



Universidade Estadual de Campinas
Instituto de Computação



Ênio de Jesus Pontes Monteiro

Adaptação Assistida de Anotações Semânticas sobre
bases RDF

CAMPINAS
2021

Ênio de Jesus Pontes Monteiro

Adaptação Assistida de Anotações Semânticas sobre bases RDF

Dissertação apresentada ao Instituto de Computação da Universidade Estadual de Campinas como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação.

Orientador: Prof. Dr. Julio Cesar dos Reis

Este exemplar corresponde à versão final da Dissertação defendida por Ênio de Jesus Pontes Monteiro e orientada pelo Prof. Dr. Julio Cesar dos Reis.

CAMPINAS
2021

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Matemática, Estatística e Computação Científica
Ana Regina Machado - CRB 8/5467

M764a Monteiro, Ênio de Jesus Pontes, 1992-
Adaptação assistida de anotações semânticas sobre bases RDF / Ênio de Jesus Pontes Monteiro. – Campinas, SP : [s.n.], 2021.

Orientador: Julio Cesar dos Reis.
Dissertação (mestrado) – Universidade Estadual de Campinas, Instituto de Computação.

1. Ontologias (Recuperação da informação). 2. Web semântica. 3. Computação semântica. I. Reis, Julio Cesar dos, 1987-. II. Universidade Estadual de Campinas. Instituto de Computação. III. Título.

Informações para Biblioteca Digital

Título em outro idioma: Assisted adaptation of semantic annotations on RDF bases

Palavras-chave em inglês:

Ontologies (Information retrieval)

Semantic web

Semantic computing

Área de concentração: Ciência da Computação

Titulação: Mestre em Ciência da Computação

Banca examinadora:

Julio Cesar dos Reis [Orientador]

Rita Cristina Galarraga Berardi

Leonardo Montecchi

Data de defesa: 13-12-2021

Programa de Pós-Graduação: Ciência da Computação

Identificação e informações acadêmicas do(a) aluno(a)

- ORCID do autor: <https://orcid.org/0000-0001-9992-864X>

- Currículo Lattes do autor: <http://lattes.cnpq.br/1120372166791871>



Universidade Estadual de Campinas
Instituto de Computação



Ênio de Jesus Pontes Monteiro

Adaptação Assistida de Anotações Semânticas sobre bases RDF

Banca Examinadora:

- Prof. Dr. Julio Cesar dos Reis
Instituto de Computação, Universidade Estadual de Campinas (UNICAMP)
- Prof. Dr. Leonardo Montecchi
Instituto de Computação, Universidade Estadual de Campinas (UNICAMP)
- Profa. Dra. Rita Cristina Galarraga Berardi
Departamento Acadêmico de Informática, Universidade Tecnológica Federal do Paraná (UTFPR)

A ata da defesa, assinada pelos membros da Comissão Examinadora, consta no SIGA/Sistema de Fluxo de Dissertação/Tese e na Secretaria do Programa da Unidade.

Campinas, 13 de dezembro de 2021

Agradecimentos

Este trabalho foi financiado pela Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) (#2019/14582-8)¹.

¹As opiniões expressas aqui não são necessariamente compartilhadas pela agência de fomento.

Resumo

O acesso e uso de metadados semanticamente definidos pode beneficiar diversos tipos de tarefas computacionais. Integração de dados entre diferentes sistemas e motores de busca podem explorar as definições semânticas associadas aos documentos e outros recursos Web para propor maneiras de combinar e encontrar informações antes não explicitamente relacionadas entre si. A geração de anotações a partir do crescente número de repositórios RDF interconectados torna disponível os metadados necessários para possibilitar aplicações semanticamente ricas na Web. Por exemplo, ferramentas semiautomáticas efetuam o reconhecimento de entidades mencionadas como pessoas e lugares em documentos Web. Contudo, a definição semântica dessas entidades tende a sofrer modificações em suas bases fonte e implica em um cenário altamente desafiador. Anotações geradas necessitam ser atualizadas em função da evolução das bases de conhecimento utilizadas na definição das anotações. A literatura existente tem apresentado metodologias para abordar esse problema, explorando principalmente a manutenção de anotações em relação às mudanças a nível de ontologia. Nossa análise da literatura não detectou investigações que abordassem a manutenção de anotações considerando sua geração no nível de instância de conceitos. Esta dissertação propõe técnicas para reconhecer as anotações afetadas pela evolução de bases RDF. Formalizamos regras que decidem o tipo de ação a ser aplicada sobre a anotação. Desenvolvemos um *framework* capaz de identificar e aplicar de forma assistida ações de adaptação em anotações afetadas pela evolução das bases de conhecimento RDF. Essa adaptação compreende a execução de três etapas: (1) detectar uma série de modificações que ocorreram em um determinado período de tempo com base em duas versões de um conjunto de dados RDF; (2) reconhecer e filtrar as anotações afetadas pelas mudanças daquelas não afetadas; (3) aplicar ações corretivas nas anotações afetadas e desatualizadas. Avaliações experimentais demonstram a aplicabilidade da proposta em bases RDF reais.

Abstract

Access to and use of semantically defined meta-data can benefit several types of computational tasks. Data integration between different systems and search engines can explore semantic definitions associated with documents and other web resources to propose ways to combine and find elements of information before not explicitly related to each other. Generating annotations from the growing number of interconnected RDF repositories makes available meta-data needed to enable semantically enriched applications on the Web. For example, semi-automatic tools perform the recognition of entities mentioned as persons and places in Web documents. However, semantic definition of these entities might undergo modifications in their source bases, which implies in a challenging scenario. In this context, generated annotations need to be updated according to the evolution of their underlying knowledge bases used in the definition of annotations. Existing literature has presented methodologies to address this problem, mainly exploring the maintenance of annotations regarding changes at the ontology level (TBox). Our literature analysis did not detect investigations that addressed the maintenance of annotations considering their generation at the level of instance of concepts (ABox). This MS.c. dissertation proposes techniques to recognize the annotations affected by the evolution of RDF datasets. We formalize rules that decide the type of action to be applied to the annotation. We develop a framework capable of identifying and applying adaptation actions to annotations affected by the evolution of RDF knowledge bases. The adaptation comprises the execution of three steps: (1) detection of a series of modifications that occurred in a given time period based on two releases of a RDF dataset; (2) recognition and filtering of annotations affected by RDF changes; (3) application of corrective actions to affected and outdated annotations in an assisted way. Experimental evaluations demonstrate the applicability of the proposal in real RDF datasets.

Lista de Figuras

1.1	Exemplo de anotação semântica usando recurso da DBpedia [2].	14
1.2	Problema da manutenção de anotação semântica.	15
1.3	Visão geral da Metodologia. Considera as Entradas, Objetivos, Resultados e Conhecimentos gerados nesta Dissertação de mestrado.	17
4.1	<i>Framework</i> AnnoLOD.	30
4.2	Exemplo de tipos de mudanças em repositórios RDF.	32
4.3	Exemplo de impacto das mudanças RDF sobre as anotações semânticas existentes.	33
4.4	Gerando lista de recursos candidatos <i>candList</i>	35
4.5	Consulta SPARQL para selecionar recursos candidatos para substituir o recurso que afetou a anotação.	36
4.6	Lista de ações realizadas em cada candidato da <i>candList</i> . Cada tipo de comparação é identificada pelos textos na cor vermelha. As extremidades das setas verdes indicam quais <i>labels</i> são comparados via cálculo de similaridade.	37
4.7	Sugestão de ações de adaptação de anotação ao usuário a partir na lista de recursos candidatos <i>candList</i>	39
4.8	Fluxo de interação do usuário no sistema de software para adaptação de anotações semânticas	41
4.9	Figura A - Tela inicial do sistema. Figura B - Interface inicial e de configuração da ferramenta.	41
4.10	Interface do painel do avaliador de anotação.	42
4.11	Interface do painel do avaliador de anotação.	43
5.1	Fragmento do arquivo RDF contendo informações de pessoas no formato <i>N-Triples</i> do conjunto de dados da <i>BBpedia</i> versão 2015-04.	47
5.2	Fragmento do arquivo RDF contendo informações de pessoas no formato <i>N-Triples</i> do subconjunto <i>YAGO 4</i>	49
5.3	Exemplo de Texto do documento 37 do <i>dataset</i> LP50.	49
5.4	Metodologia da Avaliação Experimental.	50
5.5	Metodologia conduzida para gerar os conjuntos de anotações $\mathcal{A}^j$ e $\mathcal{A}^{j+1}$	51
5.6	Consulta SPARQL para selecionar um recurso idêntico ao termo identificado no documento.	52
5.7	Exemplo de Anotação gerada pelo anotador a partir do repositório $\mathcal{R}^j$	53
5.8	Exemplo de Anotação do conjunto padrão de prata gerada pelo anotador a partir dos repositórios $\mathcal{R}^{j+1}$ e $\mathcal{R}_{ext}$	54
5.9	Exemplo de Anotação Inconsistente.	57

5.10	Lista de recursos candidatos encontrados para substituir o recurso “George W. Bush”.	60
5.11	Exemplo de Ações de Adaptação. A) substituição de objeto; B) remoção da anotação; C) não realizar ação de adaptação.	64

Lista de Tabelas

2.1	Modelo $\mathcal{ISAM}$	21
3.1	Análise do Estado da Arte.	28
5.1	Anotações $\mathcal{A}^j$	55
5.2	Anotações Padrão de Prata $\mathcal{SA}^{j+1}$	55
5.3	Lista de recursos excluídos.	57
5.4	Lista de anotações afetadas pelas mudanças $\mathcal{A}^{aff}$	58
5.5	Lista de anotações não afetadas pelas mudanças $\mathcal{A}^{unaff}$	59
5.6	Lista dos melhores recursos candidatos $bestCand$	61
5.7	Lista de anotações atualizadas $\mathcal{A}^{j+1}$	62

Sumário

1	Introdução	13
1.1	Contexto e Motivação	13
1.2	Problema e Desafios da Pesquisa	15
1.3	Objetivo e Abordagem da Pesquisa	16
1.4	Organização da Dissertação	17
2	Fundamentos	19
3	Revisão da Literatura	23
3.1	Detecção Automática e Análise de Anotações Inconsistentes	23
3.2	Correção Automática de Anotações Semânticas	25
3.3	Discussão	27
3.4	Conclusão	28
4	<i>Framework</i> AnnoLOD	29
4.1	Descrição Geral do <i>Framework</i>	29
4.2	Passo A - Detecção de Mudanças em Repositórios RDF	30
4.3	Passo B - Identificação de Anotações Inconsistentes	32
4.4	Passo C - Aplicar Ação de Adaptação de Anotação	33
4.4.1	Geração de Recursos Candidatos	34
4.4.2	Ações de Adaptação	38
4.5	Ferramenta de Software de Apoio para Adaptação Assistida de Anotação .	40
4.6	Conclusão	44
5	Avaliação Experimental	45
5.1	Conjuntos de Dados	45
5.2	Procedimentos	50
5.2.1	Gerando os Conjuntos de Anotações	50
5.3	Ilustrando uma Ocorrência de Anotação Afetada	56
5.3.1	Mudanças detectadas nos Repositórios RDF	57
5.3.2	Anotações Inconsistentes Identificadas	58
5.3.3	Aplicação das Ações de Adaptação de Anotação	59
5.4	Análise dos Resultados	63
5.5	Discussão	65
5.6	Conclusão	66
6	Conclusão	67
6.1	Contribuições	68
6.2	Disseminação dos Resultados	68

6.2.1	Artigos em Conferências	68
6.2.2	Participação em Eventos	69
6.3	Trabalhos Futuros	69
Referências Bibliográficas		70

Capítulo 1

Introdução

1.1 Contexto e Motivação

O principal objetivo da Web Semântica é permitir que a semântica de recursos descritos na Web seja interpretável por máquinas de maneira explícita e formal [3]. Esse aspecto é a base para permitir que sistemas integrem e combinem dados de maneira compartilhada. O principal elemento para aprimorar a interpretação de documentos e outros recursos como vídeos e imagens é gerar metadados (dados sobre dados) fazendo uso de bases de conhecimento que incorporam dados estruturados segundo definições expressas em ontologias. Esse tipo de metadado é denominado de anotação semântica, que consiste de rótulos sobre documentos Web para tornar interpretável seu significado para a máquina. Uma anotação é tanto dependente do documento anotado quanto da definição semântica associada. Anotações semânticas são elementos essenciais de sistemas que visam melhor interpretar, integrar e recuperar informações considerando o significado explícito compartilhado pela máquina.

Oren *et al.* [34] definiu que o termo anotação remete ao processo de anotação, bem como o resultado desse procedimento. Ainda segundo esses autores, as anotações são organizadas em três categorias. (1) Anotações informais, que não são legíveis por máquinas (por exemplo, notas adicionadas em uma página de livro); (2) Anotações formais, compreensíveis por máquina, mas não definidas por conceitos presentes em uma ontologia (por exemplo, destaques em um documento HTML); (3) Anotações ontológicas (às quais este trabalho se refere), compreensíveis pela máquina e extraídas com base em uma ontologia. Uma ontologia refere-se a um artefato formal que visa representar conhecimento em um determinado domínio, descrevendo principalmente seus conceitos, seus atributos e relacionamentos [17]. Anotações ontológicas são igualmente chamadas de anotações semânticas. Através delas, humanos e máquinas podem ter uma interpretação mais clara da semântica presente no conteúdo de um documento.

Anotações semânticas são comumente empregadas para fornecerem informações sobre um recurso ou relacionamento entre recursos Web¹. Anotar documentos para gerar metadados (dado sobre dado) vem sendo uma atividade frequente nos últimos anos, principalmente no domínio médico, de modo a tornar as informações contidas nesses docu-

¹<https://www.w3.org/TR/annotation-model/>

mentos interpretáveis por máquinas de maneira explícita e formal. Isso é normalmente realizado por meio da associação entre códigos de conceitos de uma determinada base de conhecimento (por exemplo, ontologias) com os elementos do documento. Diversas ferramentas computacionais têm sido investigadas e projetadas para a criação automática de anotações semânticas [9].

Uma anotação formal e semântica depende de elementos codificados em uma base de conhecimento [34]. As anotações são associadas aos conceitos ou instância de conceitos presentes em ontologias. Ontologias são essenciais para a criação de dados estruturados na Web descritos em *Resource Description Framework* (RDF²) [44]. A Figura 1.1 apresenta um exemplo de anotação, em que o elemento textual “*Albert Einstein*” é anotado com um recurso presente em uma base de conhecimento formalmente definindo o objeto “Albert Einstein”, descrevendo suas características de maneira interpretável pela máquina.

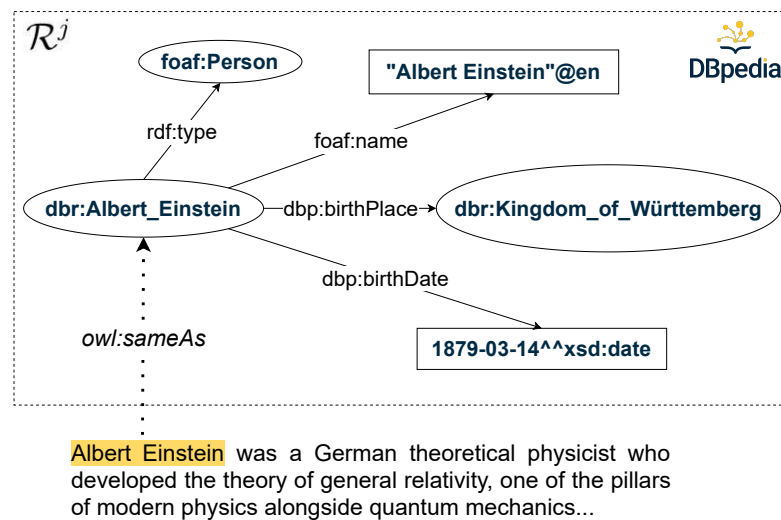


Figura 1.1: Exemplo de anotação semântica usando recurso da DBpedia [2].

Nos últimos anos surgiu um grande número de bases de conhecimento interconectadas na Web descrevendo vários tipos de recursos de maneira estruturada (RDF). Nesse contexto, *Linked Data* se refere à “[...] um conjunto de boas práticas para publicar e conectar dados estruturados na Web de uma maneira que os dados sejam legíveis por máquina, seu significado definido explicitamente, e vinculado a outros conjuntos de dados externos.” [43].

O conhecimento presente em bases de conhecimento descritos em RDF está em constante evolução. Novas descobertas e requisitos de dados em domínios específicos são necessários. Isso implica que o conteúdo de bases RDF sofram revisões e alterações ao longo do tempo. Consequentemente, mudanças em conceitos ou em valores dos atributos podem alterar a sua semântica originalmente codificada, ocasionando uma disparidade entre as versões do mesmo conceito. Esse fenômeno de evolução das bases pode impactar diretamente as anotações associadas existentes (já criadas), de forma a torná-las inválidas.

A literatura tem se interessado e apresentado metodologias para tratar esse problema. A maioria das abordagens lidam com ontologias de domínio principalmente na área bi-

²<https://www.w3.org/RDF/>

omédica, em que as anotações são criadas sobretudo no nível de classe das ontologias (cf. Seções 3.1 e 3.2). Nossa análise de literatura não detectou trabalhos abordando as anotações no nível de instância. Isso significa anotar documentos associando diretamente às instâncias de conceitos. Isso é explorado em bases RDF abertas presentes na Web. Por exemplo, a *DBpedia* [2], que é uma versão em dados RDF estruturados da *Wikipédia*. Um exemplo de anotação de instância pode ser vincular a menção “*Albert Einstein*”, presente em um determinado documento a sua definição semântica (recurso “Albert Einstein”) formalmente codificado na *DBpedia* (cf. Figura 1.1)

1.2 Problema e Desafios da Pesquisa

Esta pesquisa visa investigar anotações geradas a partir de bases de conhecimento interconectadas descritas em RDF. Atualmente, há diversos esforços para definir modelos e técnicas automáticas que visam gerar anotações de maneira (semi-) automática [34]. Definição de recursos descritos em RDF estão sujeitos a alterações. Valores de propriedades são modificados, predicados e objetos são removidos entre diversas outras operações de mudança. A Figura 1.2 apresenta a remoção do recurso “Albert Einstein” em uma base RDF e o impacto gerado na anotação dependente. As alterações podem influenciar nas anotações existentes (já geradas) e potencialmente diminuem a qualidade e consistência delas. Entendemos que as anotações semânticas são afetadas por modificações efetuadas nos conceitos e recursos que a definem. Manter a precisão das anotações é crítico pois diversas ferramentas como motores de busca e de integração de dados se fundamentam nesses metadados para seu funcionamento adequado.

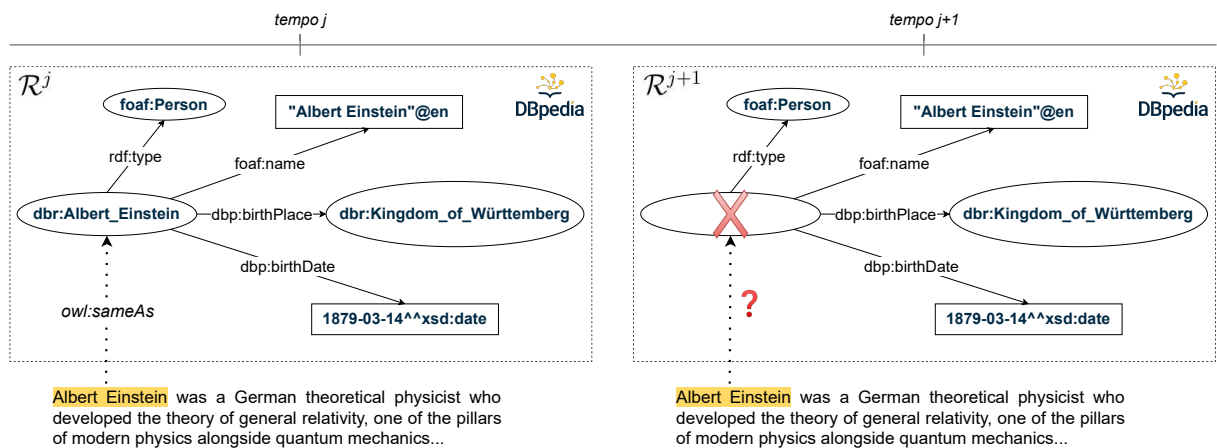


Figura 1.2: Problema da manutenção de anotação semântica.

Esta problemática é desafiadora em múltiplos sentidos e precisamos aprofundar as pesquisas para aprimorar a engenharia e confiabilidade dos métodos neste contexto. Embora a literatura tenha se interessado por esse problema, as investigações têm sido focadas em anotações referindo-se às ontologias de domínio em casos particulares (*e.g.*, biomedicina). No entanto, o estudo da influência da evolução de repositório de dados RDF nas anotações ainda não foi explorado. Esse aspecto é a ênfase desta dissertação.

Uma primeira dificuldade é lidar com dados abertos na Web, cujo ambiente é menos controlado. Assumindo que modificações em entidades das bases são as principais responsáveis por impactar as anotações, precisamos entender que tipos de modificações e como elas podem tornar as anotações desatualizadas. O principal desafio é possibilitar que a partir dos diferentes tipos de mudanças possíveis nas bases RDF (computadas de uma versão para outra do conjunto de dados RDF), elementos que compõem as anotações sejam alterados correspondentemente. Como exemplo, é necessário compreender se quando uma entidade que compõe a definição de uma anotação é removida, se tal anotação é juntamente removida, ou mais especificamente, que tipo de alteração pode ser sugerido para se manter atualizada a anotação sem efetuar sua total remoção.

A detecção e a posterior correção de anotações desatualizadas devido à nova versão das bases subjacentes é crucial para o funcionamento adequado de sistemas que dependem do uso das anotações. Devido ao grande volume de anotações, alavancadas pelo crescente número de repositórios RDF, a correção manual se torna árdua e custosa. Logo, fazer correções manuais sobre as anotações é um trabalho difícil em decorrência da possível enorme quantidade de anotações geradas pelos sistemas. Desse modo, o desenvolvimento de métodos e ferramentas computacionais capazes de auxiliar os especialistas de domínio nessa tarefa de manutenção faz-se urgente e necessário.

1.3 Objetivo e Abordagem da Pesquisa

Esta dissertação de mestrado objetiva estudar, implementar e avaliar um *framework* capaz de adaptar da forma mais automática possível anotações semânticas definidas a partir de repositórios RDF abertos na Web. O trabalho envolve conceber técnicas de detecção de anotações influenciadas pela evolução desses tipos de bases de conhecimento, assim como algoritmos de adaptação apropriados para corrigir as anotações considerando correlações com operações de mudança detectadas na evolução de bases RDF. Assumimos uma abordagem assistida, em que o processo de adaptação das anotações é feito com o auxílio de pessoas mantenedoras dos conjunto de dados e das anotações.

Nesta pesquisa, desenvolvemos técnicas para reconhecer as anotações afetadas pela evolução das bases RDF, então formalizamos regras que decidem o tipo de ação a ser aplicada sobre a anotação. A abordagem proposta destaca que modificações (adaptações) nas anotações afetadas por mudanças devem ser implementadas de acordo com os tipos de alterações identificadas nas entidades usadas como referência na anotação.

Esta pesquisa de mestrado visa alcançar os seguintes objetivos específicos:

- OBJ1: Propor e desenvolver técnicas capazes de identificar as anotações semânticas afetadas pela evolução das bases de conhecimento RDF; e
- OBJ2: Definir algoritmos adequados para transformar anotações inválidas em anotações válidas segundo a evolução de recursos de bases RDF.

A Figura 1.3 apresenta os objetivos definidos em nossa metodologia de pesquisa. Primeiro, reunimos estudos apropriados para obter uma revisão da literatura (*cf.* Capítulo 3). A partir do conhecimento adquirido com a pesquisa, projetamos nosso *framework* para

reparar anotações desatualizadas (cf. Capítulo 4). Com base nas operações de mudança calculadas a partir de versões de conjuntos de dados RDF, investigamos e implementamos um método para detectar automaticamente anotações desatualizadas (cf. Seção 4.2 e Seção 4.3). A última etapa da nossa proposta atualiza anotações inválidas (cf. Seção 4.4).

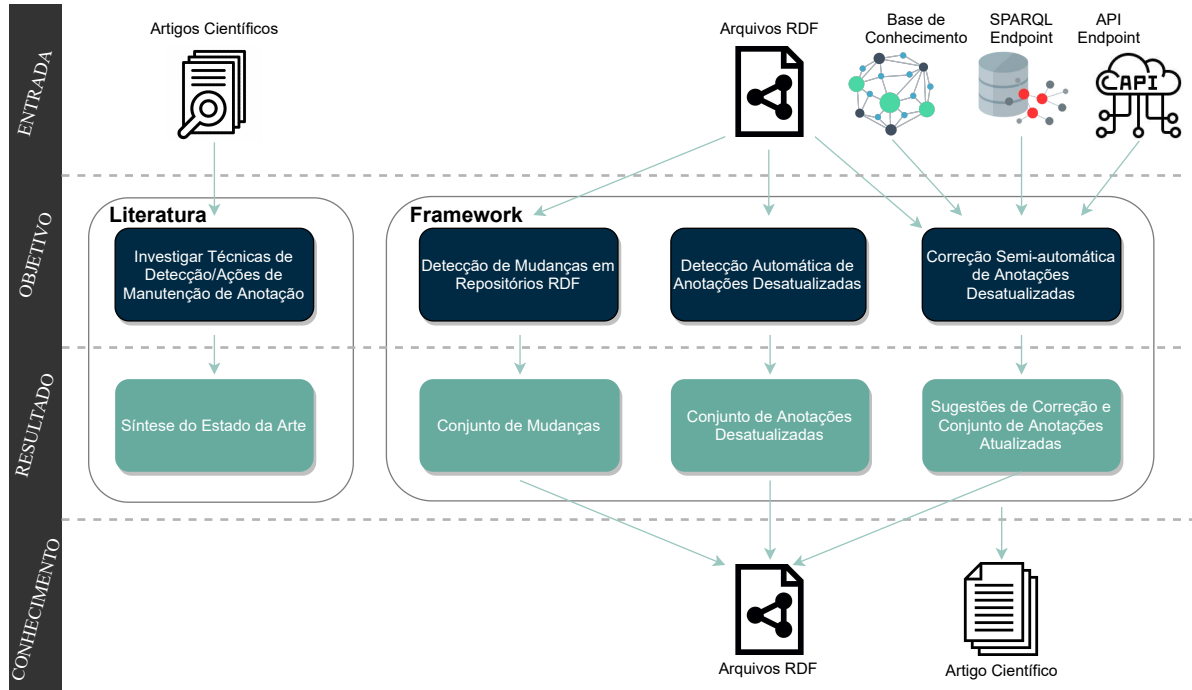


Figura 1.3: Visão geral da Metodologia. Considera as Entradas, Objetivos, Resultados e Conhecimentos gerados nesta Dissertação de mestrado.

