimento em R	RStudio (RSTUDIO, 2016)
<i>Parsing</i> (leitura) das ontologias	Biblioteca Jena (APACHE JENA, 2015) do Java (JAVA, 2016)
Técnicas de Detecção de Comunidades e métricas de topologia	Pacote igraph do R (THE R PROJECT FOR STATISTICAL COMPUTING, 2014)
Análise de Componentes Independentes	Pacote fastICA do R (THE R PROJECT FOR STATISTICAL COMPUTING, 2014)
Análise Semântica Latente	Pacote lsa do R (THE R PROJECT FOR STATISTICAL COMPUTING, 2014)
Pré-processamento de termos, Modelo do Espaço Vetorial, distâncias do cosseno, de Manhattan e euclidiana	Pacotes tm e RTextTools do R (THE R PROJECT FOR STATISTICAL COMPUTING, 2014) e funções do Matlab (MATLAB, 2016)
Técnicas de Clusterização	Pacotes cluster e fpc do R (THE R PROJECT FOR STATISTICAL COMPUTING, 2014)
Largura de silhueta	Pacote cluster do R (THE R PROJECT FOR STATISTICAL COMPUTING, 2014) e funções do Matlab (MATLAB, 2016)
Integração do Java com R (chamada de métodos do R por meio de código em Java)	Pacote Rserve do R (THE R PROJECT FOR STATISTICAL COMPUTING, 2014) e bibliotecas REngine e RserveEngine do Java (JAVA, 2016)
<i>Cluster Ensembles</i> Bayesianos	Implementação conforme Wang, Shan e Banerjee (2016)
<i>Cluster Ensembles</i> com as técnicas CSPA, HGPA e MCLA	Implementação conforme Strehl (2011)
Emparelhamento	Bibliotecas Simmetrics, Jaws WordNet e Alignment API do Java (JAVA, 2016)
Figuras com exemplos de grafos associados a ontologias	Protegé (2015)
Diagramas UML	ObjectAid (2016)

neste trabalho torná-lo-ia muito extenso, o que tenderia a não motivar muito o interesse do leitor na consulta e interpretação do código.

Os diagramas de classes permitem discernir responsabilidades conforme tarefas, que ficam distribuídas entre as classes Java. Além disso, pode-se compreender os relacionamentos entre as classes Java que foram usadas na codificação do sistema. Os diagramas de sequência servem para fornecer uma noção da sequência de passos envolvidas no fluxograma que o sistema implementa. Esta sequência de passos é materializada por meio de instâncias das classes e por

chamadas de métodos existentes nas classes, de acordo com responsabilidades a serem por elas desempenhadas no sistema.

Foi utilizada uma classe Java para representar atributos das classes das ontologias: seus identificadores, rótulos, comentários, propriedades de objetos, superclasses e instâncias. O objetivo foi utilizá-la durante a leitura das ontologias para facilitar o armazenamento desses atributos, que são usados nas fases seguintes da metodologia proposta. Outra classe importante na fase de leitura das ontologias é associada às propriedades, sendo responsável por armazenar identificadores, rótulos, comentários, classes de escopo e de domínio das propriedades contidas nas ontologias.

Foram usadas classes Java para modelar comunidades, representando coleções de classes das ontologias contidas em cada comunidade, sendo que propriedades dessas classes Java correspondem aos atributos topológicos e extensionais das comunidades e ao conjunto de termos dessas comunidades. Essas classes ficaram, então, responsáveis por armazenar dados referentes às comunidades. Foi desenvolvida também uma classe Java responsável por gerar os resultados de modularidade dos algoritmos de Detecção de Comunidades. Outra classe Java foi elaborada para calcular a quantidade de nós e de arcos dos grafos associados às ontologias.

No que tange ao alinhamento, fez-se uma classe Java para calcular as similaridades entre elementos (classes e propriedades) das comunidades de um mesmo *cluster*. Classes também foram desenvolvidas para auxiliar nas comparações do emparelhamento, impondo que apenas elementos de comunidades de ontologias diferentes fossem comparados entre si.

Para implementar a metodologia proposta, foi necessária uma orquestração dos métodos que cada uma dessas classes possui e que as permitem desempenhar suas responsabilidades. Essa orquestração foi feita de forma a respeitar e ordenar a sequência de passos necessária para a aplicação das técnicas envolvidas na metodologia proposta. A classe principal (ou “Main”, em inglês), foi desenvolvida com esse intuito e a sequência de passos que ela implementa pode ser representada por meio de um diagrama de sequência.

Os diagramas de classes e de sequência do sistema que implementa a metodologia proposta foram elaborados com base na linguagem de modelagem orientada a objeto denominada UML e encontram-se no Anexo C – Diagramas UML.

8.10 Manual de utilização do sistema

Nesta subseção, são explicadas e ilustradas a sequência de passos e as parametrizações necessárias para executar o emparelhamento de um dado par de ontologias, com base na metodologia ora proposta, usando o sistema desenvolvido neste trabalho. A execução e a parametrização envolvem a utilização das ferramentas R Studio, Eclipse e Matlab.

Inicialmente, deve-se abrir a ferramenta R Studio e executar uma rotina para instanciar um servidor, o qual estabelece um canal para integração com a ferramenta Eclipse. Isto possibilita a chamada de comandos em linguagem R diretamente pelo Eclipse, codificando instruções específicas em Java. O Eclipse exerce o papel de um cliente que solicita requisições (comandos em R), as quais são enviadas ao R Studio por meio de uma conexão e o R Studio encarrega-se de executar os comandos requisitados.

Após essa etapa de instanciação do servidor para estabelecer uma conexão entre o Eclipse e o R Studio, usa-se a ferramenta Eclipse para efetuar alguns passos preliminares. É feita a leitura das ontologias e a extração de dados associados aos elementos das ontologias, necessárias para a obtenção das comunidades das ontologias. Em seguida, o Eclipse dispara a chamada de comandos em R para o R Studio, que executa esses comandos para obter as comunidades das ontologias.

Na sequência, o Eclipse modela as comunidades conforme os aspectos terminológico, extensional e topológico. Em seguida, é utilizada novamente a ferramenta R Studio, desta vez para aplicar as técnicas de clusterização às comunidades, o que é feito para cada um dos aspectos. Também no R Studio, selecionam-se as clusterizações de maior largura de silhueta de cada aspecto.

Após essa seleção, as clusterizações selecionadas representam dados de entrada para o uso da ferramenta Matlab, que é encarregada de efetuar as clusterizações consensuais com as técnicas CSPA, HGPA, MCLA e CEB. Em seguida, volta-se à ferramenta Eclipse, usando-se as clusterizações consensuais obtidas como dados de entrada para aplicar as técnicas de emparelhamento e gerar os resultados de alinhamento das ontologias. Além de usar-se como base as clusterizações consensuais para gerar o alinhamento, pode-se também usar as clusterizações de cada aspecto como dados de entrada.

8.10.1 Instanciação de um servidor para integração entre o Eclipse e o R Studio

Inicialmente, é preciso abrir a ferramenta R Studio e executar um *script* para usar a biblioteca Rserve do R e instanciar um servidor. Este servidor estabelecerá uma conexão entre o Eclipse e o R Studio, enviando requisições de comandos em R feitas pelo Eclipse e devolvendo ao Eclipse as respostas conforme os resultados obtidos pelo R Studio. A figura 8.10 ilustra como fazer a instanciação do servidor Rserve usando o console do R Studio.

Figura 8.10 – Sequência de comandos para instanciar o servidor Rserve com o console do R Studio.

8.10.2 Parametrização das ontologias a serem emparelhadas

Após a instanciação do servidor Rserve, abre-se o projeto do sistema no Eclipse. Em seguida, executam-se as rotinas implementadas pela classe “Main”, a qual dispara a sequência de comandos e funções necessárias para ler as ontologias, extrair dados relativos às classes e propriedades das ontologias, obter as comunidades das ontologias e modelá-las de acordo com os aspectos terminológico, extensional e topológico. Para executar a sequência de rotinas da classe “Main”, é preciso clicar com o botão direito do *mouse* na classe “Main”, clicando, em seguida,

na opção “Run As” e selecionar “1 Java Application”, conforme indica a figura 8.11. O usuário precisa também criar uma pasta “texts” no diretório C:, para o Eclipse gerar arquivos com os textos contidos nas comunidades das ontologias. No mesmo diretório, é necessário criar uma pasta com o nome “topo”, para o Eclipse armazenar a matriz com as características topológicas das comunidades.

Figura 8.11 – Execução da classe “Main”, ao abrir o projeto do sistema com a ferramenta Eclipse.

Após esse passo, o console do Eclipse solicita ao usuário que especifique o caminho para o arquivo e o nome do arquivo que contém as ontologias a serem submetidas ao emparelhamento. O usuário deve entrar com esses dados e aguardar o processamento, conforme ilustra a figura 8.12. Os arquivos das ontologias têm que estar em formato RDF. Assim sendo, o usuário tem que armazenar nessas pastas as ontologias em RDF que deseja emparelhar. Um exemplo de estrutura de armazenamento é dado na figura 8.13, na qual as pastas foram numeradas de acordo com a numeração de ontologias do teste *benchmarks* da OAEI.

Figura 8.12 – Entrada de dados no console do Eclipse, com a especificação pelo usuário dos caminhos para os arquivos e nomes dos arquivos que contêm as ontologias a serem emparelhadas.

Figura 8.13 – Exemplo de estrutura de armazenamento em diretórios para armazenar as ontologias a serem emparelhadas.

8.10.3 Parametrização e execução das técnicas de clusterização dos aspectos terminológico, extensional e topológico

Mantendo o Eclipse aberto e em execução, o usuário deve voltar-se agora à interface do R Studio, para rodar os *scripts* da clusterização dos aspectos. No passo anterior, o Eclipse gerou conjuntos de dados associados às modelagens de cada aspecto e armazenou arquivos com esses dados em pastas específicas. Esses arquivos são acessados por *scripts* do R, que fazem a aplicação de técnicas de clusterização com base nos dados do Eclipse e geram os resultados. Para cada um dos aspectos, é necessário fazer uma parametrização inicial nos *scripts* do R.

- **Parametrização e execução das técnicas de clusterização do aspecto terminológico**

Devem-se abrir os *scripts* do R para o aspecto terminológico. Há dois *scripts*: uma para o ICA e outro para o LSA. Em ambos os *scripts*, deve-se especificar a quantidade de *clusters* a ser gerada pelas técnicas de clusterização, dado pelo parâmetro *kc* na figura 8.14. O valor de *kc* é a semente a ser usada pelas técnicas e precisa ser pré-definida pelo usuário. Como exemplo, foi ilustrada a parametrização para *kc* = 2.

Figura 8.14 – *Script* do R com seta indicando a parametrização que precisa ser feita quanto à quantidade de *clusters* *kc* a ser gerada pelas técnicas de clusterização.

Para rodar o *script*, o usuário precisa selecioná-lo e clicar no botão “Run”, conforme indica a figura 8.15. Pode-se fazer o mesmo usando os atalhos “Ctrl+A” para selecionar e “Ctrl+Enter” para rodar. Após isso, o usuário deverá selecionar a aba “Plots” para visualizar e analisar os gráficos de largura de silhueta de cada técnica de clusterização (figura 8.16).

O exemplo ilustrado refere-se ao *script* e resultados de clusterização do aspecto terminológico usando ICA. Com LSA, o procedimento é o mesmo. O ideal é que o usuário repita a execução do *script* de clusterização para vários valores de *kc*, tanto com ICA como com LSA, selecionando o valor de *kc* e a respectiva clusterização que gerarem a máxima largura de silhueta. Uma sugestão é utilizar o intervalo de *kc* variando, inicialmente, de 2 a 10.

Figura 8.15 – Exemplo de execução do *script* de clusterização do aspecto terminológico.

Figura 8.16 – Gráficos de largura de silhueta das clusterizações obtidos no R Studio.

Um exemplo de gráfico com resultados de clusterização retornados pelo R é dado pela figura 8.17, na qual os rótulos das comunidades são representados no eixo vertical, enquanto o eixo horizontal representa as larguras de silhueta de cada comunidade. A comunidade 1 ficou no *cluster* 1, as comunidades 2, 3, 5, 6 e 7 ficaram no *cluster* 2 e a comunidade 4 ficou no *cluster* 3.

Figura 8.17 – Exemplo de gráfico de largura de silhueta para um resultado de clusterização das comunidades das ontologias conforme o aspecto terminológico.

Após o usuário analisar visualmente os gráficos de largura de silhueta e selecionar o resultado de clusterização de maior largura, é necessário que ele crie um arquivo com extensão “txt” com os rótulos da clusterização selecionada. Para o gráfico da figura 8.17, o arquivo deve conter a seguinte sequência de caracteres na primeira linha: “1\2\2\3\2\2\2”. A interpretação desta sequência é a seguinte: há 7 comunidades no exemplo, então, a sequência deve ter, necessariamente, 7 números, sendo que cada número é o rótulo do cluster em que está a respectiva comunidade. Assim, neste exemplo, as comunidades 2, 3, 5, 6 e 7 estão no mesmo cluster (com rótulo 2) e as comunidades 1 e 4 estão isoladas, em dois diferentes clusters. A comunidade 1 está sozinha no cluster 1 e a comunidade 4 ficou no cluster 3. Para melhor esclarecer esta convenção, outra forma de expor esta rotulação das comunidades de acordo com seus *clusters* é dada a seguir:

rótulos das comunidades	1	2	3	4	5	6	7
rótulos dos <i>clusters</i>	1	2	2	3	2	2	2

- **Parametrização e execução das técnicas de clusterização do aspecto extensional**

Devem-se abrir os *scripts* do R para o aspecto extensional, que tem duas vertentes: uma baseada em ICA e outra baseada em LSA. Há, portanto, dois *scripts*, um para cada vertente. No *script* do ICA, é preciso especificar, nas primeiras linhas, o número de *clusters* que se almeja para a clusterização (parâmetro *kc*) e a quantidade de comunidades (parâmetro *nc*). No *script* do LSA, parametriza-se apenas *kc*. Um exemplo do *script* do ICA é dado na figura 8.18, na qual indicam-se a parametrização a ser feita para as variáveis *kc* (quantidade de *clusters*) e *nc* (quantidade de comunidades das ontologias). O botão “Run” fica na parte superior direita do script. Para saber a quantidade de comunidades que o Eclipse obteve para as ontologias em comparação (parâmetro *nc*), basta acessar a pasta “texts” em C: e verificar a quantidade de arquivos “txt” gerada (figura 8.19). Essa quantidade é o valor de *nc*. Para rodar o *script*, deve-se selecionar todo o *script* e clicar no botão “Run”, que está na parte superior direita da tela do *script*. No exemplo da figura 8.19, há sete arquivos, portanto, sete comunidades foram obtidas. Convencionou-se a nomenclatura dos arquivos de forma que o primeiro caractere se refere à ontologia e o segundo caractere corresponde ao número da comunidade. Assim sendo, “A2”, por exemplo, é o nome do arquivo que contém os termos da comunidade 2 da ontologia A.

Da mesma forma que no aspecto terminológico, o usuário visualiza e analisa resultados de clusterização clicando na aba “Plots” do R Studio. Isto deve ser feito após executar o *script* de clusterização extensional para vários valores de *kc*, de forma a possibilitar selecionar o valor de *kc* e a respectiva clusterização que propicie maior largura de silhueta. Analogamente ao descrito para o aspecto terminológico, o usuário precisa, para a clusterização selecionada, criar um arquivo texto com os rótulos dos *clusters* para as comunidades.

Figura 8.18 – Script do R para a clusterização do aspecto extensional baseado em ICA.

Figura 8.19 – Exemplo de coleção de arquivos texto com os conjuntos de termos de cada comunidade das ontologias.

- **Parametrização e execução das técnicas de clusterização do aspecto topológico**

Deve-se abrir o *script* do R para o aspecto topológico. Trata-se de um único *script*, pois não há redução de dimensionalidade no aspecto topológico, ou seja, não se aplicam o ICA ou o LSA. É preciso especificar, nas primeiras linhas do *script*, o número de *clusters* que se deseja obter para a clusterização (parâmetro *kc*), conforme exemplo da figura 8.20.

Analogamente ao que foi explicado para os aspectos anteriores, deve-se, então, parametrizar diversos valores de *kc*. Para cada parametrização, roda-se o *script* e analisam-se visualmente os resultados de largura de silhueta na aba “Plots”. Também de maneira análoga ao que é feito para os outros aspectos, o usuário precisa criar um arquivo txt com a rotulação dos *clusters* para as comunidades, seguindo a mesma convenção adotada conforme descrito nos aspectos anteriores.

8.10.4 Parametrização e execução das técnicas de clusterização consensual

As técnicas de clusterização consensual aplicadas na metodologia proposta são: CSPA, HGPA, MCLA e CEB. Para aplicá-las, é necessário usar a ferramenta Matlab. Os dados de entrada são os rótulos das clusterizações de cada aspecto,

Figura 8.20 – Parametrização da quantidade de *clusters* *kc* para o *script* de clusterização do aspecto topológico. No exemplo da figura, foi utilizado o valor de *kc* igual a 2.

```

1  kc <- 2
2
3
4  library(tm)
5  library(RTextTools)
6  library(lsa)
7  library(cluster)
8  library(fpc)
9
10
11 table <- read.table("C:/topo/matrixTopo.txt", header=FALSE, sep="", fill=TRUE)
12 mat1 <- as.matrix(table)
13 m1 <- mat1
14
15
16 d <- as.matrix(dist(m1, method = "euclidean"))
17
18
19 ##clusterização
20
21 #UPGMA (agnes average) - hierarq aglom
22
23 upgma_ <- agnes(d, diss = TRUE,
24                stand = FALSE, method = "average",
25                keep.diss = FALSE, keep.data = FALSE, trace.lev = 0)
26
27 upgma_si <- silhouette(cutree(upgma_, k = kc), d)
28
29 plot(upgma_si, col = c("grey"))
30
31
32
33
34 #Resumo UPGMA
35 summary(upgma_)
36

```

obtidos conforme foi descrito em 8.10.3. Há dois *scripts*: um para aplicar as técnicas CSPA, HGPA e MCLA e outro para aplicar CEB.

- **Parametrização e execução das técnicas CSPA, HGPA e MCLA**

Para as técnicas CSPA, HGPA e MCLA, inicialmente deve-se fornecer uma matriz com os rótulos das clusterizações de cada aspecto (figura 8.21). Observa-se que a matriz de rótulos do exemplo da figura 8.21 é formada por três linhas e sete

colunas. Cada linha corresponde a um dos aspectos. Por serem três aspectos, têm-se três linhas. São sete colunas na matriz, pois o exemplo usado refere-se a sete comunidades. As entradas na matriz são os rótulos das clusterizações (números de identificação dos *clusters*) para as comunidades conforme a clusterização de cada um dos aspectos. Na primeira linha da matriz da figura 8.21, por exemplo, têm-se os rótulos da clusterização do aspecto terminológico, sendo que as comunidades 2, 3, 5, 6 e 7 estão no *cluster* 2, a comunidade 1 está no *cluster* 1 e a comunidade 4 está no *cluster* 3. Trata-se da mesma convenção para rotulação das clusterizações descrita anteriormente em 8.10.3, no item sobre o aspecto terminológico. Especificada a matriz desta forma, deve-se rodar o *script*, selecionando todo o *script* e clicando no botão “Run” do Matlab, que está na parte superior central da figura 8.21.

Figura 8.21 – *Script* do Matlab para efetuar a clusterização consensual dos três aspectos.

Após rodar o *script*, a saída é um vetor de rótulos correspondente à técnica de clusterização que forneceu a maior informação mútua normalizada média (subseção 3.3). Um exemplo é dado na figura 8.22. É necessário que o usuário crie um arquivo com extensão “txt” com o vetor de rótulos obtido. Para o exemplo da figura 8.22, o arquivo deve conter a seguinte sequência de caracteres na primeira linha: “3\2\4\1\2\2\2”.

Figura 8.22 – Saída do Matlab, com a seta na figura indicando os rótulos da clusterização consensual de maior informação mútua normalizada média, dentre as obtidas com CSPA, HGPS e MCLA.

MATLAB R2016a - student use

Command Window

```
Recursive Partitioning... -----
4-way Edge-Cut: 23232323, Balance: 2.10

Timing Information -----
I/O: 0.000
Partitioning: 0.000 (PMETIS time)
Total: 0.000
*****
sgraph: partgraph0100.bat completed - loading graph1.part.4
clbtocl: delivering 4 clusters
clbtocl: average posterior prob is 0.75059
clbtocl: winning posterior probs are
    0.9999    0.5714    0.5999    0.7499    0.9999    0.6666    0.6666

clbtocl: full posterior probs are
    0.0001    0.4285    0.0000    0.7499    0.0000    0.0001    0.0001
    0.0000    0.5714    0.2000    0.2500    0.9999    0.6666    0.6666
    0.9999    0.0000    0.2000    0.0001    0.0000    0.3333    0.3333
    0.0000    0.0001    0.5999    0.0001    0.0001    0.0001    0.0000

clusterensemble: mcla at 0.7045

ans =
     3     2     4     1     2     2     2
```

←

- **Parametrização e execução da técnica de CEB**

Para parametrizar e executar a técnica de CEB, o usuário precisa, inicialmente, abrir o arquivo que contém o *script* de CEB (figura 8.23). São necessários cinco parâmetros para executar esse *script*: "base_labels", "number_baseclusterers", "true_labels", "Palpha" e "Pbeta".

Figura 8.23 – *Script* do CEB no Matlab, indicando os quatro primeiros parâmetros a serem especificados pelo usuário: "base_labels", "number_baseclusterers", "true_labels" e "Palpha".

O parâmetro “base_labels” é o primeiro a ser especificado. Ele refere-se aos rótulos das clusterizações dos aspectos e corresponde a uma matriz. Cada coluna da matriz representa um aspecto e cada linha, uma comunidade. Cada entrada na matriz é o rótulo do *cluster* para a comunidade (linha) conforme o aspecto (coluna), usando-se a rotulação convencionada em 8.10.3. Na figura 8.23, a matriz “base_labels” corresponde aos resultados das clusterizações dos aspectos (subseção 8.10.3) do exemplo adotado para o manual de utilização do sistema.

O segundo parâmetro a ser especificado é o “number_baseclusterers”, que é um vetor de dimensão igual à quantidade de aspectos e tem que ser populado com o valor igual à quantidade máxima de *clusters* dentre todas as clusterizações dos aspectos.

O terceiro parâmetro é outro vetor, representado pela variável “true_labels”, que deve ter a dimensão igual ao número de comunidades (no caso do exemplo, igual a 7). Pode-se popular esse vetor com o valor 1.

O quarto parâmetro também é um vetor, de dimensão igual ao número de *clusters* que se deseja obter para a clusterização consensual. Sugere-se populá-lo com um valor em cada entrada do vetor igual a aproximadamente o valor 1 dividido pela dimensão do vetor. No exemplo da figura 8.23, tem-se $1/3=0,333$, usando-se o vetor [0,3 ; 0,3 ; 0,4].

O quinto parâmetro do *script* do CEB é o vetor “Pbeta”, que é um vetor de matrizes (figura 8.24). Cada matriz de “Pbeta” tem que ter o número de linhas igual à quantidade de *clusters* desejada para a clusterização consensual. A quantidade de colunas de cada matriz tem que ser igual ao número máximo de *clusters* dentre todas as clusterizações dos aspectos. A soma de valores de cada linha deve ser aproximadamente igual a 1. Uma sugestão para esses valores, no caso de matrizes de quatro colunas, é dada na figura 8.24. Nessa figura, verifica-se que, no caso de uma clusterização consensual de três *clusters* e tendo-se o máximo de *clusters* das clusterizações dos aspectos igual a quatro, Pbeta deve ser populado com três matrizes, dimensionadas com três linhas e quatro colunas cada.

Após parametrizarem-se essas cinco variáveis, executa-se o *script*. Depois de rodar o *script*, a saída é um vetor de rótulos gerado pelo CEB. Um exemplo é dado na figura 8.25. É preciso que o usuário crie um arquivo com extensão “txt” com o vetor de rótulos obtido. Para o exemplo da figura 8.25, o arquivo deve conter a seguinte sequência de caracteres na primeira linha: “1\3\3\2\3\3\3”.

Figura 8.24 – *Script* do CEB no Matlab, indicando a parametrização do vetor de matrizes Pbeta

```

MATLAB R2016a
HOME PLOTS APPS EDITOR PUBLISH VIEW
New Open Save Find Files Compare Go To Find Insert Comment Indent Breakpoints Run Run and Advance
FILE NAVIGATE EDIT BREAKPOINTS RUN

C:\Users\Andre\Documents\MATLAB\bce
Command Window
runBCE_3_clusters_205.m runBCE_3_clusters_208.m runBCE_3_clusters_209.m runBCE_3_clusters_210.m
Current Folder
28 Pbeta = cell (1,3);
29
30 %Cada matriz de Pbeta precisa ter a quantidade de linhas igual ao número de
31 %clusters desejado para a solução consensual
32 Pbeta (1,1) = {[0.571429 0.142857 0.142857 0.142857;
33                0.142857 0.142857 0.571429 0.142857;
34                0.142857 0.571429 0.142857 0.142857]};
35 Pbeta (1,2) = {[0.571429 0.142857 0.142857 0.142857;
36                0.142857 0.571429 0.142857 0.142857;
37                0.142857 0.142857 0.142857 0.571429]};
38 Pbeta (1,3) = {[0.318182 0.318182 0.181818 0.181818;
39                0.181818 0.272727 0.272727 0.272727;
40                0.178571 0.571429 0.178571 0.071429]};
  
```

Figura 8.25 – Saída do Matlab após a execução do *script* de CEB, indicando-se por uma seta o vetor de rótulos da clusterização consensual obtida.

```

HOME PLOTS APPS
New Script New Open Find Files Import Save New Variable Open Variable Analyze Code
Compare Data Workspace Clear Workspace Run and Time Clear Commands
FILE VARIABLE CODE

C:\Users\Andre\Documents\MATLAB\bce
Command Window
Name Size Bytes Class Attributes
beta_tt 1x3 624 cell
t=3, error=0.021688
Name Size Bytes Class Attributes
beta_t 1x3 624 cell
Name Size Bytes Class Attributes
beta_tt 1x3 624 cell
t=4, error=0.0084915
1
3
3
2
3
3
3
  
```

8.10.5 Parametrização e execução do emparelhamento

Nesta etapa, o emparelhamento é feito com base em uma dada clusterização de comunidades. O usuário pode especificar tanto uma clusterização de um dado aspecto como uma clusterização consensual. Para tanto, deve-se retornar ao Eclipse (que estava em execução) para especificar o caminho do arquivo que contém a clusterização consensual (ou do aspecto) e especificar também o nome do arquivo que contém os rótulos da clusterização consensual (ou do aspecto). Um exemplo dessa parametrização é ilustrado por meio da figura 8.26.

Figura 8.26 – Console do Eclipse com destaque para os dados de entrada que devem ser especificados pelo usuário, relativos ao caminho e nome do arquivo com a clusterização a ser usada como base para o emparelhamento.

Após o sistema processar os dados, o usuário deve especificar o caminho e o nome do arquivo que possui o *gold standard* (referência) de alinhamento. O sistema processa essas informações e confronta com o alinhamento resultante da aplicação da metodologia, fornecendo como resposta final a métricas de precisão cobertura e *F-measure*. Este último passo é ilustrado na figura 8.27.

Figura 8.27 – Resultados de precisão cobertura e *F-measure* para o alinhamento, após o usuário fornecer o caminho e o nome do arquivo com o alinhamento de referência.

```
<terminated> Main [Java Application] C:\Program Files (x86)\Java\jre1.8.0_101\bin\javaw.exe (25 de jun de 2017 11:23:59)
Digite o caminho para o arquivo da ontologia 1 (formato C:\Pasta\Subpasta): C:\Users\Andre\benchmarks\101
Digite o nome do arquivo rdf da ontologia 1: onto
Digite o caminho para o arquivo da ontologia 2 (formato C:\Pasta\Subpasta): C:\Users\Andre\benchmarks\301
Digite o nome do arquivo rdf da ontologia 2: onto
Digite o caminho do arquivo com a clusterização consensual ou do aspecto (formato C:\Pasta\Subpasta): C:\Users\Andre
Digite o nome do arquivo txt com a clusterização consensual ou do aspecto: ceb3_301
Digite o caminho para o arquivo com o gold standard de alinhamento (formato C:\Pasta\Subpasta): C:\Users\Andre\benchmarks\301
Digite o nome do arquivo rdf com o gold standard de alinhamento: refalign

Resultados de alinhamento
Precisão: 0.8125
Cobertura: 0.22033898305084745
F-measure: 0.34666666666666666
```

The screenshot shows a Java application window titled "Java - Eclipse". The console displays a series of prompts for file paths and names, followed by the results of the alignment process. Three blue arrows are overlaid on the image, pointing to the results section. The results section contains the following text:

Resultados de alinhamento
Precisão: 0.8125
Cobertura: 0.22033898305084745
F-measure: 0.34666666666666666

9 ESTUDOS DE CASO

Neste capítulo, descrevem-se as ontologias usadas nos estudos de caso de emparelhamento. As ontologias são descritas no que concerne o seu domínio e as quantidades de seus elementos (classes, propriedades e instâncias). Ainda no que diz respeito aos estudos de caso, são apresentados neste capítulo, os resultados da aplicação da metodologia proposta às ontologias: resultados de detecção de comunidades, de redução de dimensionalidade, de clusterização e de alinhamento.

Nos estudos de caso deste trabalho, foram utilizadas as ontologias do teste *benchmarks* da OAEI. O teste de *benchmarks* é composto originalmente de 54 ontologias, rotuladas com numerações. O objetivo do teste é alinhar uma das ontologias, a ontologia 101, também denominada como ontologia-origem, com cada uma das demais 53 ontologias, denominadas como ontologias-alvo. Assim, o teste decompõe-se em 53 estudos de caso. Trata-se de ontologias relativas ao domínio acadêmico, com informações sobre referências bibliográficas, contendo dados sobre tipos de eventos nos quais as referências foram publicadas (congressos, conferências etc.), nomes e locais dos eventos, tipos de publicações, autores, suas filiações etc. Algumas das ontologias são originárias do concurso de alinhamento denominado *Evaluation of Ontologies and Ontology-based tools* (EON) e foram adaptadas para a OAEI. Outras ontologias do teste foram desenvolvidas a partir de uma parceria entre o *Massachusetts Institute of Technology* (MIT) e o *Institut National de Recherche en Informatique et en Automatique* (INRIA). Nessas ontologias, foram usados como referência os campos existentes na estrutura de dados do BibTex (BIBTEX, 2016), o qual define um padrão para enumerar e descrever referências bibliográficas em artigos acadêmicos.

Ao acessar o *sítio web* da OAEI (ONTOLOGY ALIGNMENT EVALUATION INITIATIVE, 2017), verificou-se que o *dataset* mais recente das ontologias do teste *benchmarks* refere-se ao evento da OAEI de 2016. Nesse *sítio*, apresentam-se informações sobre as características das ontologias desse *dataset* e sobre o objetivo do teste, que visa principalmente detectar pontos fortes e fracos nas soluções de emparelhamento que são submetidas ao teste. As ontologias-alvo do teste foram concebidas e adaptadas para o contexto da OAEI visando criar dificuldades. Alguns exemplos podem ser enumerados: uma das ontologias-alvo é uma generalização da

ontologia 101; outra tem mais restrições na linguagem de representação OWL; foram usadas diferentes convenções de nomenclatura para os nomes de elementos das ontologias (uso de caixa alta ou baixa, de forma diversificada); uso de sinônimos para rótulos de diferentes classes; ontologias-alvo com hierarquias expandidas ou reduzidas; exclusão de restrições de propriedades; expansões nas classes. As informações do sitio da OAEI também dão conta de que algumas ontologias estão atualmente indisponíveis, outras estão representadas na língua francesa, várias não têm instâncias ou propriedades e outras têm termos com cadeias de caracteres aleatórias e sem significado.

Tendo em vista a metodologia proposta neste trabalho, que utiliza os aspectos extensional e topológico das ontologias, não podendo, portanto, prescindir de propriedades e instâncias, escolheram-se para os estudos de caso as ontologias que tivessem esses elementos. Além disso, como foi usado na metodologia o dicionário do WordNet em inglês para auxiliar no emparelhamento, restringiram-se os estudos também às ontologias que fossem necessariamente representadas na língua inglesa. Assim sendo, restaram 12 ontologias que, ao mesmo tempo, estão disponíveis, estão representadas em inglês e que possuem, simultaneamente, propriedades e instâncias, o que permite sua aplicação no presente trabalho. Essas 12 ontologias foram selecionadas para os estudos de caso. Adicionalmente, uma das ontologias sem instâncias, a ontologia 304, foi incluída na seleção. Para viabilizar a aplicação da ontologia 304 em um dos estudos de caso e para poder simular o efeito do aspecto extensional no alinhamento, as instâncias da ontologia 101 foram inseridas na ontologia 304. Um quadro resumo com as quantidades de classes, instâncias e propriedades das ontologias do teste *benchmarks* utilizadas nos estudos de caso deste trabalho é dado na tabela 9.1. São 13 ontologias e como o teste *benchmarks* visa emparelhar a ontologia 101 (ontologia-origem) com as demais (ontologias-alvo), possuindo *gold standards* para esses emparelhamentos, resultam 12 pares de ontologias emparelháveis, perfazendo assim 12 estudos de caso. Esses foram os estudos de caso usados neste trabalho.

Nas subseções 9.1 a 9.6 a seguir, os resultados de detecção de comunidades, clusterização e alinhamento gerados pelo uso da metodologia proposta são descritos mais detalhadamente para o estudo de caso das ontologias 101 e 304. No Anexo E, estão relacionados os resultados para os demais 11 estudos de caso efetuados,

Tabela 9.1 – Tabela com as quantidades de classes, instâncias e propriedades das ontologias do teste *benchmarks* usadas nos estudos de caso.

Ontologia	Quantidade de classes	Quantidade de instâncias	Quantidade de propriedades de tipos de dados	Quantidade de propriedades de objeto	Quantidade de propriedades de herança
101	36	86	46	26	27
103	36	86	46	26	27
104	36	86	46	26	27
204	36	31	46	26	27
205	36	31	45	26	27
208	36	54	46	26	27
209	36	54	45	26	27
222	32	86	46	26	23
223	69	86	46	27	60
225	36	86	46	26	27
230	28	76	49	11	24
301	15	9	40	0	14
304	41	86	14	40	32

sempre considerando a ontologia 101 como ontologia-origem. Uma síntese com os principais resultados de todos os estudos de caso é feita na subseção 9.7.

