Pontifícia Universidade Católica
do Rio de Janeiro

Ximena Alexandra Cabrera Tapia

EnLiDa:
Enriquecimento das descrições de *Linked Data Cubes*

Dissertação de Mestrado

Dissertação apresentada como requisito parcial para obtenção do título de Mestre pelo Programa de Pós-Graduação em Informática da PUC-Rio.

Orientador: Prof. Marco Antonio Casanova

Rio de Janeiro
Agosto de 2013

PUC-Rio - Certificação Digital Nº 1121837/CA

Pontifícia Universidade Católica
do Rio de Janeiro

Ximena Alexandra Cabrera Tapia

EnLiDa:
Enriquecimento das descrições de *Linked Data Cubes*

Dissertação apresentada como requisito parcial para obtenção do título de Mestre pelo Programa de Pós-Graduação em Informática da PUC-Rio. Aprovada pela Comissão Examinadora abaixo assinada.

Prof. Marco Antonio Casanova
Orientador
Departamento de Informática – PUC-Rio

Prof. Antonio Luz Furtado
Departamento de Informática – PUC-Rio

Dra. Giseli Rabello Lopes
Departamento de Informática – PUC-Rio

Prof. José Eugenio Leal
Coordenador Setorial do Centro Técnico Cientifico – PUC-Rio

Rio de Janeiro, 30 de agosto de 2013

PUC-Rio - Certificação Digital Nº 1121837/CA

Todos os direitos reservados. É proibida a reprodução total ou parcial do trabalho sem autorização da universidade, do autor e do orientador.

Ximena Alexandra Cabrera Tapia

Graduou-se em Engenharia de Sistemas, Universidade Loyola (Bolívia), em dezembro de 2000. Ingressou no programa de mestrado do Departamento de Informática em 2011.

Ficha Catalográfica

Tapia, Ximena Alexandra Cabrera

EnLiDa: enriquecimento das descrições de Linked Data Cubes / Ximena Alexandra Cabrera Tapia ; orientador: Marco Antonio Casanova. – 2013.

83 f. : il. (color.) ; 30 cm

Dissertação (mestrado)–Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Informática, 2013.

Inclui bibliografia

1. Informática – Teses. 2. Dados interligados. 3. Propriedade owl:sameAs. 4. RDF. I. Casanova, Marco Antonio. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Informática. III. Título.

CDD: 004

PUC-Rio - Certificação Digital Nº 1121837/CA

Agradecimentos

Ao meu orientador Marco Antonio Casanova, pelo compartilhamento de seus conhecimentos durante a realização deste trabalho.

À PUC-Rio, pelo fornecimento de recursos e infraestrutura que viabilizaram a construção deste trabalho.

À CAPES pelo apoio financeiro concedido.

Aos meus pais e irmãos, pelo apoio incondicional em todos os momentos da minha vida.

Ao meu namorado, pela paciência e compreensão nesses dois anos.

Aos meus companheiros do Mestrado pela ajuda incondicional de sempre.

PUC-Rio - Certificação Digital Nº 1121837/CA

Resumo

Tapia, Ximena Alexandra Cabrera; Casanova, Marco Antonio. **EnLiDa: Enriquecimento das descrições de Linked Data Cubes**. Rio de Janeiro, 2013. 83p. Dissertação de Mestrado - Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

O termo dados interligados refere-se a conjuntos de triplas RDF organizados segundo certos princípios que facilitam a publicação e o acesso a dados por meio da infraestrutura da Web. Os princípios para organização de dados interligados são de grande importância pois oferecem uma forma de minimizar o problema de interoperabilidade entre bancos de dados expostos na Web. Este trabalho propõe enriquecer um banco de dados que contém descrições em RDF de cubos de dados, interligando seus componentes com entidades definidas em fontes de dados externas através de triplas *owl:sameAs*. O trabalho propõe uma arquitetura composta por dois componentes principais, o enriquecedor automático e o enriquecedor manual. O primeiro componente gera triplas *owl:sameAs* automaticamente enquanto que o segundo componente permite ao usuário definir manualmente as ligações. Em conjunto, estes componentes facilitam a definição de cubos de dados de acordo com os princípios de dados interligados.

PUC-Rio - Certificação Digital Nº 1121837/CA

Palavras-chave

Dados Interligados; Propriedade *owl:sameAs*; RDF.

Abstract

Tapia, Ximena Alexandra Cabrera; Casanova, Marco Antonio (Advisor). **EnLiDa: Enrichment of Linked Data Cube Descriptions.** Rio de Janeiro, 2013. 83p. MSc. Dissertation - Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

The term *Linked Data* refers to a set of RDF triples organized according to certain principles that facilitate the publishing and consumption of data using the Web infrastructure. The importance of the Linked Data principles stems from the fact that they offer a way to minimize the interoperability problem between databases exposed on the Web. This dissertation proposes to enrich a database that contains Linked Data cube descriptions by interconnecting the components of the data cubes with entities defined in external data sources, using *owl:sameAs* triples. The dissertation proposes an architecture consisting of two major components, the automatic enriching component and the manual enriching component. The first component automatically generates *owl:sameAs* triples, while the second component helps the user manually define *owl:sameAs* triples that the automatic component was not able to uncover. Together, these components therefore facilitate the definition of data cubes according to the Linked Data principles.

Keywords

Linked Data; *owl:sameAs* property; RDF.

PUC-Rio - Certificação Digital Nº 1121837/CA

Sumário

PUC-Rio - Certificação Digital Nº 1121837/CA

PUC-Rio - Certificação Digital Nº 1121837/CA

Lista de figuras

PUC-Rio - Certificação Digital Nº 1121837/CA

Lista de tabelas

PUC-Rio - Certificação Digital Nº 1121837/CA

1.

Introdução

1.1

Dados interligados

O termo *dados interligados (Linked Data)* refere-se a conjuntos de triplas RDF organizados segundo certos princípios que facilitam a publicação e o acesso a dados por meio da infraestrutura da Web(Berners-Lee T. 2009). Os princípios para organização de dados interligados são de grande importância, pois oferecem uma forma de minimizar o problema de interoperabilidade entre bancos de dados expostos na Web.

Os princípios dos dados interligados ditam o uso de RDF (*Resource Description Framework*) como modelo de dados e o uso de URIs (*Uniform Resource Identifier*) para identificar objetos do mundo real. Dados interligados usam ainda a propriedade *owl:sameAs* para indicar que duas URIs designam o mesmo indivíduo do mundo real. Triplas usando *owl:sameAs* são tradicionalmente chamadas de *interligações (links)*.

Enquanto o número de triplas em fontes de dados aumenta de forma constante, o total de interligações representa menos de 5% do número total de triplas disponíveis na Web de dados (Ngomo & Sören Auer 2011). No entanto, as interligações desempenham um papel fundamental em tarefas importantes, como o "*cross-ontology*" (Bhagdevet al. 2008; Lopez et al. 2009), inferência em grande escala (McCusker & McGuinness 2010; Urbani et al. 2010) e integração de dados (Ben-David et al. 2010; Ma et al. 2009). Além disso, enquanto o número de ferramentas para a publicação de dados interligados cresce constantemente, há poucas ferramentas eficientes para descobrir ligações

PUC-Rio - Certificação Digital Nº 1121837/CA

entre conjuntos de triplas. Dois exemplos destas ferramentas são o SILK (Auer et al. 2007) e o LIMES (Mercer 2006).

1.2

Cubos de dados interligados

Dados estatísticos são essenciais em muitas áreas. No governo, dados estatísticos mostram a anatomia da sociedade e ajudam a identificar os pontos fortes e fracos do governo, considerando-se um aspecto importante na tomada de decisões. Na ciência, dados estatísticos são um artefato essencial para provar ou refutar teorias científicas. No domínio empresarial, dados estatísticos sobre a venda de produtos ou indicadores econômicos oferecem uma contribuição crucial para a tomada de decisões estratégicas para a gestão e comercialização. No entanto, a obtenção de dados estatísticos é geralmente bastante custosa em termos de tempo e de recursos, especialmente nos casos que envolvem diferentes organizações (Salas et al. 2012).

Os dados estatísticos são frequentemente armazenados em bancos de dados relacionais. Os dados brutos são limpos, validados e armazenados em tabelas de dados, garantindo a confidencialidade dos indivíduos e entidades. Estes dados são normalmente armazenados e divulgados como estruturas multidimensionais conhecidas como cubos de dados (Cyganiak et al. 2011).

No processo de análise de dados estatísticos, algumas características são essenciais para assegurar que os dados sejam consumidos de uma forma simples e eficiente. As características principais são: (i) os dados devem ser publicados em um formato simples, complexidade poderia tornar-se um obstáculo à sua utilização de uma forma normalizada, de modo que possam ser reutilizados e processados por ferramentas automatizadas; (ii) os dados devem estar contextualizados com outros dados existentes para enriquecer a qualidade das estatísticas.

PUC-Rio - Certificação Digital Nº 1121837/CA

Neste contexto, os princípios de dados interligados podem ser eficazmente aplicados a dados estatísticos, no sentido de que os princípios oferecem uma estratégia para proporcionar a semântica dos dados. Intuitivamente, se forem seguidos, os princípios de dados interligados colocarão os dados estatísticos em contexto, ou seja, interligarão os dados estatísticos com outras fontes de dados relacionadas, criando um espaço de dados globalmente interligado que facilita uma análise dos dados mais rica (Cyganiak et al. 2011), (Ruback et al. 2013).

Em particular, a arquitetura de mediação proposta por Ruback (2013) ajuda a descrever e consumir dados estatísticos, expostos como triplas RDF, mas armazenados em bancos de dados relacionais. A arquitetura possui um Catálogo de descrições de cubos de dados, criadas de acordo com os princípios de dados interligados. O Catálogo (Manso 2013) utiliza uma descrição padronizada para cubos de dados armazenados em bases de dados estatísticos (relacionais). O mediador oferece uma interface para navegar pelas descrições dos cubos de dados e exporta os cubos de dados como triplas RDF, geradas por demanda, a partir das fontes de dados subjacentes (Ruback et al. 2013).

PUC-Rio - Certificação Digital Nº 1121837/CA

1.3

Contribuição da dissertação

Este trabalho trata do enriquecimento de descrições de cubos de dados, armazenadas no Catálogo proposto em (Manso 2013), interligando os componentes dos cubos de dados – domínios, atributos, etc. – com entidades de fontes de dados externas.

Dentro de certas condições, cada um destes componentes pode ter a sua descrição aprimorada associando o componente a um recurso externo através de *owl:sameAs*. Chamamos a este processo de *enriquecimento* do componente.

O trabalho propõe uma arquitetura composta por dois componentes principais, o *enriquecedor automático* e o *enriquecedor manual*. O primeiro componente gera triplas *owl:sameAs* automaticamente a partir de mapeamentos entre entidades locais e entidades equivalentes armazenadas em fontes externas. O segundo componente trabalha com entidades que não foram localizadas em fontes externas pelo componente automático, de maneira individual, permitindo ao usuário definir manualmente as ligações. Em conjunto, estes componentes facilitam o consumo de cubos de dados, tornando possível interligar os componentes dos cubos com entidades externas.

1.4

Organização da dissertação

Este trabalho está organizado da seguinte forma. A seção 2 apresenta uma revisão de conceitos e ferramentas utilizados no trabalho. A seção 3 descreve o módulo enriquecedor, detalhando como foi desenvolvido. A seção 4 discute os casos de uso. A seção 5 apresenta a conclusão e sugestões para trabalhos futuros.

PUC-Rio - Certificação Digital Nº 1121837/CA

2.

Revisão de conceitos e ferramentas

2.1

A propriedade *owl:sameAs*

A suposição de que nomes diferentes referem-se a indivíduos diferentes no mundo real não é plausível na Web. Por exemplo, a mesma pessoa pode ser identificada de várias maneiras diferentes (ou seja, por URIs diferentes).

OWL fornece três construções para afirmar fatos sobre a identidade dos indivíduos: *owl:sameAs* é usada para indicar que duas URIs referem-se ao mesmo indivíduo;*owl:differentFrom* é usado para indicar que duas URIs referem-se a diferentes indivíduos;*owl:allDifferent* fornece uma expressão para afirmar que os indivíduos de uma lista são todos diferentes (Klyne et al. 2004).

Em mais detalhe, a propriedade *owl:sameAs* indica que duas URIs identificam o mesmo indivíduo. Por exemplo, pode-se afirmar que as duas URIs a seguir na verdade se referem à mesma pessoa:

Exemplo 1

<rdf:Descriptionrdf:about="#William_Jefferson_Clinton">
<owl:sameAsrdf:resource="#BillClinton"/>
</rdf:Description>

Em OWL *Full*, onde uma classe pode ser tratada como uma instância de uma (meta) classe, pode-se usar *owl:sameAs* para definir igualdade de classes, ou seja,para indicar que dois conceitos têm o mesmo significado intencional.

PUC-Rio - Certificação Digital Nº 1121837/CA

Exemplo 2

<owl:Classrdf:ID="FootballTeam">

<owl:sameAsrdf:resource="http://sports.org/US#SoccerTeam"/>

</ Owl:Class>

Pode-se imaginar a declaração do Exemplo 2 como parte de uma ontologia de esporte europeu. As duas classes são tratadas aqui como indivíduos, neste caso, como instâncias da classe: *owl:class*. Isto nos permite afirmar que a classe *FootballTeam*, em alguma ontologia de esporte europeu, denota o mesmo conceito que a classe *SoccerTeam*, em alguma ontologia de esporte americano.

Observe a diferença entre Exemplo2 e o Exemplo 3 a seguir.

Exemplo 3

<footballTeam owl:equivalentClass us:soccerTeam />

O Exemplo 3 indica que as duas classes têm a mesma extensão, ou seja, o mesmo conjunto de indivíduos, mas não necessariamente representam o mesmo conceito.

2.2

Uma crítica ao uso da propriedade *owl:sameAs*

O uso de *owl:sameAs* em dados interligados é imprescindível, especialmente em ligações entre conjuntos de dados. No entanto, há a suspeita de que, dentro da comunidade de dados interligados, o uso de *owl:sameAs* pode ser de alguma forma incorreta. Na verdade, *owl:sameAs* pode ser considerado apenas um tipo de "relação de identidade", ou seja, uma relação que declara que dois itens são idênticos de alguma forma.