1.4 Organização da Dissertação

O restante desta Dissertação de mestrado está organizada da seguinte maneira:

Capítulo 2 apresenta definições formais relevantes para o entendimento da nossa proposta de pesquisa abordada no Capítulo 4.

Capítulo 3 relata uma síntese do estado da arte considerando uma revisão de trabalhos relacionados (cf. Seção 3.1 e Seção 3.2).

Capítulo 4 descreve nosso *framework* AnnoLOD para manter atualizadas anotações. Fornecemos detalhes de cada etapa da solução. Seção 4.1 apresenta uma descrição geral do *framework*; a Seção 4.2, Seção 4.3 e Seção 4.4 apresentam os detalhes da proposta. A Seção 4.4 apresenta os passos projetados de uma técnica para reparar o conjunto de anotações desatualizadas, explorando uma combinação de medidas de similaridade semântica. A Seção 4.5 apresenta a ferramenta de software desenvolvida que implementa o *framework* visando apoiar a adaptação de anotações semânticas.

Capítulo 5 retrata a avaliação experimental conduzida para analisar a efetividade do *framework* (cf. Capítulo 4). A Seção 5.1 apresenta os conjunto de dados usados e a Seção 5.2 descreve os procedimentos experimentais adotados na avaliação realizada. A Seção

5.3 apresenta um exemplo da ocorrência de um conjunto de anotações desatualizadas, e como a execução do *framework* AnnoLOD lida com isso. Efetuamos uma análise dos resultados obtidos com a aplicação do *framework* sob as anotações inválidas (cf. Seção 5.4). Complementarmente, elaboramos uma discussão sob os impactos causados pelas mudanças nos repositórios RDF sob as anotações dependentes (cf. Seção 5.5).

Capítulo 6 conclui esta Dissertação. Ele destaca as contribuições gerais de nossa investigação (cf. Seção 6.1). A Seção 6.2 demonstra como disseminamos nossos resultados. Por fim, explicamos como nossa pesquisa pode ser estendida em investigações futuras (cf. Seção 6.3).

Capítulo 2

Fundamentos

Este capítulo apresenta definições e formalizações relevantes para o entendimento da solução desenvolvida.

Ontologia. Uma ontologia $\mathcal{O}$ descreve um domínio do conhecimento em termos de conceitos, atributos e relacionamentos entre conceitos [17]. Formalmente, uma ontologia $\mathcal{O} = (\mathcal{C}_{\mathcal{O}}, \mathcal{S}_{\mathcal{O}}, \mathcal{P}_{\mathcal{O}})$ consiste de um conjunto de classes $\mathcal{C}_{\mathcal{O}}$ inter-relacionadas por relacionamentos direcionados $\mathcal{S}_{\mathcal{O}}$. Cada conceito $c \in \mathcal{C}_{\mathcal{O}}$ possui um identificador único e está associado a um conjunto de atributos $\mathcal{P}_{\mathcal{O}}(c) = \{a_1, a_2, a_3, \dots, a_n\}$.

Repositório RDF. Um repositório RDF no contexto de *Linked Open Data* é um conjunto finito de triplas RDF em um domínio [14]. Formalmente, $\mathcal{R} = (t_1, t_2, t_3, \dots, t_n)$.

Triplas. Em um repositório RDF, uma tripla associa dois nós ou recursos utilizando uma propriedade (predicado). Um recurso pode ser definido como uma instância de uma determinada classe de uma ontologia. Um recurso pode ser qualquer coisa descrita no “mundo real”, seja algo físico, como um carro, ou um conceito como “ciência da computação” [14]. Em RDF, os recursos são descritos usando *Uniform Resource Identifier* (URI)¹ para a identificação única dos recursos na Web. Uma tripla RDF refere-se a uma entidade de dados composta por sujeito, predicado e objeto definido na forma de $t = (s, p, o)$ em que:

- **Sujeito:** (s) é uma URI de referência ou um nó em branco. Nós em branco são aqueles que representam os recursos para os quais uma URI ou literal não é fornecido.
- **Predicado:** (p) é uma propriedade descrita em uma ontologia e definida por uma URI de referência que conecta o sujeito e o objeto.
- **Objeto:** (o) é descrito por uma URI de referência, um literal ou um nó em branco. Um literal é uma cadeia de caracteres combinada com um identificador de idioma ou um tipo de dado.

Como exemplo de tripla considerando a notação (s, p, o) , descrevemos a data e o local de nascimento do cientista Albert Einstein da seguinte forma:

- (Albert Einstein², *birthDate*, 1879-03-14)

¹<https://www.w3.org/wiki/URI>

²http://dbpedia.org/page/Albert_Einstein

- (Albert Einstein, *birthPlace*, Kingdom of Württemberg)

Anotação semântica. Este trabalho assume o uso de anotações ontológicas [34], definidas com base em artefatos formais como ontologias e interpretáveis por máquina. Oren *et al.* [34] define formalmente uma anotação $\mathcal{A}$ como uma tupla $\mathcal{A} = (a_s, a_p, a_o, a_c)$, em que:

- a_s : é o sujeito da anotação.
- a_p : é o predicado que define o tipo de relacionamento entre a_s e a_o .
- a_o : é o objeto da anotação.
- a_c : é o contexto na qual a anotação é criada.

Cardoso *et al.* [8] apontaram características relevantes a serem consideradas como um modelo de anotação semântica para apoiar a evolução de anotações via mecanismos de manutenção (semi-)automáticos de anotações. O modelo de Cardoso *et al.* [8] estendeu o modelo originalmente proposto por Gross *et al.* [16]. Dessa forma, Cardoso *et al.* [8] definiram um modelo de anotação semântica sendo: $\mathcal{SAM} = (I_u, ON_v, R_a, Offset, Q, H, A, SemRel, U_f)$, em que:

- $I_u = (I, j)$: é uma fonte de termos. I_u consiste de um conjunto de termos $I = \{i_1, \dots, i_n\}$ em um determinado tempo discreto $j \in \mathbb{N}$.
- $\mathcal{O}^j$: é uma ontologia em sua versão j .
- D : é um conjunto de recursos, por exemplo, um documento textual a ser anotado.
- $Offset$: é o elemento que indica a posição do item a ser anotado em um determinado recurso D .
- Q : é um conjunto de indicadores de qualidade de anotações. Os indicadores de qualidade podem ser tanto valores numéricos quanto extraídos de taxonomias de qualidade pre-estabelecidas.
- H : é um elemento para caracterizar qual atributo do conceito $c \in \mathcal{O}^j$ (por exemplo, título, sinônimo, termo preferido, etc.) foi adotado para criar a anotação $a \in \mathcal{A}$.
- $\mathcal{A}$: é um conjunto de anotações. Uma única anotação $a \in \mathcal{A}$ é definida como $a = (i, c, \{q\})$, ou seja, um item de instância $i \in I_u$ é anotado com um conceito $c \in \mathcal{O}^j$; e um conjunto de indicadores de qualidade $\{q\} \in \mathbb{Q}$.
- $SemRel$: é utilizada para caracterizar o relacionamento semântico entre o conceito presente em $\mathcal{O}^j$ e o item anotado do I_u .
- U_f : Usado para apontar para uma versão antecessora da anotação visando manter um rastreamento da evolução da anotação no tempo.

Em nossa investigação, adaptamos o modelo de anotação proposto por Cardoso *et al.* [8].

Definimos nosso modelo como $\mathcal{ISAM} = (D, \mathcal{O}^j, \mathcal{R}^j, A, SemRel, U_f)$, em que:

- D : é uma fonte de arquivos. D consiste de um conjunto de documentos $D = \{dj, \dots, dn\}$.
- $\mathcal{R}^j$: é um repositório RDF em sua versão j com seus predicados definidos na $\mathcal{O}^j$.
- $\mathcal{A}$: é um conjunto de anotações. Formalmente $\mathcal{A} = \{a_1, a_2, a_3, \dots, a_n\}$. Uma única anotação $a \in \mathcal{A}$ é definida como $a = (ne, r, d, Offset, p, v)$, ou seja, um termo (menção ou entidade mencionada) $ne \in d \subset D$ é conectado a um recurso $r \in t \subset \mathcal{R}^j \rightarrow \mathcal{O}^j$ por um predicado p ; *Offset* indica a posição inicial (*start*) e final (*end*) em que ne se encontra no documento anotado; $p \in SemRel$ descreve o tipo de relacionamento entre ne e r ; $v \in U_f$ aponta para a versão anterior da anotação.

A adaptação foi realizada, pois o modelo definido em Cardoso et al. [8] não contém todos os elementos necessários para conceber nossas ações de manutenção. Observamos a necessidade de adicionar os atributos ne , r , d , *Offset* e p na definição da anotação a . No modelo original de Cardoso et al. [8], eles foram definidos no nível do conjunto de $\mathcal{A}$. A tabela 2.1 apresenta um exemplo de anotação de instância adotando nosso modelo proposto ($\mathcal{ISAM}$). Nela observamos que a menção ao cientista “*Albert Einstein*”, presente em um determinado documento foi vinculada (anotação) à sua definição semântica (recurso “Albert Einstein”) formalmente codificado na *DBpedia* (cf. Figura 1.1).

Tabela 2.1: Modelo $\mathcal{ISAM}$

$\mathcal{A}$		D	$\mathcal{R}^j \rightarrow \mathcal{O}^j$	<i>Offset</i>		$SemRel$	U_f
a_n	ne	d	$r \in t$	<i>start</i>	<i>end</i>	p	v
	<i>Albert Einstein</i>	46	<Albert Einstein, rdf:type, Person>	1	15	<i>sameAs</i>	a_{n-1}

O vínculo de um termo a um recurso por meio de um predicado constitui uma tripla no formato $t = (s, p, o)$ [36]. Adotando com exemplo a anotação apresentada na Figura 1.1, temos que: o sujeito s é constituído pelo termo “*Albert Einstein*”; o predicado p é composto pelo relacionamento *owl:sameAs*; e o objeto o é formado pelo recurso “Albert Einstein”. Por questão de inteligibilidade, a partir de agora, quando nos referirmos a uma anotação, estaremos fazendo menção a uma tripla.

A evolução dos conjuntos de dados RDF em termos de mudanças que afetam suas triplas pode invalidar artefatos previamente associados a elas, como por exemplo as anotações semânticas. Isso dificulta a consistência das anotações ao longo do tempo. Este cenário pode derivar um conjunto de anotações inconsistentes, que não são capazes de representar seu significado pretendido como antes da evolução do repositório RDF (ou seja, o significado pretendido pelo autor da anotação). Para manter a consistência das

anotações, o relacionamento semântico deve permanecer em um estado de integridade, mesmo com a evolução dos recursos de dados RDF associados.

Neste trabalho, assumimos que uma anotação pode ficar inconsistente nos seguintes casos: 1) quando o recurso objeto da anotação for removido do conjunto fonte ($r \notin \mathcal{R}^{j+1}$); e 2) quando a semântica da tripla a qual o recurso objeto da anotação faz parte deixar de representar a intenção preliminarmente desejada pelo autor da tripla ($t^{j+1} \neq t^j$).

À medida que os conjuntos de dados RDF evoluem, é necessário definir a noção de tempo j . Para um conjunto de dados RDF, denotamos $\mathcal{R}^j$ a versão inicial do conjunto de dados $\mathcal{R}$ no momento j , $\mathcal{R}^{j+1}$ sua versão evoluída (novo lançamento) no momento $j + 1$ e Δ a diferença entre os dois conjunto de dados.

Considere uma anotação a^j gerada a partir de um repositório $\mathcal{R}$ no tempo j (cf. Figura 1.1). Modificações ocorridas no conjunto de dados envolvendo a tripla associada à anotação em um momento específico j para um momento $j + 1$ podem invalidar tal anotação. Por exemplo, um recurso r_o pode existir no tempo j ($r_o \in \mathcal{R}^j$), mas ser removido no tempo $j + 1$ ($r_o \notin \mathcal{R}^{j+1}$) (cf. Figura 1.2). Essa mudança pode tornar a anotação inválida, uma vez que o relacionamento semântico fica inconsistente. Nesse sentido, a^j deve ser atualizada considerando a nova versão do repositório $\mathcal{R}$ ($\mathcal{R}^{j+1}$). Isso pode ser realizado por meio de uma “reanotação”; por exemplo, vincular o sujeito da anotação (menção ou entidade mencionada) a um novo recurso r_n no conjunto de dados no tempo $j + 1$ ($r_n \in \mathcal{R}^{j+1}$) ou em repositório externo ($r_n \in \mathcal{R}_{ext}$), tornando a anotação válida novamente (a^{j+1}).

A anotação atualizada a^{j+1} deve ser avaliada em termos de consistência semântica para verificar se a nova anotação após a mudança permanece semanticamente consistente. Para a verificação de consistência, definimos comparações de similaridade (por exemplo, entre r_o e r_n).

Em particular, nesta dissertação adotamos a definição proposta por Euzenat e Shvaiko [13] para a definição de similaridade: dado um conjunto de strings K (no nosso caso, *labels* de recursos), a similaridade é uma dada função que associa um par de valores de K a um número real $\mathfrak{R}$, expressando quão semelhantes são os valores de K comparados (cf. Equação 2.1).

$$\text{similaridade} : K \times K \Rightarrow \mathfrak{R} \quad (2.1)$$

O valor numérico deve ser positivo (cf Equação 2.2) e a similaridade máxima dos valores K não deve ser maior entre K e ela mesma (cf Equação 2.3).

$$\forall x, y \in K, \text{sim}(x, y) \geq 0 \quad (2.2)$$

$$\forall x, y, z \in K, \text{similaridade}(x, x) \geq \text{similaridade}(y, z) \quad (2.3)$$

Capítulo 3

Revisão da Literatura

Este capítulo apresenta uma análise exploratória da literatura de modo a investigar métodos existentes para a detecção e manutenção de anotações em função da evolução de bases de conhecimento das quais elas são dependentes. Nesse contexto, buscamos por dois grupos principais de abordagens: as que realizam a detecção automática de anotações inconsistentes (*cf.* Seção 3.1); e aquelas que implementam correção automática das anotações inválidas (*cf.* Seção 3.2).

3.1 Detecção Automática e Análise de Anotações Inconsistentes

Maynar *et al.* [28] estudaram a evolução de metadados e suas correspondentes ontologias, focando em dois processos distintos: 1) como a evolução de ontologias deve ser propagada para metadados (*top-down evolution*); e 2) como a evolução dos metadados deve ser utilizada para propor mudanças em ontologias (*bottom-up evolution*). Os autores propuseram uma série de estudos de casos para a elaboração de abordagens para ambos os tipos de evolução.

Para realizar a evolução de ontologias relacionadas aos metadados, foi proposto o uso de folksonomias, de forma a alinhá-las e modificá-las para que as ontologias reflitam as mudanças nas folksonomias. Folksonomia é um neologismo criado em 2004 pelo arquiteto Thomas Vander Wal, a partir da junção das palavras inglesas *folks* (povo, pessoas, galera, turma) com *taxonomy* (taxonomia). Segundo Wal¹ “[...] o resultado da atribuição livre e pessoal de etiquetas (*tagging*) a informações ou objetos (qualquer coisa com URL), visando à sua recuperação. A atribuição de etiquetas é feita num ambiente social (compartilhado e aberto a outros)”. Neste primeiro estudo de caso, os autores investigaram como as alterações das ontologias podem ser derivadas de alterações nas folksonomias. Para essa análise, foram reutilizados os conjuntos de *tags Del.icio.us*² apresentados no trabalho de Specia e Motta [40] e *tags* do *Flickr*³ para fotos postadas entre 01/02/2004 e 01/03/2006. O primeiro passo consistiu em identificar os *clusters* que foram adicionados ou removidos.

¹<https://www.vanderwal.net/random/>

²<http://del.icio.us>

³<http://www.flickr.com>

Em seguida, para cada *cluster*, foram analisadas as mudanças internas, como adições ou exclusão de algumas *tags*. Depois que um conjunto de alterações nos metadados foi identificado, essas alterações foram usadas para sugerir atualizações correspondentes às ontologias.

O segundo estudo de caso utilizou um tipo diferente de conjunto de dados do estudo de caso anterior. Nesse caso, os autores investigaram como as ontologias podem evoluir através da aplicação de métodos de aprendizado de ontologia a dados textuais. O conjunto de dados utilizado foi uma coleção de dados textuais do FAO⁴(Organização das Nações Unidas para Agricultura e Alimentação, Departamento de Pesca). Um pré-processamento foi realizado usando as ferramentas *TermExtractor* [33] - um serviço Web para a extração de terminologia relevante para o domínio; *GATE* [12] - um conjunto de aplicativos em linguagem natural para anotação de documentos; *Text2Onto* [10] - uma ferramenta de aprendizado de ontologia. O próximo passo foi a organização conceitual, que visou derivar uma estrutura significativa entre os termos identificados. Para esta tarefa foram utilizados algoritmos de agrupamento para quando o *TermExtractor* for aplicado, levando os clusters semelhantes aos obtidos no estudo de caso 1; e os algoritmos de aprendizado de ontologia do *Text2Onto* para derivar uma estrutura da ontologia entre os termos identificados. Dependendo da estrutura conceitual derivada no passo anterior, duas abordagens de ontologia podem ser adotadas para alinhar essas estruturas às ontologias: 1) Métodos baseados em similaridade de texto foram propostos para encontrar correspondências entre os termos em um *cluster* e os rótulos dos elementos da ontologia; e 2) Métodos de correspondência estrutural para encontrar alinhamentos entre as ontologias derivadas e as ontologias que precisam ser atualizadas.

Para determinar a qualidade de anotações, Gross *et al.* [16] propuseram uma abordagem de avaliação. A metodologia considera os métodos utilizados para a geração de anotações (proveniência), evolução de ontologias, fontes de objetos, e as anotações. O método foi avaliado com um conjunto de anotações funcionais de proteínas humanas extraídas das bases de dados *Ensembl* e *Swiss-Prot*. Essas bases têm suas proteínas anotadas com conceitos da *Gene Ontology* (GO). Os autores analisaram como as anotações evoluem para diferentes tipos de proveniência, ou seja, diferentes tipos de códigos de evidência (geração manual, automática), e como mudanças nas ontologias são propagadas para as anotações. Para essa análise, foram utilizadas as versões lançadas entre março de 2014 e dezembro de 2018 dos repositórios supracitados. O trabalho constatou que a exclusão de conceitos de uma ontologia ocasiona a exclusão da anotação dependente. Adicionalmente, operações de mudanças na ontologia do tipo união de conceitos, substituição de conceitos e tornar um conceito obsoleto podem levar a uma adaptação ou a remoção completa da anotação.

Em seu trabalho, Köpke e Eder [23] investigaram se um conjunto de anotações são impactadas por alterações realizadas em ontologia de domínio usada para a criação das anotações. Para isso, eles implementaram uma abordagem que considera caminhos de anotação e a definição explícita de dependências de mudanças entre elementos de ontologia. Os *logs* de alterações e dependências de mudanças foram investigados para identificar os caminhos de anotações que necessitam de manutenção. Nessa abordagem, foram utili-

⁴ <http://www.fao.org>

zadas as ferramentas *JENA API* e o *reasoner Pellet*. O algoritmo proposto recebe como entrada um conjunto de caminhos de anotação, um conjunto de caminhos de definição de dependência e informações sobre alteração, divisão e junção de conceitos. O resultado gerado é um subconjunto do caminho de anotação que a semântica possivelmente mudou.

Cardoso *et al.* [8] realizaram uma análise empírica da evolução de anotações no domínio biomédico e sua correlação com as alterações em *Knowledge Organization Systems* (KOS) [20]. Para esse fim, foi gerado um conjunto de anotações a partir de uma coleção de documentos médicos pre-selecionados e anotados com as ontologias *ICD-9-CM* e *MeSH*. A ferramenta *COnto-Diff* [18] foi utilizada para determinar um mapeamento de evolução entre duas versões das ontologias. O estudo foi organizado em seis etapas. Primeiramente (etapa 1), foram selecionados aleatoriamente 5000 documentos do total de 733, 138 coletados; 13 versões de cada KOS lançadas entre os anos de 2004 e 2016. Na segunda etapa, as ferramentas *GATE* e *NCBO Annotator* foram empregadas para gerarem um total de 66 milhões de anotações. Os anotadores foram configurados para utilizarem uma versão específica das ontologias por vez. Ainda nesta fase, foi realizada uma filtragem para manter apenas o conceito de correspondência mais longo para uma anotação, ou seja, a palavra com o maior número de carácter. Em seguida (etapa 3), as anotações foram reagrupadas e armazenadas em um repositório. Então, calcularam a diferença simétrica entre dois conjuntos de anotações geradas com duas sucessivas versões do KOS para identificar as anotações que foram removidas, adicionadas e/ou modificadas. Na quarta etapa, cada par de versões sucessivas das ontologias foram submetidas ao *COnto-Diff* para se obter os conjuntos de mudanças ocorridas. Na etapa 5, os 13 conjuntos das anotações de cada documento foram comparadas por pares de versões dos KOS [2004–2005, 2005–2006, etc.] para detectar o que alterou nas anotações e assim encontrar correspondências com as alterações do KOS identificadas na etapa 4. Na etapa 6, os subconjuntos de alterações derivados das anotações e dos KOS foram analisados para determinar o impacto das alterações nas anotações geradas. Encontraram cinco casos de mudanças que podem causar alterações nas anotações dependentes.

3.2 Correção Automática de Anotações Semânticas

Luong e Kantz [26] apresentaram o sistema *CoSWEM* que implementa uma abordagem baseada em regras (restrições de inconsistências) para efetuar a correção de anotações. A ferramenta executa cinco funções principais. O primeiro passo consiste em transformar uma ontologia para o formato RDF(S) assim como o *log* de alterações é convertido para serem extraídas as alterações entre duas versões da ontologia. A conversão do *log* permite identificar as anotações relacionadas com as alterações da ontologia. No segundo passo, o mecanismo de busca semântica *Corese* é utilizado para recuperar possíveis anotações inconsistentes. No passo 3 são aplicadas regras para detectar as anotações verdadeiramente inconsistentes do conjunto coletado na fase anterior. No passo 4, são executadas regras de correção para a revisão das anotações identificadas na etapa antecedente. Por fim, a etapa 5 permite gerenciar versões da ontologia e conjunto de anotações semânticas no caso de armazenar diferentes versões.

Park *et al.* [35] detectaram 3 tipos de inconsistência semântica nas anotações baseadas na *Gene Ontology* (GO) [11], incluindo: anotações semanticamente redundantes; anotações inconsistentes no domínio biológico; e anotações inconsistentes na taxonomia. Para determinar essas inconsistências, eles utilizaram a estrutura hierárquica da GO e a estrutura em árvore da taxonomia *NCBI*. As anotações analisadas foram coletadas de 27 principais bases de dados biológicas, incluindo *NCBI's Gene* [4] e *Ensembl* [21].

A fim de realizar a correção dessas anotações, os autores [35] desenvolveram a ferramenta *GOChase-II*, que foi projetada para tratar os erros sintáticos não completamente solucionados pela sua antecessora *GOChase-I*. *GOChase-II* é uma ferramenta web composta de 4 módulos. *GOChase-History*, desenvolvida para tratar de todo o histórico de evolução de um ID GO. *GOChase-Species* que é responsável por realizar a classificação do uso de um termo GO em espécies distintas e exibir a organização na árvore de taxonomia. O módulo *GOChase-Correct* apresenta um termo combinado e o redireciona para o termo alvo correto no qual os termos foram associados. Para um termo obsoleto, o módulo *GOChase-Correct* apresenta os termos alternativos. *GOChase-Correct* corrige tanto anotações redundantes quanto anotações inconsistentes no domínio biológico. *GOChase-II* também fornece uma análise de enriquecimento da GO para *clusters* de expressão genes. *GOChase-II* é capaz fornecer dois modelos estatísticos, o teste hipergeométrico e o teste exato de Fisher [15], com várias correções de teste de hipóteses (correção de Bonferroni) [39].

Para apoiar a tarefa de manutenção de anotações no domínio biomédico, Cardoso *et al.* [7] projetaram um *framework* composto de 3 métodos, sendo: (1) conjunto de 7 regras (*MergeAnnot*, *IncreaseAnnot*, *ResurrectAnnot*, *PluralAnnot*, *ChangeConceptAnnot*, *SplitAnnot*, *SuperClassAnnot*); (2) *Background Knowledge* (BK) e (3) *Semantic Change Patterns* (SCP). As regras propostas foram derivadas após uma análise da evolução e adaptação de uma coleção de anotações realizada ao longo de 10 anos [8]. A abordagem proposta foi aplicada em um conjunto de documentos anotados e o resultado foi comparado com um corpus de referência (padrão de prata) representando a versão evoluída das anotações. Para compor o conjunto inicial de anotações, foram selecionadas aleatoriamente 500 anotações do conjunto produzido em [8] utilizando a versão 2009AA de quatro ontologias biomédicas (*ICD-9-CM*, *MeSH*, *NCIt* e *NOMEDCT*) 125 anotações de cada KOS. O corpus de referência foi gerado por três especialistas que validaram/corrigiram manualmente a evolução das 500 anotações selecionadas, de acordo com a versão 2010AA das ontologias correspondente. Para medir a efetividade da abordagem proposta, métricas clássicas conhecidas foram adotadas, como *Precision*, *Recall*, *F1-score*, *ROC curve*, *Accuracy*, *false/true positives/negatives*.