9.1 Resultados de Detecção de Comunidades nas Ontologias (ontologia 101 x ontologia 304)

Inicialmente, foi feita a leitura das ontologias, modelando-as conforme grafos. Alguns dos principais elementos das ontologias 101 e 304 são ilustrados nos grafos das figuras 9.1 e 9.2, respectivamente, nas quais os relacionamentos de herança entre classes são representados por linhas azuis contínuas, enquanto relações formadas por propriedades de objeto são representadas por linhas tracejadas. Legendas logo abaixo das figuras indicam as relações representadas conforme suas cores. Os grafos completos das ontologias 101 e 304, com classes, relacionamentos e instâncias, estão no Anexo A – Grafos das Ontologias. Os códigos em OWL das ontologias encontram-se no Anexo B – Ontologias em OWL.

Figura 9.1 – Grafo de parte da ontologia 101 da OAEI.
Fonte: Protegé (2015)

Legenda:

- ☒ — has subclass
- ☒ — institution (Domain>Range)

Figura 9.2 – Grafo de parte da ontologia 304 da OAEI.
Fonte: Protegé (2015)

Legenda:

- ☒ — 'in collection' (Domain>Range)
- ☒ — date (Domain>Range)
- ☒ — date(Subclass all)
- ☒ — has subclass
- ☒ — inMonography (Domain>Range)

Tabela 9.2 – Mapeamento entre identificador dos nós do grafo e nome das classes para as ontologias 101 e 304.

Ontologia 101		Ontologia 304	
ID do nó	Nome da classe	ID do nó	Nome da classe
V1	Collection	V1	Monography
V2	List	V2	InProceedings
V3	Article	V3	List
V4	InBook	V4	DateWithMonth
V5	Deliverable	V5	Book
V6	Institution	V6	MastersThesis
V7	School	V7	Thesis
V8	Academic	V8	Entry
V9	Date	V9	Title
V10	Report	V10	InCollection
V11	Monograph	V11	LectureNotes
V12	Conference	V12	Part
V13	Proceedings	V13	Institution
V14	TechReport	V14	InBook
V15	Misc	V15	Misc
V16	Publisher	V16	Manual
V17	Part	V17	Booklet
V18	Reference	V18	Deliverable
V19	PersonList	V19	Organization
V20	Journal	V20	PageRange
V21	MastersThesis	V21	Unpublished
V22	Organization	V22	Journal
V23	Informal	V23	Date
V24	Book	V24	Composite
V25	InProceedings	V25	TechReport
V26	PhdThesis	V26	PhdThesis
V27	Booklet	V27	Publisher
V28	Address	V28	Text
V29	Unpublished	V29	Informal
V30	LectureNotes	V30	Published
V31	Chapter	V31	Collection
V32	PageRange	V32	MotionPicture
V33	InCollection	V33	Article
V34	MotionPicture	V34	Chapter
V35	Manual	V35	Proceedings
		V36	Academic
		V37	School
		V38	BookPart
		V39	PersonList

Após isso, procedeu-se à obtenção de comunidades nas ontologias com uso de técnicas de Detecção de Comunidades. Conforme explanado na subseção 8.2, foram aplicadas as seguintes técnicas: passeios aleatórios (PONS; LATAPY, 2005); otimização gulosa de modularidade (CLAUSET; NEWMAN; MOORE, 2004); método de Potts (REICHARDT; BORNHOLDT, 2006); técnica de Brandes et al. (2008). Os resultados da aplicação dessas técnicas foram comparados, considerando-se os valores de modularidade da partição gerada por cada técnica. A tabela 9.3 discrimina esses valores, para cada ontologia e conforme a técnica de Detecção de Comunidades aplicada. Para a ontologia 101, a partição obtida resultou em seis comunidades. Para a ontologia 304, a partição resultou em cinco comunidades.

Tabela 9.3 – Valores de modularidade resultantes da aplicação das técnicas de Detecção de Comunidades às ontologias 101 e 304.

Técnica de Detecção de Comunidades	Modularidade	
	Ontologia 101	Ontologia 304
Técnica de Brandes	0,4812	0,5382
Método de Potts	0,4758	0,5379
Otimização gulosa de modularidade	0,4590	0,5354
Passeios aleatórios	0,4287	0,5029

Na tabela 9.4, discriminam-se os identificadores dos nós do grafo da ontologia 101, suas classes e os rótulos das comunidades às quais cada classe pertence, obtidos com a aplicação da técnica de Brandes. Na figura 9.5, ilustra-se esse resultado. Nós com a mesma cor são da mesma comunidade. Em vermelho, estão representados os arcos que conectam nós de diferentes comunidades (arcos externos) e, em preto, estão os arcos que conectam nós de uma mesma comunidade (arcos internos às comunidades). Um nó não pertence a duas ou mais comunidades simultaneamente, pertencendo a apenas uma comunidade. Não há, portanto, sobreposição de comunidades.

Na tabela 9.5, estão discriminados os identificadores dos nós do grafo da ontologia 304, as classes da ontologia e os rótulos das comunidades às quais cada classe pertence, resultantes da aplicação da técnica de Brandes. A figura 9.6 ilustra esse resultado.

Tabela 9.4 – Identificadores de nós, nomes das classes correspondentes e rótulos das comunidades as quais as classes pertencem, para representar os agrupamentos resultantes da aplicação da técnica de Brandes à ontologia 101.

Identificador do nó do grafo	Nome da Classe	Rótulo da comunidade a qual a classe pertence
V1	Collection	1
V4	InBook	1
V11	Monograph	1
V17	Part	1
V24	Book	1
V32	PageRange	1
V33	InCollection	1
V2	List	2
V15	Misc	2
V18	Reference	2
V19	PersonList	2
V31	Chapter	2
V34	MotionPicture	2
V3	Article	3
V5	Deliverable	3
V6	Institution	3
V10	Report	3
V14	TechReport	3
V16	Publisher	3
V20	Journal	3
V7	School	4
V8	Academic	4
V21	MastersThesis	4
V26	PhdThesis	4
V9	Date	5
V12	Conference	5
V13	Proceedings	5
V22	Organization	5
V25	InProceedings	5
V28	Address	5
V23	Informal	6
V27	Booklet	6
V29	Unpublished	6
V30	LectureNotes	6
V35	Manual	6

Figura 9.5 – Partição da ontologia 101 em comunidades obtida com a técnica de Brandes.

Tabela 9.5 – Identificadores de nós, respectivos nomes das classes e comunidades obtidos com a técnica de Brandes para a ontologia 304.

Identificador do nó do grafo	Nome da Classe	Rótulo da comunidade a qual a classe pertence
V1	Monography	7
V5	Book	7
V10	InCollection	7
V14	InBook	7
V31	Collection	7
V34	Chapter	7
V2	InProceedings	8
V12	Part	8
V19	Organization	8
V20	PageRange	8
V22	Journal	8
V30	Published	8
V33	Article	8
V35	Proceedings	8
V38	BookPart	8
V3	List	9
V4	DateWithMonth	9
V8	Entry	9
V9	Title	9
V23	Date	9
V24	Composite	9
V27	Publisher	9
V32	MotionPicture	9
V39	PersonList	9
V6	MastersThesis	10
V7	Thesis	10
V11	LectureNotes	10
V26	PhdThesis	10
V36	Academic	10
V37	School	10
V13	Institution	11
V15	Misc	11
V16	Manual	11
V17	Booklet	11
V18	Deliverable	11
V21	Unpublished	11
V25	TechReport	11
V28	Text	11
V29	Informal	11

Figura 9.6 – Partição da ontologia 304 em comunidades obtida com a técnica de Brandes.

9.2 Resultados de Clusterização das Comunidades Baseada no Aspecto Terminológico (ontologia 101 x ontologia 304)

Conforme explanado na metodologia (capítulo 8), a clusterização foi precedida pela redução de dimensionalidade. Para a redução, foram aplicadas e comparadas duas técnicas: ICA e LSA. Para cada técnica de redução de dimensionalidade, aplicaram-se oito técnicas de clusterização: UPGMA, *single-link*, *complete-link*, WARD, DIANA, CLARA, *K-means* e PAM. Para cada resultado de clusterização, calculou-se a largura de silhueta média, com o intuito de avaliar as clusterizações.

Ao adotar cada uma das técnicas de clusterização, foi necessário pré-definir a quantidade de *clusters*, dada por k , uma vez que as técnicas por si só não estimam o valor de k . Para escolher o melhor valor de k para uma dada técnica de clusterização, compararam-se as larguras de silhueta média obtidas com vários valores de k . Isto foi feito para cada redução de dimensionalidade (ICA e LSA). Assim, para escolher o melhor valor para k , foram comparados os respectivos resultados de largura de silhueta média para $k = 2, \dots, 10$. O valor de k escolhido foi aquele que gerou a maior

largura de silhueta média, correspondendo também ao melhor resultado de clusterização.

9.2.1 Resultados ao usar ICA

Na tabela 9.6, estão discriminados os resultados de largura de silhueta média, por técnica de clusterização e valor de k, utilizando-se ICA para a redução de dimensionalidade. Em negrito, estão indicados os valores mínimo e máximo obtidos para a largura de silhueta. Na figura 9.7, tem-se o gráfico das larguras de silhueta de cada *cluster* para a melhor clusterização, reduzindo-se a dimensionalidade com ICA. A melhor clusterização agrupou as comunidades conforme indica a tabela 9.7.

Tabela 9.6 – Resultados de largura de silhueta média das clusterizações para ICA com variação de k.

Técnica de redução de dimensionalidade: ICA									
Técnica de clusterização	Valor de k								
	2	3	4	5	6	7	8	9	10
	Largura de silhueta média								
UPGMA	0,47	0,64	0,89	0,85	0,75	0,66	0,51	0,36	0,18
<i>Single-link</i>	0,47	0,64	0,89	0,85	0,69	0,61	0,51	0,36	0,18
<i>Complete-link</i>	0,38	0,61	0,89	0,85	0,75	0,66	0,51	0,36	0,18
Ward	0,47	0,64	0,89	0,85	0,75	0,66	0,51	0,36	0,18
DIANA	0,38	0,61	0,89	0,85	0,75	0,66	0,51	0,36	0,18
CLARA	0,29	0,49	0,72	0,66	0,56	0,41	0,37	0,29	0,18
K-means	0,46	0,59	0,87	0,68	0,63	0,54	0,47	0,32	0,18
PAM	0,36	0,64	0,89	0,85	0,75	0,66	0,51	0,36	0,18
Máxima	0,47	0,64	0,89	0,85	0,75	0,66	0,51	0,36	0,18
Mínima	0,29	0,49	0,72	0,66	0,56	0,41	0,37	0,29	0,18

Figura 9.7 – Gráfico das larguras de silhueta de cada *cluster*, para a clusterização de maior largura de silhueta média, ao usar ICA.

Tabela 9.7 – Resultado de clusterização das comunidades para a máxima largura de silhueta média com ICA.

Cluster	Comunidades
1	1, 7 e 8
2	2, 5 e 9
3	3, 6 e 11
4	4 e 10

9.2.2 Resultados ao usar LSA

Na tabela 9.8, estão discriminados os resultados de largura de silhueta média, ao usar a técnica LSA para a redução de dimensionalidade. Em negrito, estão indicados os valores mínimo e máximo obtidos para a largura de silhueta. Na figura 9.8, tem-se o gráfico das larguras de silhueta de cada *cluster* para a clusterização de maior largura de silhueta média. A clusterização correspondente agrupou as comunidades conforme indica a tabela 9.9.

Tabela 9.8 – Resultados de largura de silhueta média das clusterizações para LSA com variação de k.

Técnica de redução de dimensionalidade: LSA									
Técnica de clusterização	Valor de k								
	2	3	4	5	6	7	8	9	10
	Largura de silhueta média								
UPGMA	0,51	0,61	0,80	0,77	0,71	0,69	0,53	0,35	0,18
Single-link	0,51	0,62	0,80	0,77	0,71	0,69	0,53	0,35	0,18
Complete-link	0,26	0,48	0,80	0,77	0,71	0,69	0,53	0,35	0,18
Ward	0,51	0,61	0,80	0,77	0,71	0,69	0,53	0,35	0,18
DIANA	0,37	0,61	0,80	0,77	0,71	0,69	0,53	0,35	0,15
CLARA	0,64	0,52	0,47	0,42	0,46	0,48	0,33	0,27	0,18
K-means	0,84	0,45	0,65	0,61	0,61	0,60	0,51	0,33	0,18
PAM	0,42	0,62	0,80	0,77	0,71	0,69	0,53	0,35	0,18
Máxima	0,84	0,62	0,80	0,62	0,71	0,69	0,53	0,35	0,18
Mínima	0,26	0,45	0,47	0,45	0,46	0,48	0,33	0,27	0,15

Figura 9.8 – Gráfico das larguras de silhueta de cada *cluster* para a clusterização de maior largura de silhueta média, ao usar LSA.

Tabela 9.9 – Resultado de clusterização das comunidades para a máxima largura de silhueta média com LSA.

Cluster	Comunidades
1	1,2,3,5,6,7,8,9 e 11
2	4 e 10

9.3 Resultados de Clusterização das Comunidades Baseada no Aspecto Extensional (ontologia 101 x ontologia 304)

De maneira análoga ao que foi feito no aspecto terminológico, no aspecto extensional foram aplicadas oito técnicas de clusterização: UPGMA, *single-link*, *complete-link*, Ward, DIANA, CLARA, *K-means* e PAM. Para cada uma das técnicas de clusterização, calculou-se a largura de silhueta média, objetivando avaliar as clusterizações. Previamente à aplicação das técnicas de clusterização, fez-se a redução de dimensionalidade. Isto foi feito usando-se duas abordagens: uma com ICA e outra com LSA. Assim, obtiveram-se dois resultados, um para cada abordagem.

Na aplicação das técnicas de clusterização, foi necessário estimar a quantidade de *clusters*, dada por k . Para isso, compararam-se as larguras de silhueta média obtidas com valores de k variando de 2 a 10, para uma mesma técnica de clusterização. Adotou-se o valor de k que gerou a maior largura de silhueta média, sendo este também o melhor resultado de clusterização.

9.3.1 Resultados ao usar ICA

Na tabela 9.10, apresentam-se os resultados baseados na abordagem com ICA. Em negrito, estão ressaltados os valores mínimo e máximo obtidos para a largura de silhueta. Na figura 9.9, ilustra-se o gráfico das larguras de silhueta de cada *cluster* para o maior valor de largura de silhueta média. A clusterização gerada agrupou as comunidades conforme indica a tabela 9.11.

Tabela 9.10 – Resultados de largura de silhueta média das clusterizações para ICA com variação de k.

Técnica de clusterização	Valor de k								
	2	3	4	5	6	7	8	9	10
	Largura de silhueta média								
UPGMA	0,13	0,16	0,28	0,37	0,49	0,44	0,34	0,24	0,15
Single-link	0,15	0,26	0,28	0,37	0,49	0,44	0,34	0,24	0,15
Complete-link	0,11	0,11	0,20	0,32	0,49	0,44	0,34	0,24	0,15
Ward	0,20	0,26	0,35	0,47	0,49	0,44	0,34	0,24	0,15
DIANA	0,15	0,16	0,20	0,31	0,49	0,44	0,34	0,24	0,15
CLARA	0,08	0,18	0,32	0,47	0,49	0,44	0,34	0,24	0,15
K-means	0,24	0,24	0,29	0,47	0,69	0,53	0,45	0,30	0,18
PAM	0,08	0,17	0,32	0,47	0,49	0,44	0,34	0,24	0,15
Máxima	0,24	0,26	0,35	0,47	0,69	0,53	0,45	0,30	0,18
Mínima	0,08	0,11	0,20	0,31	0,49	0,44	0,34	0,24	0,15

Figura 9.9 – Gráfico das larguras de silhueta de cada *cluster* para a clusterização de maior largura de silhueta média, ao usar a técnica ICA.

Tabela 9.11 – Resultado de clusterização das comunidades para a máxima largura de silhueta média.

Cluster	Comunidades
1	3
2	4 e 10
3	6 e 11
4	5 e 8
5	1 e 7
6	2 e 9

9.3.2 Resultados ao usar LSA

Na tabela 9.12, apresentam-se os resultados. Em negrito, estão indicados os valores mínimo e máximo obtidos para a largura de silhueta. Na figura 9.10, tem-se o gráfico das larguras de silhueta de cada *cluster* para o maior valor de largura de silhueta média. A clusterização correspondente agrupou as comunidades conforme indica a tabela 9.13.

Tabela 9.12 – Resultados de largura de silhueta média das clusterizações para LSA com variação de k.

Técnica de clusterização	Valor de k								
	2	3	4	5	6	7	8	9	10
	Largura de silhueta média								
UPGMA	0,34	0,43	0,67	0,73	0,67	0,54	0,42	0,30	0,13
<i>Single-link</i>	0,49	0,43	0,67	0,73	0,67	0,54	0,42	0,30	0,13
<i>Complete-link</i>	0,34	0,43	0,67	0,72	0,67	0,54	0,42	0,30	0,13
Ward	0,49	0,66	0,67	0,72	0,67	0,54	0,42	0,30	0,13
DIANA	0,32	0,38	0,67	0,73	0,67	0,54	0,42	0,30	0,08
CLARA	0,30	0,50	0,59	0,56	0,51	0,42	0,33	0,19	0,08
<i>K-means</i>	0,52	0,60	0,67	0,74	0,60	0,55	0,43	0,31	0,17
PAM	0,49	0,66	0,79	0,73	0,67	0,54	0,42	0,30	0,13
Máxima	0,52	0,66	0,79	0,74	0,67	0,55	0,43	0,31	0,17
Mínima	0,30	0,38	0,59	0,56	0,51	0,42	0,33	0,19	0,08

Figura 9.10 – Gráfico das larguras de silhueta de cada *cluster* para a clusterização de maior largura de silhueta média, ao usar a técnica LSA.

Tabela 9.13 – Resultado de clusterização das comunidades para a máxima largura de silhueta média.

Cluster	Comunidades
1	1, 2 e 7
2	3, 6 e 11
3	4 e 10
4	5, 8 e 9

9.4 Resultados de Clusterização das Comunidades Baseada no Aspecto Topológico (ontologia 101 x ontologia 304)

Assim como no aspecto terminológico e extensional, no aspecto topológico foram aplicadas oito técnicas de clusterização: UPGMA, *single-link*, *complete-link*, WARD, DIANA, CLARA, *K-means* e PAM. Para cada uma das técnicas de clusterização, calculou-se a largura de silhueta média, objetivando avaliar as clusterizações.

Foi preciso pré-definir a quantidade de *clusters*, dada por *k*. Para tanto, compararam-se as larguras de silhueta média obtidas com valores de *k* variando de 2

a 10, para uma mesma técnica de clusterização. Adotou-se o valor de k para o qual correspondeu a maior largura de silhueta média.

Na tabela 9.14, estão discriminados os resultados. Em negrito, estão indicados os valores mínimo e máximo obtidos para a largura de silhueta. Na figura 9.11, tem-se o gráfico das larguras de silhueta de cada *cluster* para a melhor clusterização. A clusterização correspondente agrupou as comunidades conforme indica a tabela 9.15.

Tabela 9.14 – Resultados de largura de silhueta média das clusterizações com variação de k .

Técnica de clusterização	Valor de k								
	2	3	4	5	6	7	8	9	10
	Largura de silhueta média								
UPGMA	0,61	0,66	0,85	0,66	0,74	0,67	0,50	0,34	0,16
<i>Single-link</i>	0,61	0,66	0,85	0,66	0,74	0,67	0,50	0,34	0,16
<i>Complete-link</i>	0,61	0,66	0,85	0,66	0,74	0,67	0,50	0,34	0,16
Ward	0,60	0,68	0,85	0,74	0,74	0,67	0,50	0,34	0,16
DIANA	0,60	0,68	0,85	0,69	0,59	0,67	0,50	0,34	0,16
CLARA	0,61	0,66	0,85	0,74	0,74	0,67	0,50	0,34	0,16
<i>K-means</i>	0,78	0,77	0,96	0,39	0,60	0,45	0,47	0,36	0,18
PAM	0,61	0,66	0,85	0,74	0,74	0,67	0,50	0,34	0,16
Máxima	0,78	0,77	0,96	0,74	0,74	0,67	0,50	0,36	0,18
Mínima	0,60	0,66	0,85	0,39	0,59	0,45	0,47	0,34	0,16

Figura 9.11 – Gráfico das larguras de silhueta de cada *cluster* para a clusterização de maior largura de silhueta média, ao usar o *K-means*.

Tabela 9.15 – Resultado de clusterização das comunidades para a máxima largura de silhueta média.

Cluster	Comunidades
1	1 e 9
2	2, 5, 7, 10 e 11
3	3 e 8
4	4 e 6

9.5 Resultados de Clusterização Consensual dos Aspectos (ontologia 101 x ontologia 304)

A aplicação das técnicas CSPA, HGPA e MCLA resultou em uma mesma clusterização consensual. A quantidade de *clusters* é determinada automaticamente por essas técnicas. No caso das ontologias 101 e 304, a aplicação dessas técnicas resultou em quatro *clusters* consensuais. A clusterização obtida pode ser expressa na forma de rótulos associados às comunidades, por meio do vetor V1:

$$V1 = [4 \quad 4 \quad 2 \quad 1 \quad 3 \quad 2 \quad 4 \quad 3 \quad 3 \quad 1 \quad 2] \quad (44)$$

Assim, têm-se no rótulo 4 (*cluster* 4) as comunidades 1, 2 e 7; no rótulo 2, as comunidades 3, 6 e 11; no rótulo 1, as comunidades 4 e 10; no rótulo 3, as comunidades 5, 8 e 9.

Para a técnica de CEB, usando-se três *clusters* consensuais, resultou uma clusterização consensual que pode ser expressa na forma de um vetor $V_{CEB,3}$ de rótulos associados às comunidades:

$$V_{CEB,3} = [1 \quad 2 \quad 2 \quad 3 \quad 2 \quad 2 \quad 1 \quad 1 \quad 2 \quad 3 \quad 2] \quad (45)$$

Assim, têm-se no rótulo 1 (*cluster* 1) as comunidades 1, 7 e 8; no rótulo 2, as comunidades 2, 3, 5, 6, 9 e 11; no rótulo 3, as comunidades 4 e 10.

9.6 Resultados de Alinhamento (ontologia 101 x ontologia 304)

Conforme explanado na metodologia, procedeu-se ao emparelhamento entre os elementos das ontologias (classes e propriedades) que pertençam a um mesmo

cluster consensual. Isto foi feito para as duas soluções de clusterização consensual obtidas. Além disso, fez-se o emparelhamento entre elementos de um mesmo *cluster* para cada uma das clusterizações dos aspectos considerados isoladamente. Para essas soluções, compararam-se os resultados de precisão, cobertura e *F-measure* (subseção 2.4.2), com o objetivo de avaliar o alinhamento. Usou-se como gabarito de emparelhamento a referência dada pela OAEI. Os resultados são discriminados na tabela 9.16.

Adicionalmente, no Anexo D – Alinhamentos, são tabeladas as similaridades entre os pares de elementos das ontologias, para cada uma das soluções de emparelhamento, e é discriminado também o alinhamento de referência da OAEI.

Tabela 9.16 – Resultados de precisão, cobertura e *F-measure* com base nas técnicas de clusterização.

Técnica de Clusterização	Avaliação do alinhamento		
	Precisão	Cobertura	F-measure
CSPA/HGPA/MCLA	0,8769	0,7500	0,8085
CEB	0,8529	0,7632	0,8056
Clusterização baseada no aspecto terminológico	0,8841	0,8026	0,8414
Clusterização baseada no aspecto extensional	0,8793	0,6711	0,7612
Clusterização baseada no aspecto topológico	0,6364	0,0921	0,1609

9.7 Principais Resultados de Detecção de Comunidades, de Alinhamento e de Clusterização dos Estudos de Caso

Nesta subseção, apresentam-se os principais resultados dos estudos de caso efetuados neste trabalho. Os resultados referem-se à Detecção de Comunidades, ao alinhamento e à clusterização. Um maior detalhamento dos resultados encontra-se no Anexo E.

9.7.1 Resultados de Detecção de Comunidades

Os valores de modularidade obtidos com a aplicação das técnicas de Brandes, método de Potts, otimização gulosa de modularidade e passeios aleatórios às ontologias dos estudos de caso estão discriminados na tabela 9.17. Em negrito, estão destacados os valores máximos de modularidade. A aplicação das técnicas de Detecção de Comunidades no grafo da ontologia 301 gerou modularidades iguais ou próximas de zero pois, como a ontologia tem poucas classes, não se formaram estruturas de comunidades em seu grafo, ou seja, não foram obtidas comunidades com a aplicação das técnicas. Assim, a ontologia não foi segmentada em comunidades, sendo considerada monoliticamente, de forma que a ontologia como um todo correspondeu a uma única comunidade.

Tabela 9.17 – Resultados de modularidade para as técnicas de Detecção de Comunidades aplicadas os estudos de caso.

Ontologia	Modularidade			
	Técnica de Detecção de Comunidades			
	Técnica de Brandes	Método de Potts	Otimização gulosa de modularidade	Passeios aleatórios
101	0,4812	0,4758	0,4590	0,4287
103	0,5096	0,5020	0,5062	0,4779
104	0,5551	0,5493	0,5508	0,5270
204	0,4812	0,4758	0,4590	0,4287
205	0,4771	0,4642	0,4713	0,4646
208	0,4812	0,4812	0,4590	0,4287
209	0,4771	0,4771	0,4713	0,4646
222	0,3798	0,3798	0,3590	0,3548
223	0,6749	0,6717	0,6743	0,6240
225	0,4812	0,4758	0,4590	0,4287
230	0,5421	0,5268	0,5421	0,5255
301	0,0000	-0,0026	0,0000	0,0000
304	0,5382	0,5379	0,5354	0,5029

9.7.2 Resultados de Alinhamento

Na tabela 9.18, são discriminados os resultados de precisão, cobertura e *F-measure* para todos os estudos de caso, conforme a técnica consensual ou o aspecto usados na clusterização. Em negrito, são ressaltados os valores máximos obtidos em cada estudo de caso.

Tabela 9.18 – Resultados de precisão, cobertura e F-*measure* para cada estudo de caso, conforme a técnica consensual ou o aspecto usados na solução.

Estudo de caso	CEB			CSPA, HGPA e MCLA			Extensional			Terminológico			Topológico		
Ontologia 101 x	Precisão	Cobertura	F- <i>measure</i>	Precisão	Cobertura	F- <i>measure</i>	Precisão	Cobertura	F- <i>measure</i>	Precisão	Cobertura	F- <i>measure</i>	Precisão	Cobertura	F- <i>measure</i>
103	0,8283	0,8454	0,8367	0,8851	0,7938	0,8370	0,8851	0,7938	0,8370	0,8780	0,7423	0,8045	0,8824	0,6186	0,7273
104	0,8272	0,6907	0,7528	0,9091	0,6186	0,7362	0,9091	0,6186	0,7362	0,9155	0,6701	0,7738	0,9091	0,6186	0,7362
204	0,868	0,4742	0,6133	0,8588	0,7526	0,8022	0,8588	0,7526	0,8022	0,8588	0,7526	0,8022	0,8588	0,7526	0,8022
205	0,5333	0,1649	0,2520	1,0000	0,0515	0,0980	0,5652	0,1340	0,2167	0,6207	0,1856	0,2857	0,8333	0,0515	0,0971
208	0,8400	0,4330	0,5714	0,8571	0,4330	0,5753	0,8571	0,4330	0,5753	0,8400	0,4330	0,5714	0,8478	0,4021	0,5455
209	0,6923	0,0928	0,1636	1,0000	0,0103	0,0204	0,6667	0,0825	0,1468	0,5556	0,0515	0,0943	1,0000	0,0412	0,0792
222	0,8286	0,9355	0,8788	0,8515	0,9247	0,8866	0,8515	0,9247	0,8866	0,8515	0,9247	0,8866	0,8205	0,6882	0,7485
223	0,8704	0,4845	0,6225	0,8764	0,8041	0,8387	0,8462	0,7938	0,8191	0,8764	0,8041	0,8387	0,7272	0,2474	0,3692
225	0,8349	0,9381	0,8835	0,8586	0,8763	0,8673	0,8557	0,8557	0,8557	0,8654	0,9278	0,8955	0,8586	0,8763	0,8673
230	0,8025	0,9028	0,8497	0,8429	0,8194	0,831	0,8382	0,7917	0,8143	0,8429	0,8194	0,831	0,8382	0,7917	0,8143
301	0,8125	0,2203	0,3467	0,7692	0,1695	0,2778	0,8889	0,1356	0,2353	0,8125	0,2203	0,3467	0,8000	0,2034	0,3243
304	0,8529	0,7632	0,8056	0,8769	0,75	0,8085	0,8793	0,6711	0,7612	0,8841	0,8026	0,8414	0,6364	0,0921	0,1609
médias	0,7992	0,5788	0,6314	0,8821	0,5837	0,6316	0,8252	0,5823	0,6405	0,8168	0,6112	0,6643	0,8344	0,4486	0,5227
máximas	0,8704	0,9381	0,8835	1,0000	0,9247	0,8866	0,9091	0,9247	0,8866	0,9155	0,9278	0,8955	1,0000	0,8763	0,8673

Na tabela 9.19, estão discriminadas as quantidades de vezes em que a aplicação de uma dada técnica consensual ou aspecto resultou no máximo valor para as métricas de alinhamento (precisão, cobertura ou *F-measure*), comparativamente às demais soluções, considerando-se um mesmo estudo de caso. Os valores negritados indicam qual o máximo número de vezes que isso ocorreu para cada métrica.

Tabela 9.19 – Quantidade de vezes em que a aplicação de uma dada técnica ou aspecto conseguiu atingir a máxima métrica de alinhamento em um estudo de caso.

Técnica ou aspecto	Quantidade de vezes em que a técnica ou aspecto teve a máxima métrica de alinhamento em relação às demais soluções		
	Precisão	Cobertura	<i>F-measure</i>
CEB	1	8	2
CSPA, HGPA e MCLA	7	3	5
Extensional	4	2	4
Terminológico	5	5	7
Topológico	1	1	1

Na tabela 9.20, indica-se a variação das métricas de alinhamento de acordo com a quantidade de *clusters*. Foram considerados os valores médios das métricas para cada quantidade de *clusters*, levando-se em conta todos os estudos de caso realizados. Em seguida, na figura 9.12 ilustra-se o gráfico representando o comportamento das três métricas de alinhamento conforme a quantidade de *clusters*.

Tabela 9.20 – Valores médios de precisão, cobertura e *F-measure* conforme a quantidade de *clusters*.

Quantidade de <i>clusters</i>	Valores médios		
	Precisão	Cobertura	<i>F-measure</i>
2	0,8350	0,7331	0,7465
3	0,8196	0,5178	0,5715
4	0,8037	0,4535	0,5306
5	0,8515	0,7037	0,7617
6	0,8515	0,5214	0,5807

Figura 9.12 – Gráfico indicando a variação das métricas de alinhamento conforme a quantidade de *clusters*.

9.7.3 Resultados de Clusterização

Na tabela 9.21, discriminam-se os valores médios de largura de silhueta obtidos nos estudos de caso, conforme a quantidade de *clusters* das soluções de clusterização dos aspectos. O gráfico correspondente aos valores tabelados é representado na figura 9.13.

Tabela 9.21 – Valores médios de largura de silhueta dos estudos de caso conforme a quantidade de *clusters*.

Quantidade de <i>clusters</i>	Largura de silhueta
2	0,4927
3	0,5886
4	0,7201
5	0,7233
6	0,6807
7	0,6220
8	0,5028
9	0,39
10	0,26

Figura 9.13 – Gráfico indicando a variação da largura de silhueta conforme a quantidade de *clusters*.

Na tabela 9.22, estão discriminadas as técnicas de clusterização aplicadas e a quantidade de vezes em que cada técnica atingiu a máxima largura de silhueta de um estudo de caso.

Tabela 9.22 – Quantidade de vezes em que cada técnica de clusterização obteve a máxima largura de silhueta de um estudo de caso.

Técnica de clusterização	Quantidade de vezes que atingiu a máxima largura de silhueta
K-means	22
PAM	17
Ward	15
Complete-link	14
DIANA	14
Single-link	14
UPGMA	14
CLARA	6

Na tabela 9.23, indicam-se as técnicas de redução de dimensionalidade e a quantidade de vezes que essas técnicas reduziram o espaço dimensional de atributos de forma a gerar os máximos valores de largura de silhueta. Essa contagem foi

discriminada por aspectos. Não houve redução de dimensionalidade no aspecto topológico.

Tabela 9.23 – Quantidade de vezes em que cada técnica de clusterização obteve a máxima largura de silhueta de um estudo de caso.

Aspecto	Redução de dimensionalidade	
	ICA	LSA
Extensional	5	7
Terminológico	6	6

10 DISCUSSÃO

Neste capítulo, são discutidos os resultados apresentados no capítulo 9: resultados de detecção de comunidades nas ontologias, de clusterização das comunidades e de alinhamento. Ao final, são propostos trabalhos futuros. A discussão visa ressaltar quais as técnicas que, especificamente nos estudos de caso, se sobressaíram em relação às demais, além de apresentar justificativas para esses resultados, restritas ao âmbito dos estudos de caso efetuados.

10.1 Detecção de Comunidades nas Ontologias 101 e 304

Analizando os resultados de detecção de comunidades nas ontologias 101 e 304 com base na tabela 9.3 (subseção 9.1), verifica-se que a partição obtida com a técnica de Brandes forneceu a maior modularidade para ambas ontologias. Isto ocorreu nas ontologias 101 e 304 porque, segundo o estudo de Brandes et al. (2008), algumas das outras técnicas que se baseiam em otimização da modularidade e que foram aplicadas neste trabalho, como é o caso da otimização gulosa, costumam agrupar inicialmente pares de nós com diferentes graus, para balancear a soma dos quadrados dos graus. Conforme Brandes et al. (2008), na clusterização final, esses nós acabam ficando em diferentes *clusters*, gerando pequenos *clusters* com distribuições de grau assimétricas e fazendo com que não se encontre a solução ótima em termos de modularidade.

10.2 Clusterizações das Comunidades Baseadas nos Aspectos Terminológico, Extensional e Topológico (ontologias 101 e 304)

Nas próximas subseções, discutem-se os resultados de clusterização das comunidades baseadas nos diferentes aspectos.

10.2.1 Clusterização das Comunidades baseada no Aspecto Terminológico (ontologias 101 e 304)

Na clusterização das comunidades baseada no aspecto terminológico (subseção 9.2), ao aplicar a técnica ICA para a redução de dimensionalidade, a maior

largura de silhueta média foi igual a 0,89, obtida para k igual a quatro, a qual ocorreu igualmente para as seguintes técnicas de clusterização: UPGMA, *single-link*, *complete-link*, Ward, DIANA e PAM. O menor valor de silhueta média, com base em ICA, foi igual a 0,18, para k igual a dez, obtido para todas as técnicas de clusterização aplicadas. Ao usar a técnica LSA para redução de dimensionalidade, a maior largura de silhueta média foi de 0,84, para k igual a dois, usando-se a técnica de clusterização *K-means*. Com LSA, o menor valor de silhueta média foi igual a 0,15, para k igual a dez, ao utilizar a técnica DIANA.