Halpin & Hayes (2010) delineiam quatro leituras alternativas de *owl:sameAs*:

PUC-Rio - Certificação Digital Nº 1121837/CA

- “A mesma coisa, mas com contextos diferentes”
- “A mesma coisa, mas referencialmente opaca”
- “Representa”
- “Muito semelhante”

Obviamente, a questão de como expressar as relações de identidade nos dados interligados é mais complexa do que apenas a aplicação de *owl:sameAs*. De fato, uma abordagem mais sutil que cubra as quatro possíveis leituras de *owl:sameAs* seria um passo útil para a comunidade de dados interligados.

2.3

As propriedades *equivalentClass* e *equivalentProperty*

A propriedade *owl:equivalentClass* é usada para indicar que duas classes têm exatamente as mesmas instâncias. Em OWL *Full*, pode-se usar *owl:sameAs* entre duas classes para indicar que elas são idênticas em todos os sentidos.

Exemplo 4

<owl:Classrdf:ID="Wine">
<owl:equivalentClass rdf:resource="&vin;Wine"/>
</owl:Class>

Já foi visto que as expressões de classe podem ser alvo do construtor *rdfs:subClassOf*. Elas também podem ser alvo de *owl:equivalentClass*, o que evita a necessidade de inventar nomes para cada expressão de classe.

Exemplo5

<owl:Classrdf:ID="TexasThings">
<owl:equivalentClass>

PUC-Rio - Certificação Digital Nº 1121837/CA

```
<owl:Restriction>
<owl:onProperty rdf:resource="#locatedIn" />
<owl:someValuesFrom rdf:resource="#TexasRegion" />
</owl:Restriction>
</owl:equivalentClass>
</owl:Class>
```

De forma semelhante, a propriedade *owl:equivalentProperty* é usada para indicar que duas propriedades denotam exatamente os mesmos pares de instâncias.

2.4

A ferramenta LIMES

A ferramenta LIMES implementa uma abordagem para a descoberta de ligações entre bases de conhecimento em dados interligados. A abordagem, aplicada tanto a dados sintéticos quanto a dados reais, mostra que esta ferramenta é muito eficiente quanto ao número de comparações e a tempo de execução (Ngomo & Sören Auer 2011). LIMES usa a noção de desigualdade triangular para particionar o espaço de busca. Cada partição é então representada por um modelo (Frey & Dueck 2007), que permite uma aproximação precisa da distância entre cada instância do *dataset* de origem às instâncias do *dataset* de destino. A ferramenta LIMES pode ser configurada com diferentes métricas, usadas para o processo interno de comparação (Ngomo & Sören Auer 2011).

O arquivo de configuração da ferramenta LIMES inclui a especificação de algoritmos de edição de texto, que ajudam na computação de métricas de semelhança.

De acordo com as versões do LIMES, ele funciona com determinados algoritmos. Na versão v.0.4.1 do LIMES, os algoritmos habilitados são o *levenshtein,* o *blockdistance*, o *euclidian* e o *qgrams*. Na

PUC-Rio - Certificação Digital Nº 1121837/CA

versão 0.6RC2, os algoritmos habilitados são o *trigrams*, o *cosine*, e o *jaccard*.

2.5

A ferramenta SILK

SILK -Link Discovery Framework é uma ferramenta para encontrar relações entre entidades representadas em diferentes fontes de dados. SILK apresenta uma linguagem declarativa para especificar quais tipos de ligações RDF devem ser descobertas entre as fontes de dados, bem como as condições que as entidades devem cumprir, a fim de serem interligadas. As condições de ligação podem se basear em várias métricas de similaridade e podem levar em conta o grafo em torno as entidades (Volz et al. 2009).

Em lugar de usar espaços métricos, SILK usa um índice de correspondência previamente computado para atingir complexidade de tempo quase linear. A desvantagem da abordagem de correspondência prévia é que recall não é garantido. Além disso, SILK permite a configuração manual de blocos de dados, para minimizar o tempo de execução do processo de comparação. No entanto, esta abordagem de blocos não é sem perda (Auer et al. 2007).

2.6

Algoritmos de similaridade

2.6.1

Algoritmo *Levenshtein*

O algoritmo de *Levenshtein* ou *edit distance* (distância de edição) leva o nome de seu autor. Este foi um dos primeiros algoritmos de

PUC-Rio - Certificação Digital Nº 1121837/CA

comparação de cadeias de caracteres e até hoje ainda é um dos mais utilizados.

A função de *Levenshtein* pode ser definida como “o menor número de inserções, remoções e substituições para igualar duas strings” (Navarro 01):

$$
\begin{aligned}
M(i,j) = \text{Max} \{\ & M(i-1, j) - 1, && \text{//inserção} \\
& M(i-1, j-1) + p(i, j), && \text{//match ou substituição} \\
& M(i, j-1) - 1 \} && \text{//remoção} \\
\text{Onde } p(i, j) = & +2 \text{ se } Xi = Yj && \text{//match} \\
& -1 \text{ se } Xi \neq Yj && \text{//substituição}
\end{aligned}
$$

M é a matriz de *Levenshtein* e M(0,0) = 0. A função p(i,j) é utilizada para determinar se houve igualdade entre os termos comparados (*match*) ou não (substituição). X e Y são as *strings* que estão sendo comparadas, “i” e “j” são respectivamente as posições dos caracteres destas duas strings.

O exemplo da Tabela 1 compara “medico” com “biomedicina”. Podemos observar que o escore obtido (último escore da matriz) foi 4.

	ε	b	i	o	m	e	d	i	c	i	n	a
ε	**0**	**-1**	**-2**	**-3**	-4	-5	-6	-7	-8	-9	-10	-11
m	-1	-1	-2	-3	**-1**	-2	-3	-4	-5	-6	-7	-8
e	-2	-2	-2	-3	-2	**1**	0	-1	-2	-3	-4	-5
d	-3	-3	-3	-3	-3	0	**3**	2	1	0	-1	-2
i	-4	-4	-1	-2	-3	-1	2	**5**	4	3	2	1
c	-5	-5	-2	-2	-3	-2	1	4	**7**	6	5	4
o	-6	-6	-3	0	-1	-2	0	3	6	**6**	**5**	**4**

Tabela 1 - Matriz de *Levenshtein* para o exemplo.

Este é o alinhamento obtido:

- - - m e d i c - - o

b i o m e d i c i n a

PUC-Rio - Certificação Digital Nº 1121837/CA

Quanto ao desempenho, o tempo gasto por este algoritmo é O(|x||y|), enquanto o espaço requerido é de apenas O(mínimo(|x||y|)), onde |x| e |y| são os tamanhos das strings x e y respectivamente.

2.6.2

Algoritmo *q-gram*

Um *q-gram* é o conjunto de todas as sub-*strings* que podem ser geradas a partir de uma determinada *string*, onde “q” representa o tamanho destas sub-*strings*.

Por exemplo, o *q-gram* gerado para a string “paciente” com q = 3 será:

{##p, #pa, pac, aci, cie, ien, ent, nte, te$, e$$}

Este algoritmo foi utilizado inicialmente como técnica de filtragem, cujo objetivo era descartar áreas onde não pode haver *matching* (casamento de palavras) (Foster et al. 03).

Entretanto esta técnica pode ser aplicada de forma diferente para identificar sequências de texto que possuam palavras em comum. Se tivermos as *strings* A e B, podemos gerar os *q-grams* de A e B e depois contar o número de *q-grams* idênticos. Então é possível encontrar a “*q-gramdistance*”:

| *q-gramdistance* | = | tamanho do maior *q-grams* | - | número de *q-grams* em comum |

Por exemplo, comparando-se “paciente” com “patient”, temos:

q-gram paciente {##p, #pa, pac, aci, cie, ien, ent, nte, te$, e$$}
q-gram patient {##p, #pa, pat, ati, tie, ent, nt$, t$$}
q-gams em comum {##p, #pa, ent}

Neste exemplo temos apenas 3 *q-grams* em comum, o que é um valor muito baixo, pois as *strings* comparadas são semelhantes, e a maior string possui 10 *q-grams*.

PUC-Rio - Certificação Digital Nº 1121837/CA

Aplicando a fórmula vista acima obtemos:

| *q-gramdistance* | = 10 – 3 = 7.

Isto representa uma similaridade de 3/10 = 0,3 que neste caso é muito baixa, já que as *strings* comparadas são muito semelhantes.

Sejam |x| e |y| os tamanhos das *strings* x e y que serão comparadas. O tempo gasto por este algoritmo é melhor do que o da programação dinâmica, pois como cada *q-gram* de x é comparado apenas com certa quantidade dos *q-gram* de y então o tempo gasto é O(mínimo(|x| |y|)) e a quantidade de espaço requerida é O(|x| + |y|).

2.7

Data Cube Vocabulary

O *vocabulário de cubos de dados (Data cube vocabulary*) (Cyganiak et al 2013) foi projetado para descrever cubos de dados estatísticos em RDF.Este vocabulário baseia-se em vocabulários RDF existentes:

- Simple Knowledge Organization System (SKOS) (Miles & Bechhofer 2009) – para esquemas conceituais.
- Statistical Core Vocabulary (SCOVO) (Hausenblas et al. 2012) - para estruturas centrais estatísticas.
- Dublin Core Metadata Initiative (DCMI) Metadata Terms (DCMI 2012) – para metadados.
- Vocabulary of Interlinked Datasets (VoID) (Alexander et al. 2011) – para acesso a dados.
- Friend-of-a-Friend (FOAF) (Brickley & Miller 2010) – para agentes.
- Core Organization Ontology (ORG) (Reynolds 2012) – para as organizações.

PUC-Rio - Certificação Digital Nº 1121837/CA

A Figura 1 resume as classes e propriedades do vocabulário de cubo de dados (Cyganiak et al. 2013).

Os prefixos e *namespaces* correspondentes utilizados na Figura 1 são:

qb http://purl.org/linked-data/cube#
skos http://www.w3.org/2004/02/skos/core#
rdfs http://www.w3.org/2000/01/rdf-schema#
xsd http://www.w3.org/2001/XMLSchema#
sdmx http://purl.org/linked-data/sdmx#

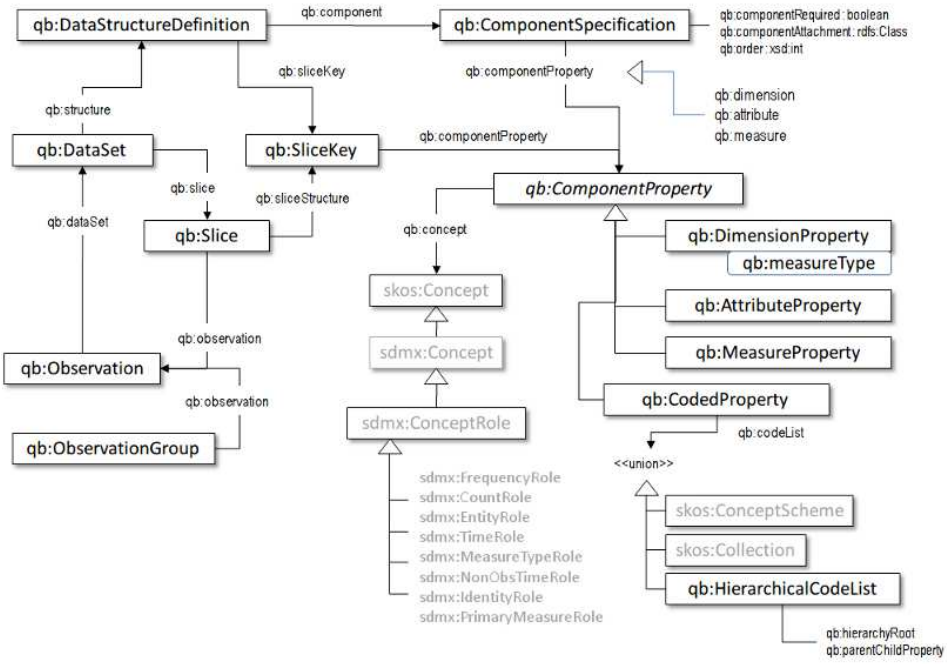


Figura 1 - Esboço do Vocabulário de Cubo de Dados (Cyganiaket al. 2013).

Para entender melhor a Figura 1, os termos do vocabulário de cubo de dados estão resumidos no que se segue:

Conjunto de dados

Classe: *qb:DataSet*Sub classe de: *qb:Attachable*Equivalente a: *scovo:Dataset*

PUC-Rio - Certificação Digital Nº 1121837/CA

Representa uma coleção de observações, possivelmente organizadas em várias partes, em conformidade com uma estrutura dimensional comum.

Observações

Classe: *qb:Observation* Sub classe de: *qb:Attachable*Equivalente a: *scovo:Item*

Uma única observação no cubo pode ter um ou mais valores de medição associados.

Propriedade: *qb:dataSet* (Domínio: *qb:Observation* -> Contradomínio: *qb:DataSet*)

Indica o conjunto de dados do qual uma observação pertence.

Propriedade: *qb:observation* (Domínio: *qb:Slice* -> Contradomínio: *qb:Observation*)

Indica que uma observação pertence a um conjunto de dados.

Subconjunto

Classe: *qb:ObservationGroup*

Um grupo de observações.

Classe: *qb:Slice* Sub classe de: *qb:Attachable*, *qb:ObservationGroup*

Denota um subconjunto de um conjunto de dados definido fixando-se um subconjunto de valores dimensionais, componente de propriedades do subconjunto.

Propriedade: *qb:slice* (Domínio: *qb:DataSet* -> Contradomínio: *qb:Slice*; sub propriedade de: *qb:observationGroup*)

Indica um subconjunto de um conjunto de dados definido fixando-se um subconjunto de valores dimensionais.

PUC-Rio - Certificação Digital Nº 1121837/CA

Propriedade: *qb:observationGroup* (Domínio: -> Contradomínio: *qb:ObservationGroup*)

Indica um grupo de observações. O domínio desta propriedade é deixado em aberto para que um grupo possa ser anexado a diferentes recursos e não precisa ser restrito a um único conjunto de dados.

Dimensões, Atributos e Medidas

Classe: *qb:Attachable*

Superclasse de tudo o que pode ter atributo e dimensão.

Classe: *qb:ComponentProperty* Sub classe de: *rdf:Property*

Superclasse de todas as propriedades que representam dimensões, atributos ou medidas.

Classe: *qb:DimensionProperty* Sub classe de: *qb:ComponentProperty*, *qb:CodedProperty*

A classe de propriedades de componentes que representam as dimensões do cubo.

Classe:*qb:AttributeProperty* Sub classe de: *qb:ComponentProperty*

A classe de propriedades de componentes que representam atributos de observações do cubo, por exemplo, unidade de medida.

Classe: *qb:MeasureProperty* Sub classe de: *qb:ComponentProperty*

A classe de propriedades de componentes que representam o valor de medição do fenômeno a ser observado.