O *framework* proposto por Cardoso *et al.* [7] foi avaliado em dois experimentos. O primeiro experimento visou avaliar se os métodos (ou uma combinação deles) fornecem qualidade satisfatória (em termos de *F1-Score*) para determinar se uma anotação impactada evoluirá ou não. O segundo experimento consistiu em aplicar os métodos para selecionar quais ações de adaptação podem fazer com que a anotação evolua corretamente e o resultado foi comparado com o padrão de prata. Os resultados dos experimentos mostraram que os métodos propostos obtiveram uma taxa adequada de confiabilidade no processo de correção/adaptação automática de anotações inconsistentes.

O *framework* proposto por Cardoso *et al.* [7] apresentou algumas limitações. Foi identificado que as regras implementadas não consideravam a forma plural dos termos. O método de BK, apresentava mapeamento desalinhado para as versões anteriores do KOS, e em alguns casos, selecionava conceitos inadequados para substituir as anotações impactadas. Por fim, o método SCP considerava apenas a vizinhança de um conceito em evolução. Para preencher as lacunas identificadas, Cardoso *et al.* [6] desenvolveram melhorias no *framework*. A fim de considerar a forma plural dos termos, foram adicionadas novas regras. Com o intuito de resolver o problema de mapeamento desalinhado apresentado pelo método BK, para todos os candidatos que satisfazem as condições especificadas no mapeamento, a proposta calcula a similaridade entre os ancestrais e o conceito de origem; então, seleciona o melhor candidato para manter a anotação. Para superar os problemas apresentados pelo método SCP, os autores adicionaram uma nova regra chamada *PartialMatch*, capaz de lidar com a evolução global de um conceito. Essa regra baseia-se em medidas de similaridade lexical (LSM) [25] e medidas de similaridade semântica (SSM) [22]. As análises experimentais conduzidas demonstraram que as alterações realizadas alcançaram melhores resultados para adaptar anotações no contexto de ontologias biomédicas.

3.3 Discussão

Declarações RDF definindo recursos do mundo real estão sujeitas às alterações para estar em linha com atualizações do domínio. Lidar com os efeitos das mudanças em bases RDF em evolução sobre as anotações dependentes continua sendo um desafio em aberto no gerenciamento de conjuntos de dados RDF. Isto é especialmente valioso por causa do potencial grande volume de anotações geradas a partir de repositórios RDF presentes na LOD.

Resultados em nossa análise de literatura indicaram que trabalhos existentes exploraram principalmente a manutenção (semi-)automática de anotações em relação às mudanças no nível de ontologia (TBOx) (cf. Tabela 3.1). No entanto, quando anotações são geradas explorando instâncias de conceitos, no melhor do nosso conhecimento, estudos existentes não foram projetados para tratar a manutenção de anotações no nível de instâncias. Esta dissertação tem como principal contribuição investigar e propor métodos para lidar com esse contexto. A Tabela 3.1 apresenta uma análise comparativa dos principais estudos encontrados e posiciona nossa investigação em relação a quatro aspectos chaves:

- **i:** Identificação de mudanças em triplas em repositórios RDF.
- **ii:** Identificação de anotações desatualizadas com base em mudanças em triplas.
- **iii:** Definição e aplicação de ações de adaptação de anotações;
- **iv:** Adaptação de anotações no nível de instância (ABOx) .

Tabela 3.1: Análise do Estado da Arte.

Trabalhos Analisados	i	ii	iii	iv
<i>Maynar et al. 2007</i>	X	X		
<i>Gross et al. 2009</i>	X	X		
<i>Köpke e Eder. 2011</i>	X	X		
<i>Cardoso et al. 2016</i>	X	X		
<i>Luong e Kantz. 2006</i>	X	X	X	
<i>Park et al. 2011</i>	X	X	X	
<i>Cardos et al. 2017</i>	X	X	X	
<i>Cardoso et al. 2018</i>	X	X	X	
<i>Esta dissertação</i>	X	X	X	X

3.4 Conclusão

Este capítulo efetuou uma análise dos principais trabalhos relacionados. Em síntese, observamos que os trabalhos analisados na literatura tratam de anotações no nível de classe, e principalmente em domínios específicos (*e.g.*, Biomedicina). Nesta dissertação abordamos anotações no nível de instância de conceitos nas ontologias correspondentes. A contribuição chave desta pesquisa consiste de um framework capaz de efetuar modificações para atualizar anotações. Desenvolvemos o *framework* AnnoLOD para apoiar a adaptação de anotações de instância em repositórios RDF (*cf.* Capítulo 4).

Capítulo 4

Framework AnnoLOD

Este capítulo apresenta a nossa proposta para adaptar anotações inconsistentes decorrentes das modificações realizadas em repositórios RDF. Para esse fim, propomos o *framework* AnnoLOD para servir como uma solução de apoio para a adaptação assistida de anotações semânticas sobre bases de dados RDF. Descrevemos a visão geral do *framework* (Seção 4.1). Organizamos seções específicas dedicadas à descrição dos seus módulos (Seção 4.2, Seção 4.3 e Seção 4.4). Apresentamos uma ferramenta de software resultante da conceitualização do *framework* AnnoLOD (Seção 4.5). Esta ferramenta ajuda a conduzir as análises sobre as anotações nas diferentes etapas do framework.

4.1 Descrição Geral do *Framework*

O *framework* AnnoLOD define ações de adaptação de anotações a serem realizadas automaticamente quando um *dataset* RDF usado para criar as anotações semânticas evolui (*i.e.*, uma nova versão é gerada). Essas ações são necessárias para manter as anotações consistentes e atualizadas ao longo do tempo. A Figura 4.1 apresenta a proposta envolvendo os principais módulos do framework e o fluxo de dados a fim de adaptar anotações desatualizadas em um determinado conjunto de dados RDF.

As entradas necessárias consistem dos conjuntos de dados iniciais interconectados, sendo $\mathcal{R}^j$, $\mathcal{R}^{j+1}$ (sua nova versão que pode afetar anotações existentes) e as anotações $\mathcal{A}^j$ existentes. O objetivo final do *framework* é obter as anotações $\mathcal{A}^{j+1}$ atualizadas conforme o novo conjunto de dados $\mathcal{R}^{j+1}$. Considerando essas entradas, o *framework* executa uma série de passos (*cf.* Figura 4.1). O Passo A é responsável por detectar e recuperar todas as alterações entre duas versões dos conjuntos de dados envolvidos. O Passo B identifica as anotações afetadas pelas alterações encontradas no dataset RDF. O Passo C realiza ações de adaptação nas anotações afetadas. Os passos são explicados em detalhes nas respectivas Seções 4.2, 4.3 e 4.4.

O Algoritmo 1 define um processo de adaptação de anotação em conformidade com o *framework* AnnoLOD. A linha 3, Algoritmo 1 retrata a Etapa A; as linhas 4, 5 e 6 remetem à Etapa B; e as linhas 7 a 10 referem-se a Etapa C (*cf.* Figura 4.1). O algoritmo necessita dos conjuntos de dados de entrada e um valor η normalizado entre 0 e 1. Esse valor η refere-se a um limite de similaridade responsável pela definição de candidatos que podem

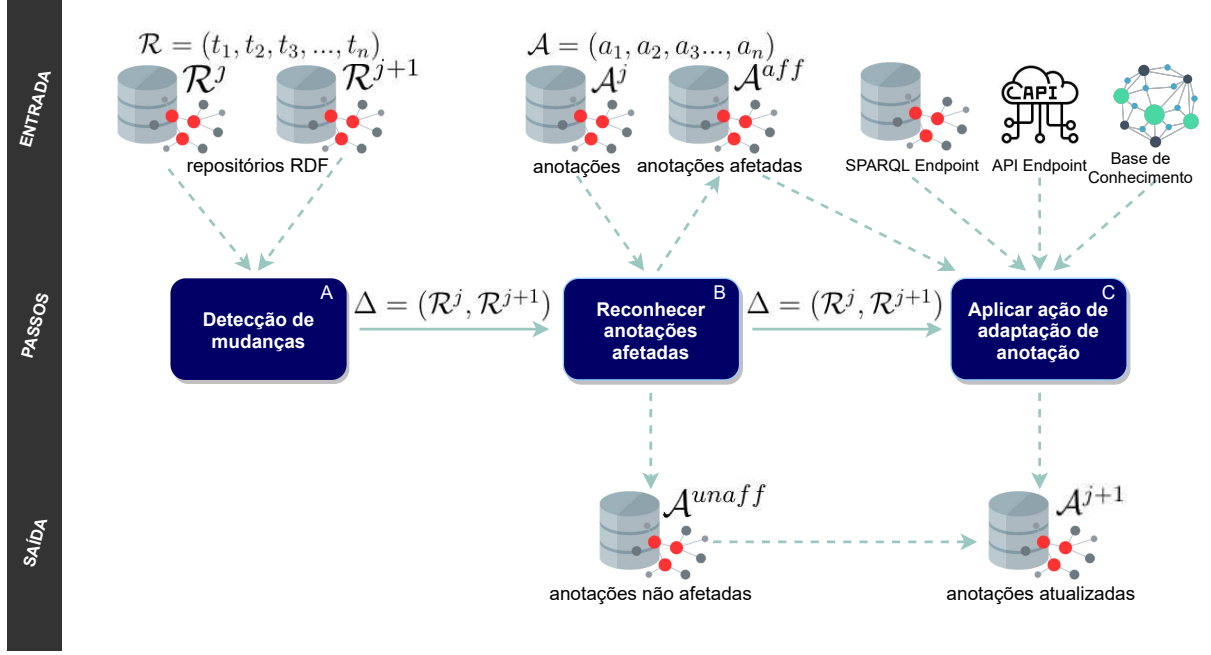


Figura 4.1: *Framework* AnnoLOD.

substituir o recurso que mudou e afetou a anotação. Recursos com valores menores do que η não são considerados candidatos adequados para a substituição. O resultado obtido após a aplicação das ações de adaptação efetuadas por intermédio do usuário refere-se as anotações atualizadas $\mathcal{A}^{j+1}$ (cf. Figura 4.1).

Algorithm 1 Algoritmo de adaptação de anotação

Require: $\mathcal{R}^j, \mathcal{R}^{j+1}, \mathcal{A}^j, \eta, candList$

- 1: $adaptationActions \leftarrow \emptyset$
 - 2: $\mathcal{A}^{j+1} \leftarrow \emptyset$
 - 3: $\Delta \leftarrow detectChanges(\mathcal{R}^j, \mathcal{R}^{j+1});$
 - 4: $\mathcal{A}^{aff} \leftarrow recAffAnnotations(\Delta, \mathcal{A}^j);$
 - 5: $\mathcal{A}^{unaff} \leftarrow \mathcal{A}^j - \mathcal{A}^{aff};$
 - 6: $\mathcal{A}^{j+1} \leftarrow \mathcal{A}^{unaff};$
 - 7: **for all** $a_i \in \mathcal{A}^{aff}$ **do**
 - 8: $candList \leftarrow Algorithm2(a_i);$
 - 9: $adaptationActions[a_i] \leftarrow Algorithm3(a_i, candList, \eta);$
 - 10: $\mathcal{A}^{j+1} \leftarrow applyAdaptationActions(adaptationActions[a_i])$
 - 11: **end for**
 - 12: **return** $\mathcal{A}^{j+1}$
-

4.2 Passo A - Detecção de Mudanças em Repositórios RDF

Esta etapa detecta uma série de modificações RDF que ocorreram em um determinado período de tempo com base em duas versões de um repositório RDF. Essa operação é

conhecida como Δ pois computa a diferença entre dois conjuntos de dados, reconhecendo elementos adicionados, removidos e não atualizados (Linha 3 do Algoritmo 1). As mudanças podem ser do tipo simples (como ações unitárias de adição ou remoção de triplas); ou operações complexas (ação de atualização) do conhecimento armazenado nos conjuntos de dados.

Para identificar as mudanças ocorridas entre versões de um repositório RDF, cada tripla da versão $\mathcal{R}^j$ é comparada com as triplas da versão $\mathcal{R}^{j+1}$. Para isso, primeiro seleciona-se uma tripla de $\mathcal{R}^j$ e outra de $\mathcal{R}^{j+1}$. Em seguida, compara-se o sujeito da tripla da versão $\mathcal{R}^j$ com o sujeito da tripla da versão $\mathcal{R}^{j+1}$. Posteriormente, são comparados os predicados, e por fim, compara-se os objetos. Logo, se pelo menos em uma das comparações entre os sujeitos, predicados e os objetos das triplas em análise, todas as partes que formam as triplas forem iguais, considera-se que não houve mudança e a tripla não foi alterada entre as versões do conjunto de dados.

Caso contrário, se após todas as comparações entre a tripla de $\mathcal{R}^j$ e todas as triplas de $\mathcal{R}^{j+1}$ serem realizadas, e na versão $\mathcal{R}^{j+1}$ não houver pelo menos uma tripla idêntica ao da versão $\mathcal{R}^j$, entende-se que houve a exclusão da tripla na versão $\mathcal{R}^{j+1}$ do repositório. Contudo, se após as comparações entre todas as triplas de $\mathcal{R}^j$ e $\mathcal{R}^{j+1}$ terem sido realizadas, e na versão $\mathcal{R}^{j+1}$ do repositório RDF houver uma ou mais triplas totalmente diferentes das encontradas na versão $\mathcal{R}^j$, tem-se o caso de triplas adicionadas.

No entanto, se na comparação entre uma tripla de $\mathcal{R}^j$ e uma outra tripla de $\mathcal{R}^{j+1}$ apenas o sujeito for diferente, tem-se o caso de mudança de sujeito. Já se apenas o predicado for diferente entre as duas triplas, tem-se o caso de mudança de predicado. Finalmente, se apenas os objetos forem diferentes, tem-se o caso de mudança de objeto da tripla.

Nesta investigação, consideramos as seguintes alterações de triplas RDF:

- $triplaAdicionada(t_k) \rightarrow (t_k \notin \mathcal{R}^j \wedge t_k \in \mathcal{R}^{j+1})$
- $triplaRemovida(t_k) \rightarrow (t_k \in \mathcal{R}^j \wedge t_k \notin \mathcal{R}^{j+1})$
- $sujeitoTriplaModificado(t_k) \rightarrow (t_k \in \mathcal{R}^j \wedge t_k \in \mathcal{R}^{j+1}, s^j \neq s^{j+1})$
- $predicadoTriplaModificado(t_k) \rightarrow (t_k \in \mathcal{R}^j \wedge t_k \in \mathcal{R}^{j+1}, p^j \neq p^{j+1})$
- $objetoTriplaModificado(t_k) \rightarrow (t_k \in \mathcal{R}^j \wedge t_k \in \mathcal{R}^{j+1}, o^j \neq o^{j+1})$

A Figura 4.2 apresenta os tipos de mudanças que podem ocorrer entre as versões $\mathcal{R}^j$ e $\mathcal{R}^{j+1}$ de um repositório RDF.

Primeiro temos um exemplo de remoção de tripla resultante da exclusão de um recurso do conjunto de dados. O recurso “XYZ” presente na versão $\mathcal{R}^j$ foi removido do repositório na sua versão $\mathcal{R}^{j+1}$ (cf. número 1 Figura 4.2) ocasionando assim a remoção completa das suas triplas dependentes $\langle XYZ, livesIn, São\ Paulo \rangle$ e $\langle XYZ, memberOf, IC \rangle$. Como exemplo de adição de tripla decorrente da inclusão de um novo recurso no conjunto de dados, temos o recurso “GHI” (cf. número 2 Figura 4.2) adicionado na versão $\mathcal{R}^{j+1}$ que deu origem a uma nova tripla $\langle GHI, memberOf, Unicamp \rangle$.

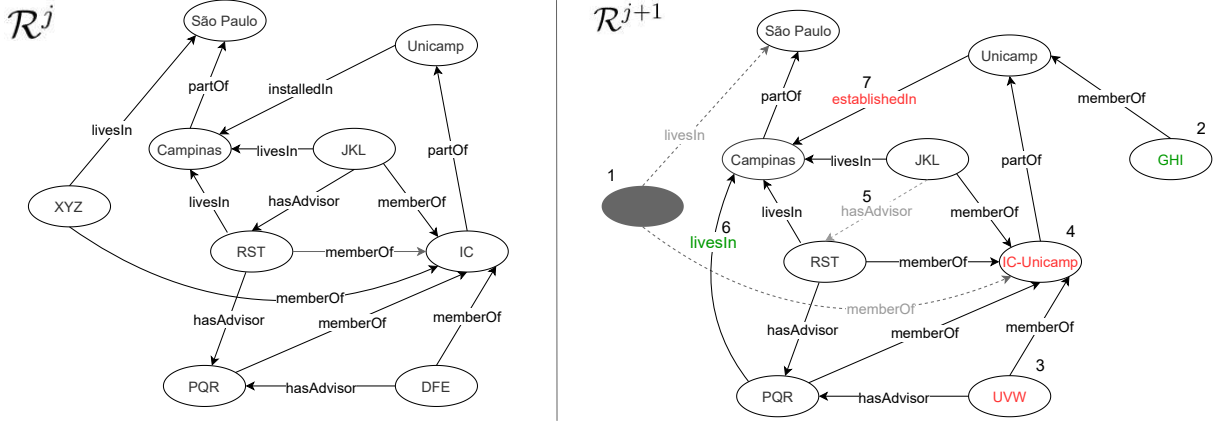


Figura 4.2: Exemplo de tipos de mudanças em repositórios RDF.

Para exemplificar uma mudança de sujeito de uma tripla, temos o recurso “DFE” sujeito da tripla $\langle DFE, hasAdvisor, PQR \rangle$ na versão $\mathcal{R}^j$ que mudou para “UVW” (cf. número 3 Figura 4.2) na versão $\mathcal{R}^{j+1}$. O recurso “IC” objeto das triplas $\langle RST, memberOf, IC \rangle$, $\langle JKL, memberOf, IC \rangle$, $\langle PQR, memberOf, IC \rangle$ e $\langle DFE, memberOf, IC \rangle$ na versão $\mathcal{R}^j$, mudou para para “IC-Unicamp” na versão $\mathcal{R}^{j+1}$ (cf. número 4 Figura 4.2), representando uma mudança de objeto.

Na versão $\mathcal{R}^j$ do conjunto de dados, o predicado *hasAdvisor* vinculava os recursos “JKL” e “RST” formando a tripla $\langle JKL, hasAdvisor, RST \rangle$, porém na versão $\mathcal{R}^{j+1}$ essa tripla foi removida (cf. número 5, Figura 4.2). Já na versão $\mathcal{R}^{j+1}$, uma nova tripla foi criada, representada pelos recursos “PQR” e “Campinas” e o predicado *livesIn* interligando-os $\langle PQR, livesIn, Campinas \rangle$ (cf. número 6, Figura 4.2). Por fim, temos a mudança de predicado de *installedIn* na versão $\mathcal{R}^j$ ($\langle Unicamp, installedIn, Campinas \rangle$) para *establishedIn* (cf. número 7, Figura 4.2) na versão $\mathcal{R}^{j+1}$ vinculando os recursos “Unicamp” e “Campinas”.

Todas essas mudanças são computadas pelo *framework* AnnoLOD para serem utilizadas nos passos B e C (cf. Seções 4.3 e 4.4) para a identificação e adaptação de anotações afetadas pelas alterações identificadas no conjunto de dados RDF.

4.3 Passo B - Identificação de Anotações Inconsistentes

Esta etapa consiste em reconhecer e filtrar as anotações afetadas pelas mudanças identificadas no Passo A daquelas não afetadas (Linhas 4 e 5, Algoritmo 1). As anotações que compartilham um sujeito (s) com uma tripla t_k na qual $t_k \in \Delta$ são consideradas afetadas pelas modificações de mudança (cf. Figura 1.2). Essas são consideradas anotações desatualizadas e ações de adaptação de anotação são aplicadas sobre elas em passos subsequentes do *framework*. Assumimos que anotações não afetadas podem ser diretamente reusadas em compor o conjunto final de anotações atualizadas e adicionadas ao conjunto final de anotações $\mathcal{A}^{j+1}$ (Linha 6, Algoritmo 1). Contudo, as anotações classificadas como afetadas devem ser tratadas pelo *framework*.

A Figura 4.3 apresenta um conjunto de anotações geradas a partir da versão $\mathcal{R}^j$ do

repositório RDF. Os termos “XYZ”, “*Instituto de Computação*”, “Unicamp”, “Campinas” e “São Paulo” foram vinculados às suas respectivas definições semânticas ($\langle \text{XYZ}, \text{sameAs}, \text{XYZ} \rangle$, $\langle \text{Instituto de Computação}, \text{sameAs}, \text{IC} \rangle$, $\langle \text{Unicamp}, \text{sameAs}, \text{Unicamp} \rangle$ e $\langle \text{São Paulo}, \text{sameAs}, \text{São Paulo} \rangle$) presentes na base de dados RDF. Uma nova versão do conjunto de dados foi lançada ($\mathcal{R}^{j+1}$) com mudanças realizadas pelo mantenedor do repositório (cf. Figura 4.2).

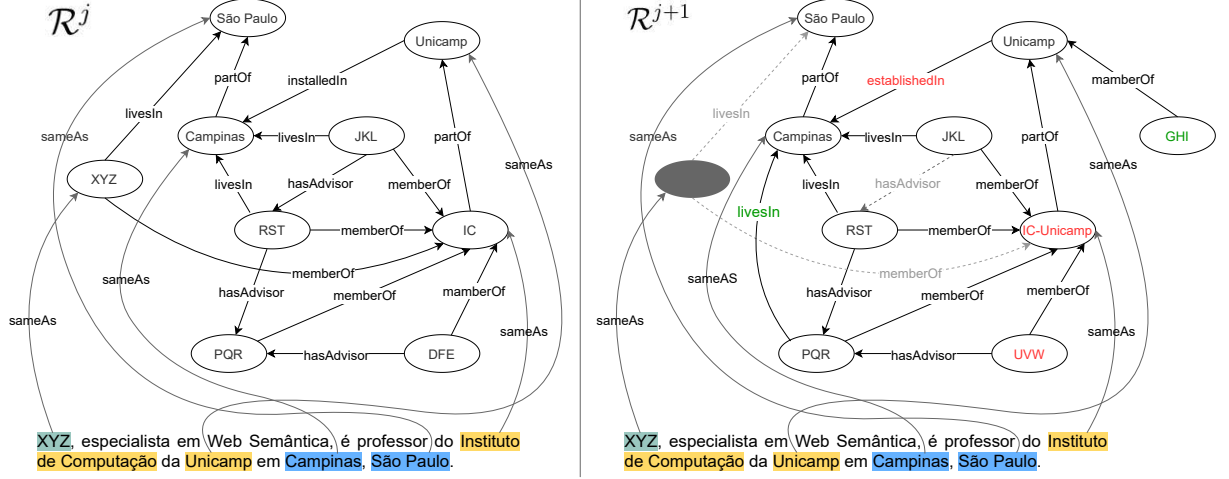


Figura 4.3: Exemplo de impacto das mudanças RDF sobre as anotações semânticas existentes.

Nesse cenário, algumas anotações são impactadas pelas mudanças. A anotação $\langle \text{XYZ}, \text{sameAs}, \text{XYZ} \rangle$ foi afetada pela remoção das triplas $\langle \text{XYZ}, \text{livesIn}, \text{São Paulo} \rangle$ e $\langle \text{XYZ}, \text{memberOf}, \text{IC} \rangle$ do conjunto de dados na versão $\mathcal{R}^{j+1}$, uma vez que, essa anotação compartilhava o recurso “XYZ” que foi juntamente removido do repositório (cf. Figura 4.3). A anotação $\langle \text{Instituto de Computação}, \text{sameAs}, \text{IC} \rangle$ também é considerada afetada, dado que, é necessário verificar se a mudança de “IC” para “IC-Unicamp” é semanticamente consistente.

O *framework* AnnoLOD considera essas anotações desatualizadas e elas são adicionadas ao conjunto das anotações afetadas ($\mathcal{A}^{aff}$). Sobre elas são aplicadas ações de adaptação de anotação (cf. Seção 4.4). As outras anotações que não compartilham um recurso com as triplas envolvidas em modificações são adicionadas ao conjunto das anotações não afetadas ($\mathcal{A}^{unaff}$) para compor o conjunto final das anotações atualizadas ($\mathcal{A}^{j+1}$) em relação à nova versão do conjunto de dados ($\mathcal{R}^{j+1}$).

4.4 Passo C - Aplicar Ação de Adaptação de Anotação

Esta etapa envolve aplicar ações de adaptação nas anotações afetadas e desatualizadas detectadas na etapa B (Linha 4, Algoritmo 1). Para realizar as ações de adaptação, nosso *framework* consulta por candidatos que possam substituir a parte inconsistente da anotação (cf. Subseção 4.4.1). A parte inconsistente da anotação é um recurso, o objeto da tripla (cf. Figura 1.2). Os candidatos são recursos no conjunto de dados na sua versão $\mathcal{R}^{j+1}$ ou em repositórios RDF externos $\mathcal{R}_{ext}$. Em seguida, o Algoritmo 1 computa

as possíveis ações de adaptação que atualiza a anotação, as linhas 7 a 11, Algoritmo 1 definem esse procedimento.

A linha 7, Algoritmo 1 é executada sobre todas as anotações afetadas. Inicialmente, o Algoritmo 1 (linha 8) recebe a lista de candidatos gerada pelo Algoritmo 2. Logo, com base nessa lista, o algoritmo 1 decide qual ação de adaptação deve ser executada após a revisão e escolha do usuário - para reconstituir o modo estável da anotação (linha 9, Algoritmo 1). O AnnoLOD apresenta a recomendação de mais de uma ação aos usuários. Ao final, apenas uma ação de adaptação é aplicada para cada operação de mudança detectada (*cf.* Seção 4.2).