Assim sendo, o melhor resultado para a largura de silhueta média foi encontrado ao usar ICA, com k igual a quatro e aplicando técnicas de clusterização particionais e hierárquicas. Esse resultado correspondeu a uma mesma clusterização para todas as técnicas, havendo, portanto, uma convergência para um mesmo resultado, o que reforça a hipótese de que a clusterização assim obtida seja a de melhor ajuste para o aspecto terminológico. Além disso, a aplicação da técnica ICA para redução de dimensionalidade contribuiu para melhorar os resultados de clusterização, pois a largura de silhueta média obtida com base na ICA foi maior do que a obtida com base na LSA, aplicando-se as mesmas técnicas de clusterização. Isto indica que a ICA superou a LSA na tarefa de reduzir a dimensionalidade dos dados, restringindo-os aos elementos das comunidades mais relevantes para a clusterização terminológica, tendo em vista a maior largura de silhueta para a solução com ICA. Esta largura maior para a solução com ICA, comparativamente à solução baseada em LSA, significa *clusters* com elementos mais compactos entre si (dentro de um mesmo *cluster*) e com maior separação entre diferentes *clusters* (subseção 3.3). Essas características são visadas para aumentar a qualidade da clusterização.

10.2.2 Clusterização das Comunidades baseada no Aspecto Extensional (ontologias 101 e 304)

Na clusterização das comunidades baseada no aspecto extensional, a maior largura de silhueta média foi de 0,79, para k igual a quatro, com a técnica de clusterização PAM e fazendo-se redução de dimensionalidade com LSA. O menor valor de largura de silhueta média foi 0,08, para k igual a dez, tanto com LSA como com ICA. O fato de que a técnica PAM forneceu a maior largura de silhueta média está indicando que as características dos dados do aspecto extensional das

comunidades são propícias à formação de *clusters* compactos, uma vez que técnicas de clusterização particional (subseção 3.2), como é o caso da PAM, habitualmente geram melhores resultados para esses tipos de *clusters* (JAIN; MURTY; FLYNN, 1999).

10.2.3 Clusterização das Comunidades baseada no Aspecto Topológico (ontologias 101 e 304)

Na clusterização das comunidades baseada no aspecto topológico, a maior largura de silhueta média obtida foi igual a 0,96, para k igual a quatro, a qual ocorreu ao usar a técnica de clusterização *K-means*. O menor valor de largura de silhueta média foi igual a 0,16, para k igual a dez, obtido com todas as técnicas de clusterização aplicadas, exceto a *K-means*, que gerou uma largura de 0,18. Não foi feita redução de dimensionalidade para o aspecto topológico, pois foram usados apenas quatro atributos para modelar as características estruturais das comunidades das ontologias. Como o *K-means* para k igual a quatro correspondeu à maior largura de silhueta média e por tratar-se de técnica do tipo particional (subseção 3.2), os dados do aspecto topológico formaram *clusters* mais compactos (JAIN; MURTY; FLYNN, 1999), a exemplo do que ocorreu com o aspecto extensional.

10.3 Alinhamento das ontologias 101 e 304

Ao analisar os resultados do alinhamento das ontologias 101 e 304 discriminados na tabela 9.16 (subseção 9.6), verifica-se que os maiores índices de cobertura, precisão e *F-measure* foram os obtidos com base na solução do aspecto terminológico. Observa-se também que, dentre as soluções baseadas em clusterização consensual, a solução baseada em CSPA, HGPA e MCLA superou a baseada em CEB. Estes melhores desempenhos resultam do fato de que as melhores soluções concentraram mais elementos similares das ontologias em um mesmo *cluster*, implicando, assim, em um emparelhamento de maior cobertura, precisão e *F-measure*.

10.4 Discussão dos Principais Resultados de Detecção de Comunidades, de Alinhamento e de Clusterização dos Estudos de Caso

Nesta subseção, é feita a discussão dos principais resultados de detecção de comunidades, de alinhamento e de clusterização para os estudos de caso efetuados. A discussão é feita com base no desempenho das técnicas, mensurado por meio da modularidade, no caso da detecção de comunidades, das métricas de cobertura, precisão e *F-measure*, para o alinhamento, e por meio da largura de silhueta, para a clusterização.

10.4.1 Detecção de Comunidades

A análise da tabela 9.17 (subseção 9.7.1) permite verificar que a técnica de Brandes foi a que preponderou em termos de modularidade, ao ser efetuada a detecção de comunidades nas ontologias dos estudos de caso. Apenas para as ontologias 208, 209 e 222 o método de Potts conseguiu o maior valor, mesmo assim foi igual ao valor da técnica de Brandes. Para a ontologia 230, a otimização gulosa de modularidade atingiu o maior valor. Ainda assim, nesse caso, a técnica de Brandes igualou-se em termos de modularidade. Ou seja, a técnica de Brandes sempre resultou no máximo valor de modularidade em todos os estudos de caso. Isto ocorreu pelos motivos já expostos na subseção 10.1.

10.4.2 Alinhamento

Da análise da tabela do Anexo E, verificam-se algumas constatações em relação ao desempenho das soluções adotadas neste trabalho, em termos de precisão, cobertura e *F-measure*, para os estudos de caso ora efetuados. Para a precisão, as técnicas que tiveram o máximo valor, na maioria dos estudos de caso, foram CSPA, HGPA e MCLA. Isto ocorreu, pois estas técnicas conseguiram agrupar comunidades, concentrando elementos das ontologias de forma que o resultado de alinhamento gerasse um maior percentual de correspondências corretas. Em termos de cobertura, a técnica de CEB conseguiu gerar os máximos valores de cobertura na maioria dos estudos de caso. Isto deveu-se ao fato de que os *clusters* gerados pelo CEB concentrou elementos das ontologias de forma tal que se conseguiu cobrir um

maior número de correspondências do alinhamento de referência. Quanto ao *F-measure*, preponderou a solução baseada no aspecto terminológico, indicando que esse aspecto foi o que melhor conseguiu agrupar os elementos de maneira a gerar um alinhamento com uma melhor ponderação final entre precisão e cobertura.

Ao analisar a tabela 9.20 e o gráfico da figura 9.12 (subseção 9.7.2), pode-se fazer algumas constatações em relação ao comportamento dos valores médios de precisão, cobertura e *F-measure*, em função da quantidade de *clusters* das soluções de alinhamento. Para os estudos de caso deste trabalho, o que se observou foi uma tendência inicial de, ao aumentar a quantidade de *clusters*, diminuir os valores das métricas de alinhamento. Isto aconteceu mais especificamente no intervalo entre dois e quatro *clusters*. Porém, em seguida, o gráfico da figura 9.12 atinge um pico, indicando um ponto ótimo de métricas de alinhamento para a quantidade de cinco *clusters*. Para seis *clusters*, as métricas de alinhamento sofrem um novo decréscimo. Assim, para os estudos de caso deste trabalho, as soluções com cinco *clusters* conseguiram agrupar as comunidades concentrando elementos das ontologias de maneira a gerar, ao mesmo tempo, alinhamentos mais precisos e cobrindo um maior percentual das correspondências dos alinhamentos de referência.

10.4.3 Clusterização

A análise da tabela 9.21 e da figura 9.13 (subseção 9.7.3) aponta que, para os estudos de caso deste trabalho, conforme aumentou-se a quantidade de *clusters* usada como semente para as técnicas de clusterização, a largura de silhueta tendeu a crescer, atingindo um pico para cinco *clusters*. Após esse máximo, as larguras apresentaram uma tendência de diminuição, conforme aumentou-se o número de *clusters*. Assim sendo, para os estudos de caso efetuados, as soluções com cinco *clusters*, comparativamente às outras soluções, foram as que melhor conseguiram reunir, dentro de cada *cluster*, comunidades mais próximas umas das outras (portanto, mais semelhantes entre si) e, ao mesmo tempo, comunidades de diferentes *clusters* ficaram o mais distante possível umas das outras (portanto, com características diferentes entre si).

Ao tabelarem-se as quantidades de vezes em que cada técnica de clusterização propiciou a máxima largura de silhueta (tabela 9.22), observou-se que as técnicas *K-means* e PAM foram as mais frequentes, com 22 e 17 vezes, seguidas

da técnica de Ward, que atingiu essa máxima por 15 vezes. As técnicas *K-means* e PAM são particionais e a de Ward é hierárquica. Este resultado está apontando que, para as maiores larguras de silhueta dos estudos de caso, preponderam *clusters* mais compactos, conforme explicado nas subseções 3.1 e 3.2.

Quanto às técnicas de redução de dimensionalidade ICA e LSA, as quais foram empregadas para reduzir a quantidade de atributos dos aspectos extensional e terminológico das comunidades, observou-se um equilíbrio na quantidade de estudos de caso em que essas técnicas conseguiram repercutir positivamente na clusterização. Conforme indica a tabela 9.23, para o aspecto extensional, verifica-se que a ICA foi por cinco vezes a base para as melhores clusterizações, enquanto a LSA teve um melhor resultado em sete vezes. Portanto, no aspecto extensional, em 42% dos estudos de caso a ICA foi a melhor, enquanto em 58% dos casos a LSA foi a melhor. Já no aspecto terminológico, essa quantidade de vezes foi igual a seis para ambas as técnicas. Conforme exposto em 4.2.2, a técnica de ICA tem melhor resultado se os dados submetidos à redução de dimensionalidade têm uma distribuição probabilística não gaussiana. Já a técnica de LSA, tem maior aderência a dados com distribuição probabilística gaussiana (subseção 4.2.1). Isto denota que, para os aspectos terminológico e extensional, o conjunto de ontologias dos estudos de caso têm características heterogêneas, balanceadas entre dados que, aproximadamente em sua metade, obedecem a distribuições não gaussianas, enquanto a outra metade segue distribuições gaussianas.

10.5 Trabalhos Futuros

Ao analisar as etapas da metodologia, vislumbram-se outros potenciais trabalhos futuros. Na modelagem de ontologias usando-se grafos, por exemplo, podem ser explorados outros relacionamentos entre elementos das ontologias, tais como relações mereológicas (do tipo “part-of”, ou “parte-de”) ou de herança entre propriedades (“subpropertyOf”). Na etapa de detecção de comunidades nos grafos associados às ontologias, outras técnicas podem ser exploradas, tais como as de Blondel et al. (2008). Além disso, na clusterização terminológica, a técnica de Latent Dirichlet Allocation (BLEI; JORDAN, 2003) pode ser aplicada e avaliada na clusterização das comunidades. Também podem ser testadas outras técnicas de

clusterização consensual, tais como as baseadas em clusterização consensual *fuzzy* (DIMITRIADOU; WEINGESSEL; HORNIK, 2002).

11 CONCLUSÕES

Este trabalho teve como objetivo desenvolver uma nova metodologia de emparelhamento de ontologias baseada em clusterização consensual de múltiplos aspectos de comunidades. Por meio da metodologia desenvolvida, foi possível criar um arcabouço metodológico, por meio do qual diferentes técnicas e aspectos podem ser incorporados e testados quanto a seu desempenho na clusterização de comunidades de ontologias e no alinhamento dessas ontologias.

Efetuuou-se, inicialmente, a detecção de comunidades nos grafos associados às ontologias e, em seguida, clusterizaram-se, de forma combinada, múltiplos aspectos das ontologias, de forma a obter uma solução única de clusterização, representativa de um consenso entre os aspectos. Isto foi feito visando reduzir o número de comparações do emparelhamento, uma vez que, na solução com clusterização, comparam-se apenas os elementos das ontologias que pertençam a um mesmo *cluster*, em vez de compararem-se todos os elementos de uma ontologia com todos os elementos da outra ontologia (solução cartesiana).

Para a detecção de comunidades, compararam-se diversas técnicas e concluiu-se que a melhor delas foi a de Brandes, usando-se a modularidade da partição gerada como métrica de comparação entre as técnicas. As comunidades assim obtidas foram consideradas quanto a três aspectos: terminológico, extensional e topológico. Foi feita a clusterização das comunidades segundo cada um desses aspectos. Para a clusterização de cada aspecto, compararam-se diferentes tipos de técnicas de clusterização (hierárquicas e particionais) e selecionou-se a clusterização de maior largura de silhueta média, a qual mede o grau de compactação dos *clusters* e o grau de separação entre *clusters*. As técnicas ICA e LSA foram usadas em caráter auxiliar na tarefa de clusterização, com o intuito de reduzir a dimensionalidade dos aspectos terminológico e extensional. No aspecto topológico, não foi necessária a redução de dimensionalidade, uma vez que foram usados apenas quatro atributos para modelar as comunidades. Obtidas as clusterizações de cada aspecto em isolado, buscou-se uma solução consensual dos aspectos. Para tanto, aplicaram-se as técnicas CEB, CSPA, HGPA e MCLA.

Ao final, procedeu-se ao emparelhamento das classes e propriedades das comunidades das ontologias que pertencessem aos mesmos *clusters*, gerando-se um alinhamento para cada caso. Isto foi feito para as soluções baseadas em clusterização

consensual e também para as soluções baseadas nas clusterizações de cada aspecto tomado individualmente. Compararam-se as métricas de alinhamento (precisão e cobertura) das diferentes soluções. Foram usadas as ontologias do teste *benchmarks* da OAEI para efetuar 12 estudos de caso de alinhamento.

Quanto ao alinhamento, nos estudos de caso deste trabalho, observou-se um destaque para as soluções consensuais. As soluções que geraram maior precisão, na maioria dos estudos de caso, foram CSPA, HGPA e MCLA. Assim, essas soluções resultaram em um maior percentual de correspondências corretas, em relação ao conjunto total de correspondências retornado por essas soluções. A solução de CEB, por sua vez, foi a que mais vezes conseguiu cobrir um maior número de correspondências do alinhamento de referência, portanto, tendo maiores valores de cobertura. Ao considerar simultaneamente a precisão e cobertura, por meio do *F-measure*, a solução baseada no aspecto terminológico foi a que preponderou, registrando um maior número de ocorrências em que teve os melhores índices dessa métrica. Além disso, observou-se um ponto ótimo para as métricas de alinhamento ao serem adotados cinco *clusters* para as clusterizações dos aspectos, denotando que essa configuração de *clusters* foi a que melhor conseguiu agrupar elementos das ontologias de forma a gerar mais precisão e cobertura nos alinhamentos dos estudos de caso efetuados.

Em termos de clusterização, observou-se nos estudos de caso uma preponderância de ocorrências em que técnicas particionais geraram as maiores larguras de silhueta, apontando que os *clusters* gerados foram, na sua maioria, mais compactos, uma vez que técnicas particionais têm melhor aderência a esses tipos de *clusters*. Verificou-se também que as máximas larguras de silhueta foram obtidas pelas soluções de cinco *clusters*, ou seja, trata-se de soluções para os estudos de caso em que as comunidades de uma mesmo *cluster* são bastante semelhantes, enquanto as comunidades de diferentes *clusters* são bastante diferentes entre si, indicando uma boa detecção de padrões. Quanto às técnicas de redução de dimensionalidade ICA e LSA, usadas nas clusterizações para reduzir os espaços dimensionais dos atributos que modelaram os aspectos terminológico e extensional, observou-se um equilíbrio nos resultados dos estudos de caso. No aspecto terminológico, em 50% dos estudos de caso, tanto o ICA quanto o LSA geraram um espaço dimensional que serviu de base para as melhores clusterizações (maiores larguras de silhueta). No aspecto extensional, em 58% dos estudos de caso houve

preponderância do ICA e em 42%, o destaque foi para o LSA. Assim sendo, a possibilidade de usar, por meio da metodologia proposta, alternativamente, tanto o ICA como o LSA, contribuíram para atingir os melhores índices de largura de silhueta.

Conclui-se que a principal contribuição deste trabalho está relacionada à metodologia proposta. A metodologia desenvolvida constitui um arcabouço, no qual novos aspectos das ontologias podem ser adicionados e novas técnicas podem ser testadas, nas diferentes fases da metodologia. O arcabouço estrutura-se principalmente por meio da implementação feita em Java, exteriorizada pela plataforma do Eclipse, e que é responsável por orquestrar e executar a sequência de passos e comandos envolvidos na metodologia. No Eclipse, também é feita a extração de características e de atributos usados na modelagem dos aspectos das comunidades, além de ser feito o emparelhamento. Os resultados da aplicação de diferentes técnicas podem ser obtidos externamente ao Eclipse, de maneira independente, com uso de *scripts* em R e no Matlab, por exemplo. Isso permite utilizar outras técnicas de detecção de comunidades, de redução de dimensionalidade, de clusterização dos aspectos e de clusterização consensual.

Além disso, deve-se ressaltar que a metodologia ora proposta pode ser aplicada em ontologias baseadas em vocabulários de outras línguas, não só a língua inglesa, como foi o caso da terminologia existente nas ontologias dos estudos de caso. Os pacotes da ferramenta R, por exemplo, oferecem suporte a outras línguas (portuguesa, francesa, alemã, italiana etc.), flexibilizando a aplicação do Modelo do Espaço Vetorial a conjuntos de termos em outras línguas. As técnicas de similaridade de caracteres usadas no emparelhamento são independentes da linguagem e existe o Wordnet em português, espanhol, italiano, francês etc.

Ao lado disso, tem-se também que a aplicação da metodologia desenvolvida neste trabalho, considerando-se apenas o aspecto terminológico, tem sua aplicabilidade também na detecção de plágio. No Modelo do Espaço Vetorial, documentos idênticos têm ângulo zero entre si, ou seja, cosseno igual a um ou distância igual a zero. Um exemplo de um trabalho específico sobre detecção de plágio com base no Modelo do Espaço Vetorial é o de Jiffriya, Jahan e Ragel (2014).

12 REFERÊNCIAS

- AIROLDI, E. M. 2007. Getting Started in Probabilistic Graphical Models. PLoS Computational Biology 3(12): e252, 2007.
- ALGERGAWY, A.; MASSMANN, S.; RAHM, E. A Clustering-Based Approach for Large-Scale Ontology Matching. Lecture Notes in Computer Science, 2011.
- ALGERGAWY, A.; MOAWED, S.; SARHAN, A.; ELDOSOUKY, A.; SAAKE, G. Improving Clustering-Based Schema Matching Using Latent Semantic Indexing. Transactions on Large-Scale Data- and Knowledge-Centered Systems XV, V. 8920, Lecture Notes in Computer Science, p. 102-123. Springer, 2015.
- APACHE JENA. A Free and Open Source Java Framework for Building Semantic Web and Linked Data Applications. Ferramenta. Disponível em <<https://jena.apache.org/>>. Acesso em 23/06/2015.
- AUER, S.; ERMILOV, I.; LEHMANN, J.; MARTIN, M. LODStats: a Statement-stream-based Approach for Gathering Comprehensive Statistics about RDF Datasets. Disponível em: <<http://stats.lod2.eu>> . Acesso em 12/11/2016.
- BAADER, F.; NUTT, W. Basic Description Logics. In: BAADER, F.; CALVANESE, D.; MCGUINNESS, D. L.; NARDI, D.; PATEL-SCHNEIDER, P.F. The Description Logic Handbook: Theory, Implementation and Applications. Cambridge University Press, p. 47-100, 2003.
- BANERJEE, A. Mixed Membership Models. São Paulo. UNICAMP, 2012. Slides da parte 10 das notas de aula do curso IA900 – Mineração de Dados Escalável e Aprendizado de Máquina para Dados Complexos, São Paulo, 2012.
- BELL, AJ; SEJNOWSKI, TJ. An Information-Maximization Approach to Blind Separation and Blind Deconvolution. In: Neural Computation (7), p. 1129-1159, 1995.
- BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The Semantic Web. In: Scientific American, 2001.
- BIBTEX. Your BibTeX Resource. Ferramenta. Disponível em: <<http://www.bibtex.org/>>. Acesso em 19/11/2016.
- BIZER, C.; HEATH, T.; BERNERS-LEE, T. Linked Data - The Story So Far. In: International Journal on Semantic Web and Information Systems, 5(3): p. 1-22, 2009.
- BLEI, D. M.; NG, A.; JORDAN, M. I. Latent Dirichlet Allocation. In: Journal of Machine Learning Research (3), p. 993-1022, 2003.
- BLONDEL, V. D.; GUILLAUME, J.; LAMBIOTTE, R.; LEFEBVRE, E. Fast Unfolding of Communities in Large Networks. In: Journal of Statistical Mechanics: Theory and Experiment (10), P10008, 2008.

BRANDES, U.; DELLING, D.; GAERTLER, M.; GORKE, R.; HOEFER, M.; NIKOLOSKE, Z.; WAGNER, D. On Modularity Clustering. In: IEEE Transactions on Knowledge and Data Engineering 20(2): p. 172-188, 2008.

BREITMAN, K. K.; CASANOVA, M. A.; TRUSZKOWSKI, W. Semantic Web: Concepts, Technologies and Applications. Springer, 2007.

CARDOSO, J. Developing an OWL Ontology for E-Tourism. In: CARDOSO, J.; SHETH, P. Semantic Web Services, Processes and Applications. Springer, p. 249-282, 2010.

CASELLA, G.; GEORGE, E. I. Explaining the Gibbs Sampler. In: The American Statistician 46 (3): p. 167, 1992.

CLAUSET, A.; NEWMAN, M. E. J.; MOORE, C. Finding Community Structure in Very Large Networks. In: Physical Review E, 70:066111, 2004.

COMON, P. Independent Component Analysis: a New Concept ? In: Signal Processing (36), p. 287–314, 1994.

COSKUN, G.; ROTHE, M.; TEYMOURIAN, K.; PASCHKE, A. Applying Community Detection Algorithms on Ontologies for Identifying Concept Groups. In: Proceedings of the 5th International Workshop on Modular Ontologies, 2011.

COVER, T. M., THOMAS, J. A. Elements of Information Theory. Wiley-Interscience New York, 1991.

DEAN, J.; GHEMAWAT, S.. MapReduce: Simplified Data Processing on Large Clusters. In: Communications of the ACM, 51(1), p. 107-113, 2008.

DIMITRIADOU, E.; WEINGESSEL, A.; HORNIK, H. A Combination Scheme for Fuzzy Clustering. In: International Journal of Pattern Recognition and Artificial Intelligence, 16: p. 901–912, 2002.

DO, H. H.; RAHM, E. Matching large schemas: Approaches and evaluation. In: Information Systems, 32(6): p.857–885, 2007.

DUNN, J. C. Well Separated Clusters and Fuzzy Partitions. In: Journal on Cybernetics, (4), p. 95-104, 1974.

ECLIPSE MARS 2 PACKAGES. Ferramenta. Disponível em: <<http://www.eclipse.org/downloads/packages/release/Mars/2>> . Acesso em 19/05/2016.

EUZENAT, J.; ALLOCA, C.; DAVID, J.; D'AQUIN, M.; DUC, C.L.; SVÁB-ZAMAZAL, O. Ontology distances for contextualization. In: Report 3.3.4, IST NeOn IP, NeOn, 2009.

EUZENAT, J.; SHVAIKO, P. Ontology Matching. Springer, 2nd edition, 2013.

FERRARA, A.; GENTA, L.; MONTANELLI, S.; CASTANO, S. Dimensional Clustering of Linked Data: Techniques and Applications. In: HAMEURLAIN, A.; KÜNG, J.; WAGNER, R.; BIANCHINI, D.; DE ANTONELLIS, V.; DE VIRGILIO, R. (eds.). Transactions on Large-Scale Data and Knowledge-Centered Systems XIX. Springer-Verlag Berlin Heidelberg, p. 55-86, 2015.

FORTUNATO, S. Community Detection in Graphs. In: Physics Reports 486 (3), p. 75-174, 2010.

GATENS, W.; KONEV, B.; WOLTER, F. Module Extraction for Acyclic Ontologies. In: Proceedings of the 7th International Workshop on Modular Ontologies, 2013.

GENESERETH, M. R.; NILSSON, N. J. Logical Foundations of Artificial Intelligence. Morgan Kaufmann Publishers Inc., 1987.

GHOSH, J.; ACHARYA, A. Cluster Ensembles: Theory and Applications. In: AGGARWAL, C. C.; REDDY, C. K. (eds.). Data Clustering: Algorithms and Applications. Chapman & Hall/CRC, p.551-570, 2014.

GRUBER, T. R. A Translational Approach to Portable Ontologies. In: Knowledge Acquisition, 5 (22): p. 199-220, 1993.

HAMDI, F.; SAFAR, B.; REYNAUD, C.; ZARGAYOUNA, H. Alignment-Based Partitioning of Large-scale Ontologies. In: Advances in Knowledge Discovery And Management, p. 251-269, 2010.

HANDL, J.; KNOWLES, J.; KELL, D. B. Computational Cluster Validation in Post-genomic Data Analysis. In: Bioninformatics, 21 (15): p. 3201-3212, 2005.

HARARY, F. Graph Theory. Reading, MA. Addison-Wesley, 1994.

HENDLER, J. Agents and the Semantic Web. IEEE Intelligent Systems, p. 30-37, 2001.

HITZLER, P.; KRÖTZSCH, M.; PARSIA, B.; PATEL-SCHNEIDER, P. F.; RUDOLPH, S. OWL 2 Web Ontology Language Primer (Second Edition). W3C Recommendation. Elaborada em 11/12/2012. Disponível em < <http://www.w3.org/TR/owl2-primer/>>. Acessado em 28/04/2015.

HOFMANN, T. Unsupervised Learning by Probabilistic Latent Semantic Analysis. In: Machine Learning, 42 (1): p. 177–296, 2001.

HONKELA, T.; HYVÄRINEN, A.; VÄYRYNEN, J.J. WordICA - Emergence of Linguistic Representations for Words by Independent Component Analysis. In: Natural Language Engineering (16), p. 277–308, 2010.

HU, B.; KALFOGLOU, Y.; ALANI, H.; DUPPLAW, D.; LEWIS, P.; SHADBOLT, N. Semantic Metrics. In: Staab, S.; Svátek, V. (eds.). EKAW 2006. LNCS (LNAI), vol. 4248, Springer, Heidelberg, p. 166–181, 2006.

HU, B.; RODRIGUES, E. M.; VIEL, E. Programmable Analytics for Linked Open Data. In: Proceedings of CEUR Workshop, vol. 1184, Linked Data on the Web (LDOW), 2014.

HU, W.; QU, Y.; CHENG, G. Matching Large Ontologies: A Divide-and-conquer Approach. In: Data & Knowledge Engineering Revue 67, p. 140-160, 2008.

HYVÄRINEN, A.; OJA, E. Independent Component Analysis: Algorithms and Applications. In: Neural Networks (13), p. 411-430, 2000.

HYVÄRINEN A. Independent Component Analysis: Recent Advances. In: Philosophical Transactions of the Real Society A 371: 20110534, 2013.

IPPOLITO, A.; ALMEIDA JÚNIOR, J. R. Ontology Matching based on Multi-Aspect Consensus Clustering of Communities. In: Hammoudi, S.; Maciaszek, L. A.; Missikoff, M.; Camp, O.; Cordeiro, J. (eds.), ICEIS 2016 – Proceedings of the 18th International Conference on Enterprise Information Systems, Volume 2, SciTePress, p. 321-326, 2016.

ISELE, R.; JENTZSCH, A.; BIZER, C. Efficient Multidimensional Blocking for Link Discovery without Losing Recall. In: International Workshop on the Web and Databases (WebDB), 2011.

JACCARD, P. Distribution de la flore alpine dans le bassin des Dranses et dans quelques régions voisines. Bulletin de la Société Vaudoise des Sciences Naturelles 37, p. 241-272, 1901.

JAIN, A. K.; MURTY, M. N.; FLYNN, P. J. Data Clustering: A Review. In: ACM Computing Surveys, Vol. 31, N° 3, 1999.

JAIN, A. K. Data Clustering: 50 Years beyond K-means. In: Pattern Recognition Letters 31, p. 651-666, 2010.

JAVA. Linguagem de Programação. Disponível em < https://www.java.com/pt_BR/ >. Acesso em 19/05/2016.

JENTZSCH, A.; CYGANIAK, R.; BIZER, C. State of the LOD Cloud. Versão 0.3. Disponível em <<http://lod-cloud.net/state/>>. Atualizada em: 19/09/2011. Acesso em: 12/11/2016.

JIFFRIYA, M. A. C.; JAHAN, M. A. C. A.; RAGEL, R. G. Plagiarism Detection on Electronic Text based Assignments using Vector Space Model. In: International Conference on Information and Automation for Sustainability, 2014.

JORDAN, M.; GHAHRAMANI, Z.; JAAKKOLA, T.; SAUL L. Introduction to Variational Methods for Graphical Models. Machine Learning 37: p. 183–233, 1999.

KARYPIS, G., AGGARWAL, R., KUMAR, V., SHEKHAR, S. Multilevel Hypergraph Partitioning: Applications in VLSI Domain. In: Proceedings of the Design and Automation Conference, 1997.

KARYPIS, G., KUMAR, V. A Fast and High Quality Multilevel Scheme for Partitioning Irregular Graphs. In: SIAM Journal on Scientific Computing, 20(1):359-392, 1998.

KAUFMAN, L.; ROUSSEEUW, P. Finding Groups in Data: An Introduction to Cluster Analysis: Wiley, 1990.

KRÖTZSCH, M.; SIMANČÍK, F.; HORROCKS, I. A Description Logic Primer. In: CoRR, abs/1201.4089, 2012.

LANDAUER, T. K.; FOLTZ, P. W.; LAHAM, D. Introduction to Latent Semantic Analysis. Discourse Processes, (25) 259-284, 1998.

LEVIN, L. A. Fundamentals of Computing. Boston University, 2016. Disponível em: <<http://www.cs.bu.edu/fac/ln/toc/z.pdf>>. Acesso em 07/06/2017.

MANNING, C. D.; RAGHAVAN, P.; SCHÜTZE, H. An Introduction to Information Retrieval. Online edition. Cambridge, 2009.

MANOLA, F.; MILLER, E.; MC BRIDE, B. RDF 1.1 - W3C Working Group Note. Elaborada em 24/06/2014. Disponível em <<http://www.w3.org/TR/rdf11-primer/>>. Acesso em 29/04/2015.

MARTIN, D.I.; BERRY, M. W. Mathematical Foundations Behind Latent Semantic Analysis. In: LANDAUER, T.; MCNAMARA, D.; DENNIS, S.; KINTSCH, W. (eds.), Latent Semantic Analysis: A Road to Meaning. Lawrence Erlbaum, 2006.

MATLAB. Ferramenta. Versão R2016a (9.0.0.341360) para estudantes. Disponível em: <<http://www.mathworks.com/products/matlab/>>. Acesso em 21/08/2016.

MC GUINNESS, D. L.; HARMELEN, F. V. OWL Web Ontology Language Overview. W3C Recommendation, 2004.

MILLER, G. A. WordNet: a Lexical Database for English. In: Communications of the ACM, Vol. 38 (11), p. 39–41, 1995.

MITCHELL, T. M. Machine Learning. M^c Graw Hill, 1997.

NADESAN, B. S. G.; ALAGARSAMY, K. Improved N-gram Extraction and Substring Removal Algorithms with Efficient Dual-Pivot Sorting in Mining Data for UML Models. In: International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 7, 2013.

NEWMAN, M. E. J. Fast Algorithm for Detecting Community Structure in Networks. In: b 69, 066133, 2004.

NEWMAN, M. E. J.; GIRVAN, M. Finding and Evaluating Community Structure in Networks. In: Physical Review E, 69:026113, 2004.

NEWMAN, M. E. J. Finding Community Structure in Networks Using the Eigenvectors of Matrices. In: Physical Review E, 74(3):036104, 2006.

ONTOLOGY ALIGNMENT EVALUATION INITIATIVE. Página web. Disponível em: <<http://oaei.ontologymatching.org/>> Acesso em 26/06/2017.

OBJECTAID. Ferramenta. Versão 1.1.11 de 11/09/2016. Disponível em: <<http://objectaid.com/>>. Acesso em 16/11/2016.

PAPADIMITRIOU, C. H.; STEIGLITZ, K. Combinatorial Optimization: Algorithms and Complexity. Dover Publications Inc., 1998.

PHAM, D-T; GARRAT, P. Blind Separation of Mixture of Independent Sources Through a Quasi-Maximum Likelihood Approach. In: IEEE Transactions on Signal Processing (45), p. 1712–1725, 1997.

PONS, P.; LATAPY, M. Computing Communities in Large Networks Using Random Walks. In: Journal of Graph Algorithms and Applications, Vol. 10, n.º 2, p. 191-218, 2005.

PORTER, M. A.; ONNELA, J.; MUCHA, P. J. Communities in Networks. In: Notices of the American Mathematical Society, Vol. 56, No. 9, p. 1-26, 2009.

POSPESEL, H. Introduction to Logic: Propositional Logic. Prentice Hall, 1979.

PROTEGÉ. Versão 5.0.0. Ferramenta. Disponível em: <<http://protege.stanford.edu/products.php>> . Acesso em 19/08/2015.

QU, Y.Z., HU, W., CHENG, G. Constructing Virtual Documents for Ontology Matching. In: Proceedings of the 15th International Conference on World Wide Web, p. 23-31, 2006.

RAHM, E. Towards Large-Scale Schema and Ontology Matching. In: BELLAHSENE, Z.; BONIFATI, A; RAHM, E. Schema Matching and Mapping. Springer, p. 3-27, 2011.

RAND, W. Objective Criteria for the Evaluation of Clustering Methods. In: Journal of the American Statistical Association, 66, p. 846–850, 1971.

REDDY, K. P.; KITSUREGAWA, M.; SREEKANTH, P.; RAO, S.S. A Graph Based Approach to Extract a Neighborhood Customer Community for Collaborative Filtering. In: Proceedings of the 2nd International Workshop on Databases in Networked Information Systems, p. 188-200. Springer-Verlag, London, 2002.

REICHARDT, J.; BORNHOLDT, S. Statistical Mechanics of Community Detection. In: Physical Review E, 74, 016110, 2006.

ROUSSEEUW, P. J. Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis. In: Journal of Computational and Applied Mathematics, (20), p.53-65, 1987.

RSTUDIO. Ferramenta. Disponível em <<https://www.rstudio.com/>>. Acesso em 19/05/2016.

SCHAEFFER, S. E. Graph Clustering. In: Computer Science Review 1, p. 27-64, 2007.

SCHMACHTENBERG, M.; BIZER, C.; PAULHEIM, H. State of the LOD Cloud 2014. Disponível em <<http://linkeddatacatalog.dws.informatik.uni-mannheim.de/state/>>. Atualizada em: 30/08/2014. Acesso em: 30/04/2015.

SEMANTIC WEB STANDARDS. Disponível em: <http://semanticweb.org/wiki/Semantic_Web_standards>. Última atualização em: 14/11/2011. Acesso em: 30/04/2015.

SHVAIKO, P. Iterative Schema-Based Semantic Matching. 248 f. Tese. Universidade de Trento, Itália, 2006a.