Classe: *qb:CodedProperty* Sub classe de: *qb:ComponentProperty*

Superclasse de todas as propriedades dos componentes codificados.

PUC-Rio - Certificação Digital Nº 1121837/CA

Propriedades reutilizáveis de componentes de uso geral

Propriedade: *qb:measureType* (Domínio: -> Contradomínio: *qb:MeasureProperty*)

Dimensão de medida genérica; o valor desta dimensão indica que medida (a partir do conjunto de medidas no *Data Structure Definition*) está sendo dado pela observação.

Definição de estrutura de dados

Classe: *qb:DataStructureDefinition* Sub classe de: *qb:ComponentSet*

Define a estrutura de um conjunto de dados ou de uma parte.

Propriedade: *qb:structure* (Domínio: *qb:DataSet* -> Contradomínio: *qb:DataStructureDefinition*)

Indica a estrutura de um conjunto de dados.

Propriedade: *qb:component* (Domínio*: qb:DataStructureDefinition* -> Contradomínio: *qb:ComponentSpecification*)

Indica uma especificação de componente que está incluída na estrutura do conjunto de dados.

Especificações de componentes para qualificar o uso do componente em uma Data Structure Definition

Classe: *qb:ComponentSpecification* Sub classe de:*qb:ComponentSet*

Usado para definir propriedades de um componente (atributo, dimensão, etc.) que são específicos para seu uso em uma *DataStructureDefinition*.

Classe: *qb:ComponentSet*

Classe abstrata dos objetos que fazem referência a um ou mais *ComponentProperties*

PUC-Rio - Certificação Digital Nº 1121837/CA

Propriedade: *qb:componentProperty* (Domínio: *qb:ComponentSet* -> Contradomínio: *qb:ComponentProperty*)

Indica um *componentProperty* (ou seja, atributo ou dimensão) esperado de um conjunto de dados, ou uma dimensão fixa em uma *SliceKey*.

Propriedade: *qb:order* (Domínio: *qb:ComponentSpecification* -> Contradomínio: *xsd:int*)

Indica uma ordem de prioridade para os componentes de conjuntos com esta estrutura, usada para guiar apresentações; números de ordem inferiores vêm antes dos números mais elevados, os componentes não numerados vêm por último.

Propriedade: *qb:componentRequired* (Domínio: *qb:ComponentSpecification* -> Contradomínio: *xsd:boolean*)

Indica se a propriedade de um componente é necessária (verdadeiro) ou opcional (falso), no contexto de uma *DataStructure Definition.* Só é aplicável a componentes correspondentes a um atributo. O *default* é falso (opcional).

Propriedade: *qb:componentAttachment* (Domínio: *qb:ComponentSpecification* -> Contradomínio: *rdfs:Class*)

Indica o nível em que a propriedade do componente deve ser anexado; pode ser um *qb:DataSet*, *qb:Slice* ou *qb:Observation*, ou um *qb:MeasureProperty*.

Propriedade: *qb:dimension* (Domínio: -> Contradomínio: *qb:DimensionProperty* ; sub propriedade de: *qb:componentProperty*)

Uma alternativa para *qb:componentProperty* que explicita que o componente é uma dimensão.

PUC-Rio - Certificação Digital Nº 1121837/CA

Propriedade: *qb:measure* (Domínio: -> Contradomínio: *qb:MeasureProperty* ; sub propriedade de: *qb:componentProperty*)

Uma alternativa para *qb:componentProperty* que explicita que o componente é uma medida.

Property: *qb:attribute* (Domínio: -> Range: *qb:AttributeProperty* ; sub propriedade de: *qb:componentProperty*)

Uma alternativa para *qb:componentProperty* que explicita que o componente é um atributo.

Propriedade: *qb:measureDimension* (Domínio: -> Contradomínio: *qb:DimensionProperty* ; sub propriedade de: *qb:componentProperty*)

Uma alternativa para *qb:componentProperty* que explicita que o componente é uma dimensão de medida.

Definições para o subconjunto

Classe: *qb:SliceKey* Sub classe de: *qb:ComponentSet*

Denota um subconjunto das propriedades dos componentes de um conjunto de dados que são fixados nas respectivas partes.

Propriedade: *qb:sliceStructure* (Domínio: *qb:Slice* -> Contradomínio: *qb:SliceKey*)

Indica a chave correspondente a esta parte.

Propriedade: qb:sliceKey (Domínio: qb:DataSet -> Contradomínio: qb:SliceKey)

Indica a chave da parte que é utilizada neste conjunto de dados.

Conceitos

Propriedade: *qb:concept* (Domínio: *qb:ComponentProperty* -> Contradomínio: *skos:Concept*)

PUC-Rio - Certificação Digital Nº 1121837/CA

Indica o conceito que está sendo medido, ou indicado por um *ComponentProperty*.

Propriedade: *qb:codeList* (Domínio: *qb:CodedProperty* -> Contradomínio: *owl:unionOf* (*skos:ConceptScheme skos:Collection qb:HierarchicalCodeList*))

Indica a lista de códigos associada a um *CodedProperty.*

Hierarquias Non-SKOS

Classe: *qb:HierarchicalCodeList*

Representa uma hierarquia dos conceitos que podem ser utilizados para a codificação. A hierarquia é definida por uma ou mais raízes, juntamente com uma propriedade que relaciona um conceito na hierarquia aos seus filhos. Um mesmo conceito pode ser membro de várias hierarquias, desde que os valores *qb:parentChildProperty* sejam usados para cada hierarquia.

Propriedade: *qb:hierarchyRoot* (Domínio: *qb:HierarchicalCodeList*)

Especifica a raiz da hierarquia. Uma hierarquia pode ter várias raízes, mas deve ter pelo menos uma.

Propriedade: *qb:parentChildProperty* (Domínio: *qb:HierarchicalCodeList* -> Contradomínio: *rdf:Property*)

Especifica uma propriedade que relaciona um conceito pai, a uma hierarquia, de um conceito filho. Note-se que um filho pode ter mais de um pai.

PUC-Rio - Certificação Digital Nº 1121837/CA

2.8

Proveniência como Metadado

Buneman et al. (2001) define proveniência de dados (*data provenance*) como a descrição da origem de um item de dado e o processo pelo qual este chegou a um banco de dados[1].

Proveniência pode ser capturada através de duas formas: retrospectiva e prospectiva (Freire et al. 2008). A forma retrospectiva captura os passos executados por uma tarefa computacional, bem como a informação sobre o ambiente utilizado para derivar um dado produto específico.

A forma prospectiva captura os passos que devem ser seguidos para a geração de um dado produto, permitindo desta forma o registro da especificação de tarefas computacionais.

A forma de captura da proveniência pode ser em maior ou menor nível de detalhe. Buneman e W. Tan (2007) classificam em dois níveis: "grão grosso" e "grão fino".

A proveniência em "grão grosso" está relacionada à proveniência de um conjunto de atividades, organizadas em um *workflow.* Ela descreve a história da execução de um *workflow* ou da derivação de um conjunto de dados.

A proveniência em "grão fino" descreve a proveniência de um item de dado pertencente a um conjunto de dados de forma a capturar a origem e a movimentação de um dado (ou seja, "onde" foi originado), que pode estar relacionado a bancos de dados integrados, e descrever a importância da presença de um item de dado na composição de uma informação ("por que" foi originado).

Podemos interpretar a proveniência dos componentes: dimensão, atributo, medida e propriedade codificada como em “grão grosso” e a proveniência das instancias de um domínio como em “grão fino”.

[1] "Sometimes called 'lineage' or 'pedigree' - is the description of the origins of a piece of data and the process by which it arrived in a database."

PUC-Rio - Certificação Digital Nº 1121837/CA

Assim como é possível a captura da proveniência de um dado, a proveniência de um processo também pode ser capturada. De acordo com Simmhan (2007), a proveniência de processos envolve a descrição da execução de um simples processo, ou seja, das tarefas que dele fazem parte.

Em determinados processos (e.g. *Workflows* Científicos), para que se tenha a informação dos dados geradores de um produto, é importante que se registre cada dado consumido pelo processo, de modo que estes dados não se percam quando for necessário investigar a proveniência de dados.

Para que as informações relacionadas a um dado sejam obtidas, é importante que os dados referentes à sua origem sejam registrados, ou seja, que a história dos fatores que envolvem o processamento e a geração desses dados seja capturada de forma a prover informações sobre a proveniência de um dado.

2.9

Resumo do capítulo

Este capítulo apresentou conceitos e ferramentas necessários ao desenvolvimento desta dissertação. Abordou a propriedade *owl:sameAs*, e seus possíveis resultados, as ferramentas LIMES e SILK, o vocabulário usado para descrever cubos de dados em RDF e o conceito da proveniência dos dados.

PUC-Rio - Certificação Digital Nº 1121837/CA

3.

Mediador para cubo de dados interligados

3.1

Arquitetura do mediador

O trabalho desta dissertação insere-se no contexto de um projeto cujo objetivo consiste em desenvolver um serviço de mediação para acesso a cubos de dados.

O serviço de mediação segue a arquitetura da Figura 2 e inclui os seguintes componentes principais: Wrappers, Catálogo, Mediador, Aplicação Cliente (Ruback et al. 2013), Recomendação das fontes de dados e Enriquecedor.

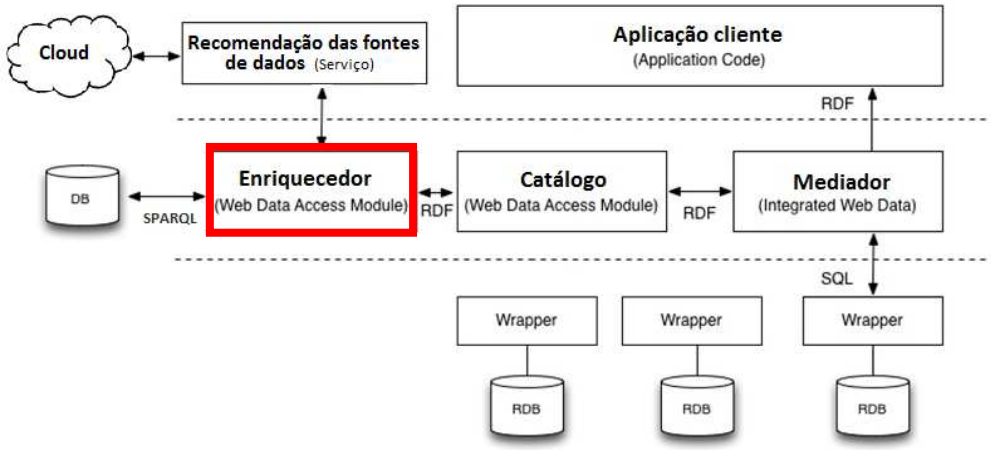


Figura 2 - Visão geral da arquitetura.

Os módulos *Wrappers*

O *Wrapper* para um banco de dados relacional subjacente fornece esquemas em forma de estrela, descrevendo dados estatísticos armazenados no banco de dados. Ressalta-se que os cubos de dados podem ser organizados no banco de dados de diversas maneiras, usando

PUC-Rio - Certificação Digital Nº 1121837/CA

várias tabelas. No entanto, o *wrapper* expõe cada cubo de dados através de um esquema em forma de estrela, cujo mapeamento para as tabelas subjacentes é interno ao *wrapper*.

O módulo Catálogo

O Catálogo contém dados públicos e privados. Dados públicos referem-se às descrições dos cubos de dados incluindo as suas dimensões e valores de dimensão, que são expostos às aplicações. Uma descrição de um cubo de dados é armazenada como um conjunto de triplas RDF.

O módulo Mediador

O mediador oferece acesso dos bancos de dados estatísticos relacionais subjacentes e expõe os cubos de dados descritos no Catálogo para os aplicativos. O Mediador permite que um aplicativo selecione a descrição de um cubo de dados interligado, armazenada no Catálogo, e aplique certas transformações ao cubo. O Mediador converte os dados (ou seja, o cubo) retornados pelo *Wrapper* para RDF, retornando as triplas para o aplicativo que fez a requisição.

O módulo de Aplicação Cliente

É um componente que interage com o Mediador enviando requisições tanto de metadados, quanto de observações dos cubos disponíveis no Catálogo.

O módulo Enriquecedor

O Enriquecedor interage internamente com o Catálogo e externamente com o serviço de Recomendação das fontes de dados.

O Enriquecedor acrescenta triplas criadas com a propriedade *owl:sameAs* da linguagem OWL interligando componentes (ver página 40 para definição de componente) das descrições dos cubos de dados com entidades externas (armazenadas em outros conjuntos de triplas). O

PUC-Rio - Certificação Digital Nº 1121837/CA

Enriquecedor usa a ferramenta LIMES (Ngomo & Sören Auer 2011) para esta tarefa.

Este módulo também é responsável pela tarefa de armazenamento da proveniência dos dados, ou seja, da proveniência das triplas *owl:sameAs* geradas com o auxílio do LIMES.

O módulo de Recomendação

O Módulo de Recomendação (Herrera et al. 2012) é responsável por retornar uma lista de recomendações de fontes RDF. Foi desenvolvido para orientar a busca por fontes de dados relevantes (Nikolov & d Aquin 2011)ao processo de enriquecimento.

PUC-Rio - Certificação Digital Nº 1121837/CA

3.2

Catálogo de Descrições de Cubos de Dados Interligados

O Catálogo de descrições de cubos de dados interligados é criado de acordo com os princípios de dados interligados (Heath & Bizer 2011). Este Catálogo contém descrições padronizadas em RDF (OECD 2006) sobre cada cubo de dados armazenado em um banco de dados estatístico (relacional) conhecido pelo ambiente de mediação. Assim, uma descrição de um cubo de dados é chamada de uma *descrição de um cubo de dados interligados* e nada mais é do que um conjunto de triplas RDF.

As triplas armazenadas no Catálogo descrevem as dimensões e os atributos de cada cubo de dados, incluindo possivelmente os valores do domínio de uma dimensão e suas conexões com bancos de dados externos. No entanto, a descrição de um cubo de dados interligado não contém triplas que capturam as observações, isto é, não é uma materialização completa de um cubo de dados em RDF; as observações do cubo de dados ainda permanecem na base de dados relacional. Portanto, o Catálogo contém metadados sobre os cubos de dados, mas não as próprias observações.

Para permitir a ligação com outras fontes de dados externas, os valores de uma dimensão também podem ser armazenados no Catálogo. Isto é essencial para indicar a semântica dos valores de uma dimensão, - seguindo os princípios de dados interligados (Heath & Bizer 2011) - ajudando assim a explicitar a semântica dos cubos de dados.