A Seção 4.4.2 explora as possíveis ações a serem aplicadas sobre as anotações. A Seção 4.4.1 apresenta o procedimento para a geração da lista de candidatos.

4.4.1 Geração de Recursos Candidatos

Esta etapa é responsável por gerar uma lista de recursos candidatos $candList = \{r_{c_1}, r_{c_2}, r_{c_3}, \dots, r_{c_n}\}$ para substituir a parte incoerente da anotação afetada. A Figura 4.4 apresenta o passo a passo da metodologia utilizada para gerar a lista de candidatos, que inicia buscando uma lista preliminar de candidatos (número 1, Figura 4.4) e finaliza gerando a lista de candidatos e seus respectivos valores de similaridade (número 8, Figura 4.4). O Algoritmo 2 descreve as etapas concebidas para a geração de uma lista de recursos candidatos.

Nossa abordagem requer como entrada dois conjuntos de dados $\mathcal{R}^{j+1}$ e $\mathcal{R}_{ext}$ para a seleção de candidatos, como também uma base de conhecimento G utilizada para realizar as comparações de similaridade e selecionar o recurso mais adequado para substituir a parte afetada de uma anotação. Qualquer operação no Passo C (*cf.* Figura 4.1) envolvendo o conjunto de dados $\mathcal{R}$ é realizada na versão evoluída (atualizada) $\mathcal{R}^{j+1}$. A versão $\mathcal{R}^j$ é usada apenas no Passo A do *framework* AnnoLOD. O resultado são os melhores candidatos selecionados que podem substituir o recurso que afetou a anotação.

O Algoritmo 2 (linha 2 – 3, Figura 4.7) busca nos repositórios $\mathcal{R}^{j+1}$ e $\mathcal{R}_{ext}$ recursos que tenham a propriedade *rdfs:label* ou *foaf:name* igual à menção (pessoa, lugar, organização, etc.) encontrada em um determinado documento e utilizada para gerar a anotação. O raciocínio é que os recursos que tenham o *label* (*rdfs:label* ou *foaf:name*) idêntico ao da menção detectada podem se referir ao mesmo objeto do mundo real. A base de dados externa ($\mathcal{R}_{ext}$) é utilizada nesta etapa com o objetivo de aumentar a probabilidade de obter recursos que possam substituir os recursos que foram removidos do conjunto de dados $\mathcal{R}$ na versão $\mathcal{R}^{j+1}$. As propriedades *rdfs:label* e *foaf:name* foram adotadas dado que elas indentificam de forma textual os recursos em repositórios RDF (*e.g.* o recurso “Albert Einstein” descrito pela URI http://dbpedia.org/resource/Albert_Einstein presente na *DBpedia* é identificado pelo *label* “Albert Einstein”). Essas propriedades também são importantes, visto que, elas são usadas pelos algoritmos de similaridade adotados pela nossa abordagem para realizarem as comparações de similaridade por meio da confrontação de *strings* (*cf.* Capítulo 2).

Nossa solução busca candidatos utilizando o menção identificada no documento na forma exata como foi detectada e também pelas partes das quais ela é composta (linha

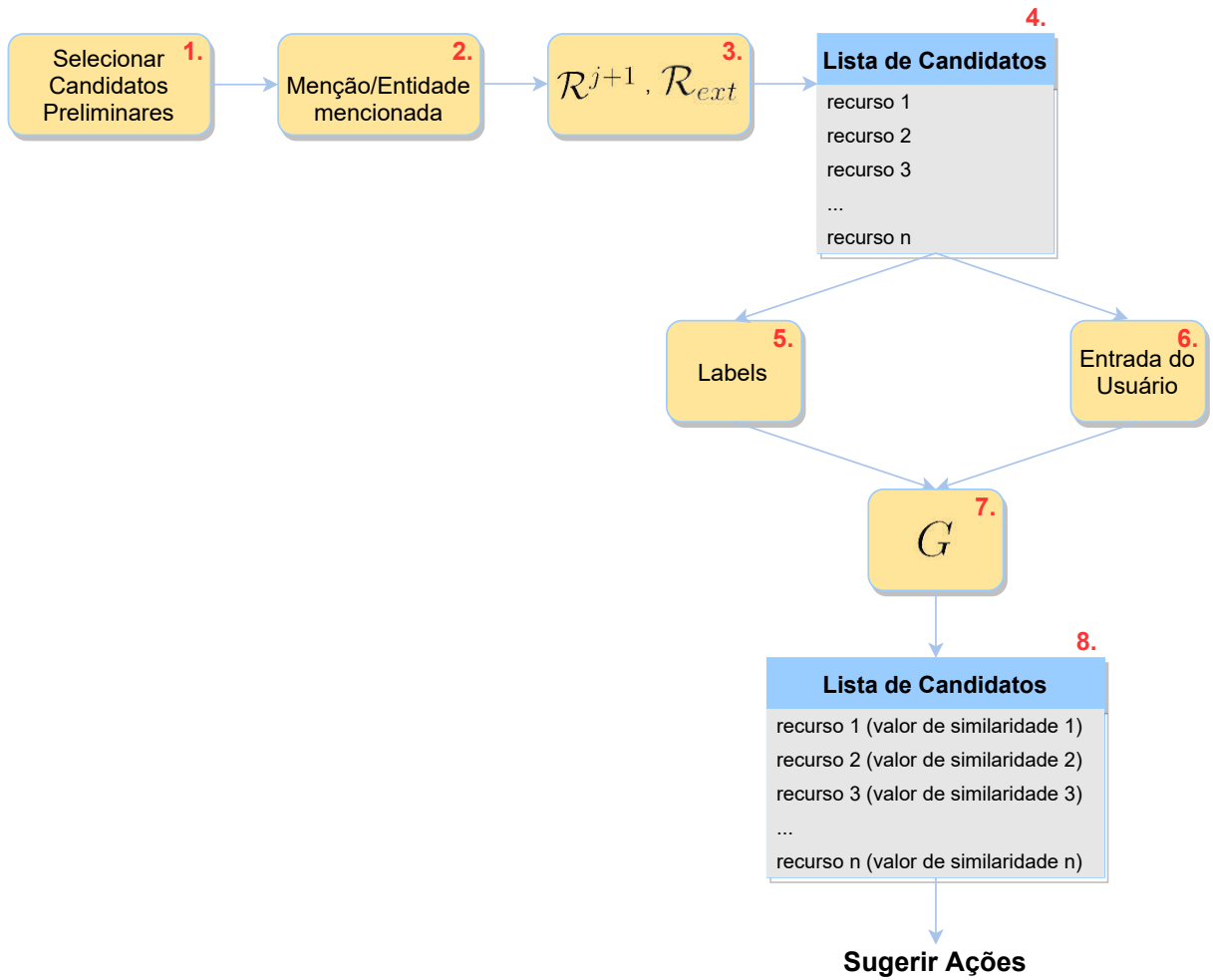


Figura 4.4: Gerando lista de recursos candidatos *candList*.

Algorithm 2 Seleção de candidatos e seus respectivos valores de similaridade.

Require: $\mathcal{R}^{j+1}, \mathcal{R}_{ext}, G, candList \leftarrow \emptyset$

- 1: **function** *getResourcesCandidates*(a_i)
 - 2: $candList \leftarrow getResourcesInitialCandidates(a_i, \mathcal{R}^{j+1}, \mathcal{R}_{ext})$
 - 3: $candList \leftarrow compareLabels(a_i, candList, G);$
 - 4: $candList \leftarrow compareUserInput(a_i, candList, G);$
 - 5: **return** $candList$
-

2, Algoritmo 2 – 2, Figura 4.4). Por exemplo, na Figura 1.1 temos a representação de uma anotação criada a partir da menção ao cientista “*Albert Einstein*”. Nesse caso, para reparar essa anotação buscamos tanto por recursos que tenham o *label* “*Albert Einstein*” quanto por aqueles que possuem de forma parcial o termo “*Albert*” ou “*Einstein*” na sua propriedade *rdfs:label* ou *foaf:name*. Adicionalmente, utilizamos a propriedade *rdf:type* para restringir que a busca retorne apenas recursos compatíveis ao tipo de menção identificada, ou seja, para a menção ao cientista “*Albert Einstein*” que se refere a uma pessoa no mundo real, seriam retornados apenas recursos que fossem instância de pessoas na base de dados RDF correspondente (cf. Figura 4.5).

A partir de agora, nos referimos a essa lista de recursos dos conjuntos de dados $\mathcal{R}^{j+1}$ e $\mathcal{R}_{ext}$ como a lista de recursos candidatos (*candList*).

```

1 ▼ PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
2 PREFIX schema: <http://schema.org/>
3 PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
4 ▼ SELECT DISTINCT * WHERE {
5     ?sub ?pred ?label.
6     ?sub rdf:type schema:Person.
7
8     FILTER ( ?pred = rdfs:label )
9     FILTER regex( str( ?label ), "termo" )
10    FILTER ( lang( ?label ) = "en" )
11 }
12 LIMIT 50

```

Figura 4.5: Consulta SPARQL para selecionar recursos candidatos para substituir o recurso que afetou a anotação.

Com base na lista de candidatos (4, Figura 4.4), as comparações (5 e 6, Figura 4.4) são efetuadas para determinar quais candidatos são de melhor potencial para substituir o recurso que impactou a anotação afetada pela mudança. Para esse fim, nossa solução realiza um conjunto de comparações entre cada candidato da lista *candList* e o recurso (objeto) da anotação afetada. As comparações (5 e 6, Figura 4.4) são:

- **Comparando *Labels* (C1):** O Algoritmo 2 compara os *labels* de cada recurso da lista de candidatos com o *label* do recurso objeto da anotação (cf. Comparação C1 Figura 4.6). O Algoritmo 2 seleciona cada recurso da *candList* que contém o valor da propriedade *rdfs:label* ou *foaf:name* para obter os *labels*. Essas propriedades são parâmetros de entrada do *framework*. Cada par de itens comparados recebe um valor normalizado ([0 - 1]) que indica o grau de similaridade. Explicamos os tipos de similaridades usadas nesta seção a seguir.
- **Comparando Entrada do Usuário (C2):** Nesta etapa, o Algoritmo 2 compara as propriedades explicitamente fornecidas pelo usuário (propriedades como *skos:altLabel* e *skos:prefLabel*) para identificar similaridade entre a lista de candidatos e o recurso inconsistente da anotação (linha 4 – Algoritmo 2). Alguns conjuntos de dados podem não apresentar as propriedades listadas na comparação anterior (*rdfs:label* ou *foaf:name*). Esta situação dificulta as comparações e o funcionamento do nosso *framework*. Como solução, nossa abordagem se beneficia da experiência

do usuário. O usuário, que na maioria dos casos é o mantenedor do conjunto de dados, pode especificar quais propriedades devem ser comparadas (*cf.* Comparação C2 Figura 4.6).

- **Os n melhores candidatos:** Neste ponto, cada recurso da *candList* obteve um valor de similaridade com base em cada uma das comparações mencionadas, gerando uma matriz de valores de similaridade.

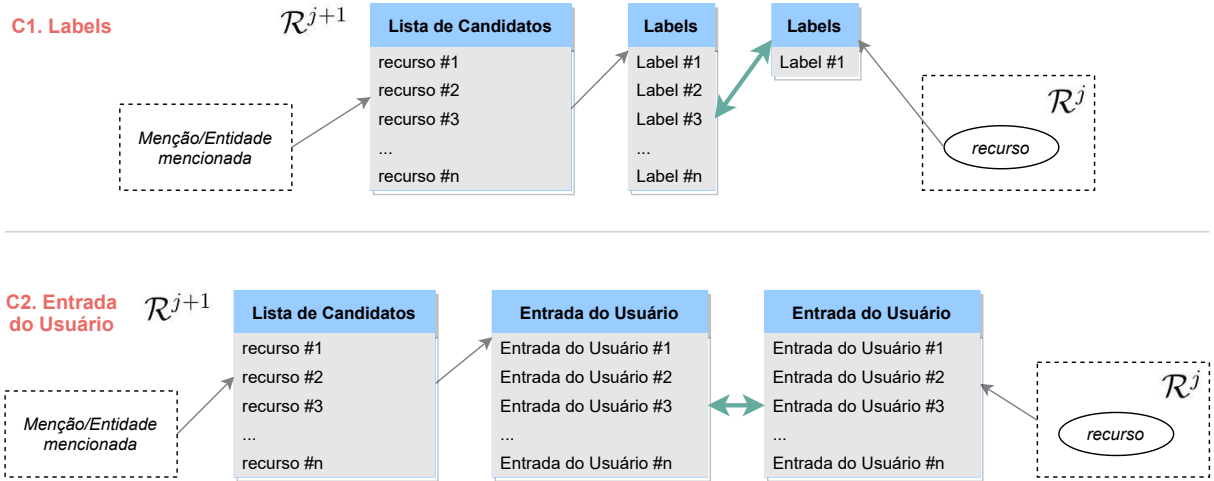


Figura 4.6: Lista de ações realizadas em cada candidato da *candList*. Cada tipo de comparação é identificada pelos textos na cor vermelha. As extremidades das setas verdes indicam quais *labels* são comparados via cálculo de similaridade.

As comparações produzem valores de similaridade usando algoritmos de similaridade semântica. A similaridade semântica computa a que ponto dois termos de entrada (cadeias de caracteres) compartilham o mesmo significado, embora eles difiram em sua forma léxica. Em geral, os algoritmos de similaridade semântica dependem de uma fonte externa de conhecimento, como vocabulários, tesouros, dicionários e ontologias para calcular a similaridade entre os itens - em nossa abordagem chamados de Base de Conhecimento.

Uma técnica comumente utilizada para computar a similaridade semântica entre termos é encontrar a distância mínima do caminho conectando-os a partir de uma entidade âncora de uma base de conhecimento; isso se aplica quando os termos em análise forem representados em uma hierarquia de conceitos [38].

Em particular, nosso framework *AnnoLOD* utiliza duas técnicas para computar a similaridade semântica. A primeira faz uso da base de conhecimento *WordNet* [30], um conjunto de dados léxico organizado em uma hierarquia de conceitos. O valor de similaridade é calculado com base em quantos nós estão entre dois termos.

A segunda técnica refere-se ao algoritmo Nasari [5], que adota como base de conhecimento a rede semântica *BabelNet* [32]. Para determinar a similaridade semântica entre dois termos, o Nasari usa *Weight Overlap* (WO). Todo termo em Nasari é denotado como um vetor de valores de similaridades semânticas [5].

A Equação 4.1 determina a similaridade semântica entre dois elementos e_1 e e_2 (termos expressos em uma cadeia de caracteres):

$$sem(e_1, e_2) = WO(v_1, v_2) \quad (4.1)$$

Na Equação 4.2 é exemplificado como *Weight Overlap* que calcula a similaridade entre os significados de dois itens léxicos determinados. Fundamentalmente:

$$WO(v_1, v_2) = \frac{\sum_{i=1}^{|S|} (r_i^1 + r_i^2)^{-1}}{\sum_{i=1}^{|S|} (2i)^{-1}} \quad (4.2)$$

Na Equação 4.2, $\mathcal{S}$ refere-se ao número de dimensões idênticas entre os vetores; r_i^1 e r_i^2 representam a classificação de dimensão i nos vetores v_1 e v_2 . O emprego da metodologia NASARI é fundamentada por melhores resultados quando comparada à Similaridade de Cossenos [5].

4.4.2 Ações de Adaptação

O resultado da metodologia apresentada na Subseção 4.4.1 aponta os recursos candidatos e seus respectivos valores de similaridade (*cf.* Figura 4.4) que podem substituir adequadamente o recurso afetado (em uma anotação em análise). Os recursos candidatos são utilizados para auxiliar na tarefa de seleção de ações a serem executadas sobre eles (recursos). A Figura 4.7 e o Algoritmo 3 apresentam as etapas necessárias na escolha da ação de adaptação de anotação.

Na linha 2, o Algoritmo 3 recebe três parâmetros: a anotação que está sendo reparada (a_i); a lista de recursos candidatos (*candList*); e um limite de similaridade definido pelo usuário (η). Caso não seja definido um limite pelo usuário, então o valor padrão de 0,5 é assumido pelo Algoritmo (linhas 3 a 5, Algoritmo 3). O valor de similaridade recuperado para cada recurso candidato é comparado ao valor limite η , gerando uma lista contendo os principais candidatos *bestCand* (linhas 6 a 10, Algoritmo 3; veja etapas 5 e 6, Figura 4.4).

A partir dos parâmetros (linha 2, Algoritmo 3) e na lista de candidatos gerada (linha 8, Algoritmo 3), o Algoritmo 3 sugere a melhor ação ao usuário. Assumindo um processo assistido, o usuário toma a decisão final. As possíveis ações são:

- **Reconexão de Objeto:** $replaceObject(a_i, r_n^{j+1}) \rightarrow (a_i \in \mathcal{A}^{j+1} \wedge r_n^{j+1} \in bestCand, r_o^j \neq r_n^{j+1})$.

Esta sugestão de ação é apresentada ao usuário junto com cada candidato melhor classificado (linha 9, Algoritmo 1 e Etapa 8, Figura 4.7). O Algoritmo 1 apresenta esta opção caso o valor de similaridade dos candidatos apresentados seja igual ou superior ao limite η (linhas 7 e 11, Algoritmo 3). Não tendo valores maiores do que o limite significa que não há candidato na *candList* (a lista está vazia *cf.* passo 7, Figura 4.7) que possa substituir o objeto parte da anotação e ser conectado ao sujeito usando o predicado *owl:sameAs*;

- **Remoção da anotação:** $removeAnnotation(a_i) \rightarrow (a_i \in \mathcal{A}^j \wedge a_i \notin \mathcal{A}^{j+1})$.

Outra ação é a remoção completa da anotação (linha 14, Algoritmo 3 e Etapa 9, Figura 4.7). Esta ação é apresentada ao usuário quando não há candidato para

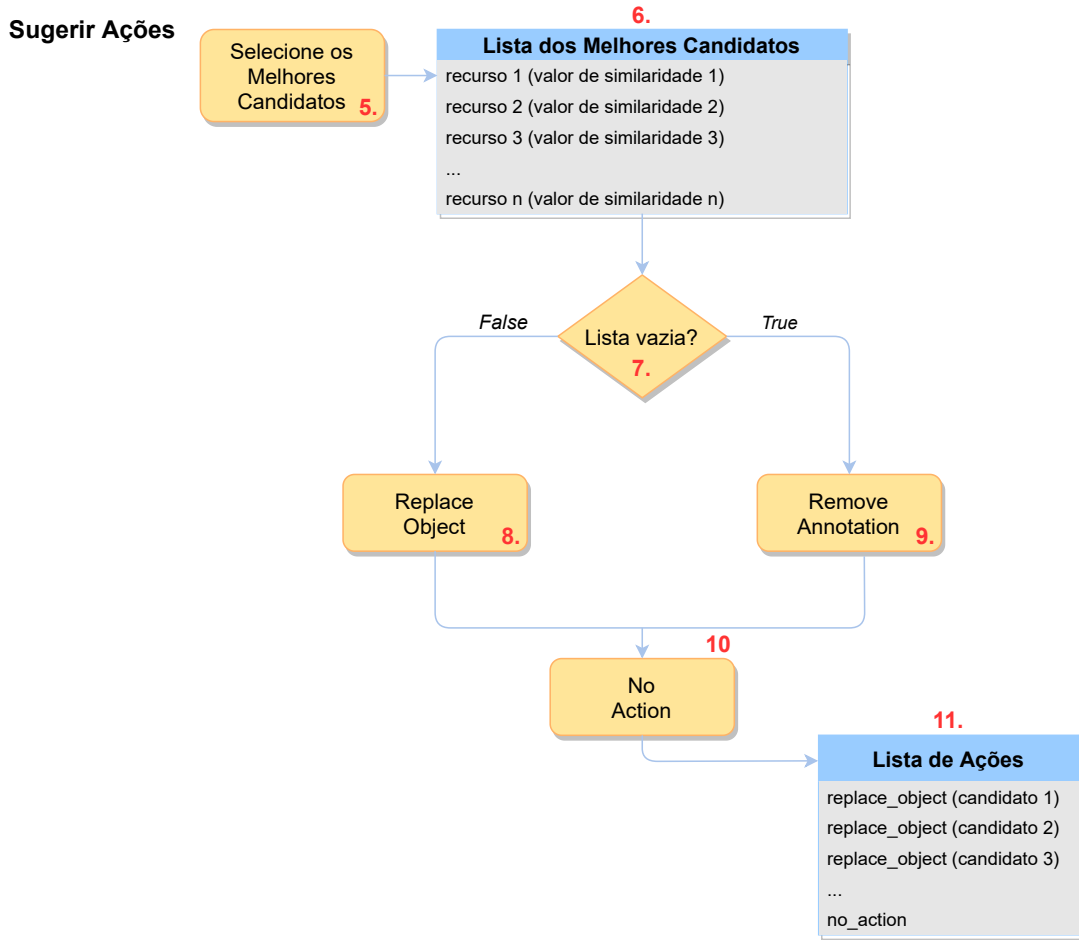


Figura 4.7: Sugestão de ações de adaptação de anotação ao usuário a partir na lista de recursos candidatos *candList*.

Algorithm 3 Obter sugestões de ações de adaptação de anotação.

```

1:  $bestCand \leftarrow \emptyset$ ;  $adaptationActions \leftarrow \emptyset$ ;  $similarityLimit \leftarrow 0,5$ 
2: function suggestAdaptationAction( $a_i$ , candList,  $\eta$ )
3:   if  $\eta = NULL$  then
4:      $\eta \leftarrow similarityLimit$ 
5:   end if
6:   for all candidate  $\in$  candList do
7:     if getSimilarityValue(candidate)  $\geq \eta$  then
8:        $bestCand \leftarrow bestCand \cup candidate$ 
9:     end if
10:  end for
11:  if  $bestCand \neq \emptyset$  then
12:     $adaptationActions \leftarrow replaceObject(a_i, bestCand)$ 
13:  else
14:     $adaptationActions \leftarrow removeAnnotation(a_i)$ 
15:  end if
16:   $adaptationActions \leftarrow adaptationActions \cup noActions()$ 
17:  return adaptationActions

```

substituir a anotação afetada (*bestCand* está vazia), com base no limite η (linhas 7 e 13, Algoritmo 3); e

- **Não realizar ação de adaptação:** $noAdaptationAction() \rightarrow (a_i \in \mathcal{A}^j \wedge a_i \in \mathcal{A}^{j+1})$.

Uma das duas ações definidas (Reconexão de objeto ou Remoção da anotação) são apresentadas ao usuário, que decide qual delas é a mais adequada em cada caso de anotação inconsistente. O *framework* pode produzir dois tipos de resultados incorretos. O primeiro caso é a categorização errada, afirmando que uma anotação estável foi afetada (falso positivo). O outro caso é a escolha errada da ação a ser tomada. Em ambos os casos, o usuário pode instruir o *framework* para não realizar nenhuma das ações e manter a anotação como está (linha 16, Algoritmo 3; Etapa 10, Figura 4.7)

Declaramos que mais de uma ação pode ser exibida para o usuário tomar a decisão. Apenas uma pode ser escolhida pelo usuário e executada pelo AnnoLOD. A linha 11, Algoritmo 1 e a Etapa 11, Figura 4.7 retornam uma lista de ações possíveis ao usuário, que escolhe a mais adequada.

4.5 Ferramenta de Software de Apoio para Adaptação Assistida de Anotação

Nesta seção apresentamos a ferramenta de software desenvolvida para apoiar a aplicação de nossa proposta. O desenvolvimento da ferramenta se fundamentou na conceitualização do *framework* AnnoLOD (*cf.* Seção 4.1) para apoiar a adaptação de anotações em conjuntos de dados RDF. Para o desenvolvimento desta ferramenta Web utilizamos a linguagem de programação *JAVA*¹, o *framework Spring Boot*² e o *framework Angular*³. Nossa ferramenta recebe duas versões diferentes do conjunto de dados RDF e detecta mudanças nas triplas de uma versão para a outra. Com base nisso, a ferramenta sugere ações de adaptação para corrigir os casos encontrados de anotações inconsistentes derivadas das mudanças ocorridas. A saída é um conjunto de anotações atualizadas. A Figura 4.8 apresenta o fluxo de interação para a realização das ações de adaptação.

No fluxo de interação da ferramenta, o usuário pode criar uma nova tarefa (“*Create a New Task*” – *cf.* A na Figura 4.9) para iniciar uma nova adaptação do conjunto de anotações; ou o usuário pode abrir uma tarefa existente (“*Open a Task*” – *cf.* B na Figura 4.9) para manter uma verificação RDF já iniciada. O menu do lado esquerdo na primeira tela permite acessar as tarefas realizadas recentemente (*cf.* D na Figura 4.9 - Figura A).

Quando o usuário seleciona a opção de criar uma nova tarefa – “*Create a New Task*” (*cf.* A, Figura 4.8 – Figura A), ele encontra a interface que descreve os dados necessários para entrada (*cf.* Figura B, Figura 4.9). O usuário informa um nome para a tarefa e inclui uma lista de parâmetros (*cf.* B, C e D, Figura 4.8).