SHVAIKO, P. Ontology Matching. In: Semantic Web Applications and Perspectives (SWAP), 2006b.

SHVAIKO, P.; EUZENAT, J. Ontology Matching: State of the Art and Future Challenges. In: IEEE Transactions on Knowledge and Data Engineering, 25(1): p. 158-176, 2013.

SOKAL, R. R.; MICHENER, C. D. A Statistical Method for Evaluating Systematic Relationships. University of Kansas Science Bulletin 38, p. 1409-1438, 1958.

STEINBACH, M.; ERTÖZ, L.; KUMAR, V. The Challenges of Clustering High-Dimensional Data. In: New Vistas in Statistical Physics: Applications in Econophysics, Bioinformatics and Pattern Recognition. Springer-Verlag, 2003.

STREHL, A.; GHOSH, J. Cluster Ensembles - A Knowledge Reuse Framework for Combining Multiple Partitions. In: Journal of Machine Learning Research, Volume 3, p. 583-617, 2003.

STREHL, A. Cluster Analysis and Cluster Ensemble Software. Disponível em: <<http://www.strehl.com/soft.html>>. Acesso em 19/11/2016. Versão 2.0 de 01/05/2011.

STUCKENSCHMIDT, H.; SPACCAPIETRA, S. Modular Ontologies: Concepts, Theories and Techniques for Knowledge Modularization. Springer, 2009.

TANG, B.; SHEPHERD M.; MILIOS, E.; HEYWOOD, M. I. Comparing and Combining Dimension Reduction Techniques for Efficient Text Clustering. In: Proc. Of the International Workshop on Feature Selection for Data Mining: Interfacing Machine Learning and Statistics, p. 17-26, 2005.

THE R PROJECT FOR STATISTICAL COMPUTING. Ferramenta. Disponível em:<<http://www.r-project.org/>>. Acesso em 16/05/2014.

TRAN, T.; WANG, H.; HAASE, P. Hermes: Data Web Search on a Pay-as-you-go Integration Infrastructure. In: Web Semantics: Science, Services and Agents on the World Wide Web 7(3), p. 189–203, 2009.

UKKONEN, E. On Approximate String Matching. In: Information Control, Volume 64, Issues 1-3, p. 100-118, Elsevier, 1983.

VAN RIJSBERGEN, C. J. Information Retrieval. 2nd Edition. Butterworth, 1979.

VESCOVO, C. D.; KLINOV, P.; PARSIA, B.; SATTler, U.; SCHNEIDER, T.; TSARKOV, D. Syntactic vs. Semantic Locality: How Good Is a Cheap Approximation? In: Proceedings of the 6th International Workshop on Modular Ontologies, 2012.

W3C OWL WORKING GROUP. OWL 2 Web Ontology Language Document Overview (Second Edition). W3C Recommendation. Elaborada em: 11/12/2012. Disponível em: <<http://www.w3.org/TR/owl2-overview/>>. Acesso em: 26/04/2015.

WANG, H; SHAN, H.; BANERJEE, A. Bayesian Cluster Ensembles. In: Statistical Analysis and Data Mining: The ASA Data Science Journal, Volume 4, Issue 1, p. 54-70. Wiley Periodicals, 2011.

WANG, H; SHAN, H.; BANERJEE, A. Code for Bayesian Cluster Ensembles. Disponível em: <http://www-users.cs.umn.edu/~shan/bce_code.html>. Acesso em 19/11/2016.

WARD, J. H. Hierarchical Grouping to Optimize an Objective Function. In: Journal of the American Statistical Association, Volume 58, Issue 301, p. 236-244, 1963.

WASSERMAN, S.; FAUST, K. Social Network Analysis: Methods and Applications. Cambridge University Press, 1994.

WEST, D. B. Introduction to Graph Theory. Upper Saddle River, NJ. Prentice Hall, 1996.

WU, Z.; PALMER, M. Verbs semantics and lexical selection. In: Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics, p. 133-138, 1994.

WU, C; POTDAR, V.; CHANG, E. Latent Semantic Analysis – The Dynamics of Semantics Web Services Discovery. In: DILLON, T.; CHANG, E.; MEERSMAN, R.; SYCARA, K. (Eds.): Advances in Web Semantics I, LNCS 4891, p. 346–373, 2008.

XU, R.; WUNSCH, D. Survey of Clustering Algorithms. In: IEEE Transactions on Neural Networks, Vol. 6, N^o 3, 2005.

YURUK, N.; METE, M.; XU, X.; SCHWEIGER, T. A. J. AHSCAN: Agglomerative Hierarchical Structural Clustering Algorithm for Networks. In: International Conference on Advances in Social Network Analysis and Mining, p. 72–77, 2009.

ZHANG, H; HU, W.; QU, Y. VDoc+: a Virtual Document Based Approach for Matching Large Ontologies using MapReduce. Journal of Zhejiang University-SCIENCE C (Computers & Electronics), 2012.

ZHAO, Y.; KARYPIS, G. Criterion Functions for Document Clustering: Experiments and Analysis. In: Technical Report TR 01-40, Department of Computer Science, University of Minnesota, 2001.

ANEXO A – Grafos das ontologias 101 e 304

Figura A1 – Grafo da ontologia 101.

Fonte: Protegé (2015)

Figura A2 – Grafo da ontologia 304.

Fonte: Protegé (2015)

Note: all rdfs:comment values for the entries come from
<http://newton.ex.ac.uk/tex/pack/bibtex/btxdoc/node6.html>.
 -->

```

<owl:Class rdf:ID="Reference">
  <rdfs:label xml:lang="en">Reference</rdfs:label>
  <rdfs:comment xml:lang="en">Base class for all entries</rdfs:comment>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#date" />
      <owl:maxCardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:maxCardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#title" />
      <owl:maxCardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:maxCardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#humanCreator" />
      <owl:maxCardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:maxCardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
</owl:Class>

<owl:Class rdf:ID="Article">
  <rdfs:subClassOf rdf:resource="#Part" />
  <rdfs:label xml:lang="en">Article</rdfs:label>
  <rdfs:comment xml:lang="en">An article from a journal or magazine.</rdfs:comment>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#author" />
      <owl:cardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:cardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#pages" />
      <owl:cardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:cardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#journal" />
      <owl:cardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:cardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#date" />
      <owl:cardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:cardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#date" />
      <owl:allValuesFrom rdf:resource="#Date" />
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#number" />
      <owl:maxCardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:maxCardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#volume" />
      <owl:maxCardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:maxCardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
</owl:Class>

<owl:Class rdf:ID="Book">
  <rdfs:subClassOf rdf:resource="#Reference" />

```

```

<rdfs:label xml:lang="en">Book</rdfs:label>
<rdfs:comment xml:lang="en">A book that may be a monograph or a collection of written
texts.</rdfs:comment>
<rdfs:subClassOf>
  <owl:Restriction>
    <owl:onProperty rdf:resource="#title" />
    <owl:cardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:cardinality>
  </owl:Restriction>
</rdfs:subClassOf>
<rdfs:subClassOf>
  <owl:Restriction>
    <owl:onProperty rdf:resource="#volume" />
    <owl:maxCardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:maxCardinality>
  </owl:Restriction>
</rdfs:subClassOf>
<rdfs:subClassOf>
  <owl:Restriction>
    <owl:onProperty rdf:resource="#publisher" />
    <owl:maxCardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:maxCardinality>
  </owl:Restriction>
</rdfs:subClassOf>
<rdfs:subClassOf>
  <owl:Restriction>
    <owl:onProperty rdf:resource="#series" />
    <owl:maxCardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:maxCardinality>
  </owl:Restriction>
</rdfs:subClassOf>
<rdfs:subClassOf>
  <owl:Restriction>
    <owl:onProperty rdf:resource="#date" />
    <owl:cardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:cardinality>
  </owl:Restriction>
</rdfs:subClassOf>
<rdfs:subClassOf>
  <owl:Restriction>
    <owl:onProperty rdf:resource="#date" />
    <owl:allValuesFrom rdf:resource="#Date" />
  </owl:Restriction>
</rdfs:subClassOf>
<rdfs:subClassOf>
  <owl:Restriction>
    <owl:onProperty rdf:resource="#author" />
    <owl:cardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:cardinality>
  </owl:Restriction>
</rdfs:subClassOf>
<rdfs:subClassOf>
  <owl:Restriction>
    <owl:onProperty rdf:resource="#edition" />
    <owl:maxCardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:maxCardinality>
  </owl:Restriction>
</rdfs:subClassOf>
</owl:Class>

<owl:Class rdf:ID="Monograph">
  <rdfs:subClassOf rdf:resource="#Book" />
  <rdfs:label xml:lang="en">Monograph</rdfs:label>
  <rdfs:comment xml:lang="en">A book that is a single entity, as opposed to a
collection.</rdfs:comment>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#chapters" />
      <owl:allValuesFrom rdf:resource="#Chapter"/>
    </owl:Restriction>
  </rdfs:subClassOf>
</owl:Class>

<owl:Class rdf:ID="Collection">
  <rdfs:subClassOf rdf:resource="#Book" />
  <rdfs:label xml:lang="en">Collection</rdfs:label>
  <rdfs:comment xml:lang="en">A book that is collection of texts or articles.</rdfs:comment>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#chapters" />
      <owl:allValuesFrom rdf:resource="#Chapter"/>
    </owl:Restriction>
  </rdfs:subClassOf>
</owl:Class>

```

```

    <owl:Restriction>
      <owl:onProperty rdf:resource="#parts" />
      <owl:allValuesFrom rdf:resource="#InCollection"/>
    </owl:Restriction>
  </rdfs:subClassOf>
</owl:Class>

<owl:Class rdf:ID="Informal">
  <rdfs:subClassOf rdf:resource="#Reference" />
  <rdfs:label xml:lang="en">Informal</rdfs:label>
  <rdfs:comment xml:lang="en">A document that was informally published or not
published.</rdfs:comment>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#title" />
      <owl:cardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:cardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
</owl:Class>

<owl:Class rdf:ID="Booklet">
  <rdfs:subClassOf rdf:resource="#Informal" />
  <rdfs:label xml:lang="en">Booklet</rdfs:label>
  <rdfs:comment xml:lang="en">A work that is printed and bound, but without a named
publisher or sponsoring institution.</rdfs:comment>
</owl:Class>

<owl:Class rdf:ID="Part">
  <rdfs:subClassOf rdf:resource="#Reference" />
  <rdfs:label xml:lang="en">Part</rdfs:label>
  <rdfs:comment xml:lang="en">A part of something (either Book or
Proceedings).</rdfs:comment>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#pages" />
      <owl:maxCardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:maxCardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#title" />
      <owl:cardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:cardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
  <!-- This is forbidden for transitive properties in OWL-DL -->
  <!--rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#isPartOf" />
      <owl:cardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:cardinality>
    </owl:Restriction>
  </rdfs:subClassOf-->
</owl:Class>

<owl:Class rdf:ID="Chapter">
  <rdfs:subClassOf rdf:resource="#Part" />
  <rdfs:label xml:lang="en">BookPart</rdfs:label>
  <rdfs:comment xml:lang="en">A chapter (or section or whatever) of a book having its own
title.</rdfs:comment>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#chapter" />
      <owl:maxCardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:maxCardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#chapter" />
      <owl:maxCardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:maxCardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
</owl:Class>

<owl:Class rdf:ID="InBook">
  <rdfs:subClassOf rdf:resource="#Part" />
  <rdfs:label xml:lang="en">InBook</rdfs:label>
  <rdfs:comment xml:lang="en">A subpart of a book given by a range of pages.</rdfs:comment>
  <rdfs:subClassOf>

```

```

    <owl:Restriction>
      <!-- author OR editor (given in the Book description) ??? -->
      <owl:onProperty rdf:resource="#author" />
      <owl:cardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:cardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
</rdfs:subClassOf>
  <owl:Restriction>
    <owl:onProperty rdf:resource="#pages" />
    <owl:minCardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:minCardinality>
  </owl:Restriction>
</rdfs:subClassOf>
</rdfs:subClassOf>
  <owl:Restriction>
    <owl:onProperty rdf:resource="#book" />
    <owl:cardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:cardinality>
  </owl:Restriction>
</rdfs:subClassOf>
</owl:Class>

<owl:Class rdf:ID="InCollection">
  <rdfs:subClassOf rdf:resource="#Part" />
  <rdfs:label xml:lang="en">Incollection</rdfs:label>
  <rdfs:comment xml:lang="en">A part of a book having its own title.</rdfs:comment>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#author" />
      <owl:cardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:cardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#collection" />
      <owl:cardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:cardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
</owl:Class>

<owl:Class rdf:ID="InProceedings"> <!-- Communication -->
  <rdfs:subClassOf rdf:resource="#Part" />
  <rdfs:label xml:lang="en">InProceedings</rdfs:label>
  <rdfs:comment xml:lang="en">An article in a conference proceedings.</rdfs:comment>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#author" />
      <owl:cardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:cardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#proceedings" />
      <owl:cardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:cardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
</owl:Class>

<owl:Class rdf:ID="LectureNotes">
  <rdfs:subClassOf rdf:resource="#Informal" />
  <rdfs:label xml:lang="en">LectureNotes</rdfs:label>
  <rdfs:comment xml:lang="en">Lecture notes.</rdfs:comment>
</owl:Class>

<owl:Class rdf:ID="Manual">
  <rdfs:subClassOf rdf:resource="#Informal" />
  <rdfs:label xml:lang="en">Manual</rdfs:label>
  <rdfs:comment xml:lang="en">Technical documentation.</rdfs:comment>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#organization" />
      <owl:maxCardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:maxCardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#edition" />
      <owl:maxCardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:maxCardinality>
    </owl:Restriction>
  </rdfs:subClassOf>

```

```

</rdfs:subClassOf>
<rdfs:subClassOf>
  <owl:Restriction>
    <owl:onProperty rdf:resource="#title" />
    <owl:cardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:cardinality>
  </owl:Restriction>
</rdfs:subClassOf>
</owl:Class>

<owl:Class rdf:ID="Academic">
  <rdfs:subClassOf rdf:resource="#Reference" />
  <rdfs:label xml:lang="en">Academic</rdfs:label>
  <rdfs:comment xml:lang="en">A Master's or PhD thesis.</rdfs:comment>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#author" />
      <owl:cardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:cardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#title" />
      <owl:cardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:cardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#school" />
      <owl:cardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:cardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#date" />
      <owl:cardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:cardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
</owl:Class>

<owl:Class rdf:ID="MastersThesis">
  <rdfs:subClassOf rdf:resource="#Academic" />
  <rdfs:label xml:lang="en">MastersThesis</rdfs:label>
  <rdfs:comment xml:lang="en">A Master's thesis.</rdfs:comment>
</owl:Class>

<owl:Class rdf:ID="PhdThesis">
  <rdfs:subClassOf rdf:resource="#Academic" />
  <rdfs:label xml:lang="en">PhdThesis</rdfs:label>
  <rdfs:comment xml:lang="en">A PhD thesis.</rdfs:comment>
</owl:Class>

<owl:Class rdf:ID="Misc">
  <rdfs:subClassOf rdf:resource="#Reference" />
  <rdfs:label xml:lang="en">Misc</rdfs:label>
  <rdfs:comment xml:lang="en">Use this type when nothing else fits.</rdfs:comment>
</owl:Class>

<owl:Class rdf:ID="Proceedings">
  <rdfs:subClassOf rdf:resource="#Book" />
  <rdfs:label xml:lang="en">Proceedings</rdfs:label>
  <rdfs:comment xml:lang="en">The proceedings of a conference.</rdfs:comment>
  <rdfs:comment xml:lang="en">A proceedings may be implicitly defined with an
inproceedings.</rdfs:comment>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#communications" />
      <owl:allValuesFrom rdf:resource="#InProceedings"/>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#event" />
      <owl:maxCardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:maxCardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>

```



```

        <owl:onProperty rdf:resource="#event" />
        <owl:allValuesFrom rdf:resource="#Conference"/>
    </owl:Restriction>
</rdfs:subClassOf>
<rdfs:subClassOf>
    <owl:Restriction>
        <owl:onProperty rdf:resource="#editor" />
        <owl:maxCardinality rdf:datatype="xsd:nonNegativeInteger">1</owl:maxCardinality>
    </owl:Restriction>
</rdfs:subClassOf>
<rdfs:subClassOf>
    <owl:Restriction>
        <owl:onProperty rdf:resource="#organization" />
        <owl:maxCardinality rdf:datatype="xsd:nonNegativeInteger">1</owl:maxCardinality>
    </owl:Restriction>
</rdfs:subClassOf>
</owl:Class>

<owl:Class rdf:ID="Report">
    <rdfs:subClassOf rdf:resource="#Reference" />
    <rdfs:label xml:lang="en">Report</rdfs:label>
    <rdfs:comment xml:lang="en">A report published by an institution with some explicit
policy.</rdfs:comment>
    <rdfs:subClassOf>
        <owl:Restriction>
            <owl:onProperty rdf:resource="#author" />
            <owl:cardinality rdf:datatype="xsd:nonNegativeInteger">1</owl:cardinality>
        </owl:Restriction>
    </rdfs:subClassOf>
    <rdfs:subClassOf>
        <owl:Restriction>
            <owl:onProperty rdf:resource="#title" />
            <owl:cardinality rdf:datatype="xsd:nonNegativeInteger">1</owl:cardinality>
        </owl:Restriction>
    </rdfs:subClassOf>
    <rdfs:subClassOf>
        <owl:Restriction>
            <owl:onProperty rdf:resource="#institution" />
            <owl:cardinality rdf:datatype="xsd:nonNegativeInteger">1</owl:cardinality>
        </owl:Restriction>
    </rdfs:subClassOf>
    <rdfs:subClassOf>
        <owl:Restriction>
            <owl:onProperty rdf:resource="#date" />
            <owl:cardinality rdf:datatype="xsd:nonNegativeInteger">1</owl:cardinality>
        </owl:Restriction>
    </rdfs:subClassOf>
    <rdfs:subClassOf>
        <owl:Restriction>
            <owl:onProperty rdf:resource="#date" />
            <owl:allValuesFrom rdf:resource="#Date" />
        </owl:Restriction>
    </rdfs:subClassOf>
    <rdfs:subClassOf>
        <owl:Restriction>
            <owl:onProperty rdf:resource="#number" />
            <owl:maxCardinality rdf:datatype="xsd:nonNegativeInteger">1</owl:maxCardinality>
        </owl:Restriction>
    </rdfs:subClassOf>
</owl:Class>

<owl:Class rdf:ID="TechReport">
    <rdfs:subClassOf rdf:resource="#Report" />
    <rdfs:label xml:lang="en">TechReport</rdfs:label>
    <rdfs:comment xml:lang="en">A report published by a school or other institution, usually
numbered within a series.</rdfs:comment>
</owl:Class>

<owl:Class rdf:ID="Deliverable">
    <rdfs:subClassOf rdf:resource="#Report" />
    <rdfs:label xml:lang="en">Deliverable report</rdfs:label>
    <rdfs:comment xml:lang="en">A report delivered for accomplishing a
contract.</rdfs:comment>
    <rdfs:subClassOf>
        <owl:Restriction>
            <owl:onProperty rdf:resource="#contract" />
            <owl:maxCardinality rdf:datatype="xsd:nonNegativeInteger">1</owl:maxCardinality>
        </owl:Restriction>
    </rdfs:subClassOf>
</owl:Class>

```

```

    </owl:Restriction>
  </rdfs:subClassOf>
</owl:Class>

<owl:Class rdf:ID="Unpublished">
  <rdfs:subClassOf rdf:resource="#Informal" />
  <rdfs:label xml:lang="en">Unpublished</rdfs:label>
  <rdfs:comment xml:lang="en">A document having an author and title, but not formally
published.</rdfs:comment>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#author" />
      <owl:cardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:cardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#title" />
      <owl:cardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:cardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#note" />
      <owl:minCardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:minCardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
</owl:Class>

<owl:Class rdf:ID="MotionPicture">
  <rdfs:subClassOf rdf:resource="#Reference" />
  <rdfs:label xml:lang="en">MotionPicture</rdfs:label>
  <rdfs:comment xml:lang="en">A film/movie/motion picture.</rdfs:comment>
</owl:Class>

<!-- SPECIAL CLASSES -->

<owl:Class rdf:ID="Journal">
  <rdfs:label xml:lang="en">Journal or magazine</rdfs:label>
  <rdfs:comment xml:lang="en">A periodical publication collecting works from different
authors.</rdfs:comment>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#name" />
      <owl:cardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:cardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#name" />
      <owl:allValuesFrom rdf:resource="&xsd:string"/>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#shortName" />
      <owl:allValuesFrom rdf:resource="&xsd:string"/>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#periodicity" />
      <owl:allValuesFrom rdf:resource="&xsd:string"/>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#publisher" />
      <owl:maxCardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:maxCardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#series" />
      <owl:maxCardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:maxCardinality>
    </owl:Restriction>
  </rdfs:subClassOf>

```

```

<rdfs:subClassOf>
  <owl:Restriction>
    <owl:onProperty rdf:resource="#firstPublished" />
    <owl:maxCardinality rdf:datatype="xsd:nonNegativeInteger">1</owl:maxCardinality>
  </owl:Restriction>
</rdfs:subClassOf>
<rdfs:subClassOf>
  <owl:Restriction>
    <owl:onProperty rdf:resource="#articles" />
    <owl:allValuesFrom rdf:resource="#Article"/>
  </owl:Restriction>
</rdfs:subClassOf>
</owl:Class>

<owl:Class rdf:ID="Conference">
  <rdfs:label xml:lang="en">The location of an event</rdfs:label>
  <rdfs:comment xml:lang="en">An event presenting work.</rdfs:comment>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#name" />
      <owl:cardinality rdf:datatype="xsd:nonNegativeInteger">1</owl:cardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#organizer" />
      <owl:allValuesFrom rdf:resource="#Institution"/>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#shortName" />
      <owl:maxCardinality rdf:datatype="xsd:nonNegativeInteger">1</owl:maxCardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#issue" />
      <owl:maxCardinality rdf:datatype="xsd:nonNegativeInteger">1</owl:maxCardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#location" />
      <owl:maxCardinality rdf:datatype="xsd:nonNegativeInteger">1</owl:maxCardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
</owl:Class>

<owl:Class rdf:ID="Address">
  <rdfs:label xml:lang="en">Address</rdfs:label>
  <rdfs:comment xml:lang="en">The street address of the location of some organization or
event.</rdfs:comment>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#country" />
      <owl:maxCardinality rdf:datatype="xsd:nonNegativeInteger">1</owl:maxCardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#country" />
      <owl:allValuesFrom rdf:resource="xsd:string"/>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#state" />
      <owl:maxCardinality rdf:datatype="xsd:nonNegativeInteger">1</owl:maxCardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#state" />
      <owl:allValuesFrom rdf:resource="xsd:string"/>
    </owl:Restriction>
  </rdfs:subClassOf>

```

```

<rdfs:subClassOf>
  <owl:Restriction>
    <owl:onProperty rdf:resource="#city" />
    <owl:maxCardinality rdf:datatype="xsd:nonNegativeInteger">1</owl:maxCardinality>
  </owl:Restriction>
</rdfs:subClassOf>
<rdfs:subClassOf>
  <owl:Restriction>
    <owl:onProperty rdf:resource="#city" />
    <owl:allValuesFrom rdf:resource="xsd:string"/>
  </owl:Restriction>
</rdfs:subClassOf>
</owl:Class>

<owl:Class rdf:ID="Institution">
  <rdfs:subClassOf rdf:resource="foaf:Organization"/>
  <rdfs:label xml:lang="en">Institution</rdfs:label>
  <rdfs:comment xml:lang="en">An institution.</rdfs:comment>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#name" />
      <owl:cardinality rdf:datatype="xsd:nonNegativeInteger">1</owl:cardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#shortName" />
      <owl:cardinality rdf:datatype="xsd:nonNegativeInteger">1</owl:cardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#address" />
      <owl:maxCardinality rdf:datatype="xsd:nonNegativeInteger">1</owl:maxCardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
</owl:Class>

<owl:Class rdf:ID="Publisher">
  <rdfs:subClassOf rdf:resource="#Institution" />
  <rdfs:label xml:lang="en">Publisher</rdfs:label>
  <rdfs:comment xml:lang="en">The publisher of books or journals.</rdfs:comment>
</owl:Class>

<!-- trouver une ontologie pour ca -->
<owl:Class rdf:ID="School">
  <rdfs:subClassOf rdf:resource="#Institution" />
  <rdfs:label xml:lang="en">School</rdfs:label>
  <rdfs:comment xml:lang="en">A school or university.</rdfs:comment>
</owl:Class>

<owl:Class rdf:ID="PersonList">
  <rdfs:subClassOf rdf:resource="rdf:List"/>
  <rdfs:label xml:lang="en">Person list</rdfs:label>
  <rdfs:comment xml:lang="en">A list of persons.</rdfs:comment>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="rdf:first" />
      <owl:cardinality rdf:datatype="xsd:nonNegativeInteger">1</owl:cardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="rdf:first" />
      <owl:allValuesFrom rdf:resource="foaf:Person" />
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="rdf:rest" />
      <owl:cardinality rdf:datatype="xsd:nonNegativeInteger">1</owl:cardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="rdf:rest" />
      <owl:allValuesFrom>

```

```

    <owl:Class>
      <owl:unionOf rdf:parseType="Collection">
        <owl:Class rdf:about="#PersonList" />
        <owl:Class>
          <owl:oneOf rdf:parseType="Collection">
            <rdf:List rdf:about="&rdf:nil" />
          </owl:oneOf>
        </owl:Class>
      </owl:unionOf>
    </owl:Class>
  </owl:allValuesFrom>
</owl:Restriction>
</rdfs:subClassOf>
</owl:Class>

<owl:Class rdf:ID="PageRange">
  <rdfs:label xml:lang="en">PageRange</rdfs:label>
  <rdfs:comment xml:lang="en">A range of pages.</rdfs:comment>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#startPage" />
      <owl:cardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:cardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#endPage" />
      <owl:cardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:cardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
</owl:Class>

<owl:Class rdf:ID="Date">
  <rdfs:label xml:lang="en">Date</rdfs:label>
  <rdfs:comment xml:lang="en">Date of a day which can be unknown (i.e., only the year is
known or only the year and month). This is for overcoming the limits of XML-Schema for wich a
date is not separable.</rdfs:comment>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#year" />
      <owl:cardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:cardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#year" />
      <owl:allValuesFrom rdf:resource="&xsd;gYear"/>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#month" />
      <owl:maxCardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:maxCardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#month" />
      <owl:allValuesFrom rdf:resource="&xsd;gMonth"/>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#day" />
      <owl:maxCardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:maxCardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#day" />
      <owl:allValuesFrom rdf:resource="&xsd;gDay"/>
    </owl:Restriction>
  </rdfs:subClassOf>
</owl:Class>

<!--
***** FIELDS *****

```

Note: all rdfs:comment values for the entries come from
<http://newton.ex.ac.uk/tex/pack/bibtex/btxdoc/node7.html>
 -->

```

<owl:DatatypeProperty rdf:ID="key">
  <rdfs:domain rdf:resource="#Reference" />
  <rdfs:range rdf:resource="&xsd:string" />
  <rdfs:label xml:lang="en">key</rdfs:label>
  <rdfs:comment xml:lang="en">The key for a particular entry. Note that the rdf:ID for each
Reference instance could be the key as well, possibly making this property
redundant.</rdfs:comment>
</owl:DatatypeProperty>

<owl:DatatypeProperty rdf:ID="reviewed">
  <rdfs:domain rdf:resource="#Reference" />
  <rdfs:range rdf:resource="&xsd:string" />
  <rdfs:label xml:lang="en">howReviewed</rdfs:label>
  <rdfs:comment xml:lang="en">The selection process overcome by the publication (full, none,
poster, invited are possible values).</rdfs:comment>
</owl:DatatypeProperty>

<owl:ObjectProperty rdf:ID="chapters">
  <rdfs:domain rdf:resource="#Reference"/>
  <rdfs:range rdf:resource="#Chapter" />
  <rdfs:label xml:lang="en">chapters</rdfs:label>
  <rdfs:comment xml:lang="en">The chapters of a book (monograph or
collection).</rdfs:comment>
</owl:ObjectProperty>

<owl:ObjectProperty rdf:ID="parts">
  <rdfs:domain rdf:resource="#Reference"/>
  <rdfs:range rdf:resource="#Part" />
  <rdfs:label xml:lang="en">parts</rdfs:label>
  <rdfs:comment xml:lang="en">The parts of some reference (chapters,
sections).</rdfs:comment>
</owl:ObjectProperty>

<owl:ObjectProperty rdf:ID="communications">
  <rdfs:domain rdf:resource="#Proceedings"/>
  <rdfs:range rdf:resource="#InProceedings" />
  <rdfs:label xml:lang="en">communications</rdfs:label>
  <rdfs:comment xml:lang="en">The communications to a conference as collected in
proceedings.</rdfs:comment>
</owl:ObjectProperty>

<owl:ObjectProperty rdf:ID="articles">
  <rdfs:domain rdf:resource="#Journal"/>
  <rdfs:range rdf:resource="#Article" />
  <rdfs:label xml:lang="en">articles</rdfs:label>
  <rdfs:comment xml:lang="en">The papers published by a journal.</rdfs:comment>
</owl:ObjectProperty>

<owl:ObjectProperty rdf:ID="address">
  <rdfs:domain rdf:resource="&owl;Thing"/>
  <rdfs:range rdf:resource="#Address" />
  <rdfs:label xml:lang="en">address</rdfs:label>
  <rdfs:comment xml:lang="en">The address of an institution or the place of a
conference.</rdfs:comment>
</owl:ObjectProperty>

<owl:ObjectProperty rdf:ID="event">
  <rdfs:domain rdf:resource="#Proceedings"/>
  <rdfs:range rdf:resource="#Conference" />
  <rdfs:label xml:lang="en">event</rdfs:label>
  <rdfs:comment xml:lang="en">The event concerned with the proceedings.</rdfs:comment>
</owl:ObjectProperty>

<owl:ObjectProperty rdf:ID="organizer">
  <rdfs:domain rdf:resource="#Conference"/>
  <rdfs:range rdf:resource="&foaf;Organization" />
  <rdfs:label xml:lang="en">organizer</rdfs:label>
  <rdfs:comment xml:lang="en">The organization taking in charge a conference.</rdfs:comment>
</owl:ObjectProperty>

<owl:DatatypeProperty rdf:ID="annotate">
  <rdfs:domain rdf:resource="#Reference" />

```

```

    <rdfs:range rdf:resource="&xsd:string" />
    <rdfs:label xml:lang="en">annotate</rdfs:label>
    <rdfs:comment xml:lang="en">An annotation. It is not used by the standard bibliography
styles, but may be used by others that produce an annotated bibliography.</rdfs:comment>
  </owl:DatatypeProperty>

  <owl:DatatypeProperty rdf:ID="periodicity">
    <rdfs:domain rdf:resource="#Journal" />
    <rdfs:range rdf:resource="&xsd:string" />
    <rdfs:label xml:lang="en">Periodicity</rdfs:label>
    <rdfs:comment xml:lang="en">The periodicity of a serial publication (yearly, biannual,
quarterly, monthly, bimonthly, weekly, biweekly, daily).</rdfs:comment>
  </owl:DatatypeProperty>

  <owl:ObjectProperty rdf:ID="contract">
    <rdfs:domain rdf:resource="#Reference" />
    <rdfs:range rdf:resource="&owl;Thing" />
    <rdfs:label xml:lang="en">Contract</rdfs:label>
    <rdfs:comment xml:lang="en">The contract in relation of which a particular reference has
been made.</rdfs:comment>
  </owl:ObjectProperty>

  <owl:DatatypeProperty rdf:ID="firstPublished">
    <rdfs:domain rdf:resource="#Chapter" />
    <rdfs:range rdf:resource="&xsd:string" />
    <rdfs:label xml:lang="en">First published</rdfs:label>
    <rdfs:comment xml:lang="en">When the reference was first published.</rdfs:comment>
  </owl:DatatypeProperty>

  <owl:DatatypeProperty rdf:ID="edition">
    <rdfs:domain>
      <owl:Class>
        <owl:unionOf rdf:parseType="Collection">
          <owl:Class rdf:about="#Book" />
          <owl:Class rdf:about="#Manual" />
        </owl:unionOf>
      </owl:Class>
    </rdfs:domain>
    <rdfs:range rdf:resource="&xsd:string" />
    <rdfs:label xml:lang="en">edition</rdfs:label>
    <rdfs:comment xml:lang="en">The edition of a book (for example, "Second"). This should be
an ordinal.</rdfs:comment>
  </owl:DatatypeProperty>

  <owl:ObjectProperty rdf:ID="humanCreator">
    <rdfs:domain rdf:resource="#Reference" />
    <rdfs:range rdf:resource="#PersonList" />
    <rdfs:label xml:lang="en">human creator</rdfs:label>
    <rdfs:comment xml:lang="en">The list of the human creators of a work.</rdfs:comment>
  </owl:ObjectProperty>

  <owl:ObjectProperty rdf:ID="author">
    <rdfs:subPropertyOf rdf:resource="#humanCreator" />
    <rdfs:label xml:lang="en">author</rdfs:label>
    <rdfs:comment xml:lang="en">The list of the author(s) of a work.</rdfs:comment>
  </owl:ObjectProperty>

  <owl:ObjectProperty rdf:ID="editor">
    <rdfs:subPropertyOf rdf:resource="#humanCreator" />
    <rdfs:label xml:lang="en">editor</rdfs:label>
    <rdfs:comment xml:lang="en">The list of persons who edited or coordinated a
work.</rdfs:comment>
  </owl:ObjectProperty>

  <owl:ObjectProperty rdf:ID="directors">
    <rdfs:subPropertyOf rdf:resource="#humanCreator" />
    <rdfs:domain rdf:resource="#MotionPicture" />
    <rdfs:label xml:lang="en">directors</rdfs:label>
    <rdfs:comment xml:lang="en">The list of directors of a motion picture or theater
play.</rdfs:comment>
  </owl:ObjectProperty>

  <owl:DatatypeProperty rdf:ID="howPublished">
    <rdfs:domain>
      <owl:Class>
        <owl:unionOf rdf:parseType="Collection">
          <owl:Class rdf:about="#Misc" />