Considerando que o Catálogo contém tipos diferentes de informação, ele é dividido em duas partes: dados públicos e dados privados. Os dados públicos referem se às descrições do cubo de dados, valores de dimensão e das triplas *owl:sameAs* (Bechhofer et al. 2004). Dados privados referem se a dados necessários internamente, ou seja, à conexão com os bancos de dados e o mapeamento dos bancos de dados relacionais subjacentes aos cubos de dados RDF. Para cada descrição de um cubo de dados interligado no Catálogo, existe pelo menos um mapeamento para um esquema em forma de estrela de um banco de dados subjacente.

Para facilitar o consumo dos dados, vocabulários conhecidos foram usados para copiar as descrições de cubo de dados interligados no Catálogo, como recomendado em (Heath & Bizer 2011). O vocabulário de cubo de dados (Cyganiak et al. 2013) foi adotado para as descrições dos cubos de dados interligados e a linguagem de mapeamento R2RML (Das et al. 2010) para os mapeamentos entre os cubos de dados interligados e os bancos de dados relacionais. Alguns termos da linguagem de mapeamento D2RQ (Cyganiak et al. 2012) também foram utilizados já que a linguagem de mapeamento R2RML não inclui termos relacionados à conexão com os bancos de dados, como o nome de usuário e senha, por exemplo.

3.3

Exemplo de Descrição de um Cubo de Dados Interligado

As tabelas seguintes contêm exemplos para ilustrar a forma como um cubo de dados interligados é descrito.

PUC-Rio - Certificação Digital Nº 1121837/CA

```
@prefix ex-resource:       <http://purl.org/GovDataCube/resources/>.
@prefixex-property:       <http://purl.org/GovDataCube/properties/>.
@prefix rdfs:                 <http://www.w3.org/2000/01/rdf-schema#>.
@prefixqb:                   <http://purl.org/linked-data/cube#>.
@prefix sdmx-attribute:  <http://purl.org/linked-data/sdmx/2009/attribute#>.

ex-resource:dataset-residents a qb:DataSet;
rdfs:label “Number of residents";
rdfs:comment "The number of residents in Brazil of each sex, race and range of age by area and
time. Dimensions: DimRace, DimSex, DimAge, DimCountry and DimYear.";
qb:structure ex-resource:dsd-residents.

ex-resource:dsd-residents a qb:DataStructureDefinition;
    # The dimensions
qb:component [qb:dimension ex-resource:dimRace];
qb:component [qb:dimension ex-resource:dimSex];
qb:component [qb:dimension ex-resource:dimAge];
qb:component [qb:dimension ex-resource:dimYear];
qb:component [qb:dimension ex-resource:dimCountry];
    # The measure
qb:component [qb:measure ex-property:numberResidents];
    # The attributes
qb:component [qb:attributes dmx-attribute:unitMeasure; qb:componentAttachment qb:DataSet;].
```

Tabela 2 - Definição do cubo “Residentes”.

A Tabela 1 apresenta a descrição de um cubo de dados, chamado no que se segue de “Residentes”. A estrutura do cubo explica como se forma o cubo e mostra seus componentes. Neste exemplo, os componentes são as cinco dimensões, uma medida e um atributo.

```
@prefix ex-resource:       <http://purl.org/GovDataCube/resources/> .
@prefix d2rq:          <http://www.wiwiss.fu-berlin.de/suhl/bizer/D2RQ/0.1#>.

ex-resource:databaseResidents a d2rq:Database;
d2rq:username "root";
d2rq:password "root";
d2rq:jdbcDSN "jdbc:mysql://localhost/project1";
d2rq:jdbcDriver "com.mysql.jdbc.Driver".
```

Tabela 3 - Definição da conexão com o *project1 RDB.*

A Tabela 2 apresenta as informações de conexão para um banco de dados relacional (RDB). A conexão apresentada neste exemplo é o *project1* RDB e é chamada *ex-resource:databaseResidents*. Esta

PUC-Rio - Certificação Digital Nº 1121837/CA

informação é usada quando o mapeamento R2RML para este RDB é estabelecido.

As Tabelas 3 e 4 apresentam os mapeamentos R2RML.

```
@prefix ex-resource:       <http://purl.org/GovDataCube/resources/> .
 @prefixex-property:      <http://purl.org/GovDataCube/properties/>.
@prefix rdf:                 <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
@prefixrr:                   <http://www.w3.org/ns/r2rml#>.
@prefix d2rq:                <http://www.wiwiss.fu-berlin.de/suhl/bizer/D2RQ/0.1#>.
@prefixqb:                  <http://purl.org/linked-data/cube#>.
@prefix sdmx-attribute:    <http://purl.org/linked-data/sdmx/2009/attribute#>.

ex-resource:TriplesMapFactResidents a rr:TriplesMap;
d2rq:dataStorage ex-resource:databaseResidents;
rr:logicalTable [ rr:sqlQuery """
SELECT DISTINCT idPerson, idCountry, idTime, numberResidents
            FROM factResidents
           WHERE numberResidents<>0 """; ];
rr:subjectMap [
rr:template
"http://purl.org/GovDataCube/resources/Observations/{idPerson}_{idArea}_{idTime}_{numberResid
ents}";
rr:classqb:Observation; ];
rr:predicateObjectMap [
rr:predicateqb:dataSet;
rr:objectMap [rr:constantex-resource:dataset-residents]; ];
rr:predicateObjectMap [
rr:predicateex-resource:dimSex;
rr:objectMap [
rr:parentTriplesMapex-resource:TriplesMapSex;
rr:joinCondition [
rr:child "idPerson";
rr:parent "idPerson"; ];];];
rr:predicateObjectMap [
rr:predicateex-resource:dimRace;
rr:objectMap [
rr:parentTriplesMapex-resource:TriplesMapRace;
rr:joinCondition [
rr:child "idPerson";
rr:parent "idPerson"; ];];];
rr:predicateObjectMap [
rr:predicateex-resource:dimAge;
rr:objectMap [
rr:parentTriplesMapex-resource:TriplesMapAge;
rr:joinCondition [
rr:child "idPerson";
rr:parent "idPerson"; ];];];
```

PUC-Rio - Certificação Digital Nº 1121837/CA

```
rr:predicateObjectMap [
rr:predicateex-resource:dimCountry;
rr:objectMap [
rr:parentTriplesMap ex-resource:TriplesMapCountry;
rr:joinCondition [
rr:child "idCountry";
rr:parent "idCountry"; ];];];
rr:predicateObjectMap [
rr:predicate ex-resource:dimYear;
rr:objectMap [
rr:parentTriplesMap ex-resource:TriplesMapYear;
rr:joinCondition [
rr:child "idTime";
rr:parent "idTime"; ];];];
rr:predicateObjectMap [
rr:predicate ex-property:numberResidents;
rr:objectMap [rr:column "numberResidents"]; ];
rr:predicateObjectMap [
rr:predicatesdmx-attribute:unitMeasure;
rr:objectMap [rr:constant<http://dbpedia.org/ontology/Person>]; ].
```

Tabela 4 - Mapeamento do cubo “Residentes”.

```
@prefix rdfs:              <http://www.w3.org/2000/01/rdf-schema#>.
@prefix rdf:               <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
@prefixrr:                 <http://www.w3.org/ns/r2rml#>.
@prefixex-class:           <http://purl.org/GovDataCube/classes/>.
@prefix ex-resource: <http://purl.org/GovDataCube/resources/>.
@prefix d2rq:          <http://www.wiwiss.fu-berlin.de/suhl/bizer/D2RQ/0.1#>.

ex-resource:TriplesMapRace a rr:TriplesMap;
d2rq:dataStorage ex-resource:databaseResidents;
rr:logicalTable [ rr:sqlQuery """
                        SELECT DISTINCTmd5(race) as race_md5, race, idPerson
                        FROM dimperson
                         WHERE race IS NOT NULL"""; ];
rr:subjectMap [
rr:template "http://purl.org/GovDataCube/resources/Race/{race_md5}";
rr:classex-class:Race;            ];
rr:predicateObjectMap [
rr:predicaterdfs:label;
rr:objectMap [ rr:column "race"; rr:language "pt" ];].
```

Tabela 5 - Mapeamento da dimensão *Race.*

Uma das características do Catálogo é a reutilização de dimensões. A reutilização das dimensões diminui o número de triplas

PUC-Rio - Certificação Digital Nº 1121837/CA

armazenadas no Catálogo e permite a padronização das triplas da mesma dimensão em diferentes cubos.

A Tabela 4 descreve o mapeamento da dimensão “*Race*” do cubo “*Residentes*” e ilustra a questão da reutilização.

É importante observar que cubos podem usar dimensões que são armazenadas em um RDB diferente do RDB onde o cubo é armazenado. Esta é a razão pela qual foi necessário usar a linguagem D2RQ.

3.4

Resumo do capítulo

Este capítulo apresentou a arquitetura do mediador para cubos de dados interligados. O capítulo detalhou o Catálogo de descrições de cubos de dados interligados e apresentou um exemplo de descrição de um cubo de dados.

PUC-Rio - Certificação Digital Nº 1121837/CA

4.

Módulo Enriquecedor de Cubos de Dados

4.1

Visão geral do Módulo Enriquecedor

4.1.1

Formas de enriquecimento

Observando a Figura 1 do Capítulo 2 na página 23, a descrição de um cubo de dados no *Data Cube vocabulary* contém diversos tipos de *componentes*, identificados pela diversas especializações da propriedade *qb:ComponentProperty*:

- Dimensão, identificada pela propriedade *qb:DimensionProperty*
- Atributo, identificada pela propriedade *qb:AttributeProperty*
- Medida, identificada pela propriedade *qb:MeasureProperty*
- Propriedade codificada, identificada pela propriedade *qb:CodedProperty.*

Além destas, consideramos também como um componente de um cubo de dados:

- As instâncias que compõem o domínio de uma dimensão

Dentro de certas condições, cada um destes componentes pode ter a sua descrição aprimorada associando o componente a um recurso

PUC-Rio - Certificação Digital Nº 1121837/CA

externo através de *owl:sameAs*. Chamamos a este processo de *enriquecimento* do componente.

Um componente pode ser enriquecido ou não de acordo com o seu *tipo*, definido da seguinte forma:

- Tipo 1: O componente:
  - É identificado por uma URI.
  - Possui um rótulo no idioma inglês e, opcionalmente, outras propriedades de tal forma que seja possível alinhá-lo automaticamente com outros recursos.

Este tipo de componente será submetida ao processo de enriquecimento automático.

- Tipo 2: O componente:
  - É identificado por uma URI
  - Não possui um rótulo no idioma inglês ou outras propriedades que permitam alinhá-lo automaticamente com outros recursos

Este tipo de componente será submetida ao processo de enriquecimento manual.

- Tipo 3: O componente:
  - É uma instância do domínio de uma dimensão.
  - Não tem uma URI que o identifique.

Este tipo de componente não pode ser enriquecido.

Como exemplos podem-se consultar as tabelas apresentadas na seção 3.3. O componente do Tipo 1é a dimensão *Country* definida na

PUC-Rio - Certificação Digital Nº 1121837/CA

Tabela 1 e na Tabela 3. A dimensão *Race* é do Tipo 2, definida na Tabela 4. Por último, as dimensões *Age* e *Year* são do Tipo 3 e são definidas explicitamente na Tabela 1.

O processo de enriquecimento está relacionado diretamente com o tipo de dado. O enriquecimento automático depende das diferentes propriedades que o componente tenha (rótulo em inglês, sinônimo, etc.), que serão consideradas pelo modulo de recomendação. No caso do enriquecimento manual, o componente a ser enriquecido carece de propriedades informativas e, por tanto, é necessária a intervenção do usuário e a experiência dele para dirigir o processo.

PUC-Rio - Certificação Digital Nº 1121837/CA

4.1.2

Componentes do Módulo Enriquecedor

O módulo Enriquecedor subdivide-se em dois módulos: Enriquecedor Automático e Enriquecedor Manual.

O Enriquecedor Automático utiliza os serviços oferecidos pela ferramenta LIMES, através da configuração de um arquivo xml que deve ser completado pelo usuário.Para completar este arquivo, o usuário tem duas opções: (1) solicitar o serviço de localização de fontes de dados, oferecido pelo módulo de recomendação de fontes de dados; (2) inserir manualmente os dados da fonte de dados a ser usada, ou seja, o *EndPoint*, o vocabulário e a classe.

O Enriquecedor Manual, como o nome indica, auxilia o usuário a enriquecer manualmente os componentes de um cubo. Os serviços de busca utilizados são o Sindice[2] e aqueles fornecidos pelo Wordnet[3] e pelo Geonames[4], mas o usuário pode escolher aquele que lhe seja mais conveniente.

[2] http://sindice.com/search

[3] http://wordnetweb.princeton.edu/perl/webwn

[4] http://www.geonames.org/

O resultado do enriquecimento, ou seja, as triplas *owl:sameAs* geradas, são armazenadas em um servidor de triplas (no caso o Virtuoso), juntamente com um arquivo de proveniência.

4.2

Enriquecedor Automático

O Enriquecedor Automático trabalha com o serviço de recomendação de fontes de dados e este é habilitado de acordo com as necessidades do usuário.

Para fins explicativos, chamaremos de Enriquecedor Automático (a) aquele que tem habilitado o serviço de recomendação de fontes de dados e Enriquecedor Automático (b) aquele onde o usuário insere os dados necessários à execução do LIMES.

As telas de apresentação da ferramenta Enriquecedor Automático (a) e (b) encontram-se no Apêndice 7.

4.2.1

Objetivos específicos, casos de uso e requisitos

Os objetivos específicos do Enriquecedor Automático são:

- Selecionar o componente que será enriquecido através da geração de triplas *owl:sameAs*. Esse dado será usado no processo de configuração da ferramenta LIMES.

- Selecionar o *EndPoint Target* adequado.
  No caso do Enriquecedor Automático (a), deverá ser escolhida uma das fontes candidatas oferecidas pelo serviço do módulo de recuperação das fontes de dados.

PUC-Rio - Certificação Digital Nº 1121837/CA

No caso do Enriquecedor Automático (b), o usuário deve inserir os dados necessários (o *EndPoint Target*, o vocabulário e a classe).

Estes dados serão usados no processo de configuração do serviço da ferramenta LIMES.

- Gerar interligações entre dados do Catálogo e bancos de dados externos, configurando e executando o serviço da ferramenta LIMES.

- Armazenar as triplas *owl:sameAs* geradas pelo serviço da ferramenta LIMES no Virtuoso.