¹<https://www.java.com/>

²<https://spring.io/projects/spring-boot>

³<https://angular.io/>

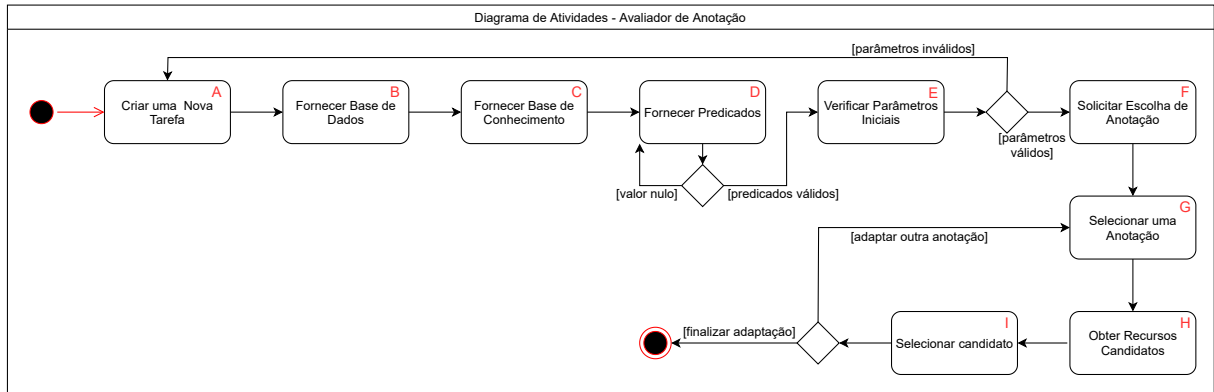


Figura 4.8: Fluxo de interação do usuário no sistema de software para adaptação de anotações semânticas

Figura A

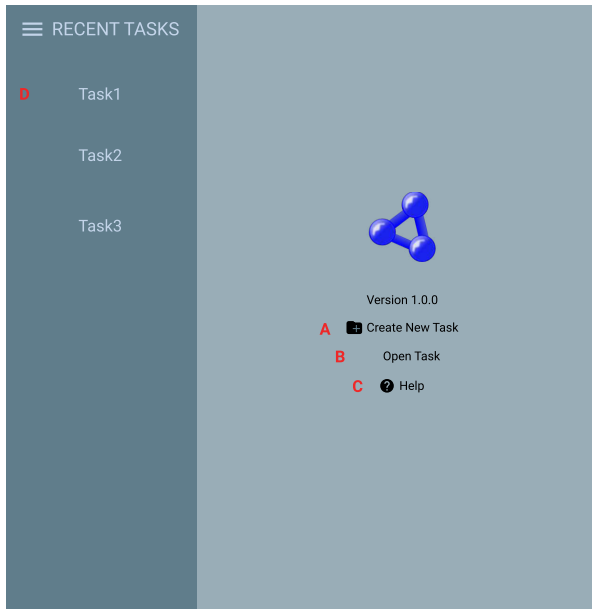


Figura B

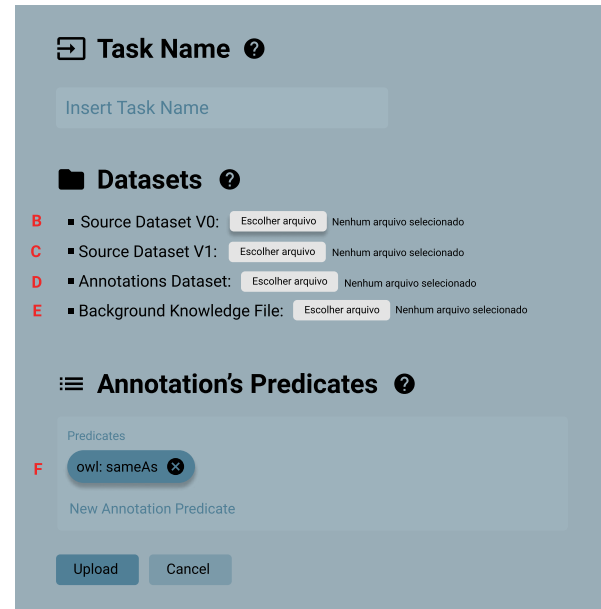


Figura 4.9: Figura A - Tela inicial do sistema. Figura B - Interface inicial e de configuração da ferramenta.

O usuário inicia com o conjunto de dados RDF fonte e sua segunda versão (*cf.* B e C, Figura 4.9 – Figura B). O usuário indica um conjunto de anotações (*cf.* D, Figura 4.9 – Figura B). Ele pode igualmente inserir um *Background Knowledge* (Base de Conhecimento) como uma rede semântica usada por nossa ferramenta para calcular similaridade semântica entre recursos RDF (*cf.* E, Figura 4.9 – Figura B). Isso é necessário para computar os recursos candidatos no processo de adaptação. O usuário pode informar os tipos de predicados das anotações que devem ser analisados (*cf.* F, Figura 4.9 – Figura B). Como padrão, consideramos “*owl:sameAs*”.

A Figura 4.10 apresenta um painel com informações relevantes sobre os conjuntos de dados de entrada. O painel permite que o usuário visualize a porcentagem detalhada das anotações identificadas e outras estatísticas relevantes sobre os conjuntos de dados. O usuário pode acessar a lista de anotações alteradas pelas mudanças ocorridas nas triplas separados por sujeito modificado ou removido, predicado e objeto através do menu do lado esquerdo (*cf.* A, B e C, Figura 4.9).

Um gráfico (*cf.* D, Figura 4.10) permite analisar a porcentagem de anotações que foram verificadas pelo usuário e o número de anotações a serem analisadas. O sistema apresenta informações sobre os conjuntos de dados envolvidos (*cf.* E, Figura 4.10), desde que as operações de alteração sejam encontradas nas triplas entre as versões RDF processadas (*cf.* F, Figura 4.10).

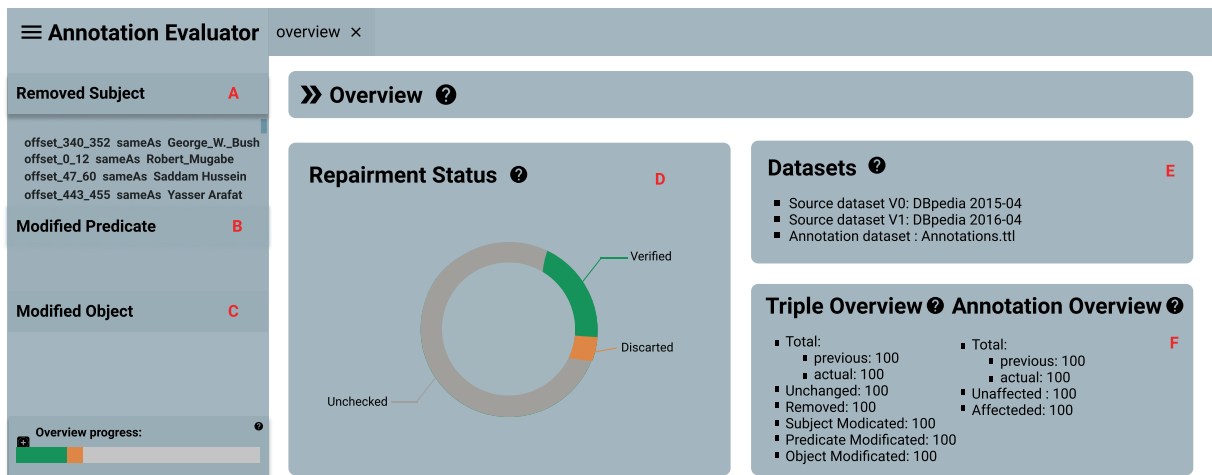


Figura 4.10: Interface do painel do avaliador de anotação.

A análise do usuário é conduzida “anotação por anotação” para concluir o processo de adaptação. Nessa fase, consideramos apenas as anotações afetadas pelas alterações RDF computadas. Isso é resultado de um procedimento de nosso sistema que categoriza uma anotação envolvida em uma mudança de tripla como afetada. Em resumo, uma anotação é considerada afetada se ela compartilha um recurso com uma tripla que sofreu algum tipo de alteração (*cf.* Seção 4.3).

O usuário pode selecionar uma das anotações afetadas (passos F e G, Figura 4.8) exibidos no menu do lado esquerdo (*cf.* A, B e C, Figura 4.10). Isso classifica as anotações com base no tipo de mudanças que as afetaram. Selecionando uma anotação, nossa ferramenta de software apresenta detalhes relativos à respectiva anotação em uma interface dedicada à sua adaptação (*cf.* Figura 4.11).

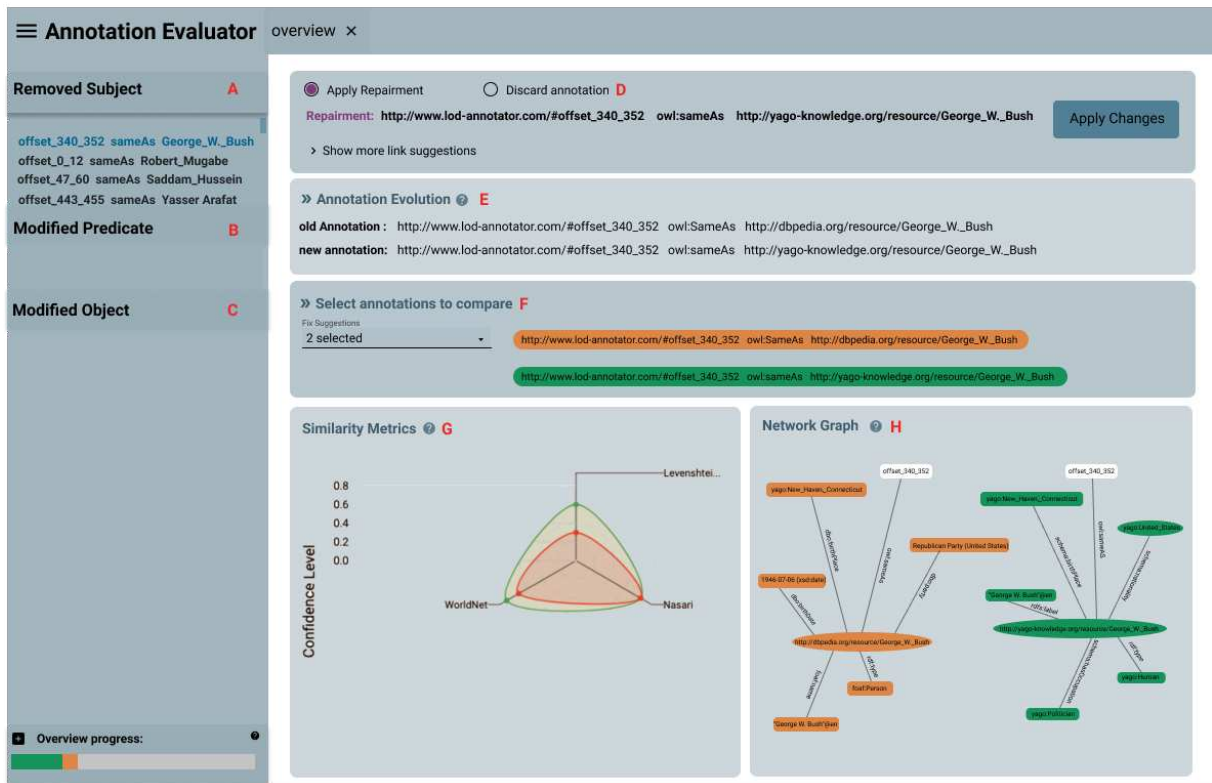


Figura 4.11: Interface do painel do avaliador de anotação.

A Figura 4.11 apresenta a interface na qual o lado esquerdo contém um menu fixo com a lista de anotações alteradas. O sistema foi projetado com diferentes painéis ilustrando informações úteis sobre a anotação selecionada. O painel (cf. E, Figura 4.11) apresenta um exemplo de anotação inconsistente relacionada às alterações ocorridas entre as versões $\mathcal{R}^j$ e $\mathcal{R}^{j+1}$ do conjunto de dados RDF em análise (*DBpedia*). Nesse exemplo, a primeira anotação (com o sujeito representado pelo *offset* 340-352 (“*George W. Bush*”)) compartilhava o recurso “George W. Bush” com uma tripla que foi removida da base de dados na versão $\mathcal{R}^{j+1}$. Após essa remoção, nossa ferramenta categorizou essa anotação como afetada.

O painel (cf. D, Figura 4.11) apresenta as ações sugeridas que o usuário pode tomar para atualizar a anotação afetada (cf. passo I, Figura 4.8). Essas ações são automaticamente computadas pelo sistema. No exemplo corrente, a primeira sugestão é a substituição do recurso com *label* “George W. Bush” da *DBpedia* pelo recurso “George W. Bush” presente no repositório RDF *YAGO*. Nesse caso, um recurso que consiste em um maior grau de similaridade com o antigo objeto da anotação.

Alternativamente, se o usuário não concordar com a sugestão da ferramenta, existe a possibilidade de remover completamente a anotação afetada do conjunto de dados (opção “*Discard annotation*” - cf. D, Figura 4.11) ou aplicar outras ações de manutenção sugeridas pela ferramenta (cf. passo I, Figura 4.8). A lista de ações sugeridas varia desde a escolha de outro recurso até a remoção total da anotação, por exemplo. Para isso, o usuário deve escolher a opção “*Show more annotation suggestions*” (cf. D, Figura 4.11).

Seguindo o processo de adaptação, dois grafos são apresentados para auxiliar o usuário

na compreensão das ações sugeridas pela ferramenta (*cf.* H, Figura 4.11). Eles ajudam as pessoas a compreenderem visualmente qual será o resultado após a aplicação de uma ou outra ação de adaptação de anotação. O painel (*cf.* G, Figura 4.11) apresenta os valores de similaridade de cada algoritmo de similaridade usado por nossa ferramenta. O exemplo (*cf.* G, Figura 4.11) ilustra dois algoritmo/base de conhecimento: *WordNet* e *Nasari*.

As anotações selecionadas (*cf.* F, Figura 4.11 - fundo verde e laranja) são comparadas usando os dois algoritmos de similaridade (*cf.* G, Figura 4.11). Valores próximos a 1 indicam que o novo objeto e o objeto antigo das anotações são mais semanticamente semelhantes quando comparados aos valores próximos a 0. A visualização do grafo apresenta as anotações afetadas e alguns de seus recursos conectados (*cf.* H, Figura 4.11).

O grafo laranja (lado esquerdo) apresenta a anotação em um estado inconsistente; o grafo verde (lado direito) apresenta a anotação após a adaptação, caso a sugestão feita pela ferramenta seja aceita pelo usuário. Ambos os painéis G e H servem de orientação ao usuário na decisão final. O usuário escolhe a ação adequada e confirma as alterações clicando em “aplicar mudança” – “*Apply change*” (*cf.* D, Figura 4.11). Se o usuário precisar analisar outra anotação, ele seleciona a anotação desejada no painel esquerdo, retornando à etapa G do sistema (*cf.* Figura 4.8).

4.6 Conclusão

Este capítulo apresentou o *framework* AnnoLOD para adaptação assistida de anotações semânticas sobre bases RDF na Web. Descrevemos as etapas envolvidas, incluindo a identificação de alterações de conjuntos de dados e a escolha de ações de adaptação de anotações. A partir da conceituação do *framework* proposto implementamos uma ferramenta de software completa para a adaptação de anotações semânticas, incluindo a interface do usuário para interagir com os recursos candidatos e ações de adaptação. O próximo capítulo conduz uma avaliação experimental de nosso estudo.

Capítulo 5

Avaliação Experimental

Este capítulo apresenta uma avaliação da aplicação do *framework*. O objetivo desta avaliação é analisar nossa solução para detecção e adaptação de anotações desatualizadas com base em mudanças identificadas em repositórios RDF do mundo real. A Seção 5.1 descreve os conjuntos de dados utilizados. A Seção 5.2 apresenta a metodologia de avaliação adotada. Os resultados obtidos com a avaliação conduzida são descritos na Seção 5.3; a Seção 5.4 efetua uma análise dos resultados alcançados. Por fim, A Seção 5.5 discute os resultados realçando o impacto das mudanças RDF sob as anotações dependentes e, a contribuição deste trabalho para o gerenciamento de conjuntos de dados RDF.

5.1 Conjuntos de Dados

Para esta avaliação, utilizamos os repositórios RDF *DBpedia* [31] e o *YAGO* [42]. Nosso critério para essa seleção foi considerar conjuntos de dados amplamente conhecidos e com características de serem bases de dados abertas, bem como sofrerem atualizações ao longo do tempo através da publicação de novas versões. A característica de evolução é importante para esta avaliação porque esses conjuntos de dados estão sujeitos a novos lançamentos periodicamente. Nossa solução é projetada para verificar a consistência das anotações quando uma nova versão do conjunto de dados das quais elas são dependentes é lançada. Além dos repositórios RDF, utilizamos uma coleção de documentos para gerarmos os conjuntos de anotações utilizados nesta avaliação (*cf.* Subseção 5.2.1). Para esse fim, adotamos o *dataset* LP50. Explicaremos o uso desses datasets em mais detalhes a seguir.

O projeto *DBpedia* extrai informações estruturadas da *Wikipédia*¹. Em nossa avaliação, selecionamos duas versões em inglês da *DBpedia* (versão 1: lançamento em Abril de 2015²; versão 2: lançamento de Abril de 2016³). A versão de Abril de 2015 em inglês da *DBpedia* descreve 5,9 milhões de recursos. No total, 4 milhões de recursos são classificados em uma ontologia consistente e consistem em 2,06 milhões de pessoas, 682 mil lugares (incluindo 455 mil lugares povoados), 376 mil trabalhos criativos (incluindo 92 mil álbuns de música, 90 mil filmes e 17 mil videogames), 188 mil organizações (incluindo 51 mil

¹<https://www.wikipedia.org/>

²<http://wikidata.dbpedia.org/dbpedia-data-set-2015-04>

³<http://wikidata.dbpedia.org/dbpedia-version-2016-04>

empresas e 33 mil instituições de ensino), 278 mil espécies e 5 mil doenças. O número total de recursos na *DBpedia* em inglês é de 15,3 milhões que, além dos 5,9 milhões de recursos, inclui 1,2 milhões de conceitos de skos (categorias), 6,83 milhões de páginas de redirecionamento, 256 mil páginas de desambiguação e 1,13 milhões de nós intermediários. Ao todo, o lançamento da *DBpedia* 2015-04 consiste em 6,9 bilhões de informações (triplas RDF).

Já a versão de Abril de 2016 em inglês da base da *DBpedia* descreve 6,0 milhões de entidades. No total, 5,2 milhões de recursos são classificados em uma ontologia consistente, consistindo em 1,5 milhões de pessoas, 810 mil lugares (incluindo 505 mil lugares povoados), 490 mil obras (incluindo 135 mil álbuns de música, 106 mil filmes e 20 mil videogames), 275 mil organizações (incluindo 67 mil empresas e 53 mil instituições de ensino), 301 mil espécies e 5 mil doenças. O número total de recursos em *DBpedia* em inglês é de 16,9 milhões que, além dos 6,0 milhões de recursos, inclui 1,7 milhões de conceitos de skos (categorias), 7,3 milhões de páginas de redirecionamento, 260 mil páginas de desambiguação e 1,7 milhões de nós intermediários. Ao todo, o lançamento da *DBpedia* 2016-04 consiste em 9,5 bilhões de informações (triplas RDF).

Devido ao grande volume de dados e limites de recursos computacionais para processamento dessa informação, usamos os subconjuntos das versões 2015-04⁴ e 2016-04⁵ da *DBpedia* relacionados a informações sobre pessoas (*persondata*); os dados estão no formato *N-Triples*⁶ agrupados em único um arquivo RDF para cada versão. A Figura 5.1 apresenta um fragmento do subconjunto da versão 2015-04. A partir de agora, nos referimos aos subconjuntos das versões 2015-04 e 2016-04 da *DBpedia* como $\mathcal{R}^j$ e $\mathcal{R}^{j+1}$ respectivamente.

YAGO é uma ontologia com ampla cobertura e qualidade. *YAGO* se baseia em entidades e relações [41]. Em nossa avaliação, selecionamos a versão 4 do repositório *YAGO*⁷, sendo esta a versão mais recente lançada deste conjunto de dados. *YAGO* 4 é baseado na *Wikidata*⁸ - a maior base de conhecimento de uso geral pública. *YAGO* é, portanto, uma versão simplificada da *Wikidata* e atualmente contém mais de 50 milhões de entidades e 2 bilhões de fatos (triplas RDF). Os fatos são automaticamente extraídos da *Wikipédia* e unificados com o *WordNet*⁹ [41].

De forma similar ao processo realizado com o repositório da *DBpedia*, utilizamos apenas o subconjunto da base de conhecimento *YAGO* relacionado aos dados sobre pessoas. Diferentemente da *DBpedia*, *YAGO* não disponibiliza essas informações em um único arquivo RDF (cf. Figura 5.1). A base de conhecimento *YAGO* 4 é distribuída usando um conjunto de arquivos *N-Triples* de texto completo independentes, que juntos constituem a base de conhecimento. Os arquivos são:

- ***Taxonomy***: a taxonomia completa das classes;
- ***Full-types***: todas as relações *rdf:type*;

⁴http://downloads.dbpedia.org/2015-04/core-i18n/en/persondata_en.ttl.bz2

⁵http://downloads.dbpedia.org/2016-04/core-i18n/en/persondata_en.ttl.bz2

⁶<https://www.w3.org/TR/n-triples/>

⁷<https://yago-knowledge.org/downloads/yago-4>

⁸<http://wikidata.org>

⁹<https://wordnet.princeton.edu/>

```

1 <http://dbpedia.org/resource/Natasha_Stott_Despoja> <http://dbpedia.org/ontology/birthPlace>
2 <http://dbpedia.org/resource/Adelaide> .
3 <http://dbpedia.org/resource/Natasha_Stott_Despoja> <http://purl.org/dc/terms/description>
4 "Australian politician"@en .
5 <http://dbpedia.org/resource/Natasha_Stott_Despoja> <http://dbpedia.org/ontology/birthDate>
6 "1969-9-9"^^<http://www.w3.org/2001/XMLSchema#date> .
7 <http://dbpedia.org/resource/Natasha_Stott_Despoja> <http://xmlns.com/foaf/0.1/name> "Natasha Stott Despoja"@en .
8 <http://dbpedia.org/resource/Natasha_Stott_Despoja> <http://xmlns.com/foaf/0.1/gender> "female"@en .
9 <http://dbpedia.org/resource/Natasha_Stott_Despoja> <http://xmlns.com/foaf/0.1/givenName> "Natasha"@en .
10 <http://dbpedia.org/resource/Natasha_Stott_Despoja> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
11 <http://xmlns.com/foaf/0.1/Person> .
12 <http://dbpedia.org/resource/Aden_Ridgeway> <http://dbpedia.org/ontology/birthPlace>
13 <http://dbpedia.org/resource/Macksville,_New_South_Wales> .
14 <http://dbpedia.org/resource/Aden_Ridgeway> <http://purl.org/dc/terms/description> "Australian politician"@en .
15 <http://dbpedia.org/resource/Aden_Ridgeway> <http://dbpedia.org/ontology/birthDate>
16 "1962-9-18"^^<http://www.w3.org/2001/XMLSchema#date> .
17 <http://dbpedia.org/resource/Aden_Ridgeway> <http://xmlns.com/foaf/0.1/name> "Aden Ridgeway"@en .
18 <http://dbpedia.org/resource/Aden_Ridgeway> <http://xmlns.com/foaf/0.1/gender> "male"@en .
19 <http://dbpedia.org/resource/Aden_Ridgeway> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
20 <http://xmlns.com/foaf/0.1/Person> .
21 <http://dbpedia.org/resource/Brian_Greig> <http://dbpedia.org/ontology/birthPlace> <http://dbpedia.org/resource/Fremantle> .
22 <http://dbpedia.org/resource/Brian_Greig> <http://purl.org/dc/terms/description> "Australian politician"@en .
23 <http://dbpedia.org/resource/Brian_Greig> <http://dbpedia.org/ontology/birthDate>
24 "1966-2-22"^^<http://www.w3.org/2001/XMLSchema#date> .
25 <http://dbpedia.org/resource/Brian_Greig> <http://xmlns.com/foaf/0.1/name> "Brian Greig"@en .
26 <http://dbpedia.org/resource/Brian_Greig> <http://xmlns.com/foaf/0.1/gender> "male"@en .
27 <http://dbpedia.org/resource/Brian_Greig> <http://xmlns.com/foaf/0.1/givenName> "Brian"@en .
28 <http://dbpedia.org/resource/Brian_Greig> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
29 <http://xmlns.com/foaf/0.1/Person> .
30 <http://dbpedia.org/resource/Brian_Greig> <http://xmlns.com/foaf/0.1/surname> "Greig"@en .

```

Figura 5.1: Fragmento do arquivo RDF contendo informações de pessoas no formato *N-Triples* do conjunto de dados da *BBpedia* versão 2015-04.

- **Labels:** todos os *labels* de entidade (*rdfs:label*, *rdfs:comment* e *schema:alternateName*);
- **Facts:** os fatos que não são *labels*;
- **Annotations:** as anotações de fatos (“fatos sobre fatos”) em RDF.
- **SameAs:** os *links owl:sameAs* para *Wikidata* e *Freebase*¹⁰.
- **Schema:** as classes e propriedades do schema.org, em OWL 2 DL¹¹; e
- **Shapes:** as restrições SHACL¹² usadas para gerar *YAGO* 4.

A versão atual do *YAGO* é distribuída em três formatos:

- **Full:** esta forma usa todos os dados do Wikidata. Portanto, é um conjunto de dados extremamente volumoso;
- **Wikipedia:** é uma versão menor do repositório *YAGO* 4 que contém apenas as instâncias que possuem um artigo da Wikipedia (em qualquer idioma); e
- **English Wikipedia:** esta é uma restrição do tipo *Wikipedia* para instâncias que têm um artigo da *Wikipedia* em inglês.

Para a nossa avaliação, selecionamos os arquivos *Full-types*¹³ e *Labels*¹⁴ fornecidos pela *English Wikipedia* lançada em Fevereiro de 2020 para constituir o nosso subconjunto de dados *YAGO* (cf. Figura 5.2). O arquivo RDF contendo as informações de pessoas foi criado manualmente. Para isso, selecionamos cada recurso do arquivo *Full-types* e seu respectivo *label* no arquivo *Labels* para formar o nosso subconjunto (cf. Figura 5.2). A partir de agora, nos referimos ao subconjunto do repositório *YAGO* 4 como $\mathcal{R}_{ext}$.

O *dataset* LP50 é composto por documentos compilados por Lee e Welsh [24]. LP50 é um conjunto composto por 50 documentos que contêm manchetes de notícias sobre vários tópicos (economia, política, etc.). Os documentos variaram em comprimento de 51 a 126 palavras e estão no formato .txt¹⁵. A Figura 5.3 descreve o conteúdo do documento 37 presente no *dataset* LP50.