```

```

        <owl:Class rdf:about="#Booklet" />
    </owl:unionOf>
</owl:Class>
</rdfs:domain>
<rdfs:range rdf:resource="&xsd:string" />
<rdfs:label xml:lang="en">published</rdfs:label>
<rdfs:comment xml:lang="en">How something strange has been published.</rdfs:comment>
</owl:DatatypeProperty>

<owl:ObjectProperty rdf:ID="institution">
    <rdfs:domain rdf:resource="#Report" />
    <rdfs:range rdf:resource="#Institution" />
    <rdfs:label xml:lang="en">institution</rdfs:label>
    <rdfs:comment xml:lang="en">The sponsoring institution of a technical
report.</rdfs:comment>
</owl:ObjectProperty>

<owl:ObjectProperty rdf:ID="isPartOf">
    <rdfs:domain rdf:resource="#Part" />
    <rdfs:label xml:lang="en">is part of</rdfs:label>
    <rdfs:type rdf:resource="&owl;TransitiveProperty" />
    <rdfs:comment xml:lang="en">The document that contains a text or article.</rdfs:comment>
</owl:ObjectProperty>

<owl:ObjectProperty rdf:ID="journal">
    <rdfs:domain rdf:resource="#Article" />
    <rdfs:range rdf:resource="#Journal" />
    <rdfs:subPropertyOf rdf:resource="#isPartOf" />
    <rdfs:label xml:lang="en">journal</rdfs:label>
    <rdfs:comment xml:lang="en">The journal in which an article is published.</rdfs:comment>
</owl:ObjectProperty>

<owl:ObjectProperty rdf:ID="book">
    <rdfs:domain rdf:resource="#InBook" />
    <rdfs:range rdf:resource="#Monograph" />
    <rdfs:subPropertyOf rdf:resource="#isPartOf" />
    <rdfs:label xml:lang="en">book</rdfs:label>
    <rdfs:comment xml:lang="en">A reference to the book that contains the
entry.</rdfs:comment>
</owl:ObjectProperty>

<owl:ObjectProperty rdf:ID="collection">
    <rdfs:domain rdf:resource="#InCollection" />
    <rdfs:range rdf:resource="#Collection" />
    <rdfs:subPropertyOf rdf:resource="#isPartOf" />
    <rdfs:label xml:lang="en">collection</rdfs:label>
    <rdfs:comment xml:lang="en">A reference to the book that contains the
entry.</rdfs:comment>
</owl:ObjectProperty>

<owl:ObjectProperty rdf:ID="proceedings">
    <rdfs:domain rdf:resource="#InProceedings" />
    <rdfs:range rdf:resource="#Proceedings" />
    <rdfs:subPropertyOf rdf:resource="#isPartOf" />
    <rdfs:label xml:lang="en">proceedings</rdfs:label>
    <rdfs:comment xml:lang="en">A reference to the proceedings in which the entry
appears.</rdfs:comment>
</owl:ObjectProperty>

<owl:ObjectProperty rdf:ID="date">
    <rdfs:domain>
        <owl:Class>
            <owl:unionOf rdf:parseType="Collection">
                <owl:Class rdf:about="#Reference" />
                <owl:Class rdf:about="#Conference" />
            </owl:unionOf>
        </owl:Class>
    </rdfs:domain>
    <rdfs:range rdf:resource="#Date" />
    <rdfs:label xml:lang="en">date</rdfs:label>
    <rdfs:comment xml:lang="en">The date when the work was published or, for an unpublished
work, in which it was written. This date can be incomplete (i.e., no known day or month) but
must contain a year.</rdfs:comment>
</owl:ObjectProperty>

<owl:DatatypeProperty rdf:ID="note">
    <rdfs:domain rdf:resource="#Reference" />

```



```

    <rdfs:range rdf:resource="&xsd:string" />
    <rdfs:label xml:lang="en">note</rdfs:label>
    <rdfs:comment xml:lang="en">Any additional information that can help the
reader.</rdfs:comment>
  </owl:DatatypeProperty>

  <owl:ObjectProperty rdf:ID="organization">
    <rdfs:domain>
      <owl:Class>
        <owl:unionOf rdf:parseType="Collection">
          <owl:Class rdf:about="#Proceedings" />
          <owl:Class rdf:about="#Manual" />
        </owl:unionOf>
      </owl:Class>
    </rdfs:domain>
    <rdfs:range rdf:resource="&foaf:Organization" />
    <rdfs:label xml:lang="en">organization</rdfs:label>
    <rdfs:comment xml:lang="en">The organization that sponsors a conference or that publishes
a manual.</rdfs:comment>
  </owl:ObjectProperty>

  <owl:ObjectProperty rdf:ID="publisher">
    <rdfs:domain>
      <owl:Class>
        <owl:unionOf rdf:parseType="Collection">
          <owl:Class rdf:about="#Reference" />
          <owl:Class rdf:about="#Journal" />
        </owl:unionOf>
      </owl:Class>
    </rdfs:domain>
    <rdfs:range rdf:resource="#Publisher" />
    <rdfs:label xml:lang="en">publisher</rdfs:label>
    <rdfs:comment xml:lang="en">The publisher of a work.</rdfs:comment>
  </owl:ObjectProperty>

  <owl:ObjectProperty rdf:ID="school">
    <rdfs:domain>
      <owl:Class>
        <owl:unionOf rdf:parseType="Collection">
          <owl:Class rdf:about="#Academic" />
          <owl:Class rdf:about="#LectureNotes" />
        </owl:unionOf>
      </owl:Class>
    </rdfs:domain>
    <rdfs:range rdf:resource="#School" />
    <rdfs:label xml:lang="en">school</rdfs:label>
    <rdfs:comment xml:lang="en">The name of the school where a thesis was
written.</rdfs:comment>
  </owl:ObjectProperty>

  <owl:DatatypeProperty rdf:ID="series">
    <rdfs:domain rdf:resource="#Reference" />
    <rdfs:range rdf:resource="&xsd:string" />
    <rdfs:label xml:lang="en">series</rdfs:label>
    <rdfs:comment xml:lang="en">The name of a series or set of books. When citing an entire
book, the title field gives its title and an optional series field gives the name of a
series or multi-volume set in which the book is published.</rdfs:comment>
  </owl:DatatypeProperty>

  <owl:DatatypeProperty rdf:ID="title">
    <rdfs:domain rdf:resource="#Reference" />
    <rdfs:range rdf:resource="&xsd:string" />
    <rdfs:label xml:lang="en">title</rdfs:label>
    <rdfs:comment xml:lang="en">The title of the work.</rdfs:comment>
  </owl:DatatypeProperty>

  <owl:DatatypeProperty rdf:ID="type">
    <rdfs:domain>
      <owl:Class>
        <owl:unionOf rdf:parseType="Collection">
          <owl:Class rdf:about="#Chapter" />
          <owl:Class rdf:about="#TechReport" />
          <owl:Class rdf:about="#Academic" />
        </owl:unionOf>
      </owl:Class>
    </rdfs:domain>
    <rdfs:range rdf:resource="&xsd:string" />

```

```

    <rdfs:label xml:lang="en">type</rdfs:label>
    <rdfs:comment xml:lang="en">The type of a technical report (for example, "Research
Note").</rdfs:comment>
  </owl:DatatypeProperty>

  <owl:DatatypeProperty rdf:ID="affiliation">
    <rdfs:domain rdf:resource="#Reference" />
    <rdfs:range rdf:resource="xsd:string" />
    <rdfs:label xml:lang="en">affiliation</rdfs:label>
    <rdfs:comment xml:lang="en">The authors affiliation.</rdfs:comment>
  </owl:DatatypeProperty>

  <owl:DatatypeProperty rdf:ID="abstract">
    <rdfs:domain rdf:resource="#Reference" />
    <rdfs:range rdf:resource="xsd:string" />
    <rdfs:label xml:lang="en">abstract</rdfs:label>
    <rdfs:comment xml:lang="en">The abstract of the work.</rdfs:comment>
  </owl:DatatypeProperty>

  <owl:DatatypeProperty rdf:ID="contents">
    <rdfs:domain rdf:resource="#Reference" />
    <rdfs:range rdf:resource="xsd:string" />
    <rdfs:label xml:lang="en">ontents</rdfs:label>
    <rdfs:comment xml:lang="en">The Table of Contents.</rdfs:comment>
  </owl:DatatypeProperty>

  <owl:DatatypeProperty rdf:ID="copyright">
    <rdfs:domain rdf:resource="#Reference" />
    <rdfs:range rdf:resource="xsd:string" />
    <rdfs:label xml:lang="en">copyright</rdfs:label>
    <rdfs:comment xml:lang="en">The Copyright information.</rdfs:comment>
  </owl:DatatypeProperty>

  <owl:DatatypeProperty rdf:ID="isbn">
    <rdfs:domain rdf:resource="#Reference" />
    <rdfs:range rdf:resource="xsd:string" />
    <rdfs:label xml:lang="en">ISBN</rdfs:label>
    <rdfs:comment xml:lang="en">The International Standard Book Number.</rdfs:comment>
  </owl:DatatypeProperty>

  <owl:DatatypeProperty rdf:ID="issn">
    <rdfs:domain rdf:resource="#Reference" />
    <rdfs:range rdf:resource="xsd:string" />
    <rdfs:label xml:lang="en">ISSN</rdfs:label>
    <rdfs:comment xml:lang="en">The International Standard Serial Number. Used to identify a
journal.</rdfs:comment>
  </owl:DatatypeProperty>

  <owl:DatatypeProperty rdf:ID="keywords">
    <rdfs:domain rdf:resource="#Reference" />
    <rdfs:range rdf:resource="xsd:string" />
    <rdfs:label xml:lang="en">keywords</rdfs:label>
    <rdfs:comment xml:lang="en">Key words used for searching or possibly for annotation (as
given by the author or publisher).</rdfs:comment>
  </owl:DatatypeProperty>

  <owl:DatatypeProperty rdf:ID="language">
    <rdfs:domain rdf:resource="#Reference" />
    <rdfs:range rdf:resource="xsd:language" />
    <rdfs:label xml:lang="en">language</rdfs:label>
    <rdfs:comment xml:lang="en">The language in which the referenced publication is written or
performed (use ISO two-letter codes).</rdfs:comment>
  </owl:DatatypeProperty>

  <owl:ObjectProperty rdf:ID="location">
    <rdfs:domain>
      <owl:Class>
        <owl:unionOf rdf:parseType="Collection">
          <owl:Class rdf:about="#Reference" />
          <owl:Class rdf:about="#Conference" />
        </owl:unionOf>
      </owl:Class>
    </rdfs:domain>
    <rdfs:range rdf:resource="#Address" />
    <rdfs:label xml:lang="en">location</rdfs:label>
    <rdfs:comment xml:lang="en">A location associated with the entry, such as the city in
which a conference took place (this is now obsolete).</rdfs:comment>
  </owl:ObjectProperty>

```

```

</owl:ObjectProperty>

<owl:DatatypeProperty rdf:ID="lccn">
  <rdfs:domain rdf:resource="#Reference" />
  <rdfs:range rdf:resource="&xsd:string" />
  <rdfs:label xml:lang="en">LCCN</rdfs:label>
  <rdfs:comment xml:lang="en">The Library of Congress Call Number (this the Congress of the
United State of America).</rdfs:comment>
</owl:DatatypeProperty>

<owl:DatatypeProperty rdf:ID="mrNumber">
  <rdfs:domain rdf:resource="#Reference" />
  <rdfs:range rdf:resource="&xsd:string" />
  <rdfs:label xml:lang="en">MR number</rdfs:label>
  <rdfs:comment xml:lang="en">The Mathematical Reviews number.</rdfs:comment>
</owl:DatatypeProperty>

<owl:DatatypeProperty rdf:ID="price">
  <rdfs:domain rdf:resource="#Reference" />
  <rdfs:range rdf:resource="&xsd:string" />
  <rdfs:label xml:lang="en">price</rdfs:label>
  <rdfs:comment xml:lang="en">The price of the document.</rdfs:comment>
</owl:DatatypeProperty>

<owl:DatatypeProperty rdf:ID="size">
  <rdfs:domain rdf:resource="#Reference" />
  <rdfs:range rdf:resource="&xsd:string" />
  <rdfs:label xml:lang="en">size</rdfs:label>
  <rdfs:comment xml:lang="en">The physical dimensions of a document.</rdfs:comment>
</owl:DatatypeProperty>

<owl:DatatypeProperty rdf:ID="url">
  <rdfs:domain rdf:resource="#Reference" />
  <rdfs:range rdf:resource="&xsd:string" />
  <rdfs:label xml:lang="en">URL</rdfs:label>
  <rdfs:comment xml:lang="en">The Universal Resource Locator that points to the item being
referenced.</rdfs:comment>
</owl:DatatypeProperty>

<owl:DatatypeProperty rdf:ID="name">
  <rdfs:range rdf:resource="&xsd:string" />
  <rdfs:label xml:lang="en">The name of an entity.</rdfs:label>
</owl:DatatypeProperty>

<owl:DatatypeProperty rdf:ID="shortName">
  <rdfs:range rdf:resource="&xsd:string" />
  <rdfs:label xml:lang="en">The name of an entity.</rdfs:label>
</owl:DatatypeProperty>

<!--
*****  "META" PROPERTIES  *****
-->

<owl:DatatypeProperty rdf:ID="chapter">
  <rdfs:domain rdf:resource="#Part" />
  <rdfs:range rdf:resource="&xsd:string" />
  <rdfs:label xml:lang="en">chapter</rdfs:label>
  <rdfs:comment xml:lang="en">The chapter (or section or whatever) number in which the entry
is found.</rdfs:comment>
</owl:DatatypeProperty>

<owl:ObjectProperty rdf:ID="pages">
  <rdfs:domain rdf:resource="#Part" />
  <rdfs:range rdf:resource="#PageRange" />
  <rdfs:label xml:lang="en">pages</rdfs:label>
  <rdfs:comment xml:lang="en">A range of page, with an upper and lower limit.</rdfs:comment>
</owl:ObjectProperty>

<owl:DatatypeProperty rdf:ID="numberOrVolume">
  <rdfs:domain>
    <owl:Class>
      <owl:unionOf rdf:parseType="Collection">
        <owl:Class rdf:about="#Reference"/>
        <owl:Class rdf:about="#Conference"/>
      </owl:unionOf>
    </owl:Class>
  </rdfs:domain>

```

```

    <rdfs:label xml:lang="en">volume or number</rdfs:label>
    <rdfs:comment xml:lang="en">The number of a journal, magazine, technical report, or of a
work in a series. An issue of a journal or magazine is usually identified by its volume and
number; the organization that issues a technical report usually gives it a number; and
sometimes books are given numbers in a named series.</rdfs:comment>
  </owl:DatatypeProperty>

  <owl:DatatypeProperty rdf:ID="number">
    <rdfs:domain rdf:resource="#Reference" />
    <rdfs:range rdf:resource="xsd:string" />
    <rdfs:subPropertyOf rdf:resource="#numberOrVolume" />
    <rdfs:label xml:lang="en">number</rdfs:label>
    <rdfs:comment xml:lang="en">The number of a journal, magazine, technical report, or of a
work in a series. An issue of a journal or magazine is usually identified by its volume and
number; the organization that issues a technical report usually gives it a number; and
sometimes books are given numbers in a named series.</rdfs:comment>
  </owl:DatatypeProperty>

  <owl:DatatypeProperty rdf:ID="issue">
    <rdfs:domain>
      <owl:Class>
        <owl:unionOf rdf:parseType="Collection">
          <owl:Class rdf:about="#Reference"/>
          <owl:Class rdf:about="#Conference"/>
        </owl:unionOf>
      </owl:Class>
    </rdfs:domain>
    <rdfs:range rdf:resource="xsd:string" />
    <rdfs:subPropertyOf rdf:resource="#numberOrVolume" />
    <rdfs:label xml:lang="en">Issue</rdfs:label>
    <rdfs:comment xml:lang="en">The issue of a conference.</rdfs:comment>
  </owl:DatatypeProperty>

  <owl:DatatypeProperty rdf:ID="volume">
    <rdfs:subPropertyOf rdf:resource="#numberOrVolume" />
    <rdfs:domain rdf:resource="#Reference"/>
    <rdfs:range rdf:resource="xsd:nonNegativeInteger" />
    <rdfs:label xml:lang="en">volume</rdfs:label>
    <rdfs:comment xml:lang="en">The volume of a journal or multivolume book.</rdfs:comment>
  </owl:DatatypeProperty>

  <owl:DatatypeProperty rdf:ID="year">
    <rdfs:domain rdf:resource="#Date" />
    <rdfs:range rdf:resource="xsd:gYear" />
    <rdfs:label xml:lang="en">year</rdfs:label>
  </owl:DatatypeProperty>

  <owl:DatatypeProperty rdf:ID="month">
    <rdfs:domain rdf:resource="#Date" />
    <rdfs:range rdf:resource="xsd:gMonth" />
    <rdfs:label xml:lang="en">month</rdfs:label>
  </owl:DatatypeProperty>

  <owl:DatatypeProperty rdf:ID="day">
    <rdfs:domain rdf:resource="#Date" />
    <rdfs:range rdf:resource="xsd:gDay" />
    <rdfs:label xml:lang="en">day</rdfs:label>
  </owl:DatatypeProperty>

  <owl:DatatypeProperty rdf:ID="city">
    <rdfs:domain rdf:resource="#Address" />
    <rdfs:range rdf:resource="xsd:string" />
    <rdfs:label xml:lang="en">city</rdfs:label>
  </owl:DatatypeProperty>

  <owl:DatatypeProperty rdf:ID="state">
    <rdfs:domain rdf:resource="#Address" />
    <rdfs:range rdf:resource="xsd:string" />
    <rdfs:label xml:lang="en">state</rdfs:label>
  </owl:DatatypeProperty>

  <owl:DatatypeProperty rdf:ID="country">
    <rdfs:domain rdf:resource="#Address" />
    <rdfs:range rdf:resource="xsd:string" />
    <rdfs:comment xml:lang="en">Usually in ISO format</rdfs:comment>
    <rdfs:label xml:lang="en">country</rdfs:label>
  </owl:DatatypeProperty>

```

```

<owl:DatatypeProperty rdf:ID="startPage">
  <rdfs:domain rdf:resource="#PageRange" />
  <rdfs:range rdf:resource="&xsd;nonNegativeInteger" />
  <rdfs:label xml:lang="en">start page</rdfs:label>
  <rdfs:comment xml:lang="en">The beginning of a range of pages.</rdfs:comment>
</owl:DatatypeProperty>

<owl:DatatypeProperty rdf:ID="endPage">
  <rdfs:domain rdf:resource="#PageRange" />
  <rdfs:range rdf:resource="&xsd;nonNegativeInteger" />
  <rdfs:label xml:lang="en">end page</rdfs:label>
  <rdfs:comment xml:lang="en">The end of a range of pages.</rdfs:comment>
</owl:DatatypeProperty>

<!-- PEOPLE -->

<foaf:Person rdf:about="#a04570373">
  <rdfs:label>John-Jules Meyer</rdfs:label>
  <foaf:name rdf:datatype="&xsd:string">John-Jules Meyer</foaf:name>
  <foaf:firstName rdf:datatype="&xsd:string">John-Jules</foaf:firstName>
  <lastName rdf:datatype="&xsd:string">Meyer</lastName>
</foaf:Person>

<foaf:Person rdf:about="#a43836633">
  <rdfs:label>Jeen Broekstra</rdfs:label>
  <foaf:name rdf:datatype="&xsd:string">Jeen Broekstra</foaf:name>
  <foaf:firstName rdf:datatype="&xsd:string">Jeen</foaf:firstName>
  <lastName rdf:datatype="&xsd:string">Broekstra</lastName>
</foaf:Person>

<foaf:Person rdf:about="#a85228505">
  <rdfs:label>Alexander Mädche</rdfs:label>
  <foaf:name rdf:datatype="&xsd:string">Alexander Mädche</foaf:name>
  <foaf:firstName rdf:datatype="&xsd:string">Alexander</foaf:firstName>
  <lastName rdf:datatype="&xsd:string">Mädche</lastName>
</foaf:Person>

<foaf:Person rdf:about="#a48552212">
  <rdfs:label>Björn Schnizler</rdfs:label>
  <foaf:name rdf:datatype="&xsd:string">Björn Schnizler</foaf:name>
  <foaf:firstName rdf:datatype="&xsd:string">Björn</foaf:firstName>
  <lastName rdf:datatype="&xsd:string">Schnizler</lastName>
</foaf:Person>

<foaf:Person rdf:about="#a971541439">
  <rdfs:label>Alberto Trombetta</rdfs:label>
  <foaf:name rdf:datatype="&xsd:string">Alberto Trombetta</foaf:name>
  <foaf:firstName rdf:datatype="&xsd:string">Alberto</foaf:firstName>
  <lastName rdf:datatype="&xsd:string">Trombetta</lastName>
</foaf:Person>

<foaf:Person rdf:about="#a11090777">
  <rdfs:label>Christine Parent</rdfs:label>
  <foaf:name rdf:datatype="&xsd:string">Christine Parent</foaf:name>
  <foaf:firstName rdf:datatype="&xsd:string">Christine</foaf:firstName>
  <lastName rdf:datatype="&xsd:string">Parent</lastName>
</foaf:Person>

<foaf:Person rdf:about="#a250331360">
  <rdfs:label>R. Schmidt</rdfs:label>
  <foaf:name rdf:datatype="&xsd:string">R. Schmidt</foaf:name>
  <foaf:firstName rdf:datatype="&xsd:string">R.</foaf:firstName>
  <lastName rdf:datatype="&xsd:string">Schmidt</lastName>
</foaf:Person>

<foaf:Person rdf:about="#a79573306">
  <rdfs:label>York Sure</rdfs:label>
  <foaf:name rdf:datatype="&xsd:string">York Sure</foaf:name>
  <foaf:firstName rdf:datatype="&xsd:string">York</foaf:firstName>
  <lastName rdf:datatype="&xsd:string">Sure</lastName>
</foaf:Person>

<foaf:Person rdf:about="#a885257047">
  <rdfs:label>M. Punceva</rdfs:label>
  <foaf:name rdf:datatype="&xsd:string">M. Punceva</foaf:name>
  <foaf:firstName rdf:datatype="&xsd:string">M.</foaf:firstName>

```

```

        <lastName rdf:datatype="&xsd:string">Punceva</lastName>
</foaf:Person>

<foaf:Person rdf:about="#a74993404">
  <rdfs:label>I. V. Levenshtein</rdfs:label>
  <foaf:name rdf:datatype="&xsd:string">I. V. Levenshtein</foaf:name>
  <foaf:firstName rdf:datatype="&xsd:string">I. V.</foaf:firstName>
  <lastName rdf:datatype="&xsd:string">Levenshtein</lastName>
</foaf:Person>

<foaf:Person rdf:about="#a71003986">
  <rdfs:label>Steffen Staab</rdfs:label>
  <foaf:name rdf:datatype="&xsd:string">Steffen Staab</foaf:name>
  <foaf:firstName rdf:datatype="&xsd:string">Steffen</foaf:firstName>
  <lastName rdf:datatype="&xsd:string">Staab</lastName>
</foaf:Person>

<foaf:Person rdf:about="#a572406328">
  <rdfs:label>Frank Boer</rdfs:label>
  <foaf:name rdf:datatype="&xsd:string">Frank Boer</foaf:name>
  <foaf:firstName rdf:datatype="&xsd:string">Frank</foaf:firstName>
  <lastName rdf:datatype="&xsd:string">Boer</lastName>
</foaf:Person>

<foaf:Person rdf:about="#a139477786">
  <rdfs:label>Maarten Menken</rdfs:label>
  <foaf:name rdf:datatype="&xsd:string">Maarten Menken</foaf:name>
  <foaf:firstName rdf:datatype="&xsd:string">Maarten</foaf:firstName>
  <lastName rdf:datatype="&xsd:string">Menken</lastName>
</foaf:Person>

<foaf:Person rdf:about="#a337716610">
  <rdfs:label>Manfred Hauswirth</rdfs:label>
  <foaf:name rdf:datatype="&xsd:string">Manfred Hauswirth</foaf:name>
  <foaf:firstName rdf:datatype="&xsd:string">Manfred</foaf:firstName>
  <lastName rdf:datatype="&xsd:string">Hauswirth</lastName>
</foaf:Person>

<foaf:Person rdf:about="#a086379337">
  <rdfs:label>Wiebe Hoek</rdfs:label>
  <foaf:name rdf:datatype="&xsd:string">Wiebe Hoek</foaf:name>
  <foaf:firstName rdf:datatype="&xsd:string">Wiebe</foaf:firstName>
  <lastName rdf:datatype="&xsd:string">Hoek</lastName>
</foaf:Person>

<foaf:Person rdf:about="#a712561038">
  <rdfs:label>Marc Ehrig</rdfs:label>
  <foaf:name rdf:datatype="&xsd:string">Marc Ehrig</foaf:name>
  <foaf:firstName rdf:datatype="&xsd:string">Marc</foaf:firstName>
  <lastName rdf:datatype="&xsd:string">Ehrig</lastName>
</foaf:Person>

<foaf:Person rdf:about="#a066600210">
  <rdfs:label>Danilo Montesi</rdfs:label>
  <foaf:name rdf:datatype="&xsd:string">Danilo Montesi</foaf:name>
  <foaf:firstName rdf:datatype="&xsd:string">Danilo</foaf:firstName>
  <lastName rdf:datatype="&xsd:string">Montesi</lastName>
</foaf:Person>

<foaf:Person rdf:about="#a093016135">
  <rdfs:label>Rogier Eijk</rdfs:label>
  <foaf:name rdf:datatype="&xsd:string">Rogier Eijk</foaf:name>
  <foaf:firstName rdf:datatype="&xsd:string">Rogier</foaf:firstName>
  <lastName rdf:datatype="&xsd:string">Eijk</lastName>
</foaf:Person>

<foaf:Person rdf:about="#a944339054">
  <rdfs:label>Frank van Harmelen</rdfs:label>
  <foaf:name rdf:datatype="&xsd:string">Frank van Harmelen</foaf:name>
  <foaf:firstName rdf:datatype="&xsd:string">Frank</foaf:firstName>
  <lastName rdf:datatype="&xsd:string">van Harmelen</lastName>
</foaf:Person>

<foaf:Person rdf:about="#a98078619">
  <rdfs:label>Philippe Cudré-Mauroux</rdfs:label>
  <foaf:name rdf:datatype="&xsd:string">Philippe Cudré-Mauroux</foaf:name>
  <foaf:firstName rdf:datatype="&xsd:string">Philippe</foaf:firstName>

```

```

        <lastName rdf:datatype="&xsd:string">Cudré-Mauroux</lastName>
</foaf:Person>

<foaf:Person rdf:about="#a39510672">
  <rdfs:label>Z. Despotovic</rdfs:label>
  <foaf:name rdf:datatype="&xsd:string">Z. Despotovic</foaf:name>
  <foaf:firstName rdf:datatype="&xsd:string">Z.</foaf:firstName>
  <lastName rdf:datatype="&xsd:string">Despotovic</lastName>
</foaf:Person>

<foaf:Person rdf:about="#a431956276">
  <rdfs:label>Stefano Spaccapietra</rdfs:label>
  <foaf:name rdf:datatype="&xsd:string">Stefano Spaccapietra</foaf:name>
  <foaf:firstName rdf:datatype="&xsd:string">Stefano</foaf:firstName>
  <lastName rdf:datatype="&xsd:string">Spaccapietra</lastName>
</foaf:Person>

<foaf:Person rdf:about="#a431956276b">
  <rdfs:label>Mike Papazoglou</rdfs:label>
  <foaf:name rdf:datatype="&xsd:string">Mike Papazoglou</foaf:name>
  <foaf:firstName rdf:datatype="&xsd:string">Mike</foaf:firstName>
  <lastName rdf:datatype="&xsd:string">Papazoglou</lastName>
</foaf:Person>

<foaf:Person rdf:about="#a431956276c">
  <rdfs:label>Zahir Tari</rdfs:label>
  <foaf:name rdf:datatype="&xsd:string">Zahir Tari</foaf:name>
  <foaf:firstName rdf:datatype="&xsd:string">Zahir</foaf:firstName>
  <lastName rdf:datatype="&xsd:string">Tari</lastName>
</foaf:Person>

<foaf:Person rdf:about="#a70955601">
  <rdfs:label>A. Datta</rdfs:label>
  <foaf:name rdf:datatype="&xsd:string">A. Datta</foaf:name>
  <foaf:firstName rdf:datatype="&xsd:string">A.</foaf:firstName>
  <lastName rdf:datatype="&xsd:string">Datta</lastName>
</foaf:Person>

<foaf:Person rdf:about="#a467748807">
  <rdfs:label>Ateret Anaby-Tavor</rdfs:label>
  <foaf:name rdf:datatype="&xsd:string">Ateret Anaby-Tavor</foaf:name>
  <foaf:firstName rdf:datatype="&xsd:string">Ateret</foaf:firstName>
  <lastName rdf:datatype="&xsd:string">Anaby-Tavor</lastName>
</foaf:Person>

<foaf:Person rdf:about="#a3105947">
  <rdfs:label>Ronny Siebes</rdfs:label>
  <foaf:name rdf:datatype="&xsd:string">Ronny Siebes</foaf:name>
  <foaf:firstName rdf:datatype="&xsd:string">Ronny</foaf:firstName>
  <lastName rdf:datatype="&xsd:string">Siebes</lastName>
</foaf:Person>

<foaf:Person rdf:about="#a29105611">
  <rdfs:label>Karl Aberer</rdfs:label>
  <foaf:name rdf:datatype="&xsd:string">Karl Aberer</foaf:name>
  <foaf:firstName rdf:datatype="&xsd:string">Karl</foaf:firstName>
  <lastName rdf:datatype="&xsd:string">Aberer</lastName>
</foaf:Person>

<foaf:Person rdf:about="#a958684218">
  <rdfs:label>Peter Mika</rdfs:label>
  <foaf:name rdf:datatype="&xsd:string">Peter Mika</foaf:name>
  <foaf:firstName rdf:datatype="&xsd:string">Peter</foaf:firstName>
  <lastName rdf:datatype="&xsd:string">Mika</lastName>
</foaf:Person>

<foaf:Person rdf:about="#a94533498">
  <rdfs:label>Peter Haase</rdfs:label>
  <foaf:name rdf:datatype="&xsd:string">Peter Haase</foaf:name>
  <foaf:firstName rdf:datatype="&xsd:string">Peter</foaf:firstName>
  <lastName rdf:datatype="&xsd:string">Haase</lastName>
</foaf:Person>

<foaf:Person rdf:about="#a900366022">
  <rdfs:label>Avigdor Gal</rdfs:label>
  <foaf:name rdf:datatype="&xsd:string">Avigdor Gal</foaf:name>
  <foaf:firstName rdf:datatype="&xsd:string">Avigdor</foaf:firstName>

```

```

        <lastName rdf:datatype="&xsd:string">Gal</lastName>
</foaf:Person>

<!-- INSTITUTIONS -->

<Journal rdf:about="#a246119474">
  <rdfs:label>Journal of Web Semantics</rdfs:label>
  <foaf:name rdf:datatype="&xsd:string">Journal of Web Semantics</foaf:name>
  <shortName rdf:datatype="&xsd:string">JWS</shortName>
</Journal>

<Publisher rdf:about="#a131020767">
  <rdfs:label>Springer-Verlag</rdfs:label>
  <name rdf:datatype="&xsd:string">Springer-Verlag</name>
  <address>
    <Address>
      <city>Heidelberg</city>
      <country>DE</country>
    </Address>
  </address>
</Publisher>

<Journal rdf:about="#a70981683">
  <rdfs:label>Cybernetics and Control Theory</rdfs:label>
  <name rdf:datatype="&xsd:string">Cybernetics and Control Theory</name>
</Journal>

<Publisher rdf:about="#a85849488">
  <rdfs:label>The MIT Press</rdfs:label>
  <name rdf:datatype="&xsd:string">The MIT Press</name>
  <address>
    <Address>
      <city rdf:datatype="&xsd:string">Cambridge</city>
      <state rdf:datatype="&xsd:string">MA</state>
      <country rdf:datatype="&xsd:string">US</country>
    </Address>
  </address>
</Publisher>

<Journal rdf:about="#a362042121">
  <rdfs:label>International journal of intelligent system</rdfs:label>
  <name rdf:datatype="&xsd:string">International journal of intelligent system</name>
  <shortName rdf:datatype="&xsd:string">IJIS</shortName>
</Journal>

<Journal rdf:about="#a674639524">
  <rdfs:label>ACM SIGMOD Record</rdfs:label>
  <name rdf:datatype="&xsd:string">ACM SIGMOD Record</name>
</Journal>

<Journal rdf:about="#a906774044">
  <rdfs:label>VLDB Journal</rdfs:label>
  <name rdf:datatype="&xsd:string">VLDB Journal</name>
</Journal>

<!-- EVENTS -->

<Conference rdf:about="#spg04">
  <rdfs:label>SemPGrid 04 Workshop</rdfs:label>
  <name>SemPGrid 04 Workshop</name>
  <location>
    <Address>
      <city>New-York</city>
      <state>NY</state>
      <country>US</country>
    </Address>
  </location>
  <date>
    <Date>
      <month rdf:datatype="&xsd:gMonth">--05</month>
      <year rdf:datatype="&xsd:gYear">2004</year>
    </Date>
  </date>
</Conference>

<Conference rdf:about="#a72192307c">