- Armazenar a configuração usada no serviço da ferramenta LIMES no Virtuoso, de acordo com o processo feito, para que capture a proveniência do processo.

A Figura 3 apresenta o diagrama de caso de uso do Enriquecedor Automático. Os objetivos específicos traduzem-se nos requisitos básicos apresentados na Tabela 5.

PUC-Rio - Certificação Digital Nº 1121837/CA

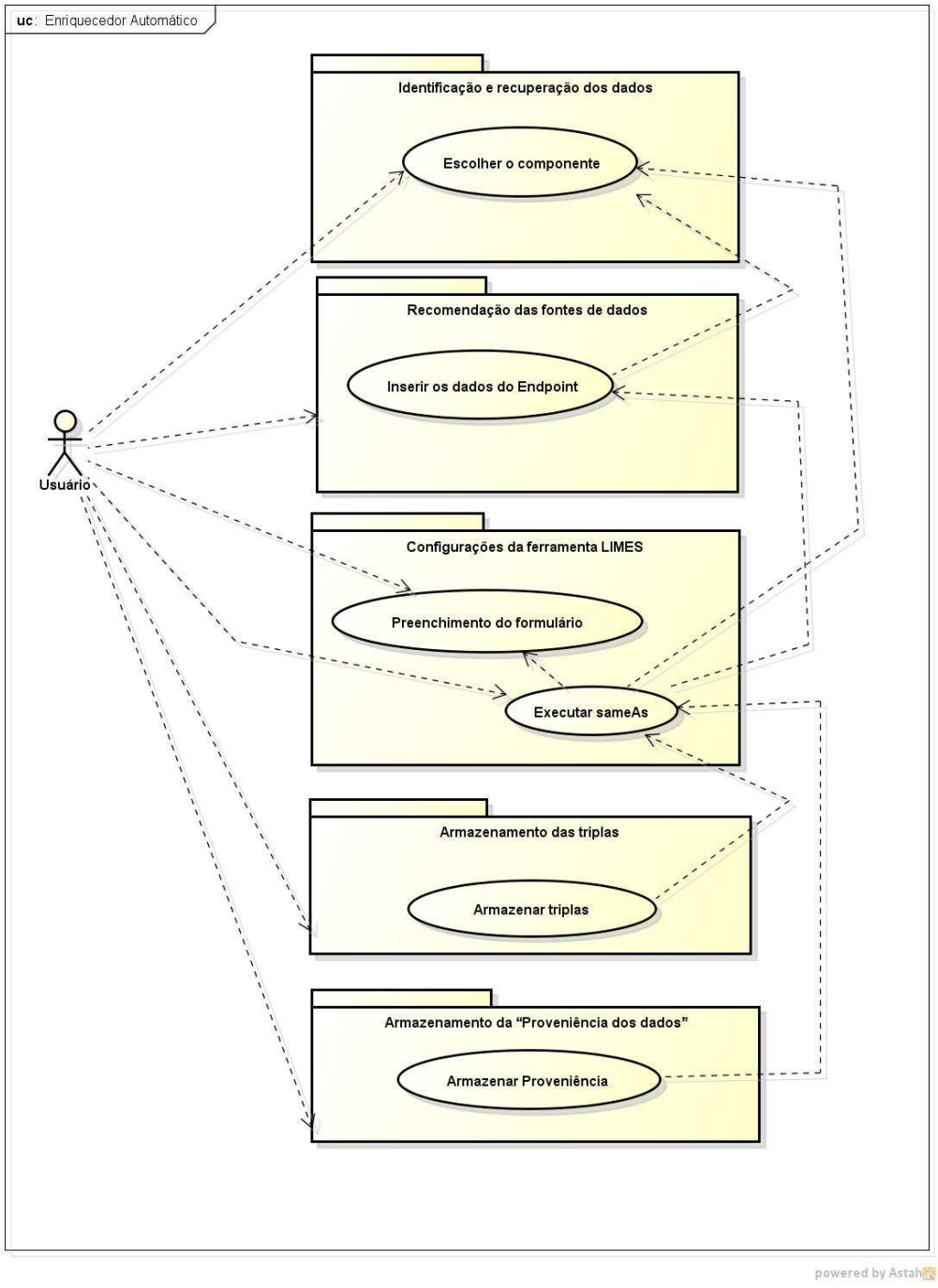


Figura 3 - Diagrama de Caso de Uso do Enriquecedor Automático.

PUC-Rio - Certificação Digital Nº 1121837/CA

	Condição de entrada		Processo	Condição de saída	
# de Processo	**Solicitação**	**Entrada**		**Saída do módulo utilizado**	**Módulo que utilizará a informação gerada pelo processo**
1	Requisição de dados ao Catálogo	--	Identificação e recuperação dos dados	Componente	Processo 2(A.1)ou 2(B) do Enriquecedor Automático
2(A.1)	Requisição de recomendação de fontes de dados ao Módulo de Recomendação	Componente.	Recomendação das fontes de dados	Fontes recomendadas	Processo 2(A.2) do Enriquecedor Automático
2(A.2)	Requisição do *owl:sameAs*	Fontes recomendadas, Algoritmo, Threshold	Configuração do serviço LIMES	Triplas *owl:sameAs* Proveniência dos dados	Processo 3 e 4do Enriquecedor Automático
2(B)	Requisição do *owl:sameAs*	Target, Vocabulário, Classe, Algoritmo, Threshold	Configuração do serviço LIMES	Triplas *owl:sameAs* Proveniência dos dados	Processo 3 e 4 do Enriquecedor Automático
3	Armazenamento das triplas *owl:sameAs*	Triplas *owl:sameAs*	Armazenamento das triplas	Grafo das Triplas *owl:sameAs*	Virtuoso
4	Armazenamento da proveniência dos dados	Proveniência dos dados	Armazenamento da proveniência dos dados	Grafo da Proveniência dos dados	Virtuoso

2(A.1) e 2(A.2): Processo obrigatório para o Enriquecedor Automático (a), com o serviço de Recomendação de Fontes de Dados.
2(B):Processo obrigatório para o Enriquecedor Automático (b), sem o serviço de Recomendação de Fontes de Dados.

Tabela 6 - Visão geral dos requisitos básicos dos processos da Ferramenta "Enriquecedor Automático".

4.2.2

Processos e arquitetura

Esta seção descreve os processos implementados pelo Enriquecedor Automático no caso (a) e no caso (b), resumidos na Figura 4.

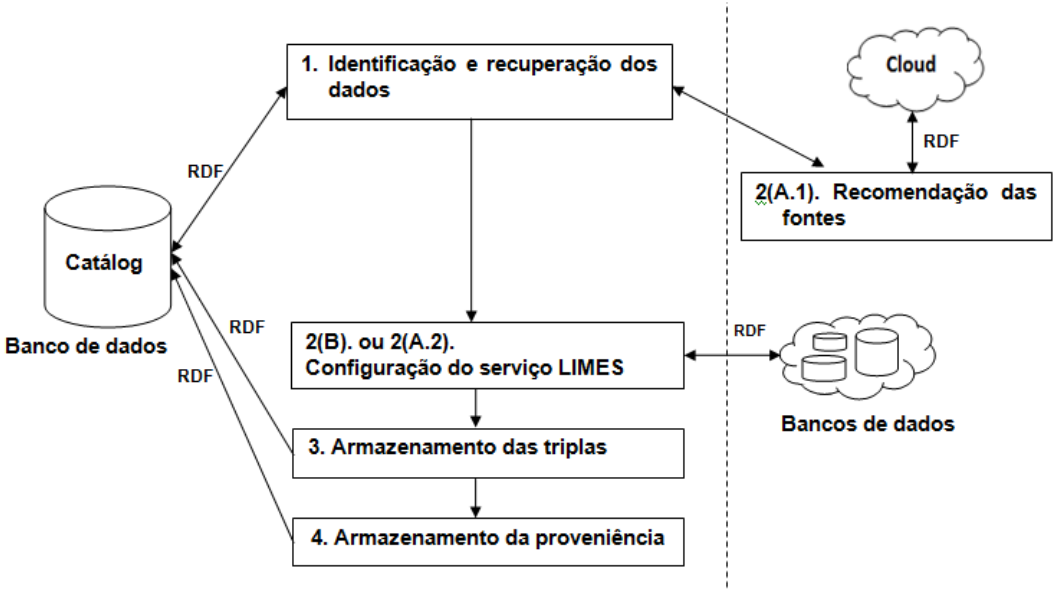


Figura 4 - Visão geral dos processos da descoberta das interligações *owl:sameAs* do Enriquecedor Automático(a) e (b).

Identificação e recuperação dos dados (1)

Processo que identificará e recuperará o componente do Catálogo com o qual será configurado o serviço da ferramenta LIMES.

A partir da execução deste processo, o usuário fará a escolha entre duas opções. A primeira que iniciará o processo de requisição de recomendações das fontes de dados no módulo de recomendação de fontes, processo chamado 2(A.1) na Figura 4. A segunda opção é aquela onde o usuário terá que inserir os dados de maneira manual, processo chamado 2(B) na Figura 4.

Recomendação das fontes de dados2(A.1)

Processo que executa o módulo de recomendação (Herrera et al. 2012) de fontes para identificar as possíveis fontes de dados com

PUC-Rio - Certificação Digital Nº 1121837/CA

as quais se fará a construção das triplas *owl:sameAs* do módulo Enriquecedor.

Configuração do serviço LIMES2(A.2)

Processo para configuração do arquivo do serviço da ferramenta LIMES, com os dados recuperados do módulo de recomendação das fontes de dados (o *EndPoint Target* e a classe a ser consultada) assim como os dados inseridos pelo usuário na tela (o algoritmo e as métricas de similaridade).

Depois de executar o serviço da ferramenta LIMES, o processo executa as tarefas de: identificação e apresentação das triplas *owl:sameAs* e identificação da informação relativa ao processo de busca (proveniência dos dados).

Configuração do serviço LIMES 2(B)

Processo para configuração do serviço da ferramenta LIMES, com os dados que serão inseridos pelo usuário na tela: (1) configurar o *endPoint source* (Dado recuperado da seleção do usuário no processo 1); (2) configurar o *EndPoint Target,* o vocabulário a ser usado e da classe do *EndPoint Target* identificada; (3) especificar o algoritmo a usar; (4) especificar as métricas de similaridade.

Internamente será feita a configuração do arquivo de saída, onde serão armazenadas as triplas *owl:sameAs* e apresentadas na tela. Também serão identificadas as informações relativas ao processo de busca (Proveniência dos dados).

Armazenamento das triplas no Virtuoso (3)

As triplas *owl:sameAs* serão armazenadas no Virtuoso.Essas triplas são o resultado da execução da ferramenta LIMES.

PUC-Rio - Certificação Digital Nº 1121837/CA

Armazenamento da proveniência dos dados (4)

Os dados informativos relacionados ao processo de comparação feito pelo serviço da ferramenta LIMES, chamados dados de proveniência, serão armazenados no Virtuoso.

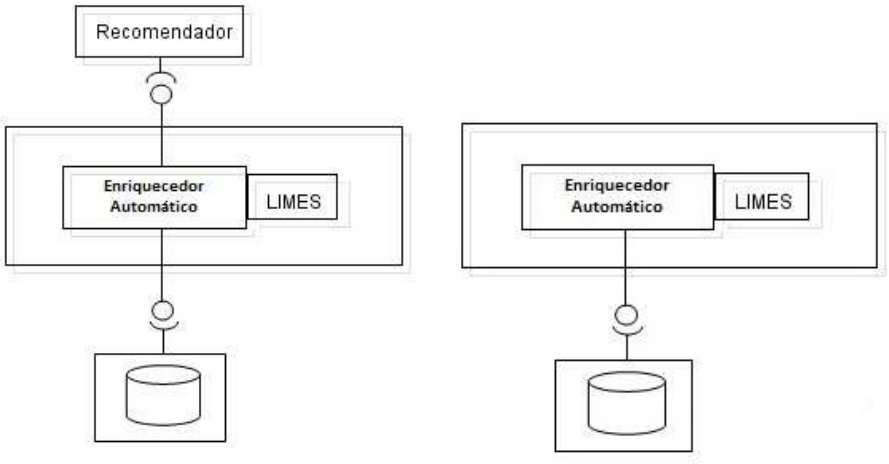


Figura 5 - Arquitetura do Enriquecedor Automático: (a), com o serviço de Recomendação; (b), sem o serviço de Recomendação.

A Figura 5 apresenta a arquitetura do Enriquecedor Automático (a) com o serviço de recomendação de fontes de dados, e do lado direito apresenta a arquitetura do Enriquecedor Automático (b) sem o serviço de recomendação de fontes de dados.

O Apêndice 5 ilustra um arquivo como proveniência de dados do exemplo *Country*.

PUC-Rio - Certificação Digital Nº 1121837/CA

4.3

Enriquecedor Manual

4.3.1

Objetivo, casos de uso e requisitos

O Enriquecedor Manual tem como objetivo principal complementar o Enriquecedor Automático, permitindo ao usuário definir manualmente triplas *owl:sameAs*.

A Figura 6 apresenta o diagrama de caso de uso do Enriquecedor Manual.

Os objetivos traduzem-se nos requisitos apresentados na Tabela 6.

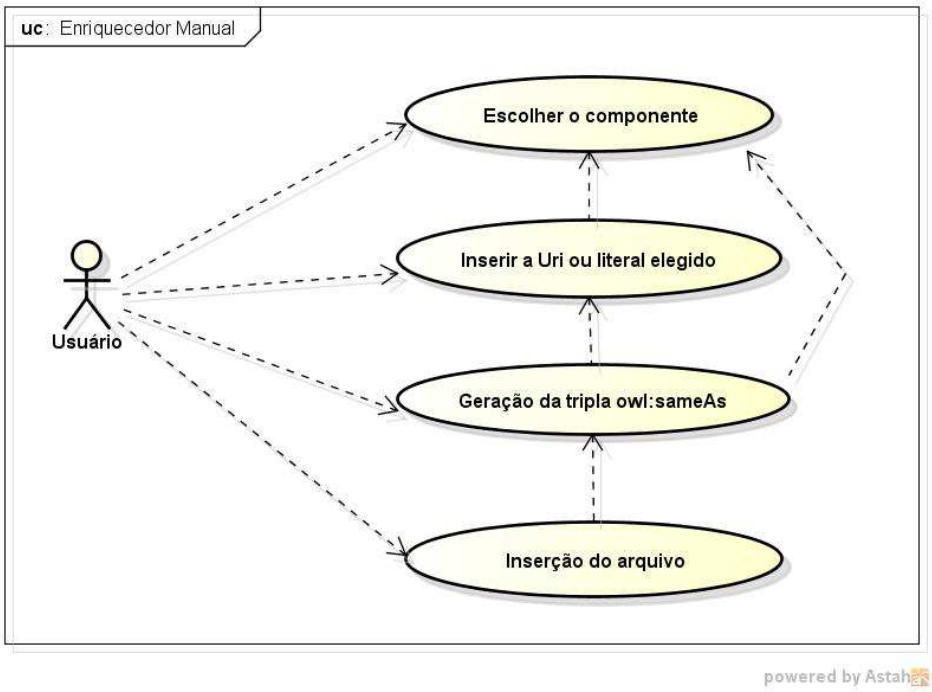


Figura 6 - Diagrama de Caso de Uso do Enriquecedor Manual.