No momento da escrita desta dissertação, o repositório para acesso aos documentos não estava mais acessível. Desta forma, disponibilizamos um *link*¹⁶ para uma pasta onde pode ser encontrado o conjunto de documentos que constitui o *dataset* LP50.

¹⁰<http://www.freebase.com>

¹¹<https://www.w3.org/ns/owl-profile/data/DL>

¹²<https://www.w3.org/TR/shacl/>

¹³<https://yago-knowledge.org/data/yago4/en/2020-02-24/yago-wd-full-types.nt.gz>

¹⁴<https://yago-knowledge.org/data/yago4/en/2020-02-24/yago-wd-labels.nt.gz>

¹⁵https://pt.wikipedia.org/wiki/Arquivo_de_texto#.txt

¹⁶<https://drive.google.com/drive/folders/1KRcDgPR41SL7hfxO4aVStkAiZa8zYcAE?usp=sharing>

```

1 <http://yago-knowledge.org/resource/Natasha_Stott_Despoja> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
2 <http://yago-knowledge.org/resource/Human> .
3 <http://yago-knowledge.org/resource/Natasha_Stott_Despoja> <http://www.w3.org/2000/01/rdf-schema#label>
4 "Natasha Stott Despoja"@en .
5 <http://yago-knowledge.org/resource/Aden_Ridgeway> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
6 <http://yago-knowledge.org/resource/Human> .
7 <http://yago-knowledge.org/resource/Aden_Ridgeway> <http://www.w3.org/2000/01/rdf-schema#label>
8 "Aden Ridgeway"@en .
9 <http://yago-knowledge.org/resource/Brian_Greig> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
10 <http://yago-knowledge.org/resource/Human> .
11 <http://yago-knowledge.org/resource/Brian_Greig> <http://www.w3.org/2000/01/rdf-schema#label>
12 "Brian Greig"@en .
13 <http://yago-knowledge.org/resource/Robert_Mugabe> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
14 <http://yago-knowledge.org/resource/Human> .
15 <http://yago-knowledge.org/resource/Robert_Mugabe> <http://www.w3.org/2000/01/rdf-schema#label>
16 "Robert Mugabe"@en .
17 <http://yago-knowledge.org/resource/Ari_Fleischer> <http://yago-knowledge.org/resource/Ari_Fleischer>
18 <http://yago-knowledge.org/resource/Human> .
19 <http://yago-knowledge.org/resource/Ari_Fleischer> <http://www.w3.org/2000/01/rdf-schema#label>
20 "Ari Fleischer"@en .
21 <http://yago-knowledge.org/resource/Colin_Powell> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
22 <http://yago-knowledge.org/resource/Human> .
23 <http://yago-knowledge.org/resource/Colin_Powell> <http://www.w3.org/2000/01/rdf-schema#label>
24 "Colin Powell"@en .
25 <http://yago-knowledge.org/resource/Osama_bin_Laden> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
26 <http://yago-knowledge.org/resource/Human> .
27 <http://yago-knowledge.org/resource/Osama_bin_Laden> <http://www.w3.org/2000/01/rdf-schema#label>
28 "Osama bin Laden"@en

```

Figura 5.2: Fragmento do arquivo RDF contendo informações de pessoas no formato *N-Triples* do subconjunto *YAGO 4*.

The Johannesburg Earth Summit is set to get under way with the promise that leaders will take action on the environment, debt and poverty. South African President Thabo Mbeki, speaking at the opening ceremony, said: "Out of Johannesburg and out of Africa must emerge something that takes the world forward." But the absence of US President George W Bush was threatening to overshadow the summit.

Figura 5.3: Exemplo de Texto do documento 37 do *dataset* LP50.

5.2 Procedimentos

A Figura 5.4 apresenta as etapas da metodologia de avaliação conduzida para analisar a efetividade do *framework* AnnoLOD. Inicialmente, criamos os conjuntos de anotações $\mathcal{A}^j$ e $\mathcal{SA}^{j+1}$ (etapa A, Figura 5.4). O conjunto $\mathcal{A}^j$ foi gerado a partir da versão $\mathcal{R}^j$ da *DBpedia*. O conjunto $\mathcal{SA}^{j+1}$ foi criado com base nos repositórios *DBpedia* ($\mathcal{R}^{j+1}$) e *YAGO* ($\mathcal{R}_{ext}$). O conjunto $\mathcal{SA}^{j+1}$ corresponde ao nosso Padrão de Prata. Seção 5.2.1 apresenta em detalhes o procedimento de geração dos conjuntos de anotações.

Na etapa B (cf. Figura 5.4), o *framework* AnnoLOD recebeu como entrada as anotações $\mathcal{A}^j$ e os conjuntos de dados da *DBpedia* nas versões $\mathcal{R}^j$ e $\mathcal{R}^{j+1}$. Primeiramente, o AnnoLOD computou todas as mudanças ocorridas entre as versões do repositório (cf. Subseção 4.2). Seguidamente, foram identificadas as anotações afetadas pelas alterações computadas (cf. Subseção 4.3), e sobre elas foram aplicadas as ações de adaptação pelo *framework* (cf. Subseção 4.4.2). Por fim, na Etapa C (cf. Figura 5.4) foi realizada uma comparação entre as anotações adaptadas ($\mathcal{A}^{j+1}$) e o conjunto padrão de prata ($\mathcal{SA}^{j+1}$) para verificar a efetividade do *framework* na detecção e manutenção das anotações desatualizadas.

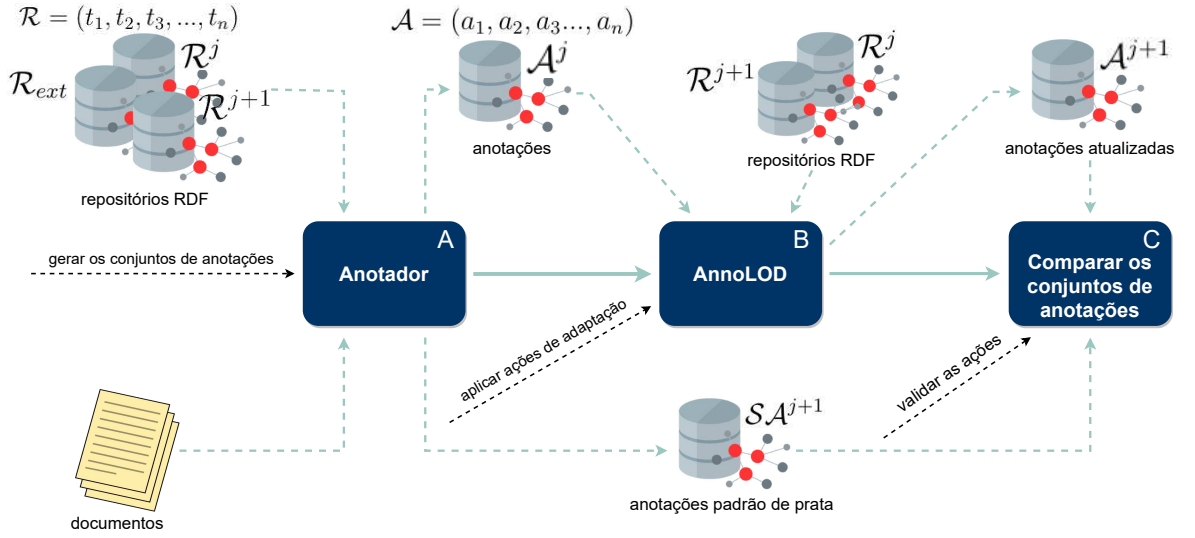


Figura 5.4: Metodologia da Avaliação Experimental.

5.2.1 Gerando os Conjuntos de Anotações

Inicialmente, buscamos na literatura conjuntos de dados sobre anotações de instância para serem adotadas nesta fase de avaliação desta pesquisa. Diversos trabalhos na literatura geraram conjunto de anotações (e.g., [16], [8]). Em uma análise minuciosa dos *datasets* existentes para fins de investigação, constatou-se que os conjuntos de dados não contêm o tipo de anotação explorada neste estudo, uma vez que as anotações foram produzidas no nível de classe e não de instância nas ontologias correspondentes.

Logo, decidimos gerar um conjunto próprio de referência de anotações no nível de instância, pois não há um padrão de ouro disponível e validado para reuso. A princípio,

as anotações seriam geradas por meio de ferramentas computacionais existentes para esse fim, como por exemplo *DBpedia Spotlight* [29] e *KIM* [37]. No entanto, essas ferramentas não permitem selecionar versões diferentes de um mesmo repositório ou elas possuem uma ontologia própria incorporada, como é caso da *KIM*. Esses fatores impossibilitaram a utilização desses anotadores no nosso processo de criação das anotações. Diante disso, desenvolvemos o nosso próprio anotador semântico.

Para desenvolver o anotador utilizamos a linguagem de programação *JAVA*, a biblioteca de leitura de arquivos RDF *Apache Jena*¹⁷, as bibliotecas *LingPipe*¹⁸ e *CoreNLP*¹⁹[27] para a remoção de *stop words*, reconhecimento e extração de entidades mencionadas, como por exemplo pessoas, lugares, organizações, etc..

Todos os documentos do conjunto LP50 foram submetidos ao anotador para que fossem identificados os termos que se referissem às pessoas; adicionalmente foram vinculados as suas respectivas definições semânticas na *DBpedia* ($\mathcal{R}^j$, $\mathcal{R}^{j+1}$) e *YAGO* ($\mathcal{R}_{ext}$) (cf. Figura 5.5) por meio do anotador. A utilização de um repositório externo foi necessária principalmente para os casos de remoção de triplas, uma vez que, quando triplas são removidas ocorre juntamente a remoção de recursos do conjunto de dados RDF. Isso diminui a probabilidade de encontrar recursos que possam substituir partes inconsistentes de anotações desatualizadas.

Para cada documento foi gerado um arquivo RDF contendo suas respectivas anotações. As anotações foram geradas no formato *N-Triples*²⁰. Os predicados utilizados na formação das anotações estão definidos nas ontologias *NLP Interchange Format* (NIF) [19] e *Web Ontology Language* (OWL) [1].

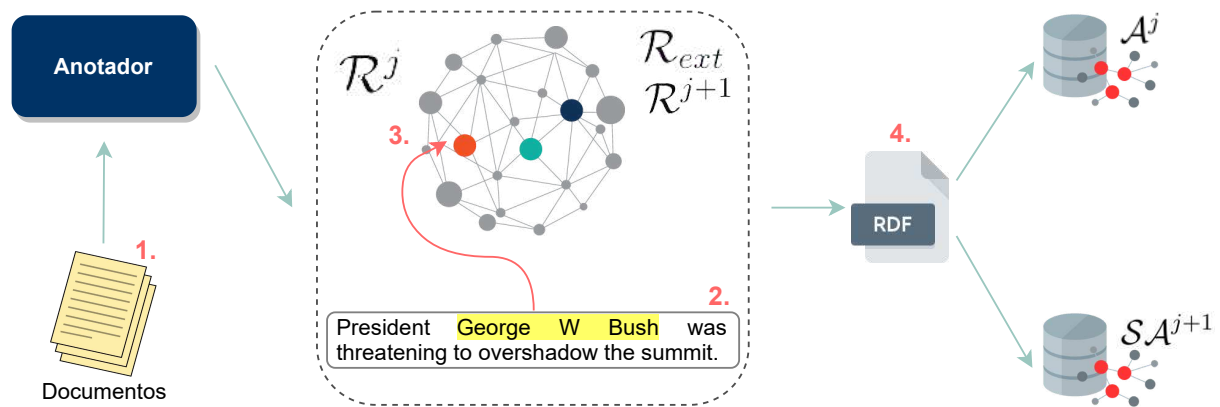


Figura 5.5: Metodologia conduzida para gerar os conjuntos de anotações $\mathcal{A}^j$ e $\mathcal{A}^{j+1}$.

Como apresentado na Figura 5.5, primeiramente o anotador carregou todos os documentos (Etapa 1 Figura 5.5). Em seguida, na etapa 2, cada documento passou pelo processo de *Named-entity recognition* - NER (Reconhecimento de Entidade Mencionada). Para esta avaliação foram considerados apenas termos que faziam menção às pessoas. Posteriormente, para cada termo identificado o anotador fez uma consulta aos repositó-

¹⁷<https://jena.apache.org/>

¹⁸<http://www.alias-i.com/lingpipe/>

¹⁹<https://stanfordnlp.github.io/CoreNLP/>

²⁰<https://www.w3.org/TR/n-triples/>

rios da *DBpedia* e *YAGO* por meio do *Protocol and RDF Query Language* (SPARQL) para selecionar o recurso correspondente ao termo identificado (cf. Figura 5.6). Havendo um recurso equivalente, o anotador criou a anotação vinculando o termo a sua definição semântica utilizando o predicado “*same as*” da OWL (*owl:sameAs*) (Etapa 3 – Figura 5.5). Por fim, na etapa 4 (cf. Figura 5.5) o anotador gerou os arquivos RDF contendo todas as anotações que formam os conjuntos $\mathcal{A}^j$ e $\mathcal{SA}^{j+1}$.

```

1 ▼ PREFIX foaf: <http://xmlns.com/foaf/0.1/>
2 PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
3 ▼ SELECT ?sub WHERE {
4     ?sub rdf:type foaf:Person;
5         foaf:name ?name.
6
7     FILTER ( ?name = "termo" )
8 }
```

Figura 5.6: Consulta SPARQL para selecionar um recurso idêntico ao termo identificado no documento.

A Figura 5.7 descreve um exemplo do arquivo contendo a anotação que foi gerada pelo anotador a partir do documento da Figura 5.3. No exemplo, foi identificado o termo “*George W. Bush*”. A menção ao ex-presidente americano “George W. Bush” foi vinculada ao recurso “George W. Bush” (sua definição semântica) presente na *DBpedia* na versão $\mathcal{R}^j$ (cf. Figura 5.1) encontrado por meio da consulta SPARQL demonstrada na Figura 5.6.

O predicado *nif-core:isString* (linha 5 – Figura 5.7) indica o texto (em vermelho) do documento anotado. O predicado *nif-core:anchorof* (linha 8 – Figura 5.7) especifica o termo (em vermelho) presente no documento que gerou a anotação. Já os predicados *nif-core:beginIndex* e *nif-core:endIndex* nas linhas 3-4 e 9-10 (cf. Figura 5.7) indicam as posições (inicial e final) onde encontram-se o texto e o termo que gerou a anotação respectivamente no documento. Na linha 14 (cf. Figura 5.7), temos a tripla que define a anotação por meio do predicado *owl:sameAs*. O sujeito da tripla é formado pela posição inicial e final (*offset*) do termo. Já o objeto é constituído pelo recurso “George W. Bush” da *DBpedia*.

A Figura 5.8 representa umas das anotações que formam o conjunto padrão de prata. As entidades mencionadas “*Abbas Khalaf*” e “*Bush*” foram vinculadas (anotadas) às suas respectivas definições semânticas nos repositórios *DBpedia* ($\mathcal{R}^{j+1}$) e *YAGO* ($\mathcal{R}_{ext}$) respectivamente. Inicialmente, a menção ao ex-presidente americano “*Bush*” (“*George W. Bush*”) tinha sido vinculada ao recurso “George W. Bush” da *DBpedia* versão $\mathcal{R}^j$ (linha 14 Figura 5.7), no entanto, ao lançarem a nova versão desse conjunto de dados ($\mathcal{R}^{j+1}$) houve a remoção do recurso “George W. Bush”. Como esse recurso deixou de existir no repositório de origem na versão $\mathcal{R}^{j+1}$, para gerar a versão atualizada dessa anotação, buscou-se na base de dados externa um recurso que fosse semanticamente idêntico ao antigo recurso excluído (linha 41 Figura 5.8). Como a menção “*Abbas Khalaf*” não teve a sua definição semântica removida conjunto de dados na versão $\mathcal{R}^{j+1}$, ela permaneceu inalterada (linha 18 Figura 5.8).

Similarmente, existe um arquivo idêntico ao das Figuras 5.7 e 5.8 para todos os do-

1	<http://www.lod-annotator.com/#offset_0_394> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#OffsetBasedString> .
2	<http://www.lod-annotator.com/#offset_0_394> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#Context> .
3	<http://www.lod-annotator.com/#offset_0_394> <http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#beginIndex> "0"^^<http://www.w3.org/2001/XMLSchema#nonNegativeInteger> .
4	<http://www.lod-annotator.com/#offset_0_394> <http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#endIndex> "394"^^<http://www.w3.org/2001/XMLSchema#nonNegativeInteger> .
5	<http://www.lod-annotator.com/#offset_0_394> <http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#isString> 'The Johannesburg Earth Summit is set to get under way with the promise that leaders will take action on the environment, debt and poverty. South African President Thabo Mbeki, speaking at the opening ceremony, said: "Out of Johannesburg and out of Africa must emerge something that takes the world forward." But the absence of US President George W Bush was threatening to overshadow the summit.' .
6	<http://www.lod-annotator.com/#offset_340_352> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#Phrase> .
7	<http://www.lod-annotator.com/#offset_340_352> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#OffsetBasedString> .
8	<http://www.lod-annotator.com/#offset_340_352> <http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#anchorOf> "George W Bush" .
9	<http://www.lod-annotator.com/#offset_340_352> <http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#beginIndex> "340"^^<http://www.w3.org/2001/XMLSchema#nonNegativeInteger> .
10	<http://www.lod-annotator.com/#offset_340_352> <http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#endIndex> "352"^^<http://www.w3.org/2001/XMLSchema#nonNegativeInteger> .
11	<http://www.lod-annotator.com/#offset_340_352> <http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#referenceContext> <http://www.lod-annotator.com/#offset_0_394> .
12	<http://www.lod-annotator.com/#offset_340_352> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://xmlns.com/foaf/0.1/Person> .
13	<http://www.lod-annotator.com/#offset_340_352> <http://www.w3.org/2005/11/its/rdf#taAnnotatorsRef> <http://www.lod-annotator.com> .
14	<http://www.lod-annotator.com/#offset_340_352> <http://www.w3.org/2002/07/owl#sameAs> <http://dbpedia.org/resource/George_W._Bush> .

Figura 5.7: Exemplo de Anotação gerada pelo anotador a partir do repositório $\mathcal{R}^j$.

cumentos anotados, tanto para $\mathcal{R}^j$ (anotações $\mathcal{A}^j$) quanto para $\mathcal{R}^{j+1}$ e $\mathcal{R}_{ext}$ (anotações $\mathcal{SA}^{j+1}$). Os conjuntos $\mathcal{A}^{j+1}$ e $\mathcal{SA}^{j+1}$ são constituídos de 67 anotações organizadas em 29 arquivos; isso porque um único arquivo pode conter uma ou mais anotações, totalizando assim 134 anotações distribuídas em 58 arquivos ao todo. Sobre o conjunto padrão de prata ($\mathcal{SA}^{j+1}$) foi realizada análise manual detalhada para verificar se os recursos vinculados a cada menção encontrada nos documentos anotados de fato correspondiam a definição semântica da menção identificada.

Ressalta-se que não foi possível gerar anotações para todos os documentos do conjunto, uma vez que, em alguns documentos não encontrou-se menções a pessoas, objeto de análise desta pesquisa. Os conjuntos de anotações gerados foram utilizados para verificar a efetividade do *framework* AnnoLOD.

A Tabela 5.1²³ e a Tabela 5.2 apresentam os conjuntos de anotações criados. Elas contêm os termos e em quais documentos eles estão presentes, bem como os recursos e os conjunto de dados correspondentes aos quais os termos foram vinculados (anotados).

²¹https://drive.google.com/drive/folders/19Me1DOO3eo_59WvqyaxIsyprLsloEYLR?usp=sharing

²²<https://drive.google.com/drive/folders/1mGSmeIc-8PK2wNFGX5iovFsPGnLPoHpT?usp=sharing>

²³Os números na frente dos termos indicam quantas vezes eles aparecem nos documentos e a quantidade de anotações geradas a partir deles

```

1 | <http://www.lod-annotator.com/#offset_0_520> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
2 | <http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#OffsetBasedString> .
3 | <http://www.lod-annotator.com/#offset_0_520> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
4 | <http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#Context> .
5 | <http://www.lod-annotator.com/#offset_0_520> <http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#beginIndex>
6 | "0"^^<http://www.w3.org/2001/XMLSchema#nonNegativeInteger> .
7 | <http://www.lod-annotator.com/#offset_0_520> <http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#endIndex>
8 | "520"^^<http://www.w3.org/2001/XMLSchema#nonNegativeInteger> .
9 | <http://www.lod-annotator.com/#offset_0_520> <http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#isString>
10 | "Iraq and Russia are close to signing a $40 billion economic cooperation plan, Iraq's ambassador said Saturday, a deal that could
11 | put Moscow at odds with the United States as it considers a military attack against Baghdad. The statement by Ambassador Abbas
12 | Khalaf came amid indications that Russia, despite its strong support for the post-Sept. 11 antiterrorism coalition, is maintaining or
13 | improving ties with Iran and North Korea, which together with Iraq are the countries President Bush has labeled the axis of evil." .
14 | <http://www.lod-annotator.com/#offset_250_261> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
15 | <http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#Phrase> .
16 | <http://www.lod-annotator.com/#offset_250_261> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
17 | <http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#OffsetBasedString> .
18 | <http://www.lod-annotator.com/#offset_250_261> <http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#anchorOf> "Abbas Khalaf" .
19 | <http://www.lod-annotator.com/#offset_250_261> <http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#beginIndex> "
20 | 250"^^<http://www.w3.org/2001/XMLSchema#nonNegativeInteger> .
21 | <http://www.lod-annotator.com/#offset_250_261> <http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#endIndex>
22 | "261"^^<http://www.w3.org/2001/XMLSchema#nonNegativeInteger> .
23 | <http://www.lod-annotator.com/#offset_250_261> <http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#referenceContext>
24 | <http://www.lod-annotator.com/#offset_0_520> .
25 | <http://www.lod-annotator.com/#offset_250_261> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://xmlns.com/foaf/0.1/Person> .
26 | <http://www.lod-annotator.com/#offset_250_261> <http://www.w3.org/2005/11/its/rdf#taAnnotatorsRef> <http://www.lod-annotator.com> .
27 | <http://www.lod-annotator.com/#offset_250_261> <http://www.w3.org/2002/07/owl#sameAs> <http://dbpedia.org/resource/Abbas_Khalaf> .
28 | <http://www.lod-annotator.com/#offset_485_488> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
29 | <http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#Phrase> .
30 | <http://www.lod-annotator.com/#offset_485_488> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
31 | <http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#OffsetBasedString> .
32 | <http://www.lod-annotator.com/#offset_485_488> <http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#anchorOf> "Bush" .
33 | <http://www.lod-annotator.com/#offset_485_488> <http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#beginIndex>
34 | "485"^^<http://www.w3.org/2001/XMLSchema#nonNegativeInteger> .
35 | <http://www.lod-annotator.com/#offset_485_488> <http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#endIndex>
36 | "488"^^<http://www.w3.org/2001/XMLSchema#nonNegativeInteger> .
37 | <http://www.lod-annotator.com/#offset_485_488> <http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#referenceContext>
38 | <http://www.lod-annotator.com/#offset_0_520> .
39 | <http://www.lod-annotator.com/#offset_485_488> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://xmlns.com/foaf/0.1/Person> .
40 | <http://www.lod-annotator.com/#offset_485_488> <http://www.w3.org/2005/11/its/rdf#taAnnotatorsRef> <http://www.lod-annotator.com> .
41 | <http://www.lod-annotator.com/#offset_485_488> <http://www.w3.org/2002/07/owl#sameAs> <http://yago-knowledge.org/resource/George_W_Bush> .

```

Figura 5.8: Exemplo de Anotação do conjunto padrão de prata gerada pelo anotador a partir dos repositórios $\mathcal{R}^{j+1}$ e $\mathcal{R}_{ext}$.

Tabela 5.1: Anotações $\mathcal{A}^j$.

* <host> é um atalho para http://dbpedia.org/resource		
Documento	Termo	Resource*
1,14,33	Natasha Stott Despoja (3)	<host>/Natasha_Stott_Despoja
14	Aden Ridgeway, Ridgeway (2)	<host>//Aden_Ridgeway
1,14	Brian Greig, Greig (3)	<host>/Brian_Greig
14	Robert Mugabe, Mugabe (4)	<host>/Robert_Mugabe
5	Ari Fleischer, Fleischer (2)	<host>/Ari_Fleischer
5	Colin Powell	<host>/Colin_Powell
7,36	Osama bin Laden, bin Laden (3)	<host>/Osama_bin_Laden
8,21	Abu Nidal (5)	<host>/Abu_Nidal
3,11,12,16,37	George W. Bush, Bush (5)	<host>/George_W._Bush
8,12,14,16	Saddam Hussein, Saddam (5)	<host>/Saddam_Hussein
14	Andrew Bartlett, Bartlett (3)	<host>/Andrew_Bartlett
21	Yasser Arafat, Arafat (2)	<host>/Yasser_Arafat
22	Mal Brough	<host>/Mal_Brough
23	John Ashcroft	<host>/John_Ashcroft
25	Olusegun Obasanjo, Obasanjo (2)	<host>/Olusegun_Obasanjo
29	Jayalath Jayawardena	<host>/Jayalath_Jayawardena
30	Gary Ellis, Ellis (2)	<host>/Gary_Ellis
32	Joe Hockey, Hockey (2)	<host>/Joe_Hockey
35	Prince William, William (2)	<host>/Prince_William,_Duke_of_Cambridge
35	Prince Harry	<host>/Prince_Harry
42	Thabo Mvuyelwa Mbeki	<host>/Thabo_Mbeki
22	Donald Henry Rumsfeld	<host>/Donald_Rumsfeld
43	John Paul II, John Paul (2)	<host>/Pope_John_Paul_II
50	Bob Brown (2)	<host>/Bob_Brown
50	Kerry Nettle	<host>/Kerry_Nettle
26	Amina Lawal, Lawal (2)	<host>/Amina_Lawal
3	Walter Kansteiner	<host>/Walter_H._Kansteiner,_III
45	Arne Rinnan	<host>/Arne_Rinnan
15	Tony Keenan	<host>/Tony_Keenan
8	Taher Jalil Habbush	<host>/Tahir_Jalil_Habbush_al-Tikriti
45	Ove Thorsheim	<host>/Ove_Thorsheim
27	David Hudak	<host>/Dăvid_Hudăk
2	Paul Batchelor	<host>/Paul_Batchelor
33	Andrew Murray	<host>/Andrew_Thomas_(Australian_politician)
11	Abbas Khalaf	<host>/Abbas_Khalaf

Tabela 5.2: Anotações Padrão de Prata $\mathcal{SA}^{j+1}$.