```



```
<rdfs:label>13th Int. Conference on Knowledge Engineering and Management (EKAW-  
2002)</rdfs:label>  
<name>Int. Conference on Knowledge Engineering and Management</name>  
<shortName>EKAW</shortName>  
<issue>13</issue>  
<!--location>  
    <Address>  
        <city></city>  
        <state></state>  
        <country></country>  
    </Address>  
</location-->  
<date>  
    <Date>  
        <month rdf:datatype="xsd:gMonth">--10</month>  
        <year rdf:datatype="xsd:gYear">2002</year>  
    </Date>  
</date>  
</Conference>
```

```
<Conference rdf:about="#a32071928c">  
    <rdfs:label>First European Semantic Web Symposium</rdfs:label>  
    <name>European Semantic Web Symposium</name>  
    <shortName>ESWS</shortName>  
    <issue>1</issue>  
    <location>  
        <Address>  
            <city rdf:datatype="xsd:string">Heraklion</city>  
            <country>GR</country>  
        </Address>  
    </location>  
    <date>  
        <Date>  
            <month rdf:datatype="xsd:gMonth">--05</month>  
            <year rdf:datatype="xsd:gYear">2004</year>  
        </Date>  
    </date>  
</Conference>
```

```
<!-- REFERENCES -->
```

```
<Proceedings rdf:about="#a060097576">  
    <rdfs:label>Proceedings of the SemPGrid 04 Workshop</rdfs:label>  
    <title>Proceedings of the SemPGrid 04 Workshop</title>  
    <date><Date><year rdf:datatype="xsd:gYear">2004</year></Date></date>  
    <event rdf:resource="#spg04"/>  
</Proceedings>
```

```
<InProceedings rdf:about="#a64263824">  
    <rdfs:label>Bibster - A Semantics-Based Bibliographic Peer-to-Peer System</rdfs:label>  
    <author>  
        <PersonList>  
            <rdf:first rdf:resource="#a43836633"/>  
            <rdf:rest>  
                <PersonList>  
                    <rdf:first rdf:resource="#a712561038"/>  
                    <rdf:rest>  
                        <PersonList>  
                            <rdf:first rdf:resource="#a94533498"/>  
                            <rdf:rest>  
                                <PersonList>  
                                    <rdf:first rdf:resource="#a944339054"/>  
                                    <rdf:rest>  
                                        <PersonList>  
                                            <rdf:first rdf:resource="#a139477786"/>  
                                            <rdf:rest>  
                                                <PersonList>  
                                                    <rdf:first rdf:resource="#a958684218"/>  
                                                    <rdf:rest>  
                                                        <PersonList>  
                                                            <rdf:first  
  
rdf:resource="#a48552212"/>  
  
                                                                <rdf:rest>  
                                                                    <PersonList>  
                                                                        <rdf:first
```

```
rdf:resource="#a3105947"/>
```

```

        <rdf:rest
        </PersonList>
        </rdf:rest>
        </PersonList>
        </rdf:rest>
        </PersonList>
        </rdf:rest>
        </PersonList>
        </rdf:rest>
        </PersonList>
        </rdf:rest>
        </PersonList>
        </rdf:rest>
        </PersonList>
        </rdf:rest>
        </PersonList>
        </author>
        <proceedings rdf:resource="#a060097576"/>
        <title>Bibster - A Semantics-Based Bibliographic Peer-to-Peer System</title>
</InProceedings>

<!-- HERE -->

<InProceedings rdf:about="#a439508789">
    <rdfs:label>Measuring Similarity between Ontologies</rdfs:label>
    <author>
        <PersonList>
            <rdf:first rdf:resource="#a85228505"/>
            <rdf:rest>
                <PersonList>
                    <rdf:first rdf:resource="#a71003986"/>
                    <rdf:rest rdf:resource="#&rdf:nil" />
                </PersonList>
            </rdf:rest>
        </PersonList>
    </author>
    <proceedings rdf:resource="#a72192307"/>
    <title>Measuring Similarity between Ontologies</title>
</InProceedings>

<Article rdf:about="#a492378321">
    <rdfs:label>{P-Grid: A Self-organizing Structured P2P System}</rdfs:label>
    <author>
        <PersonList>
            <rdf:first rdf:resource="#a29105611"/>
            <rdf:rest>
                <PersonList>
                    <rdf:first rdf:resource="#a98078619"/>
                    <rdf:rest>
                        <PersonList>
                            <rdf:first rdf:resource="#a70955601"/>
                            <rdf:rest>
                                <PersonList>
                                    <rdf:first rdf:resource="#a39510672"/>
                                    <rdf:rest>
                                        <PersonList>
                                            <rdf:first rdf:resource="#a337716610"/>
                                            <rdf:rest>
                                                <PersonList>
                                                    <rdf:first rdf:resource="#a885257047"/>
                                                    <rdf:rest>
                                                        <PersonList>
                                                            <rdf:first
rdf:resource="#a250331360"/>
                                                            <rdf:rest rdf:resource="#&rdf:nil"
/>
                                                        </PersonList>
                                                    </rdf:rest>
                                                </PersonList>
                                            </rdf:rest>
                                        </PersonList>
                                    </rdf:rest>
                                </PersonList>
                            </rdf:rest>
                        </PersonList>
                    </rdf:rest>
                </PersonList>
            </rdf:rest>
        </PersonList>
    </author>
    <proceedings rdf:resource="#a72192307"/>
    <title>Measuring Similarity between Ontologies</title>
</InProceedings>

```

```

        </PersonList>
      </rdf:rest>
    </PersonList>
  </author>
  <journal rdf:resource="#a674639524"/>
  <title>{P-Grid}: A Self-organizing Structured P2P System</title>
  <date><Date><year rdf:datatype="&xsd:gYear">2003</year></Date></date>
</Article>

<Article rdf:about="#a475526642">
  <rdfs:label>Binary Codes capable of correcting deletions, insertions, and
reversals</rdfs:label>
  <author>
    <PersonList>
      <rdf:first rdf:resource="#a74993404"/>
      <rdf:rest rdf:resource="&rdf:nil" />
    </PersonList>
  </author>
  <journal rdf:resource="#a70981683"/>
  <title>Binary Codes capable of correcting deletions, insertions, and reversals</title>
  <date><Date><year rdf:datatype="&xsd:gYear">1996</year></Date></date>
</Article>

<InBook rdf:about="#a71568377">
  <rdfs:label>Database integration: the key to data interoperability</rdfs:label>
  <author>
    <PersonList>
      <rdf:first rdf:resource="#a11090777"/>
      <rdf:rest>
        <PersonList>
          <rdf:first rdf:resource="#a431956276"/>
          <rdf:rest rdf:resource="&rdf:nil" />
        </PersonList>
      </rdf:rest>
    </PersonList>
  </author>
  <book rdf:resource="#a108048723"/>
  <title>Database integration: the key to data interoperability</title>
  <editor>
    <PersonList>
      <rdf:first rdf:resource="#a431956276"/>
      <rdf:rest>
        <PersonList>
          <rdf:first rdf:resource="#a431956276b"/>
          <rdf:rest>
            <PersonList>
              <rdf:first rdf:resource="#a431956276c"/>
              <rdf:rest rdf:resource="&rdf:nil" />
            </PersonList>
          </rdf:rest>
        </PersonList>
      </rdf:rest>
    </PersonList>
  </editor>
</InBook>

<Proceedings rdf:about="#a72192307">
  <rdfs:label>Proc. Of the 13th Int. Conference on Knowledge Engineering and Management
(EKAW-2002)</rdfs:label>
  <publisher rdf:resource="#a131020767"/>
  <title>Proc. Of the 13th Int. Conference on Knowledge Engineering and Management (EKAW-
2002)</title>
  <event rdf:resource="#a72192307c"/>
  <date><Date><year rdf:datatype="&xsd:gYear">2002</year></Date></date>
</Proceedings>

<Proceedings rdf:about="#a32071928">
  <rdfs:label>Proceedings of the First European Semantic Web Symposium</rdfs:label>
  <publisher rdf:resource="#a131020767"/>
  <event rdf:resource="#a32071928c"/>
  <title>Proceedings of the First European Semantic Web Symposium</title>
  <date><Date><year rdf:datatype="&xsd:gYear">2004</year></Date></date>
</Proceedings>

<Misc rdf:about="#a140583454">
  <rdfs:label>{QOM} - Quick Ontology Mapping</rdfs:label>
  <author>

```

```

    <PersonList>
      <rdf:first rdf:resource="#a712561038"/>
      <rdf:rest>
        <PersonList>
          <rdf:first rdf:resource="#a71003986"/>
          <rdf:rest rdf:resource="&rdf:nil" />
        </PersonList>
      </rdf:rest>
    </PersonList>
  </author>
  <title>{QOM} - Quick Ontology Mapping</title>
  <note>submitted to the ISWC 04</note>
  <date><Date><year rdf:datatype="&xsd:gYear">2004</year></Date></date>
</Misc>

<InProceedings rdf:about="#a11065952">
  <rdfs:label>Ontology Mapping - An Integrated Approach</rdfs:label>
  <author>
    <PersonList>
      <rdf:first rdf:resource="#a712561038"/>
      <rdf:rest>
        <PersonList>
          <rdf:first rdf:resource="#a79573306"/>
          <rdf:rest rdf:resource="&rdf:nil" />
        </PersonList>
      </rdf:rest>
    </PersonList>
  </author>
  <proceedings rdf:resource="#a32071928"/>
  <title>Ontology Mapping - An Integrated Approach</title>
  <url>http://www.aifb.uni-
karlsruhe.de/WBS/meh/publications/ehrig04ontology_ESWS04.pdf</url>
</InProceedings>

<Article rdf:about="#a80299267">
  <rdfs:label>Start making sense: The Chatty Web approach for global semantic
agreements</rdfs:label>
  <author>
    <PersonList>
      <rdf:first rdf:resource="#a29105611"/>
      <rdf:rest>
        <PersonList>
          <rdf:first rdf:resource="#a98078619"/>
          <rdf:rest>
            <PersonList>
              <rdf:first rdf:resource="#a337716610"/>
              <rdf:rest rdf:resource="&rdf:nil" />
            </PersonList>
          </rdf:rest>
        </PersonList>
      </rdf:rest>
    </PersonList>
  </author>
  <journal rdf:resource="#a246119474"/>
  <title>Start making sense: The Chatty Web approach for global semantic
agreements</title>
  <date><Date><month rdf:datatype="&xsd:gMonth">--12</month><year
rdf:datatype="&xsd:gYear">2003</year></Date></date>
</Article>

<Monograph rdf:about="#a108048723">
  <rdfs:label>Object-Oriented Data Modeling</rdfs:label>
  <publisher rdf:resource="#a85849488"/>
  <title>Object-Oriented Data Modeling</title>
  <date><Date><year rdf:datatype="&xsd:gYear">2000</year></Date></date>
</Monograph>

<Article rdf:about="#a456080390">
  <rdfs:label>On dynamically generated ontology translators in agent
communication</rdfs:label>
  <author>
    <PersonList>
      <rdf:first rdf:resource="#a093016135"/>
      <rdf:rest>
        <PersonList>
          <rdf:first rdf:resource="#a572406328"/>
          <rdf:rest>

```

```

        <PersonList>
          <rdf:first rdf:resource="#a086379337"/>
          <rdf:rest>
            <PersonList>
              <rdf:first rdf:resource="#a04570373"/>
              <rdf:rest rdf:resource="&rdf:nil" />
            </PersonList>
          </rdf:rest>
        </PersonList>
      </rdf:rest>
    </PersonList>
  </rdf:rest>
</author>
<journal rdf:resource="#a362042121"/>
<title>On dynamically generated ontology translators in agent communication</title>
<pages>
  <PageRange>
    <startPage rdf:datatype="&xsd;nonNegativeInteger">587</startPage>
    <endPage rdf:datatype="&xsd;nonNegativeInteger">607</endPage>
  </PageRange>
</pages>
<date><Date><month rdf:datatype="&xsd:gMonth">--12</month><year
rdf:datatype="&xsd:gYear">2001</year></Date></date>
</Article>

<Article rdf:about="#a846015923">
  <rdfs:label>A Framework for Modeling and Evaluating Automatic Semantic
Reconciliation</rdfs:label>
  <author>
    <PersonList>
      <rdf:first rdf:resource="#a900366022"/>
      <rdf:rest>
        <PersonList>
          <rdf:first rdf:resource="#a467748807"/>
          <rdf:rest>
            <PersonList>
              <rdf:first rdf:resource="#a971541439"/>
              <rdf:rest>
                <PersonList>
                  <rdf:first rdf:resource="#a066600210"/>
                  <rdf:rest rdf:resource="&rdf:nil" />
                </PersonList>
              </rdf:rest>
            </PersonList>
          </rdf:rest>
        </PersonList>
      </rdf:rest>
    </PersonList>
  </author>
  <journal rdf:resource="#a906774044"/>
  <title>A Framework for Modeling and Evaluating Automatic Semantic
Reconciliation</title>
  <note>to appear</note>
  <date><Date><year rdf:datatype="&xsd:gYear">2004</year></Date></date>
</Article>
</rdf:RDF>

```

2) Ontologia 304 da OAEI

```
<?xml version="1.0" encoding="utf-8"?>

<!-- NE PAS OUBLIER DE MODIFIER DES URIREFS -->
<!DOCTYPE rdf:RDF [
    <!ENTITY xsd          "http://www.w3.org/2001/XMLSchema#" >
    <!ENTITY rdf          "http://www.w3.org/1999/02/22-rdf-syntax-ns#" >
    <!ENTITY rdfs         "http://www.w3.org/2000/01/rdf-schema#" >
    <!ENTITY dc           "http://purl.org/dc/elements/1.1/" >
    <!ENTITY owl        "http://www.w3.org/2002/07/owl#" >
    <!ENTITY foaf         "http://xmlns.com/foaf/0.1/" >
    <!ENTITY units       "http://visus.mit.edu/fontomri/0.01/units.owl#" >
    <!ENTITY wot         "http://xmlns.com/wot/0.1/" >
    <!ENTITY bibtex       "http://purl.org/net/nknouf/ns/bibtex#" >
    <!ENTITY dcterms      "http://purl.org/dc/terms/" >
    <!ENTITY dctype       "http://purl.org/dc/dcmitype/" > ]>

<rdf:RDF
    xmlns="http://oaei.ontologymatching.org/2011/benchmarks/304/onto.rdf#"
    xml:base="http://oaei.ontologymatching.org/2011/benchmarks/304/onto.rdf#"
    xmlns:units    ="&units;"
    xmlns:foaf     ="&foaf;"
    xmlns:rdf      ="&rdf;"
    xmlns:xsd      ="&xsd;"
    xmlns:rdfs     ="&rdfs;"
    xmlns:owl      ="&owl;"
    xmlns:wot      ="&wot;"
    xmlns:dc       ="&dc;"
    xmlns:dcterms  ="&dcterms;"
    xmlns:dctype   ="&dctype;"
    xmlns:bibtex   ="&bibtex;"

    <owl:Ontology rdf:about="">
        <dc:creator>Nick Knouf &lt;nknouf@mit.edu&gt;</dc:creator>
        <dc:contributor>Antoine Zimmermann
&lt;antoine.zimmermann@inrialpes.fr&gt;</dc:contributor>
        <dc:description>Transformation of bibTeX into an OWL ontology</dc:description>
        <dc:date>2004-01-10</dc:date>
        <dc:subject>fMRI, ontology, fMRI analysis, foo, bar</dc:subject>
        <rdfs:label>bibtex ontology</rdfs:label>
        <dc:title>bibtex.owl: an ontology for bibTeX entries</dc:title>
        <dc:identifier
rdf:datatype="&xsd:anyURI">http://visus.mit.edu/bibtex/0.01/bibtex.owl</dc:identifier>
        <rdfs:comment>Possible ontology to describe bibTeX entries.</rdfs:comment>
        <owl:versionInfo>0.01</owl:versionInfo>
    </owl:Ontology>

    <!-- foaf extensions -->
    <owl:Class rdf:about="&foaf;Person" />
    <owl:Class rdf:about="&foaf;Organization" />
    <owl:DatatypeProperty rdf:about="&foaf;firstName" />
    <owl:DatatypeProperty rdf:ID="lastName" />
    <owl:DatatypeProperty rdf:about="&foaf;name" />

    <!--
    ***** ENTRIES *****

    Entries form the basis of a bibTeX database and are categorized by their type, such as a
    book, journal article, conference proceedings, etc. Each entry type has a specific set of
    fields that are required in bibTeX; as a first pass, these fields will also be required in the
    OWL ontology.

    Note: all rdfs:comment values for the entries come from
    http://newton.ex.ac.uk/tex/pack/bibtex/btxdoc/node6.html.
    -->
    <owl:Class rdf:ID="Entry">
        <rdfs:label xml:lang="en">Entry</rdfs:label>
        <rdfs:comment xml:lang="en">Base class for all entries</rdfs:comment>
        <rdfs:subClassOf>
            <owl:Restriction>
                <owl:onProperty rdf:resource="#date" />
                <owl:maxCardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:maxCardinality>
            </owl:Restriction>

```

```

</rdfs:subClassOf>
<rdfs:subClassOf>
  <owl:Restriction>
    <owl:onProperty rdf:resource="#title" />
    <owl:maxCardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:maxCardinality>
  </owl:Restriction>
</rdfs:subClassOf>
<rdfs:subClassOf>
  <owl:Restriction>
    <owl:onProperty rdf:resource="#author" />
    <owl:maxCardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:maxCardinality>
  </owl:Restriction>
</rdfs:subClassOf>
</owl:Class>

<!--
  Trouver un autre nom por cette classe.
  Regroupe Book (Monography+Collection), InBook (InMonography, InCollection), Article
-->
<owl:Class rdf:ID="Published">
  <rdfs:subClassOf rdf:resource="#Entry" />
  <rdfs:label xml:lang="en">Published</rdfs:label>
  <rdfs:comment xml:lang="en">.</rdfs:comment>
</owl:Class>

<owl:Class rdf:ID="Article">
  <rdfs:subClassOf rdf:resource="#Published" />
  <rdfs:subClassOf rdf:resource="#Part" />
  <rdfs:label xml:lang="en">Article</rdfs:label>
  <rdfs:comment xml:lang="en">An article from a journal or magazine.</rdfs:comment>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#author" />
      <owl:cardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:cardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#title" />
      <owl:cardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:cardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#inJournal" />
      <owl:cardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:cardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#date" />
      <owl:cardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:cardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#date" />
      <owl:allValuesFrom rdf:resource="#DateWithMonth" />
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#number" />
      <owl:maxCardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:maxCardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#volume" />
      <owl:maxCardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:maxCardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
</owl:Class>

<owl:Class rdf:ID="Composite">
  <rdfs:subClassOf rdf:resource="#Entry" />

```

```

<rdfs:label xml:lang="en">Book</rdfs:label>
<rdfs:comment xml:lang="en">A work that may be composed of several parts or chapter. May
be a monography, a collection or a proceedings.</rdfs:comment>
<rdfs:subClassOf>
  <owl:Restriction>
    <owl:onProperty rdf:resource="#title" />
    <owl:cardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:cardinality>
  </owl:Restriction>
</rdfs:subClassOf>
<rdfs:subClassOf>
  <owl:Restriction>
    <owl:onProperty rdf:resource="#numberOrVolume" />
    <owl:maxCardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:maxCardinality>
  </owl:Restriction>
</rdfs:subClassOf>
<rdfs:subClassOf>
  <owl:Restriction>
    <owl:onProperty rdf:resource="#publisher" />
    <owl:maxCardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:maxCardinality>
  </owl:Restriction>
</rdfs:subClassOf>
<rdfs:subClassOf>
  <owl:Restriction>
    <owl:onProperty rdf:resource="#series" />
    <owl:maxCardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:maxCardinality>
  </owl:Restriction>
</rdfs:subClassOf>
<rdfs:subClassOf>
  <owl:Restriction>
    <owl:onProperty rdf:resource="#title" />
    <owl:cardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:cardinality>
  </owl:Restriction>
</rdfs:subClassOf>
<rdfs:subClassOf>
  <owl:Restriction>
    <owl:onProperty rdf:resource="#date" />
    <owl:cardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:cardinality>
  </owl:Restriction>
</rdfs:subClassOf>
<rdfs:subClassOf>
  <owl:Restriction>
    <owl:onProperty rdf:resource="#date" />
    <owl:allValuesFrom rdf:resource="#DateWithMonth" />
  </owl:Restriction>
</rdfs:subClassOf>
</owl:Class>

<owl:Class rdf:ID="Book">
  <rdfs:subClassOf rdf:resource="#Composite" />
  <rdfs:subClassOf rdf:resource="#Published" />
  <rdfs:label xml:lang="en">Book</rdfs:label>
  <rdfs:comment xml:lang="en">A single book that may be a monography or a collection of
written texts.</rdfs:comment>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#humanCreator" />
      <owl:cardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:cardinality>
      <dc:description>One thing to figure out is whether or not this tells us what we want;
that is, does this imply that we must have at least one property of humanCreator, which could
be either author or editor?</dc:description>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#edition" />
      <owl:maxCardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:maxCardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
</owl:Class>

<owl:Class rdf:ID="Monography">
  <rdfs:subClassOf rdf:resource="#Book" />
  <rdfs:label xml:lang="en">Monography</rdfs:label>
  <rdfs:comment xml:lang="en">A book that is a single entity, as opposed to a
collection.</rdfs:comment>
</owl:Class>

```



```

<owl:Class rdf:ID="Collection">
  <rdfs:subClassOf rdf:resource="#Book" />
  <rdfs:label xml:lang="en">Collection</rdfs:label>
  <rdfs:comment xml:lang="en">A book that is collection of texts or articles.</rdfs:comment>
</owl:Class>

<owl:Class rdf:ID="Informal">
  <rdfs:subClassOf rdf:resource="#Entry" />
  <rdfs:label xml:lang="en">Informal</rdfs:label>
  <rdfs:comment xml:lang="en">A document that that was informally published or not
published.</rdfs:comment>
</owl:Class>

<owl:Class rdf:ID="Booklet">
  <rdfs:subClassOf rdf:resource="#Informal" />
  <rdfs:label xml:lang="en">Booklet</rdfs:label>
  <rdfs:comment xml:lang="en">A work that is printed and bound, but without a named
publisher or sponsoring institution.</rdfs:comment>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#title" />
      <owl:cardinality rdf:datatype="xsd:nonNegativeInteger">1</owl:cardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
</owl:Class>

<!--
<owl:Class rdf:ID="Conference">
  <rdfs:subClassOf rdf:resource="#Part" />
  <owl:equivalentClass rdf:resource="#InProceedings" />
  <rdfs:label xml:lang="en">Conference</rdfs:label>
  <rdfs:comment xml:lang="en">The same as INPROCEEDINGS, included for Scribe
compatibility.</rdfs:comment>
</owl:Class>
-->

<owl:Class rdf:ID="Deliverable">
  <rdfs:subClassOf rdf:resource="#Informal" />
  <rdfs:label xml:lang="en">Conference</rdfs:label>
  <rdfs:comment xml:lang="en">TROUVER UNE DESCRIPTION POUR CA.</rdfs:comment>
</owl:Class>

<owl:Class rdf:ID="Part">
  <rdfs:subClassOf rdf:resource="#Entry" />
  <rdfs:label xml:lang="en">Part</rdfs:label>
  <rdfs:comment xml:lang="en">A part of something (either Book or
Proceedings).</rdfs:comment>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#pages" />
      <owl:maxCardinality rdf:datatype="xsd:nonNegativeInteger">1</owl:maxCardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#title" />
      <owl:cardinality rdf:datatype="xsd:nonNegativeInteger">1</owl:cardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#isPartOf" />
      <owl:cardinality rdf:datatype="xsd:nonNegativeInteger">1</owl:cardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
</owl:Class>

<owl:Class rdf:ID="BookPart">
  <rdfs:subClassOf rdf:resource="#Part" />
  <rdfs:subClassOf rdf:resource="#Published" />
  <rdfs:label xml:lang="en">BookPart</rdfs:label>
  <rdfs:comment xml:lang="en">May be a chapter (or section or whatever), a range of pages or
a part of a book having its own title.</rdfs:comment>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#chapter" />
      <owl:maxCardinality rdf:datatype="xsd:nonNegativeInteger">1</owl:maxCardinality>
    </owl:Restriction>
  </rdfs:subClassOf>

```

```

    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#pagesOrChapter" />
      <owl:maxCardinality rdf:datatype="&xsd;nonNegativeInteger">2</owl:maxCardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
</owl:Class>

<owl:Class rdf:ID="InBook">
  <rdfs:subClassOf rdf:resource="#Chapter" />
  <rdfs:label xml:lang="en">InBook</rdfs:label>
  <rdfs:comment xml:lang="en">May be a chapter (or section or whatever) and/or a range of
pages.</rdfs:comment>
  <rdfs:subClassOf>
    <owl:Restriction>
      <!-- author OR editor (given in the Book description) ??? -->
      <owl:onProperty rdf:resource="#author" />
      <owl:cardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:cardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#pagesOrChapter" />
      <owl:minCardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:minCardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#inMonography" />
      <owl:cardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:cardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
</owl:Class>

<owl:Class rdf:ID="InCollection">
  <rdfs:subClassOf rdf:resource="#Chapter" />
  <rdfs:label xml:lang="en">Incollection</rdfs:label>
  <rdfs:comment xml:lang="en">A part of a book having its own title.</rdfs:comment>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#author" />
      <owl:cardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:cardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#inCollection" />
      <owl:cardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:cardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
</owl:Class>

<owl:Class rdf:ID="InProceedings">
  <rdfs:subClassOf rdf:resource="#Part" />
  <rdfs:label xml:lang="en">InProceedings</rdfs:label>
  <rdfs:comment xml:lang="en">An article in a conference proceedings.</rdfs:comment>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#author" />
      <owl:cardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:cardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#inProceedings" />
      <owl:cardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:cardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
</owl:Class>

<owl:Class rdf:ID="LectureNotes">
  <rdfs:subClassOf rdf:resource="#Informal" />
  <rdfs:subClassOf rdf:resource="#Academic" />
  <rdfs:label xml:lang="en">LectureNotes</rdfs:label>
  <rdfs:comment xml:lang="en">Lecture notes.</rdfs:comment>

```

```

</owl:Class>

<owl:Class rdf:ID="Manual">
  <rdfs:subClassOf rdf:resource="#Informal" />
  <rdfs:label xml:lang="en">Manual</rdfs:label>
  <rdfs:comment xml:lang="en">Technical documentation.</rdfs:comment>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#organization" />
      <owl:maxCardinality rdf:datatype="xsd:nonNegativeInteger">1</owl:maxCardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#edition" />
      <owl:maxCardinality rdf:datatype="xsd:nonNegativeInteger">1</owl:maxCardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#title" />
      <owl:cardinality rdf:datatype="xsd:nonNegativeInteger">1</owl:cardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
</owl:Class>

<owl:Class rdf:ID="Academic">
  <rdfs:subClassOf rdf:resource="#Entry" />
  <rdfs:label xml:lang="en">Academic</rdfs:label>
  <rdfs:comment xml:lang="en">A Master's or PhD thesis.</rdfs:comment>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#author" />
      <owl:cardinality rdf:datatype="xsd:nonNegativeInteger">1</owl:cardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#title" />
      <owl:cardinality rdf:datatype="xsd:nonNegativeInteger">1</owl:cardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#school" />
      <owl:cardinality rdf:datatype="xsd:nonNegativeInteger">1</owl:cardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
</owl:Class>

<owl:Class rdf:ID="Thesis">
  <rdfs:subClassOf rdf:resource="#Academic" />
  <rdfs:label xml:lang="en">Thesis</rdfs:label>
  <rdfs:comment xml:lang="en">A thesis (either Master or PhD).</rdfs:comment>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#date" />
      <owl:cardinality rdf:datatype="xsd:nonNegativeInteger">1</owl:cardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#date" />
      <owl:allValuesFrom rdf:resource="#DateWithMonth" />
    </owl:Restriction>
  </rdfs:subClassOf>
</owl:Class>

<owl:Class rdf:ID="MastersThesis">
  <rdfs:subClassOf rdf:resource="#Thesis" />
  <rdfs:label xml:lang="en">MastersThesis</rdfs:label>
  <rdfs:comment xml:lang="en">A Master's thesis.</rdfs:comment>
</owl:Class>

<owl:Class rdf:ID="Misc">
  <rdfs:subClassOf rdf:resource="#Informal" />
  <rdfs:label xml:lang="en">Misc</rdfs:label>

```

```

    <rdfs:comment xml:lang="en">Use this type when nothing else fits.</rdfs:comment>
  </owl:Class>

  <owl:Class rdf:ID="PhdThesis">
    <rdfs:subClassOf rdf:resource="#Thesis" />
    <rdfs:label xml:lang="en">PhdThesis</rdfs:label>
    <rdfs:comment xml:lang="en">A PhD thesis.</rdfs:comment>
  </owl:Class>

  <owl:Class rdf:ID="Proceedings">
    <rdfs:subClassOf rdf:resource="#Composite" />
    <rdfs:label xml:lang="en">Proceedings</rdfs:label>
    <rdfs:comment xml:lang="en">The proceedings of a conference.</rdfs:comment>
    <rdfs:comment xml:lang="en">A proceedings may be implicitly defined with an
inproceedings.</rdfs:comment>
    <rdfs:subClassOf>
      <owl:Restriction>
        <owl:onProperty rdf:resource="#publisher" />
        <owl:maxCardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:maxCardinality>
      </owl:Restriction>
    </rdfs:subClassOf>
    <rdfs:subClassOf>
      <owl:Restriction>
        <owl:onProperty rdf:resource="#editor" />
        <owl:maxCardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:maxCardinality>
      </owl:Restriction>
    </rdfs:subClassOf>
    <rdfs:subClassOf>
      <owl:Restriction>
        <owl:onProperty rdf:resource="#organization" />
        <owl:maxCardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:maxCardinality>
      </owl:Restriction>
    </rdfs:subClassOf>
  </owl:Class>

  <owl:Class rdf:ID="TechReport">
    <rdfs:subClassOf rdf:resource="#Informal" />
    <rdfs:label xml:lang="en">TechReport</rdfs:label>
    <rdfs:comment xml:lang="en">A report published by a school or other institution, usually
numbered within a series.</rdfs:comment>
    <rdfs:subClassOf>
      <owl:Restriction>
        <owl:onProperty rdf:resource="#author" />
        <owl:cardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:cardinality>
      </owl:Restriction>
    </rdfs:subClassOf>
    <rdfs:subClassOf>
      <owl:Restriction>
        <owl:onProperty rdf:resource="#title" />
        <owl:cardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:cardinality>
      </owl:Restriction>
    </rdfs:subClassOf>
    <rdfs:subClassOf>
      <owl:Restriction>
        <owl:onProperty rdf:resource="#institution" />
        <owl:cardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:cardinality>
      </owl:Restriction>
    </rdfs:subClassOf>
    <rdfs:subClassOf>
      <owl:Restriction>
        <owl:onProperty rdf:resource="#date" />
        <owl:cardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:cardinality>
      </owl:Restriction>
    </rdfs:subClassOf>
    <rdfs:subClassOf>
      <owl:Restriction>
        <owl:onProperty rdf:resource="#date" />
        <owl:allValuesFrom rdf:resource="#DateWithMonth" />
      </owl:Restriction>
    </rdfs:subClassOf>
    <rdfs:subClassOf>
      <owl:Restriction>
        <owl:onProperty rdf:resource="#number" />
        <owl:maxCardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:maxCardinality>
      </owl:Restriction>
    </rdfs:subClassOf>
  </owl:Class>

```

```

<owl:Class rdf:ID="Unpublished">
  <rdfs:subClassOf rdf:resource="#Informal" />
  <rdfs:label xml:lang="en">Unpublished</rdfs:label>
  <rdfs:comment xml:lang="en">A document having an author and title, but not formally
published.</rdfs:comment>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#author" />
      <owl:cardinality rdf:datatype="xsd:nonNegativeInteger">1</owl:cardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#title" />
      <owl:cardinality rdf:datatype="xsd:nonNegativeInteger">1</owl:cardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#note" />
      <owl:minCardinality rdf:datatype="xsd:nonNegativeInteger">1</owl:minCardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
</owl:Class>

<owl:Class rdf:ID="MotionPicture">
  <rdfs:subClassOf rdf:resource="#Entry" />
  <rdfs:label xml:lang="en">MotionPicture</rdfs:label>
  <rdfs:comment xml:lang="en">A film/movie/motion picture.</rdfs:comment>
</owl:Class>

<!--
***** FIELDS *****

Fields can be considered the properties of a bibTeX entry. Each entry type (described
above) has a different set of required and optional fields. The fields will all be datatype
properties for the time being.

Note: all rdfs:comment values for the entries come from
http://newton.ex.ac.uk/tex/pack/bibtex/btxdoc/node7.html
-->

<owl:DatatypeProperty rdf:ID="key">
  <rdfs:domain rdf:resource="#Entry" />
  <rdfs:range rdf:resource="xsd:string" />
  <rdfs:label xml:lang="en">key</rdfs:label>
  <rdfs:comment xml:lang="en">The key for a particular bibTeX entry. Note that the rdf:ID
for each Entry instance could be the bibTeX key as well, possibly making this property
redundant.</rdfs:comment>
</owl:DatatypeProperty>

<!--

Les adresses seront indiquées dans les descriptions des objets adéquats

<owl:DatatypeProperty rdf:ID="address">
  <rdfs:domain rdf:resource="#Entry" />
  <rdfs:range rdf:resource="xsd:string" />
  <rdfs:label xml:lang="en">address</rdfs:label>
  <rdfs:comment xml:lang="en">Usually the address of the publisher or other type of
institution. For major publishing houses, van Leunen recommends omitting the information
entirely. For small publishers, on the other hand, you can help the reader by giving the
complete address.</rdfs:comment>
</owl:DatatypeProperty>
-->

<owl:ObjectProperty rdf:ID="annote">
  <rdfs:domain rdf:resource="#Entry" />
  <rdfs:range rdf:resource="#Text" />
  <rdfs:label xml:lang="en">annote</rdfs:label>
  <rdfs:comment xml:lang="en">An annotation. It is not used by the standard bibliography
styles, but may be used by others that produce an annotated bibliography.</rdfs:comment>
</owl:ObjectProperty>

<owl:ObjectProperty rdf:ID="author">

```

```

<rdfs:domain rdf:resource="#Entry" />
<rdfs:range rdf:resource="#PersonList" />
<rdfs:subPropertyOf rdf:resource="#humanCreator" />
<rdfs:label xml:lang="en">author</rdfs:label>
<rdfs:comment xml:lang="en">The name(s) of the author(s), in the format described in the
LaTeX book.</rdfs:comment>
<dc:description>This is tricky due to the fact that order is not (generally) preserved in
RDF documents. The problem arises when you want to have an author list where the order is
_extremely_ important. How shall we do that? Perhaps we want to define "hasPrimaryAuthor",
"hasSecondaryAuthor", "hasTertiaryAuthor", and "hasRemainingAuthors", or something of that
sort. This will be have to given more thought.</dc:description>
</owl:ObjectProperty>

<owl:ObjectProperty rdf:ID="chapter">
<rdfs:domain rdf:resource="#Chapter" />
<rdfs:range rdf:resource="#Text" />
<rdfs:subPropertyOf rdf:resource="#pagesOrChapter" />
<rdfs:label xml:lang="en">chapter</rdfs:label>
<rdfs:comment xml:lang="en">A chapter (or section or whatever) number.</rdfs:comment>
</owl:ObjectProperty>

<owl:ObjectProperty rdf:ID="edition">
<rdfs:domain>
<owl:Class>
<owl:unionOf rdf:parseType="Collection">
<owl:Class rdf:about="#Book" />
<owl:Class rdf:about="#Manual" />
</owl:unionOf>
</owl:Class>
</rdfs:domain>
<rdfs:range rdf:resource="#Text" />
<rdfs:label xml:lang="en">edition</rdfs:label>
<rdfs:comment xml:lang="en">The edition of a book--for example, "Second". This should be
an ordinal, and should have the first letter capitalized, as shown here; the standard styles
convert to lower case when necessary.</rdfs:comment>
</owl:ObjectProperty>

<owl:ObjectProperty rdf:ID="editor">
<rdfs:domain rdf:resource="#Composite" />
<rdfs:range rdf:resource="#PersonList" />
<rdfs:subPropertyOf rdf:resource="#humanCreator" />
<rdfs:label xml:lang="en">editor</rdfs:label>
<rdfs:comment xml:lang="en">Name(s) of editor(s), typed as indicated in the LaTeX book. If
there is also an author field, then the editor field gives the editor of the book or
collection in which the reference appears.</rdfs:comment>
<dc:description>Again, the same issues that arose with the "author" property apply
here.</dc:description>
</owl:ObjectProperty>

<owl:ObjectProperty rdf:ID="howPublished">
<rdfs:domain>
<owl:Class>
<owl:unionOf rdf:parseType="Collection">
<owl:Class rdf:about="#Misc" />
<owl:Class rdf:about="#Booklet" />
</owl:unionOf>
</owl:Class>
</rdfs:domain>
<rdfs:range rdf:resource="#Text" />
<rdfs:label xml:lang="en">published</rdfs:label>
<rdfs:comment xml:lang="en">How something strange has been published. The first word
should be capitalized.</rdfs:comment>
</owl:ObjectProperty>

<owl:ObjectProperty rdf:ID="isPartOf">
<rdfs:domain rdf:resource="#Part" />
<rdfs:label xml:lang="en">is part of</rdfs:label>
<rdf:type rdf:resource="#owl:TransitiveProperty" />
<rdfs:comment xml:lang="en">The document that contains a text or article.</rdfs:comment>
<dc:description>This is an object property that refers to another entry an external
objects.</dc:description>
</owl:ObjectProperty>

<owl:ObjectProperty rdf:ID="institution">
<rdfs:domain rdf:resource="#TechReport" />
<!-- trouver une ontologie avec institution ??? -->
<rdfs:range rdf:resource="#Institution" />