PUC-Rio - Certificação Digital Nº 1121837/CA

	Condição de entrada		Processo	Condição de saída	
# de Processo	**Solicitação**	**Entrada**		**Saída do módulo utilizado**	**Módulo que utilizará a informação gerada pelo processo**
1	Requisição do componente	--	Escolha do componente	Componente	Processo (2) do Enriquecedor Manual
2	Requisição da URI, na Web http://sindice.com/search/	Componente	Inserção do URI ou literal	URI ou literal inserido	Processo (3) do Enriquecedor Manual
3	Inserção das triplas *owl:sameAs*	Triplas *owl:sameAs por ser criada*	Inserção da tripla *owl:sameAs*	Lista de triplas *owl:sameAs*	Processo (4) do Enriquecedor Manual
4	Armazenamento das triplas *owl:sameAs*	Lista de triplas *owl:sameAs*	Geração do arquivo *sameAsManual*	Grafo de Triplas *owl:sameAs*	Virtuoso

Tabela 7 - Visão geral dos requisitos básicos dos processos da Ferramenta "Enriquecedor Manual".

4.3.2

Processos e arquitetura

Esta seção descreve os processos implementados pelo Enriquecedor Manual, resumidos na Figura 7. A tela de apresentação do Enriquecedor Manual encontra-se no Apêndice 8.

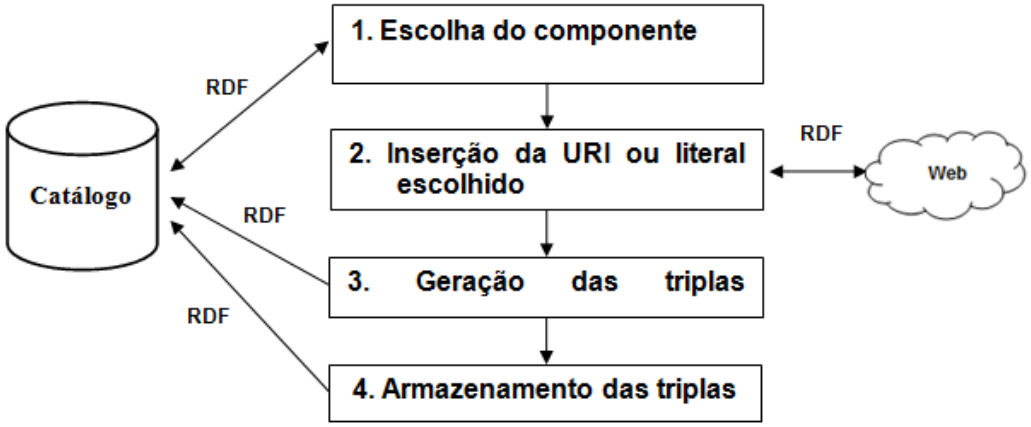


Figura 7 - Visão geral dos processos da descoberta das interligações *owl:sameAs* do Enriquecedor Manual.

Escolha do componente (1)

Processo que identificará o componente a ser eleito pelo usuário na construção das triplas *owl:sameAs*.

Inserção da URI ou literal escolhido (2)

Para a construção do *owl:sameAs* o usuário deverá consultar diferentes recursos na Web[5].

Geração da tripla owl:sameAs (3)

Quando o usuário terminar de construir a tripla *owl:sameAs*, poderá inserir a mesma na lista *sameAsManual*, que será construída de acordo com os requerimentos do usuário na criação de novas triplas com as instâncias da lista apresentada na tela.

[5]http://sindice.com/search/, http://wordnetweb.princeton.edu/perl/webwn, http://www.geonames.org/

PUC-Rio - Certificação Digital Nº 1121837/CA

Inserção do arquivo sameAsManual (4)

Quando o usuário determinar que todas as instâncias já foram enriquecidas, será possível criar o arquivo *sameAsManual*.O arquivo criado será inserido no Virtuoso para futuras consultas.

4.4

Resumo do capítulo

Este capítulo apresentou uma visão geral do módulo enriquecedor, as formas de enriquecimento que podem ser feitas, assim como a descrição de cada um dos componentes que dele fazem parte. Em seguida, detalhou os componentes.

PUC-Rio - Certificação Digital Nº 1121837/CA

5.

Exemplos de Uso do Enriquecedor de Cubos de Dados

5.1

Enriquecimento automático

Recorde do capítulo 3 que a descrição de um cubo de dados no *Data Cube vocabulary* contém diversos tipos de componentes, identificados pelas especializações da propriedade *qb:ComponentProperty*:

- Dimensão, identificada pela propriedade *qb:DimensionProperty*
- Domínio de uma dimensão, contendo as instâncias de uma propriedade
- Atributo, identificada pela propriedade *qb:AttributeProperty*
- Medida, identificada pela propriedade *qb:MeasureProperty*
- Propriedade codificada, identificada pela propriedade *qb:CodedProperty*

De acordo com a discussão da Seção 4.1.1, um componente de um cubo de dados, para ser enriquecido automaticamente, deve ter as seguintes características:

- É identificado por uma URI.
- Possui um rótulo no idioma inglês.
- Possui, opcionalmente, outras propriedades.

Estas características são necessárias para executar o serviço LIMES para gerar o *owl:sameAs*.

PUC-Rio - Certificação Digital Nº 1121837/CA

O resto desta seção descreve os passos do processo de enriquecimento automático com o serviço de recomendação de fontes de dados para as dimensões com as características de enriquecimento do Tipo1, definido na seção 4.1.1.

5.1.1

Passos do processo de enriquecimento automático

Passo 1 - Identificação e recuperação do componente a ser enriquecido.

O usuário escolhe o componente do cubo a ser enriquecido, ou seja, para o qual se deseja criar triplas *owl:sameAs* com recursos externos.

Suponha no que se segue que o componente escolhido seja uma dimensão, representada por uma classe e suas instâncias. Suponha ainda que o usuário já tenha escolhido as propriedades das instâncias a serem utilizadas pelo Módulo de Recomendação de Fontes.

O Enriquecedor Automático apresenta então na tela a etiqueta (*label)* e as URIs das instâncias a serem entregues ao Módulo de Recomendação de Fontes de Dados.

O resultado do passo 1 é um arquivo com os enriquecimentos encontrados para o componente escolhido. Por exemplo, o Apêndice 1 mostra um arquivo com as instâncias da dimensão *Country* enriquecidas com o rotulo no idioma inglês.

Passo 2(A.1) - Recomendação das fontes de dados.

A classe e as instâncias da classe, com as propriedades escolhidas, são entregues ao Módulo de Recomendação de Fontes.

O Módulo de Recomendação fará a busca das possíveis fontes de dados a serem utilizadas.

O resultado de passo 2(A.1) é um arquivo, conforme gerado pelo Módulo de Recomendação de fontes de dados. O Apêndice2 contém um exemplo.

PUC-Rio - Certificação Digital Nº 1121837/CA

Passo 2(A.2) - Configuração e execução do serviço da ferramenta LIMES.

As fontes de dados recomendadas pelo Módulo de Recomendação são carregadas no *combobox* da aplicação e apresentadas ao usuário. O usuário deve selecionar uma delas.

O usuário passa então a especificar o algoritmo a ser usado pelo serviço da ferramenta LIMES para comparar cadeias de caracteres, bem como dois parâmetros ou métricas de similaridade (*threshold parameters*), um para a geração do arquivo *Acceptance* e outro para a geração do arquivo *Review*. Os parâmetros devem estar entre 0 e 1.

A classe (ou seja, a dimensão) escolhida, com as suas instâncias, a fonte de dados selecionada, o algoritmo e os parâmetros inseridos, são passadas para o serviço da ferramenta LIMES, que gera então triplas *owl:sameAs* em dois arquivos, chamados *Acceptance* e *Review*. O arquivo *Acceptance* é considerado como a resposta do serviço e apresentado na tela.

O usuário pode ver as equivalências encontradas, armazenadas no arquivo *Acceptance*, validando-as ou não.

Caso não seja uma resposta adequada, o usuário pode chamar novamente o serviço da ferramenta LIMES com uma nova configuração.

O Apêndice 3 contém um exemplo do arquivo gerado por este processo e o arquivo de configuração do serviço da ferramenta LIMES, para a dimensão *Country* com o algoritmo *Levenshtein* e as métricas de similaridade (*threshold parameters*) de 1 e de 0.95.

Passo 3 - Armazenamento das triplas no Virtuoso.

O arquivo *Acceptance*, após validação, será armazenado no Virtuoso para depois ser consultado a partir de uma consulta em SPARQL.

O Apêndice 4 ilustra um arquivo *Acceptance*, assim como o grafo gerado pelo processo de armazenamento no Virtuoso.

Passo 4 - Armazenamento da proveniência dos dados.

PUC-Rio - Certificação Digital Nº 1121837/CA

Este processo armazena no Virtuoso a configuração do serviço da ferramenta LIMES.

Por exemplo, o Apêndice 5ilustra um arquivo como proveniência dos dados do exemplo *Country*. Este arquivo contém, sob forma de triplas, os dados resumidos na Tabela 8.

Variáveis do arquivo	Conteúdo das variáveis
Criado por	Limes
Data	2013-06-16T06:00TZD
Descrição	"Calculo de similaridade, entre dois bancos de dados"
Componente	Country
EndPoint Target	<http://dbpedia.org/sparql>
EndPoint Source	<http://purl.org/GovDataCube/listaSourceI1I2>
Métrica (Algoritmo)	Levenshtein
Parâmetro de aceitação (Threshold Acceptance)	1
Parâmetro de revisão (Thresold Review)	0.95

Tabela 8 - Dados que são apresentados no arquivo proveniência de dados.

As seguintes propriedades *rdf:type erdf:label* do vocabulário rdf:(*Resource Description Framework*) e as propriedades *dc:creator*, *dc:date*, *dc:description* e *dc:relation* do vocabulário dc: (*Dublin Core*) serão utilizadas para representar proveniência.

Para o caso de eleger a opção de não usar o Módulo de Recomendação de Fontes de Dados, o passo 2(A.1) e 2(A.2) se reduzem a um passo só, chamado 2(B). Este passo, com ajuda do usuário na tarefa de inserção dos dados, vai gerar como resultado o arquivo de configuração do serviço da ferramenta LIMES.

PUC-Rio - Certificação Digital Nº 1121837/CA

5.1.2

Exemplos de uso do Enriquecedor Automático

Esta seção apresenta exemplos do resultado obtido pelo Enriquecedor Automático com várias configurações do serviço da ferramenta LIMES, variando:

- O algoritmo para comparação de cadeias de caracteres (o Algoritmo de *Levenshtein* ou o algoritmo *Qgram*).
- Os parâmetros de aceitação e de revisão ou métricas de similaridade (*threshold parameters*).

5.1.2.1

Exemplos utilizando o Algoritmo *Levenshtein*

A dimensão a ser analisada é *Country*, com 69 instâncias e o *EndPoint Target* utilizado é dbpedia.org.

Parâmetros inseridos no LIMES (Threshold)		**Dados aceitos**	**Dados para revisar**	**Tempo**
Parâmetro de aceitação	**Parâmetro de revisão**			
1	0.95	42	0	12.823 seg.
0.9	0.85	53	0	17.023 seg.
0.8	0.75	54	3	12.662 seg.
0.7	0.65	65	22	15.443 seg.
0.6	0.55	146	109	14.673 seg.

Tabela 9 - Casos de Uso – Tipo 1. Geração do *owl:sameAs* Automático com variação dos parâmetros de aceitação e revisão (*Threshold).*

5.1.2.2

Exemplos utilizando o Algoritmo *Qgrams*

A dimensão a ser analisada é *Country*, com 69 instâncias e o *EndPoint Target* utilizado é dbpedia.org.

PUC-Rio - Certificação Digital Nº 1121837/CA

Parâmetros inseridos no LIMES (Threshold)		Dados aceitos	Dados para revisar	Tempo
Parâmetro de aceitação	**Parâmetro de revisão**			
0,7	0.65	11	8	10.830 seg.
0.6	0.55	32	11	15.716 seg.
0.5	0.45	52	2	10.010 seg.
0.48	0.38	52	16	10.321 seg.
0.4	0.35	63	18	10.255 seg.

Tabela 10 - Casos de Uso – Tipo 1. Geração do *owl:sameAs* Automático com variação dos parâmetros de aceitação e revisão (*Threshold).*

5.1.3

Análise do uso do Enriquecedor Automático

O objetivo desta seção é identificar os parâmetros ideais em cada um dos casos analisado, ilustrando assim o uso do Enriquecedor Automático. Os indicadores usados para a análise são:

- Número de triplas pertencentes à dimensão
- Número de acertos (no arquivo *Acceptance*)
- Taxa de acerto
- Tempo de execução do módulo

CP	pa	pr	Total	LIMES		t	Válidos	Erro	Revocação	Precisão
				Aceitos	Revisão					
a	1	0.95	69	42	0	12,823	42	0	0,608695652	0
b	0.9	0.85	69	53	0	17,023	53	0	0,768115942	0
c	0.8	0.75	69	54	3*	12,662	54	0	0,782608696	0
d	0.7	0.65	69	65	22	15,443	55	10	0,797101449	0,144927536
f	0.6	0.55	69	146	109	14,673	56	90	0,811594203	1,304347826

3* Os 3 com erro
CP:Combinações dos parâmetros
pa: Parâmetro de Aceitação
pr: Parâmetro de revisão
t: Tempo
Válidos: Quantidade de triplas válidas das triplas aceitas
Erro: Quantidade de triplas erradas das triplas aceitas
Combinação aceitável

Country
Levenshtein

Tabela 11 - Identificação dos parâmetros ideais para execução do Enriquecedor Automático para a classe *Country* com o algoritmo *Levenshtein*.