* <host> é um atalho para http://yago-knowledge.org/resource		
Documento	Termo	Resource*
1,14,33	Natasha Stott Despoja (3)	<host>/Natasha_Stott_Despoja
14	Aden Ridgeway, Ridgeway (2)	<host>//Aden_Ridgeway
1,14	Brian Greig, Greig (3)	<host>/Brian_Greig
14	Robert Mugabe, Mugabe (4)	<host>/Robert_Mugabe
5	Ari Fleischer, Fleischer (2)	<host>/Ari_Fleischer
5	Colin Powell	<host>/Colin_Powell
7,36	Osama bin Laden, bin Laden (3)	<host>/Osama_bin_Laden
8,21	Abu Nidal (5)	<host>/Abu_Nidal
3,11,12,16,37	George W. Bush, Bush (5)	<host>/George_W._Bush

Continuação da Tabela 5.2		
Documento	Label	Resource*
8,12,14,16	Saddam Hussein, Saddam (5)	<host>/Saddam_Hussein
14	Andrew Bartlett, Bartlett (3)	<host>/Andrew_Bartlett
21	Yasser Arafat, Arafat (2)	<host>/Yasser_Arafat
22	Mal Brough	<host>/Mal_Broug
23	John Ashcroft	<host>/John_Ashcroft
25	Olusegun Obasanjo, Obasanjo (2)	<host>/Olusegun_Obasanjo
29	Jayalath Jayawardena	<host>/Jayalath_Jayawardena
30	Gary Ellis, Ellis (2)	<host>/Gary_Ellis
32	Joe Hockey, Hockey (2)	<host>/Joe_Hockey
35	Prince William, William (2x)	<host>/Prince_William,_Duke_of_Cambridge
35	Prince Harry	<host>/Prince_Harry,_Duke_of_Sussex
42	Thabo Mvuyelwa Mbeki	<host>/Thabo_Mbeki
22	Donald Henry Rumsfeld	<host>/Donald_Rumsfeld
43	John Paul II, John Paul (2)	<host>/Pope_John_Paul_II
50	Bob Brown (2)	<host>Bob_Brown
50	Kerry Nettle	<host>/Kerry_Nettle
26	Amina Lawal, Lawal (2)	<host>/Amina_Lawal
3	Walter Kansteiner	<host>/Walter_H._Kansteiner,_III
45	Arne Rinnan	<host>/Arne_Rinnan
* <host> é um atalho para http://dbpedia.org/resource		
15	Tony Keenan	<host>/Tony_Keenan
8	Taher Jalil Habbush	<host>/Tahir_Jalil_Habbush_al-Tikriti
45	Ove Thorsheim	<host>/Ove_Thorsheim
27	David Hudak	<host>/DÃavid_HudÃjk
2	Paul Batchelor	<host>/Paul_Batchelor
33	Andrew Murray	<host>/Andrew_Thomas_(Australian_politician)
11	Abbas Khalaf	<host>/Abbas_Khalaf

A Seção 5.3 apresenta em detalhes os resultados obtidos nas etapas 2 e 3 da avaliação realizada.

5.3 Ilustrando uma Ocorrência de Anotação Afetada

Esta Seção apresenta um exemplo da ocorrência de um conjunto de anotações inconsistentes. Descrevemos brevemente a primeira e a segunda etapa de funcionamento do nosso *framework* e nos concentramos na terceira (*cf.* Figura 4.1). Os conjuntos de dados utilizados aqui são aqueles descritos na Seção 5.1 e na Subseção 5.2.1. A Figura 5.9 descreve a anotação e o repositório usados como exemplo. $\mathcal{R}^j$ e $\mathcal{R}^{j+1}$ representam a *DBpedia* nas versões de Abril de 2015 e 2016, respectivamente. A anotação pertence ao conjunto $\mathcal{A}^j$.

Selecionamos uma anotação relacionada ao termo “George W. Bush” que foi vinculado pelo predicado “*sameAs*” a sua definição semântica (recurso “George W. Bush”) formalmente codificado na *DBpedia* no tempo j . Uma nova versão $j+1$ da *DBpedia* foi lançada, e algum mantenedor removeu o recurso “George W. Bush”. A anotação anteriormente estabelecida (“*sameAs*”) tornou-se inconsistente porque o recurso (“George W. Bush”) não pode ser mais encontrado no repositório de origem.

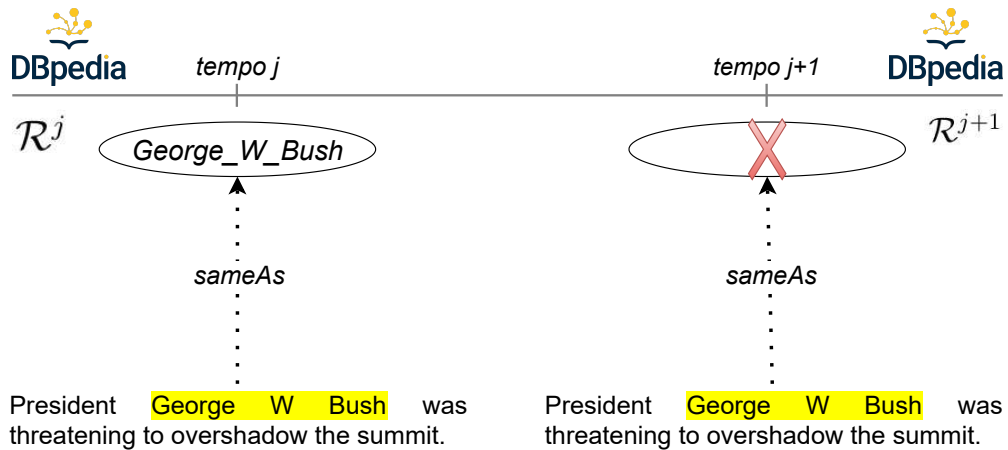


Figura 5.9: Exemplo de Anotação Inconsistente.

5.3.1 Mudanças detectadas nos Repositórios RDF

Determinado o contexto apresentado na Figura 5.9, ilustramos como o AnnoLOD lida com essa ocorrência de anotação inconsistente. Primeiramente, o *framework* recebeu como entrada os conjuntos de dados $\mathcal{R}^j$ e $\mathcal{R}^{j+1}$ e computou todas as mudanças ocorridas entre as versões dos repositórios da *DBpedia* nos tempos j e $j + 1$ (Δ). Dentre as mudanças detectadas, o *framework* encontra a remoção de várias triplas, incluindo a tripla com recurso “George W. Bush”. O *framework* processou todas as alterações e gerou uma lista de mudanças a serem utilizadas nas próximas etapas. A Tabela 5.3 apresenta o conjunto de recursos que foram removidos do repositório da *DBpedia* na sua versão $\mathcal{R}^{j+1}$. Por questão de inteligibilidade, descrevemos apenas as modificações de triplas que estavam envolvidas nas anotações do conjunto $\mathcal{A}^j$.

Tabela 5.3: Lista de recursos excluídos.

* <host> é um atalho para http://dbpedia.org/resource		
ID	Label	Resource*
1	Natasha Stott Despoja	<host>/Natasha_Stott_Despoja
2	Aden Ridgeway	<host>/Aden_Ridgeway
3	Brian Greig	<host>/Brian_Greig
4	Robert Gabriel Mugabe	<host>/Robert_Mugabe
5	Ari Fleischer	<host>/Ari_Fleischer
6	Colin Powell	<host>/Colin_Powell
7	Osama bin Laden	<host>/Osama_bin_Laden
8	Abu Nidal	<host>/Abu_Nidal
9	George Walker Bush	<host>/George_W._Bush
10	Saddam Hussein	<host>/Saddam_Hussein
11	Andrew Bartlett	<host>/Andrew_Bartlett
12	Yasser Arafat	<host>/Yasser_Arafat
13	Malcolm “Mal” Thomas Brough	<host>/Mal_Brough
14	John Ashcroft	<host>/John_Ashcroft
15	Olusegun Obasanjo	<host>/Olusegun_Obasanjo
16	Jayalath Jayawardena	<host>/Jayalath_Jayawardena
17	Gary Ellis	<host>/Gary_Ellis
18	Joe Hockey	<host>/Joe_Hockey
19	Duke of Cambridge Prince William	<host>/Prince_William,_Duke_of_Cambridge

Continuação da Tabela 5.3		
ID	Label	Resource*
20	Prince Harry	<host>/Prince_Harry
21	Thabo Mvuyelwa Mbeki	<host>/Thabo_Mbeki
22	Donald Henry Rumsfeld	<host>/Donald_Rumsfeld
23	Pope Saint John Paul II	<host>/Pope_John_Paul_II
24	Robert James Brown"	<host>/Bob_Brown
25	Kerry Nettle	<host>/Kerry_Nettle
26	Amina Lawal	<host>/Amina_Lawal
27	Walter H. III Kansteiner	<host>/Walter_H._Kansteiner,_III
28	Arne Rinnan	<host>/Arne_Rinnan

5.3.2 Anotações Inconsistentes Identificadas

A segunda etapa categorizou todas as alterações descobertas na etapa anterior como possíveis fontes de anotações inconsistentes (lista de anotações afetadas) ou alterações sem efeitos nas anotações existentes (lista de anotações não afetadas). No exemplo (*cf.* Figura 5.9), a exclusão do recurso “George W. Bush” criou uma inconsistência na anotação. Com base nisso, essa anotação foi incluída na lista de anotações afetadas. Para a identificação de anotações inconsistentes, o *framework* verifica se um determinado recurso (sujeito) de uma tripla pertencente ao conjunto das triplas modificadas está vinculado a alguma anotação. Isso sendo verdade, essa anotação é considerada afetada e adicionada ao conjunto de anotações inconsistentes. As linhas 4 e 5 do Algoritmo 1 apresentam a criação de ambas as listas de anotações afetadas e não afetadas, com base nas alterações encontradas no Passo A (*cf.* Subseção 5.3.1).

A partir da lista de modificações geradas na etapa anterior (*cf.* Tabela 5.3), o *framework* indentificou um total de 60 anotações afetadas das 67 que formam o conjunto $\mathcal{A}^j$. A Tabela 5.4 apresenta as anotações afetadas. Sobre essa lista aplicamos ações de adaptação de anotações (*cf.* Subseção 5.3.3).

Tabela 5.4: Lista de anotações afetadas pelas mudanças $\mathcal{A}^{aff}$.

* <host> é um atalho para http://dbpedia.org/resource		
Documento	Termo	Resource*
1,14,33	Natasha Stott Despoja (3)	<host>/Natasha_Stott_Despoja
14	Aden Ridgeway, Ridgeway (2)	<host>/Aden_Ridgeway
1,14	Brian Greig, Greig (3)	<host>/Brian_Greig
14	Robert Mugabe, Mugabe (4)	<host>/Robert_Mugabe
5	Ari Fleischer, Fleischer (2)	<host>/Ari_Fleischer
5	Colin Powell	<host>/Colin_Powell
7,36	Osama bin Laden, bin Laden (3)	<host>/Osama_bin_Laden
8,21	Abu Nidal (5)	<host>/Abu_Nidal
3,11,12,16,37	George W. Bush, Bush (5)	<host>/George_W._Bush
8,12,14,16	Saddam Hussein, Saddam (5)	<host>/Saddam_Hussein
14	Andrew Bartlett, Bartlett (3)	<host>/Andrew_Bartlett
21	Yasser Arafat, Arafat (2)	<host>/Yasser_Arafat
22	Mal Brough	<host>/Mal_Brough
23	John Ashcroft	<host>/John_Ashcroft
25	Olusegun Obasanjo, Obasanjo (2)	<host>/Olusegun_Obasanjo
29	Jayalath Jayawardena	<host>/Jayalath_Jayawardena

Continuação da Tabela 5.4		
Documento	Label	Resource*
30	Gary Ellis, Ellis (2)	<host>/Gary_Ellis
32	Joe Hockey, Hockey (2)	<host>/Joe_Hockey
35	Prince William, William (2)	<host>/Prince_William,_Duke_of_Cambridge
35	Prince Harry	<host>Prince_Harry
42	Thabo Mvuyelwa Mbeki	<host>/Thabo_Mbeki
22	Donald Henry Rumsfeld	<host>/Donald_Rumsfeld
43	John Paul II, John Paul (2)	<host>/Pope_John_Paul_II
50	Bob Brown (2)	<host>Bob_Brown
50	Kerry Nettle	<host>/Kerry_Nettle
26	Amina Lawal, Lawal (2)	<host>/Amina_Lawal
3	Walter Kansteiner	<host>/Walter_H._Kansteiner,_III
45	Arne Rinnan	<host>/Arne_Rinnan

Tabela 5.5: Lista de anotações não afetadas pelas mudanças $\mathcal{A}^{unaff}$.

* <host> é um atalho para http://dbpedia.org/resource		
Documento	Termo	Resource*
15	Tony Keenan	<host>/Tony_Keenan
8	Taher Jalil Habbush	<host>/Tahir_Jalil_Habbush_al-Tikriti
45	Ove Thorsheim	<host>/Ove_Thorsheim
27	David Hudak	<host>/DÃ¡vid_HudÃ¡k
2	Paul Batchelor	<host>/Paul_Batchelor
33	Andrew Murray	<host>/Andrew_Thomas_(Australian_politician)
11	Abbas Khalaf	<host>/Abbas_Khalaf

5.3.3 Aplicação das Ações de Adaptação de Anotação

A última etapa recebeu como entrada a lista de anotações afetadas (*cf.* Tabela 5.4) e o *framework* apresentou uma ou mais ações a serem realizadas. Por exemplo, se na *DBpedia* ou *YAGO* existir um recurso semanticamente similar a “George W. Bush”, a reconexão deve ser realizada.

A primeira etapa do procedimento de adaptação foi obter os recursos semanticamente relacionados à menção que deu origem a cada anotação listada na Tabela 5.4. Para esse fim, o Algoritmo 2 pesquisou nas bases *DBpedia* ($\mathcal{R}^{j+1}$) e *YAGO* ($\mathcal{R}_{ext}$) os recursos que contêm essa menção associada aos seus *labels* (propriedade *rdfs:label* ou *foaf:name*). Isso gerou uma lista de candidatos (*candList*) que podem potencialmente substituir os recursos que afetaram as anotações ($\mathcal{A}^{aff}$). A Figura 5.10 apresenta um exemplo de lista de recursos e seus respectivos *labels* associados. Essa lista refere-se especificamente aos recursos que podem substituir o recurso “George W. Bush” (recursos candidatos).

As comparações realizadas (C1 e C2 da Figura 4.6) geraram os seguintes resultados:

- **Valores de similaridade em comparação com os *labels*.** Cada *label* da Tabela 4.7 é comparada semanticamente com a *label* “George W. Bush”;
- **Valores de similaridade relativos às propriedades de entrada do usuário.** Nesta avaliação não houve a necessidade de usar propriedades fornecidas pelo usuá-

	sub	pred	label
1	http://yago-knowledge.org/resource/Anthony_Bushell	rdfs:label	"Anthony Bushell"@en
2	http://yago-knowledge.org/resource/Bushra_al-Assad	rdfs:label	"Bushra al-Assad"@en
3	http://yago-knowledge.org/resource/Bushra_Bibi	rdfs:label	"Bushra Bibi"@en
4	http://yago-knowledge.org/resource/Barbara_Bush	rdfs:label	"Barbara Bush"@en
5	http://yago-knowledge.org/resource/Claudia_Lauper_Bushman	rdfs:label	"Claudia Lauper Bushman"@en
6	http://yago-knowledge.org/resource/Dorothy_Bush_Koch	rdfs:label	"Dorothy Bush Koch"@en
7	http://yago-knowledge.org/resource/Dorothy_Walker_Bush	rdfs:label	"Dorothy Walker Bush"@en
8	http://yago-knowledge.org/resource/Columba_Bush	rdfs:label	"Columba Bush"@en
9	http://yago-knowledge.org/resource/George_H._W._Bush	rdfs:label	"George H. W. Bush"@en
10	http://yago-knowledge.org/resource/Grand_L._Bush	rdfs:label	"Grand L. Bush"@en
11	http://yago-knowledge.org/resource/George_W._Bush	rdfs:label	"George W. Bush"@en
12	http://yago-knowledge.org/resource/Jeb_Bush	rdfs:label	"Jeb Bush"@en
13	http://yago-knowledge.org/resource/Laura_Bush	rdfs:label	"Laura Bush"@en
14	http://yago-knowledge.org/resource/Lauren_Bush	rdfs:label	"Lauren Bush"@en
15	http://yago-knowledge.org/resource/Neil_Bush	rdfs:label	"Neil Bush"@en
16	http://yago-knowledge.org/resource/Prescott_Bush	rdfs:label	"Prescott Bush"@en
17	http://yago-knowledge.org/resource/Pauline_Bush_(actress)	rdfs:label	"Pauline Bush"@en
18	http://yago-knowledge.org/resource/Richard_Bushman	rdfs:label	"Richard Bushman"@en
19	http://yago-knowledge.org/resource/Sophia_Bush	rdfs:label	"Sophia Bush"@en
20	http://yago-knowledge.org/resource/Sarah_Bush_Lincoln	rdfs:label	"Sarah Bush Lincoln"@en

Figura 5.10: Lista de recursos candidatos encontrados para substituir o recurso “George W. Bush”.

rio, uma vez que, todos os repositórios utilizados possuem as propriedades requerida na comparação C1.

De forma similar, foram geradas as listas de recursos candidatos e realizadas as comparações para todas as anotações identificadas como afetadas ($\mathcal{A}^{aff}$) na etapa anterior (cf. Tabela 5.4). Por questão de inteligibilidade, a Tabela 5.6 apresenta apenas os melhores recursos candidatos (linha 8 Algoritmo 3) para cada anotação afetada, ou seja, os recursos que obtiveram um valor de similaridade maior ou igual o limite de $\eta = 0,5$ (linha 7, Algoritmo 3) na comparação C1 (cf. Subseção 4.4.2). O objetivo dos n principais candidatos (*bestCand*) é definir quais recursos são mais similares aos recursos que afetaram as anotações. Com base nos melhores candidatos (cf. Tabela 5.6), o Algoritmo 3 formulou as ações a serem tomadas para adaptar as anotações em análise.

A Figura 5.11) descreve as ações de adaptação sugeridas pelo *framework* para atualizar as anotações inconsistentes encontradas. A primeira ação de adaptação (cf. A Figura 5.11) refere-se à substituição de objeto (reanotação), onde o recurso que afetou a anotação é substituído por um novo recurso semanticamente idêntico. A segunda ação de adaptação (cf. B Figura 5.11) diz respeito à exclusão de anotação, na qual a anotação é completamente removida quando não há pelo menos um recurso que possa substituir o recurso que afetou a anotação. Por fim, a ultima ação de adaptação (cf. C Figura 5.11) corresponde a não execução das ações anteriores. Neste caso, o usuário após analisar a anotação decide mantê-la no mesmo estado que ela se encontra.

Tabela 5.6: Lista dos melhores recursos candidatos *bestCand*.

ID	Label	C1 $\geq \eta$
1	Natasha Stott Despoja	0,5
2	Aden Ridgeway	0,8
3	Brian Greig	0,5
4	Robert Mugabe	0,7
5	Ari Fleischer	0,5
6	Colin Powell	0,8
7	Osama bin Laden	0,8
8	Abu Nidal	0,5
9	George W. Bush	0,8
10	Saddam Hussein	1
11	Andrew Bartlett	1
12	Yasser Arafat	0,8
13	Yasir Arafat	0,8
14	Mal Brough	-
15	John Ashcroft	0,8
16	Olusegun Obasanjo	0,5
17	Jayalath Jayawardena	0,5
18	Gary Ellis	0,8
19	Joe Hockey	0,8
20	Prince William Duke of Cambridge	-
21	Prince William of SchaumburgLippe	0,6
22	Prince William Henry Duke of Gloucester and Edinburgh	0,6
23	Prince William Frederick Duke of Gloucester and Edinburgh	0,6
24	Prince William of HesseKassel	0,6
25	Prince Harry Duke of Sussex	-

Continuação da Tabela 5.6		
ID	Label	C1 $\geq \eta$
26	Donald Rumsfeld	0,5
27	Pope John Paul II	0,7
28	Bob Brown	-
29	Kerry Nettle	0,8
30	Amina Lawal	0,5
31	Walter H Kansteiner III	0,8
32	Walter Humphreys Jr	0,6
33	Walter Hodgson Bevan Graham	0,6
34	Arne Rinnan	0,5

A Tabela 5.6 indica que para três recursos não foi possível determinar um valor de similaridade, identificados pelo (-) na terceira coluna da respectiva tabela. Embora eles fossem os recursos adequados para a substituição, que são os casos dos recursos identificados pelos *labels* “*Prince Harry Duke of Sussex*”, “*Prince William Duke of Cambridge*” e “*Bob Brown*”. Nesse contexto, as anotações dependentes foram completamente removidas, dado que os recursos adequados para atualizá-las não foram inclusos na lista dos melhores recursos candidatos.

Para substituir o recurso “Duke of Cambridge Prince William” da *DBpedia* (cf. Tabela 5.3) foram erradamente selecionados quatro recursos na base de dados *YAGO* (IDs 20-24 na Tabela 5.6), já que, todos eles se referem à pessoas diferentes no mundo real em relação ao “Prince William” em questão. Esse cenário também levou a remoção completa da anotação dependente, pois não havia na lista dos melhores recursos candidatos pelo menos um recurso que pudesse substituir o recurso “Duke of Cambridge Prince William”.

Em dois casos de anotação inconsistente, envolvendo os recursos “Yasser Arafat” e “Walter H. III Kansteiner” da *DBpedia*, ocorreu a seleção dos recursos corretos (IDs 12 e 31 na Tabela 5.6) no repositório *YAGO* acompanhados de outros recursos que não correspondem a “Yasser Arafat” e “Walter H. III Kansteiner” no mundo real, identificados pelos IDs 13, 32 e 33 (cf. Tabela 5.6). Nesse caso, as anotações dependentes foram atualizadas, já que havia na lista dos melhores recursos candidatos pelo menos um recurso que poderia substituir os recursos que afetaram as anotações.

Com os valores de similaridade computados, obteve-se o(s) candidato(s) mais adequado(s) (*bestCand*) para atualizar cada anotação inconsistente (Linha 8, Algoritmo 2). A Tabela 5.7 apresenta o conjunto atualizado das anotações obtido após a aplicação das ações de adaptação sugeridas pelo *framework*. O “X” indica as anotações que foram removidas. O conjunto das anotações atualizadas ($\mathcal{A}^{j+1}$) é composto pelas anotações que passaram pelo processo de adaptação ($\mathcal{A}^{aff}$) e pelas anotações que não foram afetadas pelas mudanças identificadas ($\mathcal{A}^{unaff}$).

Tabela 5.7: Lista de anotações atualizadas $\mathcal{A}^{j+1}$.

* <host> é um atalho para http://yago-knowledge.org/resource		
Documento	Termo	Resource*
1,14,33	Natasha Stott Despoja (3)	<host>/Natasha_Stott_Despoja
14	Aden Ridgeway, Ridgeway (2)	<host>/Aden_Ridgeway
1,14	Brian Greig, Greig (3)	<host>/Brian_Greig
14	Robert Mugabe, Mugabe (4)	<host>/Robert_Mugabe

Continuação da Tabela 5.7		
Documento	Label	Resources*
5	Ari Fleischer, Fleischer (2)	<host>/Ari_Fleischer
5	Colin Powell	<host>/Colin_Powell
7,36	Osama bin Laden, bin Laden (3)	<host>/Osama_bin_Laden
8,21	Abu Nidal (5)	<host>/Abu_Nidal
3,11,12,16,37	George W. Bush, Bush (5)	<host>/George_W._Bush
8,12,14,16	Saddam Hussein, Saddam (5)	<host>/Saddam_Hussein
14	Andrew Bartlett, Bartlett (3)	<host>/Andrew_Bartlett
21	Yasser Arafat, Arafat (2)	<host>/Yasser_Arafat (usuário)
22	Mal Brough	X
23	John Ashcroft	<host>/John_Ashcroft
25	Olusegun Obasanjo, Obasanjo (2)	<host>/Olusegun_Obasanjo
29	Jayalath Jayawardena	<host>/Jayalath_Jayawardena
30	Gary Ellis, Ellis (2)	<host>/Gary_Ellis
32	Joe Hockey, Hockey (2)	<host>/Joe_Hockey
35	Prince William, William (2x)	X
35	Prince Harry	X
42	Thabo Mvuyelwa Mbeki	<host>/Thabo_Mbeki
22	Donald Henry Rumsfeld	<host>/Donald_Rumsfeld
43	John Paul II, John Paul (2)	<host>/Pope_John_Paul_II
50	Bob Brown (2)	X
50	Kerry Nettle	<host>/Kerry_Nettle
26	Amina Lawal, Lawal (2)	<host>/Amina_Lawal
3	Walter Kansteiner	<host>/Walter_H._Kansteiner,_III (usuário)
45	Arne Rinnan	<host>/Arne_Rinnan
* <host> é um atalho para http://dbpedia.org/resource		
15	Tony Keenan	<host>/Tony_Keenan
8	Taher Jalil Habbush	<host>/Tahir_Jalil_Habbush_al-Tikriti
45	Ove Thorsheim	<host>/Ove_Thorsheim
27	David Hudak	<host>/DÃ vid_HudÃ k
2	Paul Batchelor	<host>/Paul_Batchelor
33	Andrew Murray	<host>/Andrew_Thomas_(Australian_politician)
11	Abbas Khalaf	<host>/Abbas_Khalaf

5.4 Análise dos Resultados

Após a execução do Passo B pelo *framework* AnnoLOD (*cf.* Subseção 5.3.2) sobre o conjunto de anotações $\mathcal{A}^j$ foram identificadas um total de 60 anotações afetadas ($\mathcal{A}^{aff}$) pelas mudanças detectadas no Passo A entre as versões $\mathcal{R}^j$ e $\mathcal{R}^{j+1}$ da base de dados analisada (*cf.* Subseção 5.3.1), e um grupo de 7 anotações não afetadas ($\mathcal{A}^{unaff}$). No Passo C (*cf.* Subseção 5.3.3) foram aplicadas as ações de adaptação para atualizar as anotações $\mathcal{A}^{aff}$ em relação a nova versão do conjunto de dados ($\mathcal{R}^{j+1}$) ou acerca do repositório externo ($\mathcal{R}_{ext}$).