```

```

    <rdfs:label xml:lang="en">institution</rdfs:label>
    <rdfs:comment xml:lang="en">The sponsoring institution of a technical
report.</rdfs:comment>
    <dc:description>This could be an object property that refers to an external set of
institution instances.</dc:description>
  </owl:ObjectProperty>

  <owl:ObjectProperty rdf:ID="inJournal">
    <rdfs:domain rdf:resource="#Article" />
    <!-- trouver une ontologie avec journal -->
    <rdfs:range rdf:resource="#Journal" />
    <rdfs:subPropertyOf rdf:resource="#isPartOf" />
    <rdfs:label xml:lang="en">in journal</rdfs:label>
    <rdfs:comment xml:lang="en">A journal name. Abbreviations are provided for many journals;
see the Local Guide</rdfs:comment>
    <dc:description>This could optionally be an object property, whereby the range would refer
to an external set of journal instances, thus providing standardized abbreviations for
different bibliographic styles.</dc:description>
  </owl:ObjectProperty>

  <owl:ObjectProperty rdf:ID="date">
    <rdfs:domain rdf:resource="#Entry" />
    <rdfs:range rdf:resource="#Date" />
    <rdfs:label xml:lang="en">date</rdfs:label>
    <rdfs:comment xml:lang="en">The date when the work was published or, for an unpublished
work, in which it was written.</rdfs:comment>
  </owl:ObjectProperty>

  <owl:ObjectProperty rdf:ID="note">
    <rdfs:domain rdf:resource="#Entry" />
    <rdfs:range rdf:resource="#Text" />
    <rdfs:label xml:lang="en">note</rdfs:label>
    <rdfs:comment xml:lang="en">Any additional information that can help the reader. The first
word should be capitalized.</rdfs:comment>
  </owl:ObjectProperty>

  <owl:DatatypeProperty rdf:ID="number">
    <rdfs:domain>
      <owl:Class>
        <owl:unionOf rdf:parseType="Collection">
          <owl:Class rdf:about="#Article" />
          <owl:Class rdf:about="#Composite" />
        </owl:unionOf>
      </owl:Class>
    </rdfs:domain>
    <rdfs:range rdf:resource="xsd:string" />
    <rdfs:subPropertyOf rdf:resource="#numberOrVolume" />
    <rdfs:label xml:lang="en">number</rdfs:label>
    <rdfs:comment xml:lang="en">The number of a journal, magazine, technical report, or of a
work in a series. An issue of a journal or magazine is usually identified by its volume and
number; the organization that issues a technical report usually gives it a number; and
sometimes books are given numbers in a named series.</rdfs:comment>
  </owl:DatatypeProperty>

  <owl:ObjectProperty rdf:ID="organization">
    <rdfs:domain>
      <owl:Class>
        <owl:unionOf rdf:parseType="Collection">
          <owl:Class rdf:about="#Proceedings" />
          <owl:Class rdf:about="#Manual" />
        </owl:unionOf>
      </owl:Class>
    </rdfs:domain>
    <rdfs:range rdf:resource="#Organization" />
    <rdfs:label xml:lang="en">organization</rdfs:label>
    <rdfs:comment xml:lang="en">The organization that sponsors a conference or that publishes
a manual.</rdfs:comment>
    <dc:description>This could be an object property that refers to an external set of
organization instances.</dc:description>
  </owl:ObjectProperty>

  <owl:ObjectProperty rdf:ID="pages">
    <rdfs:domain>
      <owl:Class>
        <owl:unionOf rdf:parseType="Collection">
          <owl:Class rdf:about="#Part" />
        </owl:unionOf>

```

```

    </owl:Class>
  </rdfs:domain>
  <rdfs:range rdf:resource="#PageRange" />
  <rdfs:subPropertyOf rdf:resource="#pagesOrChapter" />
  <rdfs:label xml:lang="en">pages</rdfs:label>
  <rdfs:comment xml:lang="en">A range of page, with an upper and lower limit.</rdfs:comment>
</owl:ObjectProperty>

<owl:ObjectProperty rdf:ID="publisher">
  <rdfs:domain rdf:resource="#Composite" />
  <!-- trouver une ontologie avec Publisher ??? -->
  <rdfs:range rdf:resource="#Publisher" />
  <rdfs:label xml:lang="en">publisher</rdfs:label>
  <rdfs:comment xml:lang="en">The publisher's name.</rdfs:comment>
  <dc:description>This is a case where an ObjectProperty might be a better choice, where the
range is some set of publisher names defined in another ontology. That would allow all of the
metadata for the publisher to be incorporated as needed.</dc:description>
</owl:ObjectProperty>

<owl:ObjectProperty rdf:ID="school">
  <rdfs:domain>
    <owl:Class>
      <owl:unionOf rdf:parseType="Collection">
        <owl:Class rdf:about="#Thesis" />
        <owl:Class rdf:about="#LectureNotes" />
      </owl:unionOf>
    </owl:Class>
  </rdfs:domain>
  <!-- trouver une ontologie avec school ??? -->
  <rdfs:range rdf:resource="#School" />
  <rdfs:label xml:lang="en">school</rdfs:label>
  <rdfs:comment xml:lang="en">The name of the school where a thesis was
written.</rdfs:comment>
  <dc:description>As with "publisher", this could be an ObjectProperty that refers to an
external set of school instances.</dc:description>
</owl:ObjectProperty>

<owl:ObjectProperty rdf:ID="series">
  <rdfs:domain rdf:resource="#Composite" />
  <rdfs:range rdf:resource="#Title" />
  <rdfs:label xml:lang="en">series</rdfs:label>
  <rdfs:comment xml:lang="en">The name of a series or set of books. When citing an entire
book, the the title field gives its title and an optional series field gives the name of a
series or multi-volume set in which the book is published.</rdfs:comment>
</owl:ObjectProperty>

<owl:ObjectProperty rdf:ID="title">
  <rdfs:domain rdf:resource="#Entry" />
  <rdfs:range rdf:resource="#Title" />
  <rdfs:label xml:lang="en">title</rdfs:label>
  <rdfs:comment xml:lang="en">The work's title, typed as explained in the LaTeX
book.</rdfs:comment>
</owl:ObjectProperty>

<owl:ObjectProperty rdf:ID="type">
  <rdfs:domain>
    <owl:Class>
      <owl:unionOf rdf:parseType="Collection">
        <owl:Class rdf:about="#Chapter" />
        <owl:Class rdf:about="#TechReport" />
        <owl:Class rdf:about="#Thesis" />
      </owl:unionOf>
    </owl:Class>
  </rdfs:domain>
  <rdfs:range rdf:resource="#Text" />
  <rdfs:label xml:lang="en">type</rdfs:label>
  <rdfs:comment xml:lang="en">The type of a technical report--for example, "Research
Note".</rdfs:comment>
</owl:ObjectProperty>

<owl:DatatypeProperty rdf:ID="volume">
  <rdfs:domain>
    <owl:Class>
      <owl:unionOf rdf:parseType="Collection">
        <owl:Class rdf:about="#Article" />
        <owl:Class rdf:about="#Composite" />
      </owl:unionOf>
    </owl:Class>
  </rdfs:domain>
  <rdfs:range rdf:resource="#Text" />
  <rdfs:label xml:lang="en">volume</rdfs:label>
  <rdfs:comment xml:lang="en">The volume number of a work, if it is part of a series or multi-volume set.
</rdfs:comment>
</owl:DatatypeProperty>

```



```

    </owl:Class>
  </rdfs:domain>
  <rdfs:subPropertyOf rdf:resource="#numberOrVolume" />
  <rdfs:range rdf:resource="xsd:nonNegativeInteger" />
  <rdfs:label xml:lang="en">volume</rdfs:label>
  <rdfs:comment xml:lang="en">The volume of a journal or multivolume book.</rdfs:comment>
</owl:DatatypeProperty>

<owl:ObjectProperty rdf:ID="affiliation">
  <rdfs:domain rdf:resource="#Entry" />
  <rdfs:range rdf:resource="#Text" />
  <rdfs:label xml:lang="en">affiliation</rdfs:label>
  <rdfs:comment xml:lang="en">The authors affiliation.</rdfs:comment>
</owl:ObjectProperty>

<owl:ObjectProperty rdf:ID="abstract">
  <rdfs:domain rdf:resource="#Entry" />
  <rdfs:range rdf:resource="#Text" />
  <rdfs:label xml:lang="en">abstract</rdfs:label>
  <rdfs:comment xml:lang="en">An abstract of the work.</rdfs:comment>
</owl:ObjectProperty>

<owl:ObjectProperty rdf:ID="contents">
  <rdfs:domain rdf:resource="#Entry" />
  <rdfs:range rdf:resource="#Text" />
  <rdfs:label xml:lang="en">ontents</rdfs:label>
  <rdfs:comment xml:lang="en">A Table of Contents.</rdfs:comment>
</owl:ObjectProperty>

<owl:ObjectProperty rdf:ID="copyright">
  <rdfs:domain rdf:resource="#Entry" />
  <rdfs:range rdf:resource="#Text" />
  <rdfs:label xml:lang="en">copyright</rdfs:label>
  <rdfs:comment xml:lang="en">Copyright information.</rdfs:comment>
</owl:ObjectProperty>

<owl:ObjectProperty rdf:ID="isbn">
  <rdfs:domain rdf:resource="#Entry" />
  <rdfs:range rdf:resource="#Text" />
  <rdfs:label xml:lang="en">ISBN</rdfs:label>
  <rdfs:comment xml:lang="en">The International Standard Book Number.</rdfs:comment>
</owl:ObjectProperty>

<owl:ObjectProperty rdf:ID="issn">
  <rdfs:domain rdf:resource="#Entry" />
  <rdfs:range rdf:resource="#Text" />
  <rdfs:label xml:lang="en">ISSN</rdfs:label>
  <rdfs:comment xml:lang="en">The International Standard Serial Number. Used to identify a
journal.</rdfs:comment>
</owl:ObjectProperty>

<owl:ObjectProperty rdf:ID="keywords">
  <rdfs:domain rdf:resource="#Entry" />
  <rdfs:range rdf:resource="#Text" />
  <rdfs:label xml:lang="en">keywords</rdfs:label>
  <rdfs:comment xml:lang="en">Key words used for searching or possibly for
annotation.</rdfs:comment>
</owl:ObjectProperty>

<owl:ObjectProperty rdf:ID="language">
  <rdfs:domain rdf:resource="#Entry" />
  <rdfs:range rdf:resource="#Text" />
  <rdfs:label xml:lang="en">language</rdfs:label>
  <rdfs:comment xml:lang="en">The language the document is in.</rdfs:comment>
</owl:ObjectProperty>

<owl:ObjectProperty rdf:ID="location">
  <rdfs:domain rdf:resource="#Entry" />
  <rdfs:range rdf:resource="#Text" />
  <rdfs:label xml:lang="en">location</rdfs:label>
  <rdfs:comment xml:lang="en">A location associated with the entry, such as the city in
which a conference took place.</rdfs:comment>
</owl:ObjectProperty>

<owl:ObjectProperty rdf:ID="lccn">
  <rdfs:domain rdf:resource="#Entry" />
  <rdfs:range rdf:resource="#Text" />

```

```

    <rdfs:label xml:lang="en">LCCN</rdfs:label>
    <rdfs:comment xml:lang="en">The Library of Congress Call Number.</rdfs:comment>
</owl:ObjectProperty>

<owl:ObjectProperty rdf:ID="mrNumber">
  <rdfs:domain rdf:resource="#Entry" />
  <rdfs:range rdf:resource="#Text" />
  <rdfs:label xml:lang="en">MR number</rdfs:label>
  <rdfs:comment xml:lang="en">The Mathematical Reviews number.</rdfs:comment>
</owl:ObjectProperty>

<owl:ObjectProperty rdf:ID="price">
  <rdfs:domain rdf:resource="#Entry" />
  <rdfs:range rdf:resource="#Text" />
  <rdfs:label xml:lang="en">price</rdfs:label>
  <rdfs:comment xml:lang="en">The price of the document.</rdfs:comment>
</owl:ObjectProperty>

<owl:ObjectProperty rdf:ID="size">
  <rdfs:domain rdf:resource="#Entry" />
  <rdfs:range rdf:resource="#Text" />
  <rdfs:label xml:lang="en">size</rdfs:label>
  <rdfs:comment xml:lang="en">The physical dimensions of a work.</rdfs:comment>
</owl:ObjectProperty>

<owl:ObjectProperty rdf:ID="url">
  <rdfs:domain rdf:resource="#Entry" />
  <rdfs:range rdf:resource="#Text" />
  <rdfs:label xml:lang="en">URL</rdfs:label>
  <rdfs:comment xml:lang="en">The WWW Universal Resource Locator that points to the item
being referenced. This often is used for technical reports to point to the ftp or web site
where the postscript source of the report is located.</rdfs:comment>
</owl:ObjectProperty>

<owl:ObjectProperty rdf:ID="inMonography">
  <rdfs:domain rdf:resource="#InBook" />
  <rdfs:range rdf:resource="#Monography" />
  <rdfs:subPropertyOf rdf:resource="#isPartOf" />
  <rdfs:label xml:lang="en">inMonography</rdfs:label>
  <rdfs:comment xml:lang="en">A reference to the book that contains the entry. The entry
must be an InBook end the book a Monography.</rdfs:comment>
</owl:ObjectProperty>

<owl:ObjectProperty rdf:ID="inCollection">
  <rdfs:domain rdf:resource="#InCollection" />
  <rdfs:range rdf:resource="#Collection" />
  <rdfs:subPropertyOf rdf:resource="#isPartOf" />
  <rdfs:label xml:lang="en">in collection</rdfs:label>
  <rdfs:comment xml:lang="en">A reference to the book that contains the entry. The entry
must be an Inbook or Incollection.</rdfs:comment>
</owl:ObjectProperty>

<owl:ObjectProperty rdf:ID="inProceedings">
  <rdfs:domain rdf:resource="#InProceedings" />
  <rdfs:range rdf:resource="#Proceedings" />
  <rdfs:subPropertyOf rdf:resource="#isPartOf" />
  <rdfs:label xml:lang="en">in proceedings</rdfs:label>
  <rdfs:comment xml:lang="en">A reference to the proceedings in which the entry appears. The
entry must be an InProceedings.</rdfs:comment>
</owl:ObjectProperty>

<!-- Properties for motion picture -->
<owl:ObjectProperty rdf:ID="directors">
  <rdfs:subPropertyOf rdf:resource="#humanCreator" />
  <rdfs:domain rdf:resource="#MotionPicture" />
  <rdfs:range rdf:resource="#PersonList" />
  <rdfs:label xml:lang="en">directors</rdfs:label>
  <rdfs:comment xml:lang="en">The list of persons who directed the motion
picture.</rdfs:comment>
</owl:ObjectProperty>

<!--
*****  "META"  PROPERTIES  *****

We define here some meta properties that the fields above may use as superproperties.
This is necessary in order to define cardinality constraints properly.
-->

```

```

<owl:ObjectProperty rdf:ID="humanCreator">
  <rdfs:domain rdf:resource="#Entry" />
  <rdfs:range rdf:resource="#PersonList" />
  <rdfs:label xml:lang="en">human creator</rdfs:label>
  <rdfs:comment xml:lang="en">A generic human creator category, necessary in order to
contain both author and editor.</rdfs:comment>
</owl:ObjectProperty>

<owl:ObjectProperty rdf:ID="pagesOrChapter">
  <rdfs:domain>
    <owl:Class>
      <owl:unionOf rdf:parseType="Collection">
        <owl:Class rdf:about="#Part" />
      </owl:unionOf>
    </owl:Class>
  </rdfs:domain>
  <rdfs:label xml:lang="en">page and/or chapter data</rdfs:label>
  <rdfs:comment xml:lang="en">A generic property to hold page and/or chapter
data.</rdfs:comment>
</owl:ObjectProperty>

<owl:DatatypeProperty rdf:ID="numberOrVolume">
  <rdfs:domain>
    <owl:Class>
      <owl:unionOf rdf:parseType="Collection">
        <owl:Class rdf:about="#Article" />
        <owl:Class rdf:about="#Composite" />
      </owl:unionOf>
    </owl:Class>
  </rdfs:domain>
  <rdfs:label xml:lang="en">volume or number</rdfs:label>
  <rdfs:comment xml:lang="en">The number of a journal, magazine, technical report, or of a
work in a series. An issue of a journal or magazine is usually identified by its volume and
number; the organization that issues a technical report usually gives it a number; and
sometimes books are given numbers in a named series.</rdfs:comment>
</owl:DatatypeProperty>

<!--
***** OTHER CLASSES *****

  Other useful classes.
-->

<!-- trouver une ontologie pour ca -->
<owl:Class rdf:ID="Institution">
  <rdfs:label xml:lang="en">Institution</rdfs:label>
  <rdfs:comment xml:lang="en">An institution.</rdfs:comment>
  <!-- A COMPLETER
  <owl:equivalentClass rdf:resource="" />
  -->
</owl:Class>

<!-- trouver une ontologie pour ca -->
<owl:Class rdf:ID="School">
  <rdfs:label xml:lang="en">School</rdfs:label>
  <rdfs:comment xml:lang="en">A school or university.</rdfs:comment>
  <!-- A COMPLETER
  <owl:equivalentClass rdf:resource="" />
  -->
</owl:Class>

<!-- trouver une ontologie pour ca -->
<owl:Class rdf:ID="Journal">
  <rdfs:label xml:lang="en">Journal</rdfs:label>
  <rdfs:comment xml:lang="en">A journal.</rdfs:comment>
  <!-- A COMPLETER
  <owl:equivalentClass rdf:resource="" />
  -->
</owl:Class>

<!-- trouver une ontologie pour ca -->
<owl:Class rdf:ID="Organization">
  <rdfs:label xml:lang="en">Organization</rdfs:label>
  <rdfs:comment xml:lang="en">An organization.</rdfs:comment>
  <!-- A COMPLETER
  <owl:equivalentClass rdf:resource="" />
  -->

```

```

</owl:Class>

<!-- trouver une ontologie pour ca -->
<owl:Class rdf:ID="Publisher">
  <rdfs:label xml:lang="en">Publisher</rdfs:label>
  <rdfs:comment xml:lang="en">A publisher.</rdfs:comment>
  <!-- A COMPLETER
  <owl:equivalentClass rdf:resource="" />
  -->
</owl:Class>

<!-- Les personnes sont des foaf:Person -->

<owl:Class rdf:ID="PersonList">
  <rdfs:subClassOf rdf:resource="&rdf;List"/>
  <rdfs:label xml:lang="en">Person list</rdfs:label>
  <rdfs:comment xml:lang="en">A list of persons.</rdfs:comment>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="&rdf;first" />
      <owl:cardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:cardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="&rdf;first" />
      <owl:allValuesFrom rdf:resource="&foaf;Person" />
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="&rdf;rest" />
      <owl:cardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:cardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="&rdf;rest" />
      <owl:allValuesFrom>
        <owl:Class>
          <owl:unionOf rdf:parseType="Collection">
            <owl:Class rdf:about="#PersonList" />
            <owl:Class>
              <owl:oneOf rdf:parseType="Collection">
                <rdf:List rdf:about="&rdf:nil" />
              </owl:oneOf>
            </owl:Class>
          </owl:unionOf>
        </owl:Class>
      </owl:allValuesFrom>
    </owl:Restriction>
  </rdfs:subClassOf>
</owl:Class>

<!-- redefinition des proprietes de rdf:List (pour jena) -->
<owl:Class rdf:about="&rdf;List" />
<owl:ObjectProperty rdf:about="&rdf;first">
  <rdfs:domain rdf:resource="&rdf;List" />
</owl:ObjectProperty>
<owl:ObjectProperty rdf:about="&rdf;rest">
  <rdfs:domain rdf:resource="&rdf;List" />
</owl:ObjectProperty>

<owl:Class rdf:ID="PageRange">
  <rdfs:label xml:lang="en">PageRange</rdfs:label>
  <rdfs:comment xml:lang="en">A range of pages.</rdfs:comment>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#startPage" />
      <owl:cardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:cardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>

```

```

    <owl:Restriction>
      <owl:onProperty rdf:resource="#endPage" />
      <owl:cardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:cardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
</owl:Class>

<owl:DatatypeProperty rdf:ID="startPage">
  <rdfs:domain rdf:resource="#PageRange" />
  <rdfs:range rdf:resource="&xsd;nonNegativeInteger" />
  <rdfs:label xml:lang="en">start page</rdfs:label>
  <rdfs:comment xml:lang="en">The beginning of a range of pages.</rdfs:comment>
</owl:DatatypeProperty>

<owl:DatatypeProperty rdf:ID="endPage">
  <rdfs:domain rdf:resource="#PageRange" />
  <rdfs:range rdf:resource="&xsd;nonNegativeInteger" />
  <rdfs:label xml:lang="en">end page</rdfs:label>
  <rdfs:comment xml:lang="en">The end of a range of pages.</rdfs:comment>
</owl:DatatypeProperty>

<owl:Class rdf:ID="Date">
  <rdfs:label xml:lang="en">Date</rdfs:label>
  <rdfs:comment xml:lang="en">Date of a day, month or year.</rdfs:comment>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#year" />
      <owl:cardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:cardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#month" />
      <owl:maxCardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:maxCardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
</owl:Class>

<owl:Class rdf:ID="DateWithMonth">
  <rdfs:label xml:lang="en">Date</rdfs:label>
  <rdfs:comment xml:lang="en">Date of a day, month or year.</rdfs:comment>
  <rdfs:subClassOf rdf:resource="#Date"/>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#month" />
      <owl:cardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:cardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#day" />
      <owl:maxCardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:maxCardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
</owl:Class>

<owl:DatatypeProperty rdf:ID="year">
  <rdfs:domain rdf:resource="#Date" />
  <rdfs:range rdf:resource="&xsd;nonNegativeInteger" />
  <rdfs:label xml:lang="en">year</rdfs:label>
</owl:DatatypeProperty>
<owl:DatatypeProperty rdf:ID="month">
  <rdfs:domain rdf:resource="#Date" />
  <rdfs:range rdf:resource="&xsd;nonNegativeInteger" />
  <rdfs:label xml:lang="en">month</rdfs:label>
</owl:DatatypeProperty>
<owl:DatatypeProperty rdf:ID="day">
  <rdfs:domain rdf:resource="#DateWithMonth" />
  <rdfs:range rdf:resource="&xsd;nonNegativeInteger" />
  <rdfs:label xml:lang="en">day</rdfs:label>
</owl:DatatypeProperty>

<!-- definir ces attributs comme des objets -->
<owl:Class rdf:ID="Text">
  <rdfs:label xml:lang="en">Text</rdfs:label>
  <rdfs:comment xml:lang="en">A simple text with possibly other information such as
language, etc.</rdfs:comment>

```

```

    <rdfs:comment xml:lang="en">Domain for affiliation, type, edition, chapter, howpublished,
number, copyright, contents, isbn, issn, language, lccn, price, size, location,
url.</rdfs:comment>
  </owl:Class>
  <owl:DatatypeProperty rdf:ID="lang">
    <!-- peut etre utiliser simplement xml:lang ? -->
    <rdfs:domain rdf:resource="#Text" />
    <rdfs:range rdf:resource="&xsd:string" />
    <rdfs:label xml:lang="en">lang</rdfs:label>
    <rdfs:comment xml:lang="en">The language in which the words are written.</rdfs:comment>
  </owl:DatatypeProperty>
  <owl:Class rdf:ID="Title">
    <rdfs:label xml:lang="en">Title</rdfs:label>
    <rdfs:comment xml:lang="en">A title.</rdfs:comment>
    <rdfs:subClassOf rdf:resource="#Text" />
  </owl:Class>
  <owl:DatatypeProperty rdf:ID="short">
    <rdfs:domain rdf:resource="#Title" />
    <rdfs:range rdf:resource="&xsd:string" />
    <rdfs:label xml:lang="en">short</rdfs:label>
    <rdfs:comment xml:lang="en">A shorten version of the title.</rdfs:comment>
  </owl:DatatypeProperty>

<!--

A FAIRE:

- ecrire les criteres d'unification en un pseudo langage formel
- mettre a jour BibTeX2RDF **rapidement**

-->

</rdf:RDF>

```

ANEXO C – Diagramas UML do sistema

1) Diagrama de classes

2) Diagrama de sequência

ANEXO D - Alinhamentos das ontologias 101 e 304

Por convenção, classes iniciam com letra maiúscula e propriedades, com minúscula.

Tabela D1 - Alinhamento e similaridades com base nas técnicas CSPA, HGPA e MCLA.

Classe ou propriedade da ontologia 101	Classe ou propriedade da ontologia 304	Similaridade
Part	Part	1,0000
series	series	1,0000
Institution	Institution	1,0000
numberOrVolume	numberOrVolume	1,0000
price	price	1,0000
lccn	lccn	1,0000
rest	rest	1,0000
date	date	1,0000
book	volume	1,0000
language	language	1,0000
size	size	1,0000
InBook	InBook	1,0000
type	type	1,0000
isbn	isbn	1,0000
PersonList	PersonList	1,0000
startPage	startPage	1,0000
mrNumber	mrNumber	1,0000
chapter	chapter	1,0000
Unpublished	Unpublished	1,0000
endPage	endPage	1,0000
PageRange	PageRange	1,0000
InCollection	InCollection	1,0000
edition	edition	1,0000
Booklet	Booklet	1,0000
TechReport	TechReport	1,0000
institution	institution	1,0000
abstract	abstract	1,0000
key	key	1,0000
issn	issn	1,0000
note	note	1,0000
Collection	InCollection	0,8013
InCollection	Collection	0,8013
annotate	annotate	1,0000
collection	inCollection	0,7212
copyright	copyright	1,0000

Classe ou propriedade da ontologia 101	Classe ou propriedade da ontologia 304	Similaridade
howPublished	howPublished	1,0000
School	School	1,0000
humanCreator	humanCreator	1,0000
List	List	1,0000
affiliation	affiliation	1,0000
directors	directors	1,0000
MastersThesis	MastersThesis	1,0000
pages	pages	1,0000
location	location	1,0000
number	number	1,0000
Book	Book	1,0000
volume	volume	1,0000
publisher	publisher	1,0000
url	url	1,0000
title	title	1,0000
Informal	Informal	1,0000
organization	organization	1,0000
Manual	Manual	1,0000
isPartOf	isPartOf	1,0000
Monograph	Monography	0,8413
Deliverable	Deliverable	1,0000
first	first	1,0000

Tabela D2 - Alinhamento e similaridades com base em CEB com três *clusters*.

Classe ou propriedade da ontologia 1	Classe ou propriedade da ontologia 2	Similaridade
Part	Part	1,0000
note	note	1,0000
Deliverable	Deliverable	1,0000
issue	number	1,0000
mrNumber	mrNumber	1,0000
price	price	1,0000
pages	pages	1,0000
humanCreator	humanCreator	1,0000
size	size	1,0000
isbn	isbn	1,0000
keywords	keywords	1,0000
InBook	InBook	1,0000
school	school	1,0000
type	type	1,0000
List	List	1,0000
Unpublished	Unpublished	1,0000

Classe ou propriedade da ontologia 101	Classe ou propriedade da ontologia 304	Similaridade
url	url	1,0000
institution	institution	1,0000
Academic	Academic	1,0000
Manual	Manual	1,0000
PageRange	PageRange	1,0000
InCollection	InCollection	1,0000
organization	organization	1,0000
School	School	1,0000
Institution	Institution	1,0000
location	location	1,0000
series	series	1,0000
abstract	abstract	1,0000
volume	volume	1,0000
copyright	copyright	1,0000
contents	contents	1,0000
Collection	InCollection	0,8013
InCollection	Collection	0,8013
title	title	1,0000
numberOrVolume	numberOrVolume	1,0000
chapter	chapter	1,0000
Collection	Collection	1,0000
Booklet	Booklet	1,0000
directors	directors	1,0000
annotate	annotate	1,0000
language	language	1,0000
number	number	1,0000
LectureNotes	LectureNotes	1,0000
PersonList	PersonList	1,0000
lccn	lccn	1,0000
PhdThesis	PhdThesis	1,0000
chapters	chapter	1,0000
Chapter	Chapter	1,0000
edition	edition	1,0000
publisher	publisher	1,0000
Informal	Informal	1,0000
key	key	1,0000
first	first	1,0000
endPage	endPage	1,0000
Book	Book	1,0000
startPage	startPage	1,0000
rest	rest	1,0000
affiliation	affiliation	1,0000
issn	issn	1,0000

Classe ou propriedade da ontologia 101	Classe ou propriedade da ontologia 304	Similaridade
TechReport	TechReport	1,0000
date	date	1,0000
howPublished	howPublished	1,0000
book	volume	1,0000
MastersThesis	MastersThesis	1,0000
collection	inCollection	0,7212
Monograph	Monography	0,8413
MotionPicture	MotionPicture	1,0000
isPartOf	isPartOf	1,0000

Tabela D3 - Alinhamento e similaridades com base no aspecto terminológico.

Classe ou propriedade da ontologia 101	Classe ou propriedade da ontologia 304	Similaridade
PhdThesis	PhdThesis	1,0000
Unpublished	Unpublished	1,0000
Monograph	Monography	0,8413
organization	organization	1,0000
lccn	lccn	1,0000
issue	number	1,0000
annotate	annotate	1,0000
affiliation	affiliation	1,0000
Deliverable	Deliverable	1,0000
note	note	1,0000
collection	inCollection	0,7212
howPublished	howPublished	1,0000
first	first	1,0000
PageRange	PageRange	1,0000
edition	edition	1,0000
numberOrVolume	numberOrVolume	1,0000
price	price	1,0000
School	School	1,0000
startPage	startPage	1,0000
volume	volume	1,0000
institution	institution	1,0000
abstract	abstract	1,0000
publisher	publisher	1,0000
Informal	Informal	1,0000
Institution	Institution	1,0000
InBook	InBook	1,0000
issn	issn	1,0000
keywords	keywords	1,0000
year	year	1,0000

Classe ou propriedade da ontologia 101	Classe ou propriedade da ontologia 304	Similaridade
title	title	1,0000
series	series	1,0000
InCollection	InCollection	1,0000
List	List	1,0000
mrNumber	mrNumber	1,0000
Book	Book	1,0000
Booklet	Booklet	1,0000
pages	pages	1,0000
number	number	1,0000
Part	Part	1,0000
MotionPicture	MotionPicture	1,0000
rest	rest	1,0000
chapter	chapter	1,0000
location	location	1,0000
InCollection	Collection	0,8013
Manual	Manual	1,0000
Collection	Collection	1,0000
isPartOf	isPartOf	1,0000
size	size	1,0000
copyright	copyright	1,0000
Date	Date	1,0000
date	date	1,0000
Collection	InCollection	0,8013
TechReport	TechReport	1,0000
endPage	endPage	1,0000
month	month	1,0000
isbn	isbn	1,0000
book	volume	1,0000
directors	directors	1,0000
day	day	1,0000
contents	contents	1,0000
humanCreator	humanCreator	1,0000
language	language	1,0000
key	key	1,0000
MastersThesis	MastersThesis	1,0000
school	school	1,0000
PersonList	PersonList	1,0000
type	type	1,0000
url	url	1,0000
Academic	Academic	1,0000

Tabela D4 - Alinhamento e similaridades com base no aspecto topológico.

Classe ou propriedade da ontologia 101	Classe ou propriedade da ontologia 304	Similaridade
Article	Article	1,0000
book	volume	1,0000
type	type	1,0000
Journal	Journal	1,0000
edition	edition	1,0000
Misc	Misc	1,0000
Publisher	Published	0,8081
Chapter	Chapter	1,0000
howPublished	howPublished	1,0000
chapters	chapter	1,0000
organization	organization	1,0000

Tabela D5 - Alinhamento e similaridades com base no aspecto extensional.

Classe ou propriedade da ontologia 101	Classe ou propriedade da ontologia 304	Similaridade
Booklet	Booklet	1,0000
directors	directors	1,0000
rest	rest	1,0000
TechReport	TechReport	1,0000
type	type	1,0000
size	size	1,0000
edition	edition	1,0000
affiliation	affiliation	1,0000
first	first	1,0000
Collection	InCollection	0,8013
keywords	keywords	1,0000
issue	number	1,0000
PersonList	PersonList	1,0000
mrNumber	mrNumber	1,0000
contents	contents	1,0000
School	School	1,0000
Institution	Institution	1,0000
isbn	isbn	1,0000
Deliverable	Deliverable	1,0000
InBook	InBook	1,0000
PhdThesis	PhdThesis	1,0000
volume	volume	1,0000
abstract	abstract	1,0000
Unpublished	Unpublished	1,0000

Classe ou propriedade da ontologia 101	Classe ou propriedade da ontologia 304	Similaridade
organization	organization	1,0000
chapter	chapter	1,0000
note	note	1,0000
numberOrVolume	numberOrVolume	1,0000
copyright	copyright	1,0000
MotionPicture	MotionPicture	1,0000
institution	institution	1,0000
InCollection	Collection	0,8013
List	List	1,0000
Academic	Academic	1,0000
humanCreator	humanCreator	1,0000
publisher	publisher	1,0000
InCollection	InCollection	1,0000
Collection	Collection	1,0000
annotate	annotate	1,0000
MastersThesis	MastersThesis	1,0000
howPublished	howPublished	1,0000
Book	Book	1,0000
Monograph	Monography	0,8413
lccn	lccn	1,0000
key	key	1,0000
language	language	1,0000
collection	inCollection	0,7212
location	location	1,0000
price	price	1,0000
school	school	1,0000
issn	issn	1,0000
title	title	1,0000
Manual	Manual	1,0000
Informal	Informal	1,0000
date	date	1,0000
series	series	1,0000
number	number	1,0000
url	url	1,0000

Tabela D6 - Alinhamento de referência da OAEI.