PUC-Rio - Certificação Digital Nº 1121837/CA

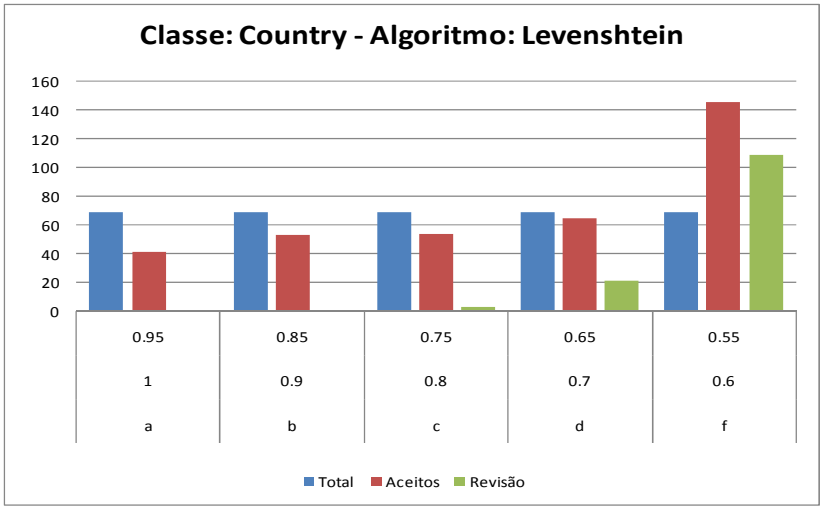


Figura 8- Apresentação dos resultados da geração do *owl:sameAs* da classe *Country* e o algoritmo *Levenshtein*.

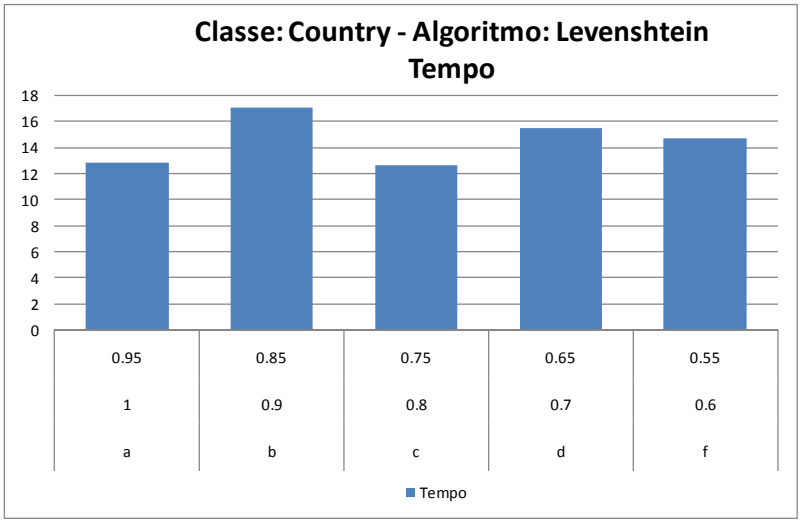


Figura 9 - Apresentação dos tempos usados para a geração do *owl:sameAs* da classe *Country* e o algoritmo *Levenshtein*.

CP	pa	pr	Total	LIMES		t	Válidos	Erro	Revocação	Precisão
				Aceitos	Revisão					
a	0.7	0.65	69	11	8	10,83000	11	0	0,15942029	0
b	0.6	0.55	69	32	11*	15,71600	32	0	0,463768116	0
c	0.5	0.45	69	52	2*	10,01000	52	0	0,753623188	0
d	0.48	0.38	69	52	16*	10,32100	52	0	0,753623188	0
f	0.4	0.35	69	63	18	10,25500	56	7	0,811594203	0,101449275

11* Todos ok
2* Os 2 com erro
16* Tem 9 URIS não repetidas, 2 ok e 14 erros
CP:Combinações dos parâmetros
pa: Parâmetro de Aceitação
pr: Parâmetro de revisão
t: Tempo
Válidos: Quantidade de triplas válidas das triplas aceitas
Erro: Quantidade de triplas erradas das triplas aceitas
Combinação aceitável

Country
q-gram

Tabela 12 - Identificação dos parâmetros ideais para execução do Enriquecedor Automático para a classe *Country* com o algoritmo *Qgram*.

PUC-Rio - Certificação Digital Nº 1121837/CA

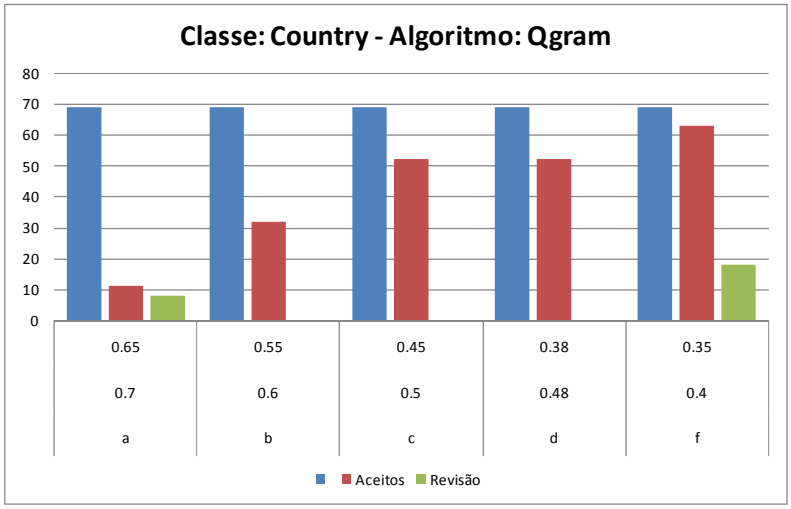


Figura 10 - Apresentação dos resultados da geração do *owl:sameAs* da classe *Country* e o algoritmo *Qgram*.

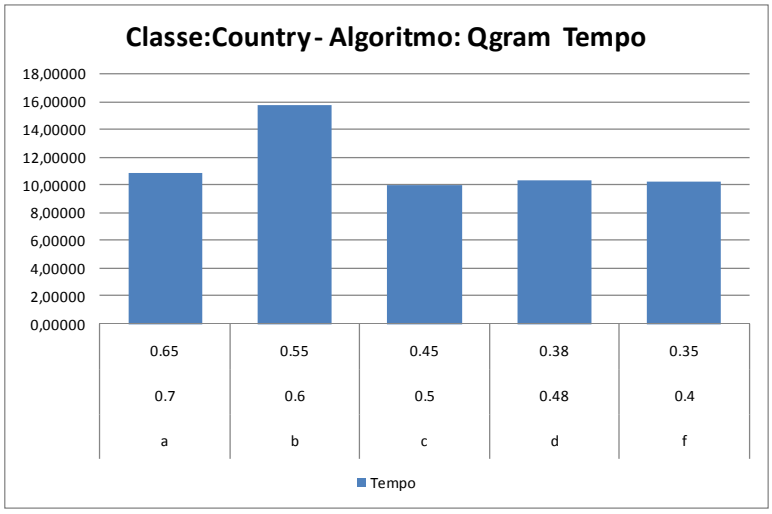


Figura 11 - Apresentação dos tempos usados para a geração do *owl:sameAs* da classe *Country* e o algoritmo *Qgram.*

PUC-Rio - Certificação Digital Nº 1121837/CA

5.2

Enriquecimento manual

O enriquecimento manual pode ser feito para dois casos específicos.

De acordo com a discussão da Seção 4.1.1, um componente de um cubo de dados deve ser enriquecido manualmente se possui as seguintes características:

- É identificado por uma URI
- Não possui um rótulo no idioma inglês ou outras propriedades que permitam alinhá-lo automaticamente com outros recursos

Alternativamente, quando é primordial completar o processo feito pela ferramenta automática de enriquecimento e ainda se tem instâncias sem a geração da tripla *owl:sameAs*, o enriquecimento manual deve ser utilizado.

O resto desta seção descreve os passos do processo de enriquecimento manual para as dimensões com as características de enriquecimento do Tipo 2, apresentada na seção 4.1.1.

PUC-Rio - Certificação Digital Nº 1121837/CA

5.2.1

Passos do processo de enriquecimento manual

Passo 1 - Escolha do componente.

O usuário escolhe o componente do cubo a ser enriquecido.

O Enriquecedor Manual apresenta na tela a etiqueta (*label)* e a URI das instâncias do componente escolhido.

Passo 2 - Inserção da URI ou literal escolhido.

O usuário deve eleger um elemento. A escolha deve ser feita sobre a lista apresentada na tela.

Para este passo a interface principal do Enriquecedor Manual tem a estrutura de uma tripla: a instância escolhida é apresentada de um lado do termo "*owl:sameAs*" e do outro, um *textbox* vazio.

O usuário deve completar o *textbox* vazio com uma URI ou um literal, consultando diferentes recursos na Web, para completar a tripla *owl:sameAs*. Um exemplo deste tipo de serviço é o SINDICE (disponível em http://sindice.com/search)

Passo 3 - Geração da tripla *owl:sameAs*

Este processo, insere a tripla *owl:sameAs* gerada em uma lista chamada *sameAsManual*.

O usuário pode repetir o processo para outras instâncias da dimensão.

Passo 4 – Inserção do arquivo *sameAsManual*

Quando o usuário terminar o processo de inserir as triplas na lista, as triplas criadas serão armazenadas no Virtuoso.

O Apêndice 6 apresenta exemplos do Enriquecimento Manual dos componentes *Race* e *Country*.

5.2.2

Exemplos de uso do Enriquecedor Manual

Nestes exemplos, a taxa de acertos é a quantidade de triplas *owl:sameAs* que são geradas pelo usuário, levando em conta o conceito que elas representam e ligando as instâncias com informação relevante.

A Tabela 11 mostra o componente *Race que pertence ao Tipo 2* e mais outros dois componentes *Religion* e *Sex*. Casos de uso para o cálculo dos acertos e dos tempos.

Caso de uso Nº	**1**	**2**	**3**
Dimensão	*Race*	*Religion*	*Sex*
# instâncias	6	74	2
Serviço	dbpedia.org	dbpedia.org Wordnet	dbpedia.org
Acertos	6	38	2
Sem informação	0	36	0
Tempo de execução*	8 min.15seg.	105 min. 28seg	1 min.20seg.

(*): O tempo depende da habilidade do usuário.

Tabela 13 - Caso de Uso - Tipo 2. Geração do *sameAsManual.*

No caso de utilizar o Enriquecedor Manual e aplicar ao caso de uso do Tipo 1, obteve-se os seguintes resultados:

PUC-Rio - Certificação Digital Nº 1121837/CA

Dimensão:	Country
# instâncias:	69
Serviço:	dbpedia.org
Acertos:	64
Sem informação:	5
Tempo de execução*:	63 min. 34 seg.

(*): O tempo depende da habilidade do usuário

Tabela 14 - Caso de Uso do Tipo 1 executado na Ferramenta "Enriquecedor Manual".

No caso de complementar ao processo automático feito pela ferramenta Enriquecedor Automático e considerando a combinação aceitável dos parâmetros de aceitação e revisão apresentadas nas Tabelas 9 e 10 respectivamente, se obteve os seguintes resultados.

Algoritmo:	*Levenshtein*	*Qgram*
# total de instâncias:	69	69
# instâncias trabalhadas:	54	52
# instâncias restantes a serem trabalhadas:	15	17
Serviço:	dbpedia.org	dbpedia.org
Acertos:	10	12
Sem informação:	5	5
Tempo de execução*:	11 min. 15 seg.	13 min. 10 seg.

(*): O tempo depende da habilidade do usuário

Tabela 15- Caso de Uso do Tipo 1 executado na Ferramenta "Enriquecedor Manual" para complementar o processo feito pela ferramenta "Enriquecedor Automático".

5.3

Comentários sobre o processo de enriquecimento

O módulo Catálogo faz parte do Mediador como é apresentado no capítulo 3. Todos os dados que são consultados pelo usuário estão armazenadas no Catálogo. Os dados que fazem parte do Catálogo ajudam a encontrar outros dados dentro do banco de dados local acessado via*Wrappers*.Assim, o Catálogo provê para o mediador a informação necessária a ser retornada ao usuário.

PUC-Rio - Certificação Digital Nº 1121837/CA

O processo de enriquecimento tem como objetivo adicionar novas informações relativas e relevantes ao Catálogo. Essas novas informações podem ser considerados atributos que são disponibilizados dentro do processo geral de consulta no Mediador.

O Enriquecedor encontra na Web os dados que podem ser relevantes para enriquecer a semântica das descrições de cubos, consistentemente com os princípios de dados interligados (*linked data).*Desta forma, a execução do enriquecedor ajuda o usuário a interpretar os cubos de dados.

Tão importante quanto os dados produzidos a partir de uma ação gerada sobre um banco de dados é a sua proveniência. Portanto, para atribuir maior valor a um dado produzido é necessário efetuar a ligação entre o resultado e a sua origem. Em outras palavras, capturar a sua proveniência.

Registrar a proveniência dos dados para uso futuro é necessário para interpretar os resultados, verificar a corretude do processo e rastrear a origem dos dados.

PUC-Rio - Certificação Digital Nº 1121837/CA

5.4

Resumo do capítulo

Este capítulo descreveu em detalhe a implementação do módulo Enriquecedor, com casos de uso de acordo com o tipo de componente a ser enriquecido. Estes casos de uso permitirão uma melhor compreensão do funcionamento do módulo em geral e dos processos que o compõem.

6.

Conclusão e trabalhos futuros

6.1

Conclusões e contribuições

Esta dissertação posiciona-se em um contexto mais amplo em que conjuntos de triplas distintos são interligados através do uso da propriedade *owl:sameAs*.

A dissertação apresentou uma ferramenta para facilitar a interligação de componentes de descrições de cubos de dados, criadas utilizando o *Data Cube vocabulary*, com recursos externos. As interligações são expressas através da propriedade *owl:sameAs* do vocabulário OWL.

O Enriquecedor Automático realiza o trabalho de busca de recursos semelhantes ao componente de um cubo de dados utilizando o serviço LIMES e o módulo de recomendação de fontes de dados. Para tal, o componente deve necessariamente ter propriedades que ajudem a busca. Por exemplo, para melhorar a resposta do serviço de recomendação, pode-se trabalhar na tradução para o inglês dos termos usados nos bancos de dados fonte. O uso de sinônimos também pode ser importante para a identificação de um conjunto de triplas candidatas.

É primordial, também, eleger os melhores parâmetros de semelhança (parâmetro de aceitação e parâmetro de revisão – *Threshold* do LIMES) para obter uma resposta adequada de comparação, considerando uma margem menor de erro que tenha a maior quantidade de triplas com semelhanças corretas.