O AnnoLOD sugeriu 54 reconexão de objeto (reanotação - linha 12, Algoritmo 2 e Caso A, Figura 5.11). Essa ação foi apresentada porque havia pelo menos um recurso candidato que atingiu o valor de similaridade igual ou acima do limite η estabelecido (0,5).

Complementarmente, o AnnoLOD sugestionou 6 remoções (Linha 14, Algoritmo 2

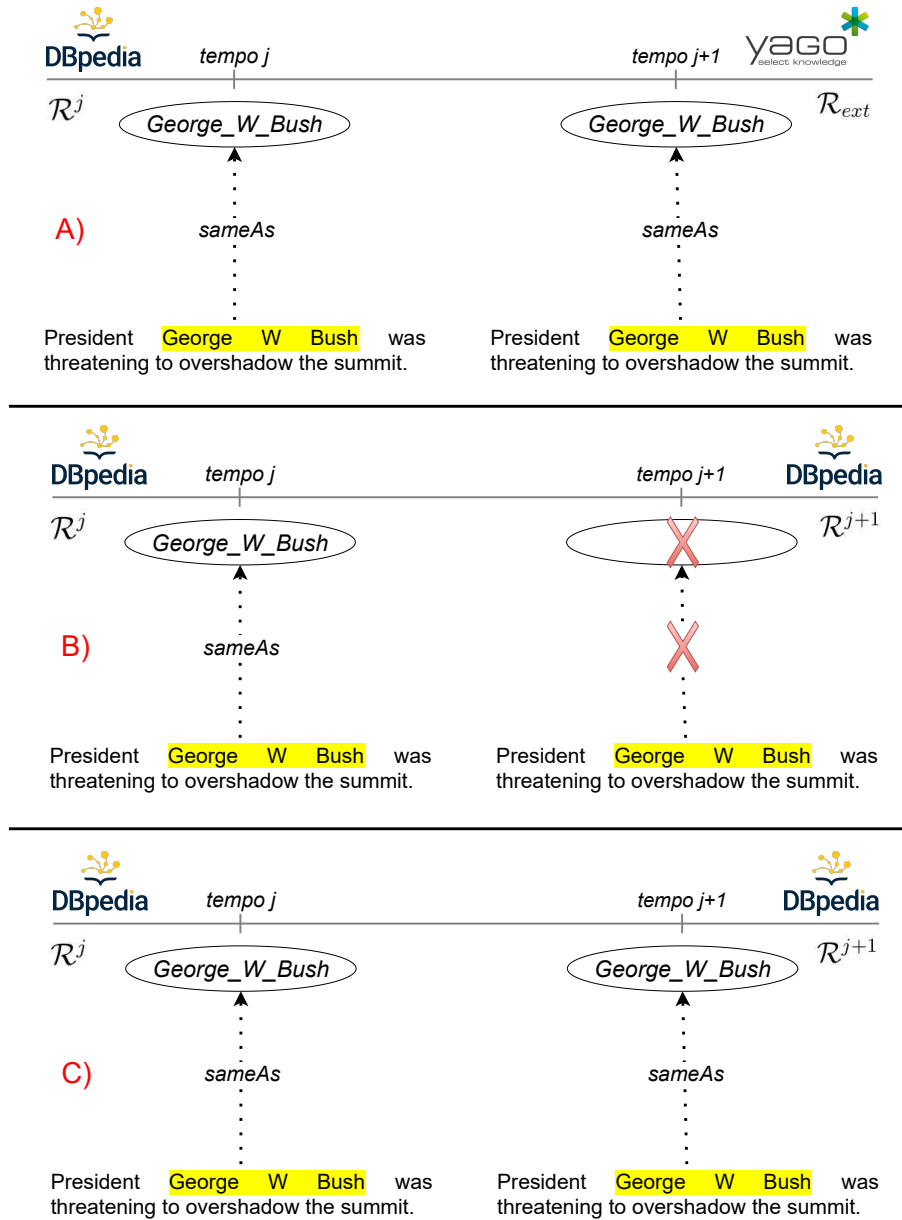


Figura 5.11: Exemplo de Ações de Adaptação. A) substituição de objeto; B) remoção da anotação; C) não realizar ação de adaptação.

e Caso B, Figura 5.11). Esta ação foi apresentada porque não houve candidatos que obtivessem o valor de similaridade igual ou maior que o limite η , ou após a análise dos recursos candidatos verificou-se que eles não correspondiam a mesma pessoa no mundo real em relação ao antigo recurso analisado.

Suplementarmente, identificamos que houve um caso em que o Algoritmo 1 sugeriu uma reconexão de objeto de forma errada, relacionado ao recurso “Prince William Duke of Cambridge” que foi removido da *DBpedia*. Nesse caso, foram apresentados 4 recursos com valores de similaridade igual 0,6 que poderiam substituir o recurso removido. Contudo, os candidatos sugeridos não se referiam à mesma pessoa do mundo real, o que levou à remoção total da anotação.

Ressaltamos que não houve categorização errada de anotação pelo *framework* AnnoLOD, ou seja, não encontramos casos de anotações estáveis classificadas como afetadas (falso positivo) ou anotações afetadas que não foram identificadas (falso negativo) na avaliação realizada.

Com os resultados obtidos, observamos que o método de adaptação de anotação adotado pelo *framework* AnnoLOD utilizando algoritmos de similaridade semântica alcançou bons resultados quando comparamos o conjunto de anotações atualizadas $\mathcal{A}^{j+1}$ ao conjunto de anotações padrão de prata $\mathcal{SA}^{j+1}$ (cf. Tabelas 5.7 e 5.2). Contudo, ocorreram casos onde os algoritmos de similaridade não conseguiram determinar um valor de similaridade aos recursos adequados para substituir os recursos de algumas anotações afetada, mesmo quando eles estavam presentes na lista de recursos candidatos (cf. Tabela 5.6), resultando na remoção de várias anotações (cf. Tabela 5.7). As comparação entre os conjuntos $\mathcal{A}^{j+1}$ e $\mathcal{SA}^{j+1}$ é realizada confrontando a primeira anotação do conjunto $\mathcal{A}^{j+1}$ com a primeira anotação do conjunto $\mathcal{SA}^{j+1}$, e assim sucessivamente até alcançar a última anotação de cada conjunto.

Nosso estudo forneceu um exemplo ilustrativo que descreve completamente a execução do *framework* AnnoLOD para destacar as potencialidades e desafios de nossa solução para adaptar anotações semanticamente inconsistentes. Trabalhos futuros envolvem outras avaliações experimentais para analisar melhor a execução do *framework*.

5.5 Discussão

Lidar com os impactos da evolução de bases RDF sobre anotações dependentes é essencial para manter a consistência e os benefícios desse artefato semântico ao longo do tempo. Isso é especialmente valioso devido ao potencial grande volume de anotações geradas por meio de ferramentas computacionais que vinculam automaticamente termos contidos em documentos Web às suas definições semânticas presentes nos conjuntos de dados LOD.

Este capítulo descreveu a avaliação do *framework* AnnoLOD que pode rastrear e adaptar casos de anotações inconsistentes semanticamente. Concebemos e desenvolvemos uma metodologia que pode ser usada para melhorar a qualidade das anotações previamente existentes. Em particular, avaliamos casos reais de anotações inconsistentes identificadas por nosso *framework*. Consideramos o exemplo do ex-presidente “George W. Bush” como uma fonte completa para descrever os procedimentos concebidos e implementados

no *framework* como um todo. Primeiro, o *framework* detectou que o recurso “George W. Bush” foi removido entre as versões do conjunto de dados RDF. Em seguida, o *framework* identificou que a mudança afetou negativamente a semântica da anotação dependente. Posteriormente, a solução buscou recursos que pudessem substituir o recurso removido (objeto e definição semântica da anotação) e considerar outras alternativas além da remoção completa da anotação. Na metodologia do *framework*, as ações recomendadas são apresentadas ao usuário, que escolhe a melhor ação e efetua as alterações. Com base nisso, o *framework* gera a anotação final atualizada.

Optamos por não deixar o Algoritmo 1 totalmente automático em sua última etapa (ações de adaptação) porque entendemos que a decisão final do especialista e a curadoria são relevantes para decidir a consistência do conjunto de anotações em análise. Mesmo que nossa solução encontre anotações inconsistentes e apresente soluções possíveis, cabe ao mantenedor das anotações a decisão final.

O *framework* visa auxiliar o mantenedor humano na tarefa de manter a consistência das anotações. Se o mantenedor tivesse que encontrar essas anotações, ele enfrentaria uma tarefa muito árdua e demorada. O mesmo se aplica a encontrar o melhor candidato para substituição. Ao restringir o número de candidatos possíveis usando nossas comparações propostas, nossa solução reduz e otimiza a escolha apenas para recursos candidatos e com maior potencial de serem consistentes; e não para qualquer recurso em todo o conjunto de dados. Demonstramos a efetividade do AnnoLOD considerando grandes conjuntos de dados que mudam constantemente.

Para trabalhos futuros em termos experimentais, a análise de dados textuais (contendo mais de um *label* e palavras) pode ser uma fonte de dados a serem usados como um novo método de comparação (similaridade). Propriedades como *rdfs:comments*, *dbo:abstract*, *dct:description*, contendo informações sobre o recurso podem ser úteis para ajudar a encontrar similaridade entre recursos, complementando as comparações realizadas pelo AnnoLOD.

5.6 Conclusão

Neste capítulo realizamos análises experimentais completas com conjuntos de dados do mundo real para verificar a efetividade do *framework* AnnoLOD na detecção e adaptação de anotações inconsistentes. O próximo capítulo efetua o fechamento desta dissertação.

Capítulo 6

Conclusão

Bases de dados abertas na Web que não são atualizadas corretamente de acordo com as alterações de seu domínio estão sujeitas a se tornarem obsoletas. Os sistemas que consultam conjuntos de dados desatualizados podem encontrar dados incorretos. A atualização e manutenção semi-automáticas de conjuntos de dados RDF são de alta relevância porque o tempo necessário para sua avaliação manual tende a ser maior do que o uso de ferramentas semi-automáticas. No entanto, as alterações em triplas RDF impactam artefatos dependentes, como as anotações consistentes preexistentes.

Entendemos que o estudo e desenvolvimento de métodos para depuração de anotações, verificação de consistência e adaptação são essenciais para preservar a qualidade desse artefato semântico crucial para sistemas computacionais que fazem uso do significado explícito compartilhado pela máquina. O valor real dos sistemas computacionais semanticamente enriquecidos se fundamenta principalmente na confiabilidade das anotações semânticas.

O problema de anotações inconsistentes foi abordado pela adaptação de anotações como o objetivo principal desta pesquisa. Esta dissertação estudou como manter anotações baseadas em RDF atualizadas de acordo com a evolução dos repositórios RDF. Esta dissertação propôs um *framework* para adaptação (semi-)automática de anotações semânticas afetadas pela evolução de dados RDF. Nosso algoritmo de adaptação definido funciona com base em operações de mudança automaticamente identificadas na evolução de conjuntos de dados RDF. Investigamos as ações de adaptação, sua formalização e aplicabilidade. Posteriormente, implementamos uma ferramenta de software para a adaptação assistida de anotações semânticas sobre bases RDF.

No melhor do nosso conhecimento, a nossa solução é original quando comparada com a literatura existente. Derivamos uma metodologia que lida com anotações inconsistentes com base na evolução do conjunto de dados de origem. Nossa originalidade reside em uma estratégia de adaptação sobre as anotações desatualizadas de forma semi-automática no nível de instância dos conceitos, em vez da estratégia de detecção e manutenção apenas no nível de ontologia adotada por abordagens encontradas no estado da arte.

6.1 Contribuições

Consideramos que esta pesquisa de mestrado alcançou resultados significativos para o estado da arte. Nossa principal contribuição consiste em um *framework* completo que dado como entrada uma nova versão de um conjunto de dados RDF, o *framework* atualiza as anotações dependentes. Nossos resultados obtidos são úteis para detectar e adaptar anotações desatualizadas. De forma resumida, sumarizamos as principais contribuições deste trabalho da seguinte forma:

1. Estudo, concepção, implementação e avaliação de um *framework* para adaptação de anotações.
2. Um método semi-automático que aplica ações de adaptação em anotações desatualizadas, sugerindo ações a serem tomadas e recursos candidatos que podem ser escolhidos para substituir partes inconsistentes de anotações afetadas por alterações RDF.
3. Todas as bibliotecas, algoritmos e dados de pesquisa produzidos nesta investigação estão disponíveis *online*¹.

Complementarmente a contribuição científica obtida nesta dissertação, ressaltamos a contribuição tecnológica proporcionada por nossa ferramenta de software de adaptação de anotações. Isso pode ser usado por especialistas de domínio e mantenedores de conjunto de dados para verificar como as alterações RDF realizadas entre as versões de lançamento afetam as anotações dependentes.

6.2 Disseminação dos Resultados

Os resultados de pesquisa obtidos levaram a publicações científicas no período da dissertação. Parte do material apresentado nesta dissertação apareceu em artigos de conferências publicados no decorrer desta dissertação. A seguir, listamos os artigos publicados resultantes desta dissertação.

6.2.1 Artigos em Conferências

- MONTEIRO, E. D. J. P., and DOS REIS, J. C. 2020. *A Framework for the Automatic Adaptation of RDF-based Semantic Annotations*. Em Managing the Evolution and Preservation of the Data Web (MEPDaW) co-located with 19th International Semantic Web Conference (ISWC). Evento Virtual, Vol. 2821 of CEUR Workshop Proceedings, pp. 59-63, 2020;
- REGINO, A. G., MONTEIRO, E. J. P., dos SANTOS, A. C; DOS REIS, J. C. 2021. *UpLOD: A Tool for Inconsistent Links Repairment in the LOD*. Em Managing the Evolution and Preservation of the Data Web (MEPDaW) co-located with 20th International Semantic Web Conference (ISWC).

¹<https://gitlab.ic.unicamp.br/jreis/evoLOD-analysis>

6.2.2 Participação em Eventos

- Artigo: *A Framework for the Automatic Adaptation of RDF-based Semantic Annotations*. Em Workshop Managing the Evolution and Preservation of the Data Web (MEPDaW) co-located with 19th International Semantic Web Conference (ISWC). apresentação virtual do artigo nesta conferência internacional.

6.3 Trabalhos Futuros

Esta seção fornece perspectivas sobre extensões futuras desta pesquisa.

- Analisar as propriedades textuais. O AnnoLOD realiza várias comparações entre os elementos de dados, incluindo *rdfs:label* e *rdf:name*, cujos valores representam outros recursos ou palavras curtas. Seria interessante uma pesquisa para incluir outras comparações envolvendo textos contendo parágrafos em vez de algumas palavras ou recursos. As propriedades a serem comparadas podem ser *rdfs:comment*, *dbo:abstract* e *dct:description*, por exemplo. O desafio aqui é a descoberta de um método de similaridade que tome como entrada dois textos longos e produza o grau de relacionamento entre eles. Isso poderia melhorar a precisão obtida pelo AnnoLOD. Ambos os algoritmos de detecção e correção de anotações inconsistentes distinguem entre recursos homônimos (por exemplo, “Washington” pode se referir a uma cidade e a uma pessoa) é um exemplo de benefício adquirido pela exploração de propriedades textuais.
- Explorar algoritmos de aprendizado de máquina para produzir limites de similaridade melhores. Na terceira etapa do AnnoLOD, o algoritmo escolhe candidatos a recursos com base em um limite η . Apenas candidatos com valores de similaridade maiores que η são mostrados ao usuário. Uma extensão deste trabalho poderia criar e aplicar um modelo de aprendizado de máquina que avalia anotações e gera um valor limite mais refinado de acordo com conjuntos de dados específicos em análise. Valores mais altos de limite reduzem o número de recursos candidatos e aumentam o valor de similaridade semântica necessário para os recursos comparados. Valores menores aumentam o número de recursos candidatos, mas também aumentam o número de recursos não relacionados. O refinamento do limiar seria valioso para atingir o valor que mais respeita a compensação entre o número de recursos e a qualidade.
- Averiguar mais medidas de similaridades semânticas. Explorar tesouros, ontologias e base de conhecimento de domínio específico pode melhorar os resultados obtidos no cálculo de similaridade. Essa análise minimizaria a geração de casos de falsos positivos (anotações estáveis que são considerados afetadas pelo algoritmo) e falsos negativos (anotações afetadas que o algoritmo não conseguiu detectar).

Referências Bibliográficas

- [1] Grigoris Antoniou and Frank Van Harmelen. Web ontology language: Owl. In *Handbook on ontologies*, pages 67–92. Springer, 2004.
- [2] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. Dbpedia: A nucleus for a web of open data. In *The semantic web*, pages 722–735. Springer, 2007.
- [3] Tim Berners-Lee, James Hendler, and Ora Lassila. The semantic web. *Scientific american*, 284(5):34–43, 2001.
- [4] Garth R Brown, Vichet Hem, Kenneth S Katz, Michael Ovetsky, Craig Wallin, Olga Ermolaeva, Igor Tolstoy, Tatiana Tatusova, Kim D Pruitt, Donna R Maglott, et al. Gene: a gene-centered information resource at ncbi. *Nucleic acids research*, 43(D1):D36–D42, 2015.
- [5] José Camacho-Collados, Mohammad Taher Pilehvar, and Roberto Navigli. Nasari: a novel approach to a semantically-aware representation of items. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 567–577, 2015.
- [6] Silvio Cardoso, Chantal Reynaud-Delaître, Marcos Da Silveira, Ying-Chi Lin, Anika Gross, Erhard Rahm, and Cédric Pruski. Evolving semantic annotations through multiple versions of controlled medical terminologies. *Health and Technology*, 8(5):361–376, 2018.
- [7] Silvio Domingos Cardoso, Reynaud-Delaître Chantal, Marcos Da Silveira, and Cédric Pruski. Combining rules, background knowledge and change patterns to maintain semantic annotations. In *AMIA Annual Symposium Proceedings*, volume 2017, page 505. American Medical Informatics Association, 2017.
- [8] Silvio Domingos Cardoso, Cédric Pruski, Marcos Da Silveira, Ying-Chi Lin, Anika Groß, Erhard Rahm, and Chantal Reynaud-Delaître. Leveraging the impact of ontology evolution on semantic annotations. In *European Knowledge Acquisition Workshop*, pages 68–82. Springer, 2016.
- [9] Victor Christen, Ying-Chi Lin, Anika Groß, Silvio Domingos Cardoso, Cédric Pruski, Marcos Da Silveira, and Erhard Rahm. A learning-based approach to combine medical annotation results. In *International Conference on Data Integration in the Life Sciences*, pages 135–143. Springer, 2018.

- [10] Philipp Cimiano and Johanna Völker. A framework for ontology learning and data-driven change discovery. In *Proceedings of the 10th International Conference on Applications of Natural Language to Information Systems (NLDB)*, pages 227–238. Springer, 2005.
- [11] Gene Ontology Consortium. The gene ontology project in 2008. *Nucleic acids research*, 36(suppl_1):D440–D444, 2008.
- [12] H. CUNNINGHAM. Gate : A framework and graphical development environment for robust nlp tools and applications. *Proc. 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002)*, pages 168–175, 2002.
- [13] Jérôme Euzenat, Pavel Shvaiko, et al. *Ontology matching*, volume 18. Springer, 2007.
- [14] Sidra Faisal, Kemele M Endris, Saeedeh Shekarpour, Sören Auer, and Maria-Esther Vidal. Co-evolution of rdf datasets. In *International Conference on Web Engineering*, pages 225–243. Springer, 2016.
- [15] Ronald Aylmer Fisher. Statistical methods for research workers. In *Breakthroughs in statistics*, pages 66–70. Springer, 1992.
- [16] Anika Gross, Michael Hartung, Toralf Kirsten, and Erhard Rahm. Estimating the quality of ontology-based annotations by considering evolutionary changes. In *International Workshop on Data Integration in the Life Sciences*, pages 71–87. Springer, 2009.
- [17] Thomas R Gruber. Toward principles for the design of ontologies used for knowledge sharing? *International journal of human-computer studies*, 43(5-6):907–928, 1995.
- [18] Michael Hartung, Anika Groß, and Erhard Rahm. Conto-diff: generation of complex evolution mappings for life science ontologies. *Journal of biomedical informatics*, 46(1):15–32, 2013.
- [19] Sebastian Hellmann, Jens Lehmann, and Sören Auer. Nif: An ontology-based and linked-data-aware nlp interchange format. *Working draft*, page 252, 2012.
- [20] Gail Hodge. Systems of knowledge organization for digital libraries. *Digital Library Federation, USA*, 2000.
- [21] Kevin L Howe, Premanand Achuthan, James Allen, Jamie Allen, Jorge Alvarez-Jarreta, M Ridwan Amode, Irina M Armean, Andrey G Azov, Ruth Bennett, Jyothish Bhai, et al. Ensembl 2021. *Nucleic acids research*, 49(D1):D884–D891, 2021.
- [22] Jay J Jiang and David W Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. *arXiv preprint cmp-lg/9709008*, 1997.
- [23] Julius Köpke and Johann Eder. Semantic invalidation of annotations due to ontology evolution. In *OTM Confederated International Conferences On the Move to Meaningful Internet Systems*, pages 763–780. Springer, 2011.

- [24] Michael D Lee, Brandon Pincombe, and Matthew Welsh. An empirical evaluation of models of text document similarity. In *Proceedings of the annual meeting of the cognitive science society*, volume 27, 2005.
- [25] Ying-Chi Lin, Victor Christen, Anika Groß, Silvio Domingos Cardoso, Cédric Pruski, Marcos Da Silveira, and Erhard Rahm. Evaluating and improving annotation tools for medical forms. In *International Conference on Data Integration in the Life Sciences*, pages 1–16. Springer, 2017.
- [26] Phuc-Hiep Luong and Rose Dieng-Kuntz. A rule-based approach for semantic annotation evolution in the coswem system. In *Canadian Semantic Web*, pages 103–120. Springer, 2006.
- [27] Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60, 2014.
- [28] Diana Maynard, Wim Peters, Mathieu d’Aquin, and Marta Sabou. Change management for metadata evolution. In *Proceedings of the International Workshop on Ontology Dynamics (IWOD-07)*, volume 2740, 2007.
- [29] Pablo N Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. Dbpedia spotlight: shedding light on the web of documents. In *Proceedings of the 7th international conference on semantic systems*, pages 1–8, 2011.
- [30] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.
- [31] Mohamed Morsey, Jens Lehmann, Sören Auer, Claus Stadler, and Sebastian Hellmann. Dbpedia and the live extraction of structured data from wikipedia. *Program*, 2012.
- [32] Roberto Navigli and Simone Paolo Ponzetto. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial intelligence*, 193:217–250, 2012.
- [33] Roberto Navigli and Paola Velardi. Learning domain ontologies from document warehouses and dedicated web sites. *Computational Linguistics*, 30(2):151–179, 2004.
- [34] Eyal Oren, Knud Möller, Simon Scerri, Siegfried Handschuh, and Michael Sintek. What are semantic annotations. *Relatório técnico. DERI Galway*, 9:62, 2006.
- [35] Yu Rang Park, Jihun Kim, Hye Won Lee, Young Jo Yoon, and Ju Han Kim. Gochase-ii: correcting semantic inconsistencies from gene ontology-based annotations for gene products. *BMC bioinformatics*, 12(1):S40, 2011.

- [36] Fernando Pech, Alicia Martinez, Hugo Estrada, and Yasmin Hernandez. Semantic annotation of unstructured documents using concepts similarity. *Scientific Programming*, 2017, 2017.
- [37] Borislav Popov, Atanas Kiryakov, Angel Kirilov, Dimitar Manov, Damyan Ognyanoff, and Miroslav Goranov. Kim-semantic annotation platform. In *International Semantic Web Conference*, pages 834–849. Springer, 2003.
- [38] Roy Rada, Hafeedh Mili, Ellen Bicknell, and Maria Blettner. Development and application of a metric on semantic nets. *IEEE transactions on systems, man, and cybernetics*, 19(1):17–30, 1989.
- [39] Philip Sedgwick. Multiple significance tests: the bonferroni correction. *Bmj*, 344, 2012.
- [40] Lucia Specia and Enrico Motta. Integrating folksonomies with the semantic web. In *European semantic web conference*, pages 624–639. Springer, 2007.
- [41] Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*, pages 697–706, 2007.
- [42] Thomas Pellissier Tanon, Gerhard Weikum, and Fabian Suchanek. Yago 4: A reasonable knowledge base. In *European Semantic Web Conference*, pages 583–596. Springer, 2020.
- [43] Julius Volz, Christian Bizer, Martin Gaedke, and Georgi Kobilarov. Discovering and maintaining links on the web of data. In *International Semantic Web Conference*, pages 650–665. Springer, 2009.
- [44] Xiang Zhang, Gong Cheng, and Yuzhong Qu. Ontology summarization based on rdf sentence graph. In *Proceedings of the 16th international conference on World Wide Web*, pages 707–716, 2007.