Classe ou propriedade da ontologia 101	Classe ou propriedade da ontologia 304
InBook	InBook
Informal	Informal
MotionPicture	MotionPicture
editor	editor
Collection	Collection
book	inMonography
MastersThesis	MastersThesis
location	location
type	type
price	price
affiliation	affiliation
size	size
author	author
Chapter	InBook
PageRange	PageRange
Manual	Manual
contents	contents
Reference	Entry
series	series
key	key
abstract	abstract
humanCreator	humanCreator
day	day
startPage	startPage
lccn	lccn
LectureNotes	LectureNotes
InCollection	InCollection
url	url
annotate	annotate
keywords	keywords
howPublished	howPublished
proceedings	inProceedings
Part	Part
isPartOf	isPartOf
Article	Article
address	location
year	year
numberOrVolume	numberOrVolume
Publisher	Publisher
school	school
Deliverable	Deliverable

Classe ou propriedade da ontologia 101	Classe ou propriedade da ontologia 304
organization	organization
number	number
Misc	Misc
institution	institution
mrNumber	mrNumber
note	note
Monograph	Monography
issn	issn
date	date
PhdThesis	PhdThesis
Date	Date
InProceedings	InProceedings
PersonList	PersonList
title	title
journal	inJournal
Unpublished	Unpublished
Booklet	Booklet
Proceedings	Proceedings
TechReport	TechReport
volume	volume
month	month
endPage	endPage
copyright	copyright
collection	inCollection
School	School
publisher	publisher
edition	edition
directors	directors
chapter	chapter
Academic	Academic
Journal	Journal
pages	pages
isbn	isbn
Book	Book
Institution	Institution

ANEXO E - Resultados de clusterização e de alinhamento das demais ontologias do teste *benchmarks*

No emparelhamento, o par de ontologias a ser alinhado é composto de uma ontologia origem e outra destino. No teste *benchmarks*, a ontologia 101 é a ontologia de origem. Assim, na primeira coluna da tabela está indicada a ontologia alvo no emparelhamento. Os resultados máximos de largura de silhueta média e das métricas de alinhamento estão destacados em negrito.

Tabela E1 – Resultados de clusterização e alinhamento para as demais ontologias do teste de *benchmarks*.

Ontologia	Aspecto/Clusterização consensual	Redução de dimensionalidade	Técnica de clusterização	Valor de k										Alinhamento		
				2	3	4	5	6	7	8	9	10	Precisão	Cobertura	F-measure	
				Largura de silhueta média												
103	Terminológico	ICA	UPGMA	0,52	0,60	0,77	0,88	0,78	0,62	0,54	0,36	0,18	0,8780	0,7423	0,8045	
			Single-link	0,52	0,60	0,77	0,88	0,78	0,62	0,54	0,36	0,18				
			Complete-link	0,36	0,61	0,77	0,88	0,78	0,62	0,54	0,36	0,18				
			Ward	0,52	0,61	0,77	0,88	0,78	0,62	0,54	0,36	0,18				
			DIANA	0,34	0,60	0,83	0,72	0,78	0,62	0,54	0,36	0,18				
			CLARA	0,26	0,45	0,67	0,72	0,67	0,58	0,47	0,31	0,15				
			K-means	0,41	0,59	0,82	0,89	0,82	0,65	0,48	0,36	0,18				
			PAM	0,29	0,60	0,83	0,88	0,78	0,62	0,54	0,36	0,18				
	Extensional	LSA	UPGMA	0,60	0,80	0,87	0,71	0,56	0,42	0,36	0,33	0,18	0,8851	0,7938	0,8370	
			Single-link	0,60	0,80	0,87	0,71	0,56	0,42	0,36	0,33	0,18				
			Complete-link	0,60	0,80	0,87	0,71	0,56	0,42	0,36	0,33	0,18				
			Ward	0,60	0,80	0,87	0,71	0,56	0,42	0,36	0,33	0,18				
			DIANA	0,60	0,80	0,87	0,71	0,56	0,42	0,36	0,33	0,18				
			CLARA	0,52	0,80	0,87	0,71	0,56	0,42	0,36	0,33	0,18				
			K-means	0,75	0,90	0,97	0,79	0,63	0,54	0,53	0,35	0,18				
			PAM	0,52	0,80	0,87	0,71	0,56	0,42	0,36	0,33	0,18				
	Topológico	Não se aplica	UPGMA	0,51	0,65	0,82	0,69	0,56	0,42	0,28	0,18	0,17	0,8824	0,6186	0,7273	
			Single-link	0,56	0,65	0,82	0,69	0,54	0,41	0,28	0,18	0,17				
			Complete-link	0,56	0,65	0,82	0,69	0,56	0,42	0,28	0,18	0,17				
			Ward	0,63	0,70	0,82	0,69	0,55	0,42	0,28	0,26	0,17				
			DIANA	0,63	0,70	0,82	0,69	0,56	0,42	0,28	0,18	0,17				
			CLARA	0,48	0,56	0,82	0,69	0,54	0,39	0,27	0,26	0,17				
			K-means	0,81	0,82	0,61	0,83	0,71	0,56	0,40	0,36	0,18				
			PAM	0,48	0,56	0,82	0,69	0,55	0,40	0,28	0,18	0,17				
	CSPA, HGPA e MCLA		Não se aplica	Não se aplica										0,8851	0,7938	0,8370
	CEB		Não se aplica	Não se aplica										0,8283	0,8454	0,8367

Ontologia	Aspecto/Clusterização consensual	Redução de dimensionalidade	Técnica de clusterização	Valor de k									Alinhamento		
				2	3	4	5	6	7	8	9	10	Precisão	Cobertura	F-measure
				Largura de silhueta média											
104	Terminológico	ICA	UPGMA	0,52	0,68	0,91	0,93	0,80	0,76	0,65	0,49	0,33	0,9155	0,6701	0,7738
			Single-link	0,52	0,68	0,91	0,93	0,89	0,76	0,65	0,49	0,33			
			Complete-link	0,44	0,63	0,91	0,93	0,80	0,76	0,65	0,49	0,33			
			Ward	0,52	0,68	0,91	0,93	0,80	0,76	0,65	0,49	0,33			
			DIANA	0,52	0,59	0,91	0,93	0,80	0,76	0,65	0,49	0,33			
			CLARA	0,36	0,51	0,73	0,69	0,68	0,61	0,54	0,45	0,30			
			K-means	0,44	0,55	0,38	0,27	0,12	0,54	0,52	0,44	0,27			
			PAM	0,52	0,68	0,91	0,93	0,80	0,76	0,65	0,49	0,33			
	Extensional	ICA	UPGMA	0,24	0,41	0,57	0,74	0,89	0,79	0,65	0,50	0,33	0,9091	0,6186	0,7362
			Single-link	0,24	0,41	0,57	0,74	0,89	0,79	0,65	0,50	0,33			
			Complete-link	0,17	0,32	0,48	0,66	0,89	0,79	0,65	0,50	0,33			
			Ward	0,20	0,41	0,57	0,74	0,89	0,79	0,65	0,50	0,33			
			DIANA	0,17	0,32	0,48	0,66	0,89	0,79	0,65	0,50	0,33			
			CLARA	0,25	0,42	0,59	0,74	0,89	0,79	0,65	0,50	0,33			
			K-means	0,26	0,43	0,56	0,73	0,96	0,79	0,63	0,50	0,33			
			PAM	0,25	0,42	0,59	0,74	0,89	0,79	0,65	0,50	0,33			
	Topológico	Não se aplica	UPGMA	0,61	0,54	0,50	0,72	0,59	0,51	0,35	0,30	0,28	0,9091	0,6186	0,7362
			Single-link	0,34	0,12	0,50	0,72	0,64	0,51	0,35	0,30	0,28			
			Complete-link	0,61	0,54	0,50	0,72	0,59	0,42	0,35	0,30	0,28			
			Ward	0,61	0,65	0,68	0,72	0,59	0,51	0,35	0,30	0,28			
			DIANA	0,57	0,47	0,68	0,72	0,59	0,42	0,35	0,30	0,28			
			CLARA	0,51	0,65	0,69	0,72	0,64	0,50	0,49	0,32	0,28			
			K-means	0,75	0,59	0,77	0,65	0,60	0,51	0,35	0,41	0,24			
			PAM	0,51	0,65	0,69	0,72	0,64	0,51	0,35	0,30	0,28			
	CSPA, HGPA e MCLA		Não se aplica	Não se aplica									0,9091	0,6186	0,7362
	CEB		Não se aplica	Não se aplica									0,8272	0,6907	0,7528

Ontologia	Aspecto/Clusterização consensual	Redução de dimensionalidade	Técnica de clusterização	Valor de k									Alinhamento		
				2	3	4	5	6	7	8	9	10	Precisão	Cobertura	F-measure
				Largura de silhueta média											
204	Terminológico	LSA	UPGMA	0,71	0,97	0,85	0,97	0,95	0,79	0,65	0,50	0,33	0,8588	0,7526	0,8022
			Single-link	0,71	0,97	0,85	0,97	0,95	0,79	0,65	0,50	0,33			
			Complete-link	0,70	0,97	0,85	0,97	0,95	0,79	0,65	0,50	0,33			
			Ward	0,72	0,97	0,85	0,97	0,95	0,79	0,65	0,50	0,33			
			DIANA	0,70	0,97	0,85	0,97	0,95	0,79	0,65	0,50	0,33			
			CLARA	0,65	0,83	0,81	0,72	0,83	0,69	0,54	0,40	0,31			
			K-means	0,78	0,95	0,95	0,80	0,63	0,78	0,62	0,49	0,33			
			PAM	0,71	0,97	0,85	0,97	0,95	0,79	0,65	0,50	0,33			
	Extensional	ICA	UPGMA	0,25	0,42	0,59	0,76	0,98	0,82	0,66	0,50	0,33	0,8588	0,7526	0,8022
			Single-link	0,26	0,42	0,59	0,76	0,98	0,82	0,66	0,50	0,33			
			Complete-link	0,24	0,36	0,58	0,76	0,98	0,82	0,66	0,50	0,33			
			Ward	0,20	0,37	0,55	0,76	0,98	0,82	0,66	0,50	0,33			
			DIANA	0,24	0,36	0,58	0,75	0,98	0,82	0,66	0,50	0,33			
			CLARA	0,26	0,42	0,59	0,61	0,98	0,82	0,66	0,50	0,33			
			K-means	0,26	0,43	0,60	0,78	0,61	0,83	0,27	0,50	0,33			
			PAM	0,26	0,42	0,59	0,61	0,98	0,82	0,66	0,50	0,33			
	Topológico	Não se aplica	UPGMA	0,59	0,72	0,94	0,98	1,00	0,83	0,67	0,50	0,33	0,8588	0,7526	0,8022
			Single-link	0,59	0,72	0,94	0,98	1,00	0,83	0,67	0,50	0,33			
			Complete-link	0,59	0,72	0,94	0,98	1,00	0,83	0,67	0,50	0,33			
			Ward	0,65	0,76	0,94	0,98	1,00	0,83	0,67	0,50	0,33			
			DIANA	0,65	0,76	0,94	0,98	1,00	0,83	0,67	0,50	0,33			
			CLARA	0,59	0,72	0,94	0,98	1,00	0,83	0,67	0,50	0,33			
			K-means	0,49	0,82	0,82	1,00	1,00	0,83	0,67	0,50	0,33			
			PAM	0,59	0,72	0,94	0,98	1,00	0,83	0,67	0,50	0,33			
	CSPA, HGPA e MCLA		Não se aplica	Não se aplica									0,8588	0,7526	0,8022
	CEB		Não se aplica	Não se aplica									0,868	0,4742	0,6133

Ontologia	Aspecto/Clusterização consensual	Redução de dimensionalidade	Técnica de clusterização	Valor de k									Alinhamento		
				2	3	4	5	6	7	8	9	10	Precisão	Cobertura	F-measure
				Largura de silhueta média											
205	Terminológico	ICA	UPGMA	0,49	0,65	0,89	0,95	0,89	0,73	0,58	0,37	0,33	0,6207	0,1856	0,2857
			Single-link	0,49	0,65	0,89	0,95	0,89	0,73	0,58	0,37	0,33			
			Complete-link	0,39	0,69	0,89	0,95	0,89	0,73	0,52	0,37	0,33			
			Ward	0,44	0,69	0,89	0,95	0,89	0,73	0,58	0,37	0,33			
			DIANA	0,45	0,64	0,89	0,95	0,89	0,73	0,52	0,37	0,33			
			CLARA	0,41	0,54	0,71	0,69	0,71	0,59	0,47	0,42	0,28			
			K-means	0,43	0,59	0,88	0,85	0,86	0,70	0,56	0,48	0,32			
			PAM	0,45	0,62	0,89	0,95	0,89	0,73	0,52	0,37	0,33			
	Extensional	ICA	UPGMA	0,27	0,25	0,41	0,55	0,74	0,70	0,59	0,47	0,33	0,5652	0,1340	0,2167
			Single-link	0,27	0,28	0,41	0,55	0,74	0,70	0,59	0,47	0,33			
			Complete-link	0,27	0,20	0,37	0,52	0,74	0,70	0,59	0,47	0,33			
			Ward	0,25	0,28	0,47	0,55	0,74	0,70	0,59	0,47	0,33			
			DIANA	0,27	0,25	0,31	0,53	0,74	0,70	0,59	0,47	0,33			
			CLARA	0,25	0,31	0,43	0,66	0,74	0,70	0,59	0,47	0,33			
			K-means	0,21	0,35	0,53	0,60	0,87	0,71	0,64	0,48	0,33			
			PAM	0,25	0,31	0,43	0,66	0,74	0,70	0,59	0,47	0,33			
	Topológico	Não se aplica	UPGMA	0,68	0,63	0,79	0,76	0,62	0,41	0,38	0,37	0,33	0,8333	0,0515	0,0971
			Single-link	0,35	0,63	0,79	0,64	0,62	0,41	0,41	0,37	0,33			
			Complete-link	0,35	0,63	0,79	0,76	0,62	0,41	0,38	0,37	0,33			
			Ward	0,67	0,74	0,79	0,76	0,55	0,41	0,38	0,37	0,33			
			DIANA	0,67	0,62	0,79	0,76	0,62	0,41	0,38	0,37	0,33			
			CLARA	0,63	0,74	0,79	0,76	0,75	0,61	0,41	0,37	0,33			
			K-means	0,81	0,83	0,80	0,62	0,58	0,66	0,49	0,29	0,24			
			PAM	0,63	0,64	0,79	0,76	0,62	0,41	0,41	0,37	0,33			
	CSPA, HGPA e MCLA		Não se aplica	Não se aplica									1,0000	0,0515	0,0980
	CEB		Não se aplica	Não se aplica									0,5333	0,1649	0,2520

Ontologia	Aspecto/Clusterização consensual	Redução de dimensionalidade	Técnica de clusterização	Valor de k									Alinhamento		
				2	3	4	5	6	7	8	9	10	Precisão	Cobertura	F-measure
				Largura de silhueta média											
208	Terminológico	LSA	UPGMA	0,48	0,64	0,84	0,70	0,71	0,58	0,53	0,43	0,26	0,8400	0,4330	0,5714
			Single-link	0,43	0,64	0,84	0,57	0,71	0,58	0,53	0,43	0,26			
			Complete-link	0,34	0,55	0,84	0,77	0,71	0,67	0,53	0,43	0,26			
			Ward	0,48	0,64	0,84	0,77	0,71	0,58	0,53	0,43	0,26			
			DIANA	0,43	0,59	0,84	0,70	0,71	0,67	0,53	0,43	0,26			
			CLARA	0,52	0,54	0,56	0,60	0,40	0,30	0,34	0,13	0,08			
			K-means	0,62	0,55	0,61	0,66	0,50	0,35	0,40	0,26	0,13			
			PAM	0,45	0,64	0,84	0,70	0,71	0,58	0,53	0,43	0,26			
	Extensional	ICA	UPGMA	0,26	0,42	0,59	0,76	0,97	0,82	0,66	0,49	0,33	0,8571	0,4330	0,5753
			Single-link	0,26	0,42	0,59	0,76	0,97	0,82	0,66	0,49	0,33			
			Complete-link	0,24	0,36	0,58	0,75	0,97	0,82	0,66	0,49	0,33			
			Ward	0,26	0,42	0,59	0,76	0,97	0,82	0,66	0,49	0,33			
			DIANA	0,24	0,36	0,58	0,75	0,97	0,82	0,66	0,49	0,33			
			CLARA	0,26	0,42	0,59	0,61	0,97	0,82	0,66	0,49	0,33			
			K-means	0,26	0,43	0,60	0,78	0,61	0,83	0,67	0,50	0,33			
			PAM	0,26	0,42	0,59	0,61	0,97	0,82	0,66	0,49	0,33			
	Topológico	Não se aplica	UPGMA	0,61	0,71	0,94	0,88	0,82	0,83	0,67	0,50	0,33	0,8478	0,4021	0,5455
			Single-link	0,61	0,71	0,94	0,88	0,82	0,83	0,67	0,50	0,33			
			Complete-link	0,61	0,71	0,94	0,88	0,82	0,83	0,67	0,50	0,33			
			Ward	0,65	0,76	0,94	0,88	0,82	0,83	0,67	0,50	0,33			
			DIANA	0,65	0,76	0,94	0,88	0,82	0,83	0,67	0,50	0,33			
			CLARA	0,61	0,71	0,94	0,88	0,90	0,83	0,67	0,50	0,33			
			K-means	0,43	0,82	0,74	0,67	0,64	0,83	0,67	0,50	0,33			
			PAM	0,61	0,71	0,94	0,88	0,82	0,83	0,67	0,50	0,33			
	CSPA, HGPA e MCLA		Não se aplica	Não se aplica									0,8571	0,4330	0,5753
	CEB		Não se aplica	Não se aplica									0,8400	0,4330	0,5714

Ontologia	Aspecto/Clusterização consensual	Redução de dimensionalidade	Técnica de clusterização	Valor de k									Alinhamento		
				2	3	4	5	6	7	8	9	10	Precisão	Cobertura	F-measure
				Largura de silhueta média											
209	Terminológico	LSA	UPGMA	0,52	0,58	0,77	0,67	0,63	0,57	0,60	0,48	0,33	0,5556	0,0515	0,0943
			Single-link	0,52	0,58	0,77	0,69	0,63	0,57	0,60	0,48	0,33			
			Complete-link	0,58	0,77	0,77	0,67	0,63	0,57	0,60	0,48	0,33			
			Ward	0,58	0,77	0,77	0,67	0,63	0,57	0,60	0,48	0,33			
			DIANA	0,52	0,77	0,77	0,67	0,63	0,57	0,60	0,48	0,33			
			CLARA	0,62	0,59	0,63	0,65	0,33	0,39	0,35	0,20	0,14			
			K-means	0,72	0,62	0,67	0,67	0,40	0,47	0,43	0,32	0,20			
			PAM	0,52	0,58	0,77	0,67	0,63	0,57	0,60	0,48	0,33			
	Extensional	ICA	UPGMA	0,27	0,25	0,41	0,55	0,74	0,70	0,59	0,47	0,33	0,6667	0,0825	0,1468
			Single-link	0,27	0,28	0,41	0,55	0,74	0,70	0,59	0,47	0,33			
			Complete-link	0,27	0,20	0,37	0,52	0,74	0,70	0,59	0,47	0,33			
			Ward	0,25	0,28	0,41	0,55	0,74	0,70	0,59	0,47	0,33			
			DIANA	0,27	0,25	0,37	0,53	0,74	0,70	0,59	0,47	0,33			
			CLARA	0,25	0,31	0,43	0,66	0,74	0,70	0,59	0,47	0,33			
			K-means	0,21	0,35	0,53	0,60	0,87	0,71	0,64	0,48	0,33			
			PAM	0,25	0,31	0,43	0,66	0,74	0,70	0,59	0,47	0,33			
	Topológico	Não se aplica	UPGMA	0,68	0,63	0,79	0,76	0,62	0,41	0,38	0,37	0,33	1,0000	0,0412	0,0792
			Single-link	0,35	0,63	0,79	0,64	0,62	0,41	0,41	0,37	0,33			
			Complete-link	0,35	0,63	0,79	0,76	0,62	0,41	0,38	0,37	0,33			
			Ward	0,67	0,74	0,79	0,76	0,55	0,41	0,38	0,37	0,33			
			DIANA	0,67	0,62	0,79	0,76	0,62	0,41	0,38	0,37	0,33			
			CLARA	0,63	0,74	0,79	0,76	0,75	0,61	0,41	0,37	0,33			
			K-means	0,81	0,83	0,80	0,62	0,58	0,66	0,49	0,29	0,24			
			PAM	0,63	0,74	0,79	0,76	0,62	0,41	0,41	0,37	0,33			
	CSPA, HGPA e MCLA		Não se aplica	Não se aplica									1,0000	0,0103	0,0204
	CEB		Não se aplica	Não se aplica									0,6923	0,0928	0,1636

Ontologia	Aspecto/Clusterização consensual	Redução de dimensionalidade	Técnica de clusterização	Valor de k									Alinhamento		
				2	3	4	5	6	7	8	9	10	Precisão	Cobertura	F-measure
				Largura de silhueta média											
222	Terminológico	LSA	UPGMA	0,55	0,42	0,59	0,68	0,57	0,47	0,39	0,24	0,18	0,8515	0,9247	0,8866
			Single-link	0,55	0,42	0,36	0,68	0,58	0,47	0,39	0,24	0,18			
			Complete-link	0,49	0,75	0,74	0,68	0,57	0,47	0,39	0,24	0,18			
			Ward	0,58	0,75	0,74	0,68	0,57	0,47	0,39	0,24	0,18			
			DIANA	0,47	0,75	0,74	0,68	0,57	0,47	0,39	0,24	0,18			
			CLARA	0,77	0,60	0,54	0,56	0,39	0,28	0,33	0,22	0,18			
			K-means	0,87	0,72	0,61	0,68	0,53	0,55	0,40	0,24	0,07			
			PAM	0,49	0,60	0,59	0,68	0,58	0,47	0,39	0,26	0,18			
	Extensional	LSA	UPGMA	0,53	0,46	0,47	0,59	0,58	0,46	0,30	0,28	0,12	0,8515	0,9247	0,8866
			Single-link	0,53	0,30	0,47	0,59	0,58	0,46	0,30	0,28	0,12			
			Complete-link	0,53	0,30	0,47	0,59	0,58	0,46	0,30	0,28	0,12			
			Ward	0,53	0,46	0,55	0,59	0,58	0,46	0,43	0,28	0,12			
			DIANA	0,53	0,30	0,47	0,35	0,58	0,46	0,30	0,28	0,12			
			CLARA	0,53	0,46	0,55	0,59	0,58	0,46	0,30	0,28	0,12			
			K-means	0,71	0,38	0,64	0,70	0,51	0,33	0,33	0,20	0,16			
			PAM	0,53	0,46	0,55	0,59	0,58	0,46	0,30	0,28	0,12			
	Topológico	Não se aplica	UPGMA	0,49	0,55	0,53	0,58	0,48	0,41	0,18	0,13	0,04	0,8205	0,6882	0,7485
			Single-link	0,49	0,29	0,25	0,58	0,48	0,41	0,18	0,13	0,04			
			Complete-link	0,49	0,55	0,53	0,58	0,48	0,41	0,18	0,13	0,04			
			Ward	0,56	0,60	0,62	0,58	0,48	0,41	0,18	0,13	0,04			
			DIANA	0,56	0,41	0,62	0,58	0,48	0,41	0,18	0,13	0,04			
			CLARA	0,55	0,55	0,53	0,58	0,48	0,41	0,18	0,13	0,04			
			K-means	0,56	0,70	0,73	0,69	0,48	0,40	0,18	0,21	0,14			
			PAM	0,74	0,55	0,53	0,58	0,46	0,40	0,29	0,13	0,04			
	CSPA, HGPA e MCLA		Não se aplica	Não se aplica									0,8515	0,9247	0,8866
	CEB		Não se aplica	Não se aplica									0,8286	0,9355	0,8788

Ontologia	Aspecto/Clusterização consensual	Redução de dimensionalidade	Técnica de clusterização	Valor de k									Alinhamento		
				2	3	4	5	6	7	8	9	10	Precisão	Cobertura	F-measure
				Largura de silhueta média											
223	Terminológico	ICA	UPGMA	0,50	0,59	0,83	0,96	0,88	0,78	0,68	0,60	0,50	0,8764	0,8041	0,8387
			Single-link	0,35	0,65	0,83	0,96	0,88	0,78	0,68	0,60	0,50			
			Complete-link	0,48	0,60	0,83	0,96	0,88	0,78	0,68	0,60	0,50			
			Ward	0,48	0,64	0,83	0,96	0,88	0,78	0,68	0,60	0,50			
			DIANA	0,46	0,60	0,83	0,96	0,88	0,78	0,68	0,60	0,50			
			CLARA	0,27	0,45	0,52	0,62	0,69	0,64	0,58	0,54	0,45			
			K-means	0,37	0,59	0,31	0,59	0,70	0,72	0,65	0,52	0,38			
			PAM	0,46	0,65	0,83	0,96	0,88	0,78	0,68	0,60	0,50			
	Extensional	LSA	UPGMA	0,30	0,38	0,49	0,62	0,55	0,47	0,46	0,43	0,33	0,8462	0,7938	0,8191
			Single-link	0,29	0,38	0,48	0,62	0,55	0,47	0,36	0,43	0,33			
			Complete-link	0,30	0,38	0,49	0,62	0,57	0,52	0,46	0,43	0,33			
			Ward	0,28	0,38	0,49	0,62	0,57	0,52	0,46	0,43	0,33			
			DIANA	0,30	0,38	0,48	0,62	0,57	0,52	0,46	0,43	0,33			
			CLARA	0,30	0,37	0,49	0,43	0,57	0,52	0,46	0,43	0,33			
			K-means	0,44	0,35	0,57	0,82	0,77	0,74	0,62	0,54	0,41			
			PAM	0,30	0,37	0,49	0,62	0,57	0,52	0,46	0,43	0,33			
	Topológico	Não se aplica	UPGMA	0,64	0,70	0,83	0,74	0,59	0,53	0,43	0,36	0,23	0,7272	0,2474	0,3692
			Single-link	0,46	0,70	0,83	0,74	0,67	0,58	0,45	0,30	0,18			
			Complete-link	0,64	0,70	0,83	0,74	0,59	0,53	0,43	0,36	0,23			
			Ward	0,64	0,70	0,83	0,74	0,59	0,53	0,43	0,36	0,23			
			DIANA	0,46	0,59	0,69	0,74	0,59	0,53	0,43	0,36	0,23			
			CLARA	0,54	0,66	0,83	0,74	0,59	0,53	0,43	0,34	0,21			
			K-means	0,81	0,82	0,96	0,87	0,75	0,65	0,55	0,47	0,37			
			PAM	0,54	0,66	0,83	0,74	0,55	0,48	0,39	0,34	0,21			
	CSPA, HGPA e MCLA		Não se aplica	Não se aplica									0,8764	0,8041	0,8387
	CEB		Não se aplica	Não se aplica									0,8704	0,4845	0,6225

Ontologia	Aspecto/Clusterização consensual	Redução de dimensionalidade	Técnica de clusterização	Valor de k									Alinhamento		
				2	3	4	5	6	7	8	9	10	Precisão	Cobertura	F-measure
				Largura de silhueta média											
225	Terminológico	ICA	UPGMA	0,48	0,64	0,81	0,83	0,75	0,83	0,66	0,50	0,33	0,8654	0,9278	0,8955
			Single-link	0,48	0,64	0,81	0,83	0,75	0,83	0,66	0,50	0,33			
			Complete-link	0,37	0,59	0,81	0,83	0,75	0,83	0,66	0,50	0,33			
			Ward	0,47	0,58	0,81	0,83	0,90	0,83	0,66	0,50	0,33			
			DIANA	0,41	0,64	0,81	0,83	0,75	0,83	0,66	0,50	0,33			
			CLARA	0,43	0,57	0,74	0,75	0,87	0,77	0,62	0,46	0,32			
			K-means	0,34	0,55	0,85	0,87	0,71	0,83	0,66	0,50	0,33			
			PAM	0,48	0,64	0,81	0,83	0,90	0,83	0,66	0,50	0,33			
	Extensional	LSA	UPGMA	0,67	0,76	0,97	0,98	1,00	0,83	0,67	0,50	0,33	0,8557	0,8557	0,8557
			Single-link	0,67	0,76	0,97	0,98	1,00	0,83	0,67	0,50	0,33			
			Complete-link	0,67	0,76	0,97	0,98	1,00	0,83	0,67	0,50	0,33			
			Ward	0,67	0,76	0,97	0,98	1,00	0,83	0,67	0,50	0,33			
			DIANA	0,67	0,76	0,97	0,98	1,00	0,83	0,67	0,50	0,33			
			CLARA	0,67	0,76	0,97	0,98	1,00	0,83	0,67	0,50	0,33			
			K-means	0,79	0,82	1,00	1,00	1,00	0,83	0,67	0,50	0,33			
			PAM	0,67	0,76	0,97	0,98	1,00	0,83	0,67	0,50	0,33			
	Topológico	Não se aplica	UPGMA	0,61	0,71	0,94	0,88	0,82	0,83	0,67	0,50	0,33	0,8586	0,8763	0,8673
			Single-link	0,61	0,71	0,94	0,88	0,82	0,83	0,67	0,50	0,33			
			Complete-link	0,61	0,71	0,94	0,88	0,82	0,83	0,67	0,50	0,33			
			Ward	0,65	0,76	0,94	0,88	0,82	0,83	0,67	0,50	0,33			
			DIANA	0,65	0,76	0,94	0,88	0,82	0,83	0,67	0,50	0,33			
			CLARA	0,61	0,71	0,94	0,88	0,90	0,83	0,67	0,50	0,33			
			K-means	0,43	0,82	0,67	0,91	0,85	0,83	0,67	0,50	0,33			
			PAM	0,61	0,71	0,94	0,88	0,82	0,83	0,67	0,50	0,33			
	CSPA, HGPA e MCLA		Não se aplica	Não se aplica									0,8586	0,8763	0,8673
	CEB		Não se aplica	Não se aplica									0,8349	0,9381	0,8835

Ontologia	Aspecto/Clusterização consensual	Redução de dimensionalidade	Técnica de clusterização	Valor de k									Alinhamento		
				2	3	4	5	6	7	8	9	10	Precisão	Cobertura	F-measure
				Largura de silhueta média											
230	Terminológico	LSA	UPGMA	0,54	0,68	0,94	0,79	0,62	0,49	0,32	0,23	0,18	0,8429	0,8194	0,831
			Single-link	0,54	0,69	0,94	0,77	0,62	0,45	0,32	0,23	0,18			
			Complete-link	0,42	0,66	0,94	0,79	0,62	0,49	0,32	0,23	0,18			
			Ward	0,54	0,68	0,94	0,79	0,62	0,49	0,32	0,23	0,18			
			DIANA	0,44	0,64	0,94	0,79	0,62	0,49	0,32	0,23	0,18			
			CLARA	0,50	0,52	0,67	0,67	0,56	0,46	0,32	0,12	0,04			
			K-means	0,69	0,63	0,85	0,83	0,66	0,53	0,38	0,31	0,17			
			PAM	0,54	0,69	0,94	0,77	0,62	0,45	0,32	0,23	0,18			
	Extensional	LSA	UPGMA	0,41	0,60	0,77	0,86	0,71	0,55	0,40	0,22	0,18	0,8382	0,7917	0,8143
			Single-link	0,41	0,60	0,77	0,86	0,71	0,55	0,40	0,22	0,18			
			Complete-link	0,41	0,57	0,77	0,86	0,68	0,53	0,37	0,22	0,18			
			Ward	0,41	0,60	0,77	0,86	0,71	0,55	0,37	0,22	0,18			
			DIANA	0,41	0,60	0,77	0,86	0,68	0,53	0,37	0,22	0,18			
			CLARA	0,41	0,60	0,77	0,86	0,71	0,55	0,40	0,22	0,18			
			K-means	0,52	0,72	0,92	0,97	0,79	0,60	0,42	0,24	0,18			
			PAM	0,41	0,60	0,77	0,86	0,71	0,55	0,40	0,22	0,18			
	Topológico	Não se aplica	UPGMA	0,44	0,59	0,77	0,63	0,58	0,40	0,34	0,28	0,18	0,8382	0,7917	0,8143
			Single-link	0,44	0,59	0,77	0,63	0,58	0,40	0,34	0,28	0,18			
			Complete-link	0,62	0,59	0,77	0,72	0,58	0,40	0,34	0,28	0,18			
			Ward	0,63	0,74	0,77	0,72	0,58	0,40	0,34	0,28	0,18			
			DIANA	0,63	0,74	0,77	0,72	0,58	0,40	0,34	0,28	0,18			
			CLARA	0,62	0,74	0,77	0,72	0,58	0,40	0,34	0,28	0,18			
			K-means	0,80	0,85	0,60	0,84	0,77	0,59	0,45	0,35	0,18			
			PAM	0,62	0,74	0,77	0,72	0,58	0,40	0,34	0,28	0,18			
	CSPA, HGPA e MCLA		Não se aplica	Não se aplica									0,8429	0,8194	0,831
	CEB		Não se aplica	Não se aplica									0,8025	0,9028	0,8497

Ontologia	Aspecto/Clusterização consensual	Redução de dimensionalidade	Técnica de clusterização	Valor de k										Alinhamento		
				2		3	4	5	6	7	8	9	10	Precisão	Cobertura	F-measure
				Largura de silhueta média												
301	Terminológico	LSA	UPGMA	0,55	0,61	0,35	0,26	0,23					0,8125	0,2203	0,3467	
			Single-link	0,55	0,61	0,35	0,26	0,23								
			Complete-link	0,57	0,61	0,36	0,26	0,23								
			Ward	0,57	0,61	0,35	0,26	0,23								
			DIANA	0,55	0,61	0,35	0,26	0,23								
			CLARA	0,53	0,54	0,51	0,30	0,06								
			K-means	0,66	0,66	0,65	0,35	0,24								
			PAM	0,48	0,61	0,35	0,26	0,23								
	Extensional	LSA	UPGMA	0,50	0,43	0,64	0,37	0,29					0,8889	0,1356	0,2353	
			Single-link	0,50	0,43	0,64	0,37	0,29								
			Complete-link	0,50	0,43	0,64	0,37	0,29								
			Ward	0,50	0,43	0,64	0,37	0,29								
			DIANA	0,50	0,43	0,64	0,37	0,29								
			CLARA	0,50	0,53	0,64	0,37	0,29								
			K-means	0,65	0,18	0,70	0,41	0,15								
			PAM	0,50	0,53	0,64	0,37	0,29								
	Topológico	Não se aplica	UPGMA	0,40	0,41	0,45	0,49	0,25					0,8000	0,2034	0,3243	
			Single-link	0,40	0,41	0,45	0,49	0,25								
			Complete-link	0,40	0,41	0,45	0,49	0,25								
			Ward	0,54	0,50	0,45	0,49	0,25								
			DIANA	0,54	0,25	0,54	0,49	0,25								
			CLARA	0,36	0,41	0,45	0,49	0,25								
			K-means	0,77	0,63	0,63	0,56	0,28								
			PAM	0,36	0,41	0,45	0,49	0,25								
	CSPA, HGPA e MCLA		Não se aplica	Não se aplica										0,7692	0,1695	0,2778
	CEB		Não se aplica	Não se aplica										0,8125	0,2203	0,3467