O Enriquecedor Automático também armazena a proveniência das triplas *owl:sameAs* geradas, aplicando os princípios de dados interligados.

O Enriquecedor Manual auxilia o usuário a identificar recursos semelhantes e definir as interligações manualmente. Para o usuário construir uma tripla *owl:sameAs* adequada, o usuário deve ter conhecimento prévio do recurso a ser procurado na Web.

PUC-Rio - Certificação Digital Nº 1121837/CA

6.2

Trabalhos futuros

Como trabalho futuro sugere-se comparar os resultados obtidos utilizando o serviço LIMES com o serviço SILK, para o caso do Enriquecedor Automático. Os parâmetros a serem analisados seriam os tempos de execução e a qualidade das interligações geradas.

Sugere-se ainda utilizar um tradutor automático para o idioma inglês, para que se possam efetuar as buscas da semelhança com maior facilidade, evitando erros pela ausência da tradução da palavra para o idioma inglês. Esta ferramenta terá que ser executada no banco de dados local, para adicionar informação relevante ao dado e facilitar a busca do semelhante na Web.

Para gerar triplas *owl:sameAs* através do Enriquecedor Manual, recomenda-se acrescentar e configurar os serviços de buscas também dentro deste módulo.

Por fim, sugere-se dotar o Enriquecedor Manual de um dicionário de ontologias mais populares, por domínio de aplicação, que facilite a tarefa do usuário.

PUC-Rio - Certificação Digital Nº 1121837/CA

Referências Bibliográficas

[Alexander et al. 2011] Alexander, K. Cyganiak, R. Hausenblas, M. and Zhao, J. **Describing linked datasets with the void vocabulary.** W3C interest group note. Retirado em maio 17, 2013. http://www.w3.org/TR/void/, 2011.

[Auer et al. 2007] Auer, S. Bizer, C. Kobilarov, G. Lehmann, J. Cyganiak, R. and Ives, Z**. DBpedia: A nucleus for a web of open data**. The Semantic Web. 4825:722–735, 2007.

[Bhagdev et al. 2008] R. Bhagdev, S. Chapman, F. Ciravegna, V. Lanfranchi, and D. Petrelli. **Hybrid search: Effectively combining keywords and semantic searches**. In ESWC 2008, paginas 554-568, 2008.

[Bechhofer et al. 2004] Bechhofer, S. Van Harmelen, F. Hendler, J. Horrocks, I. McGuinness, D. L. Patel-Schneider, P. F. Stein, L. A. Dean, M. and Schreiber, G. **Owl web ontology language reference, w3c recommendation**. Retirado em maio 17, 2013. http://www.w3.org/TR/owl-ref/, 2004.

[Ben-David et al. 2010] D. Ben-David, T. Domany, and A. Tarem. **Enterprise data classication using semantic web technologies**. In ISWC2010, 2010.

[Berners-Lee 2009] Berners-Lee, T. **Linked data: Design issues**. Retirado em dezembro 10, 2010. http://www.w3.org/DesignIssues/LinkedData.html, 2009.

[Brickley & Miller 2010] Brickley, D. and Miller, L. **Foaf vocabulary specification 0.98. Namespace document**. Retirado em maio 17, 2013. http://xmlns.com/foaf/spec/, 2010. Marco Polo Edition.

[Buneman et al. 2001] Buneman, P. Khanna, S. E Tan W.-C. **Why and where: A characterization of data provenance**. Proc. 2001 International Conference on Database Theory, pags.4-6, 2001.

[Buneman e W. Tan, 2007] Buneman, P. and W.Tan. **Provenance in databases**. In Proceedings of ACM SIGMOD, 10:1171-1173, 2007.

[Cyganiak et al. 2011] Cyganiak, R. Hausenblas, M. and Mccuirc, E. **Official statistics and the practice of data fidelity.** Linking Government Data, p.135-151, Springer New York, 2011.

[Cyganiak et al. 2012] Cyganiak, R. Bizer, C. Garbers, J. Maresch, O. and Becker, C. **The d2rq mapping language**. Retirado em novembro 10, 2012. http://d2rq.org/d2rq-language/, 2012.

PUC-Rio - Certificação Digital Nº 1121837/CA

[Cyganiak et al. 2013] Cyganiak, R. Reynolds, D. and Tennison, J.. **The rdf data cube vocabulary**. W3C working draft. Retirado em maio 17, 2013. http://www.w3.org/TR/vocab-data-cube/, 2013.

[Das et al. 2010] Das, S. Sundara, S. and Cyganiak, R. **R2rml: Rdb to rdf mapping language**. W3C recommendation. Retirado em novembro 15, 2012. http://www.w3.org/TR/r2rml/, 2012.

[DCMI 2012] **Dcmi metadata terms**. Retirado em junho 10, 2013. http://dublincore.org/documents/dcmi-terms/, 2012.

[Freire et al., 2008] Freire, J. Koop, D. Santos, E. and Silva, C. **Provenance for computational tasks: A survey**. Computing in Science & Engineering, 10, 2008.

[Frey & Dueck 2007] B. J. Frey and D. Dueck. **Clustering by passing messages between data points.** Science. 315, 972-976. 2007.

[Foster et al. 03] Foster Jr, Blair and Evans, Dr Patricia. **Comparative q-gram Analysis of Gene Promoter Regions.** New Brunswick, 2003

[Halpin & Hayes 2010] Halpin, H. & Hayes, P. J. **When owl: sameAs isn't the same: An analysis of identity links on the semantic web Linked Data on the Web.** WWW2010 Workshop (LDOW2010), 2010

[Hausenblas et al. 2009] Hausenblas, M. Halb, W. Raimond, Y. Feigenbaum, L. and Ayers, D. **Scovo: Using statistics on the web of data**. In: Proceedings of the 6th European Semantic Web Conference on the Semantic Web: Research and Applications, ESWC 2009 Heraklion, p. 708- 722, Berlin, Heidelberg, 2009. Springer-Verlag.

[Heath & Bizer 2011] Heath, T. and Bizer, C. **Linked data: Evolving the web into a global data space**. Synthesis Lectures on the Semantic Web: Theory and Technology, p.1-136, 2011.

[Herrera et al.2012] Herrera, José Eduardo Breitman, Karin. **Arquitetura para Recomendação de Fontes de Dados RDF**. Rio de Janeiro, 2012. Dissertação de Mestrado. Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

[Klyne et al. 2004] Klyne, G. Carroll, J. J. and McBride, B. **Resource description framework (rdf): Concepts and abstract syntax**. W3C recommendation. Retirado em maio 18, 2013, fromhttp://www.w3.org/ TR/rdf-concepts/, 2004.

[Lopez et al. 2009] V. Lopez, V. Uren, M. R. Sabou, And E. Motta. **Cross ontology query answering on the semantic web: an initial evaluation**. In K-CAP '09, paginas 17-24, New York, NY, USA, 2009. ACM.

PUC-Rio - Certificação Digital Nº 1121837/CA

[Ma et al. 2009] L. Ma, X. Sun, F. Cao, C. Wang, and X. Wang. **Semantic enhancement for enterprise data management**. In ISWC2009, 2009.

[Manso et al. 2013] Ribeiro Manso de Abreu e Silva, Sofia Casanova, Marco Antonio. **Catalogue of Linked Data Cube Descriptions**. Rio de Janeiro, 2013. Dissertação de Mestrado. Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

[McCusker & McGuinness 2010] McCusker, J. and McGuinness, D. **Towards identity in *Linked Data*.** In Proceedings of OWL. Experiences and Directions Seventh Annual Workshop, 2010.

[Mercer 2006] Mercer, D. **Drupal: Creating Blogs, Forums, Portals, and Community Websites**. Packt Publishing, 2006.

[Miles & Bechhofer 2009] Miles, A. and Bechhofer, S. **Skos simple knowledge organization system reference**. W3C recommendation. Retirado em maio 17, 2013. http://www.w3.org/TR/skos-reference/, 2009.

[Navarro 01] Navarro, Gonzalo. **A Guided Tour to Approximate String Matching**. University of Chile. ACM Computing Surveys, Vol. 33, No. 1, March 2001, pp. 31-88.

[Ngomo & Auer 2011] Ngomo, A.-C. N. and Auer, S. **Limes: a time-efficient approach for large-scale link discovery on the web of data**. In: Proceedings of the Twenty-second International Joint Conference on Artificial Intelligence - volume three, IJCAI'11, p. 2312-2317. AAAI Press, 2011.

[Nikolov & d Aquin 2011] Nikolov, A. and d Aquin, M. **Identifying relevant sources for data linking using a semantic web index**. WWW2011 Workshop: *Linked Data* on the Web (LDOW 2011) at 20th International World Wide Web Conference (WWW 2011). [S.l.: s.n.], 2011.

[OECD 2006] **Management of statistical metadata at the OECD**. Retirado em setembro 11, 2011. http://www.oecd.org/dataoecd/26/33/33869551.pdf/, 2006.

[Reynolds 2012] Reynolds, D. **An organization ontology**. W3C working draft. Retirado em junho 10, 2013, http://www.w3.org/TR/vocab-org/, 2012.

[Rubacket al. 2013] Ruback, L. Manso, S. Salas, P. E. R. Pesce, M. Ortiga, S. and Casanova, M. A. **A mediator for statisticallinked data**. Proceedings of the 28th Annual ACM Symposium on Applied Computing, SAC '13, p. 339-341, New York, NY, USA, 2013. ACM.

PUC-Rio - Certificação Digital Nº 1121837/CA

[Salas et al. 2012] Salas, P. Martin, M. Mota, F. Auer, S. Breitman, K. and Casanova, M. **Publishing statistical data on the web**. Proceddings IEEE ICSC, 2012.

[Simmhan, 2007] Simmhan, Y. L. **Provenance framework in support of data quality estimation**. Master thesis, Indiana University, 2007.

[Smith & McGuinness, 2004] Michael K. Smith, Chris Welty and Deborah L. McGuinness. **OWL Web Ontology Language Guide**. W3C Recommendation 10 February 2004. Retirada em abril 8, 2013. http://www.w3.org/TR/2004/REC-owl-guide 20040210/#equivalentClass1.

[Urbani et al. 2010] Urbani, S. Kotoulas, J. Maassen, F. Van Harmelen, A and Bal, H. **Owl reasoning with webpie: calculating the closure of 100 billion triples**. ESWC2010, 2010.

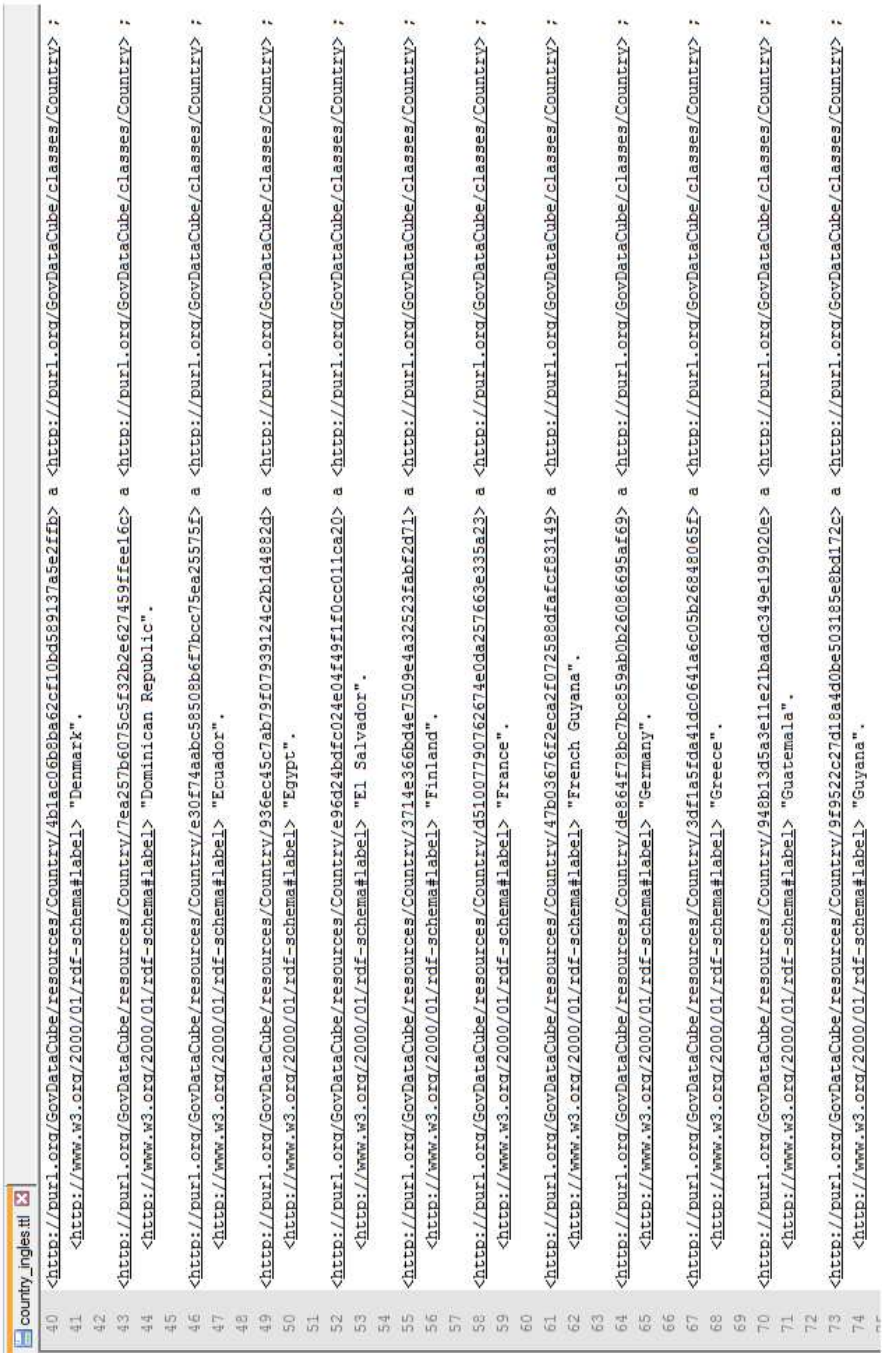
[Volz et al. 2009] Volz, J. Bizer, C. Gaedke, M. and Kobilarov, G. **Silk Discovering and maintaining links on the web of data**. SWC 2009, paginas 650-665. Springer, 2009.

PUC-Rio - Certificação Digital Nº 1121837/CA

Apêndices

Apêndice 1 - Arquivo do componente enriquecido na língua inglesa

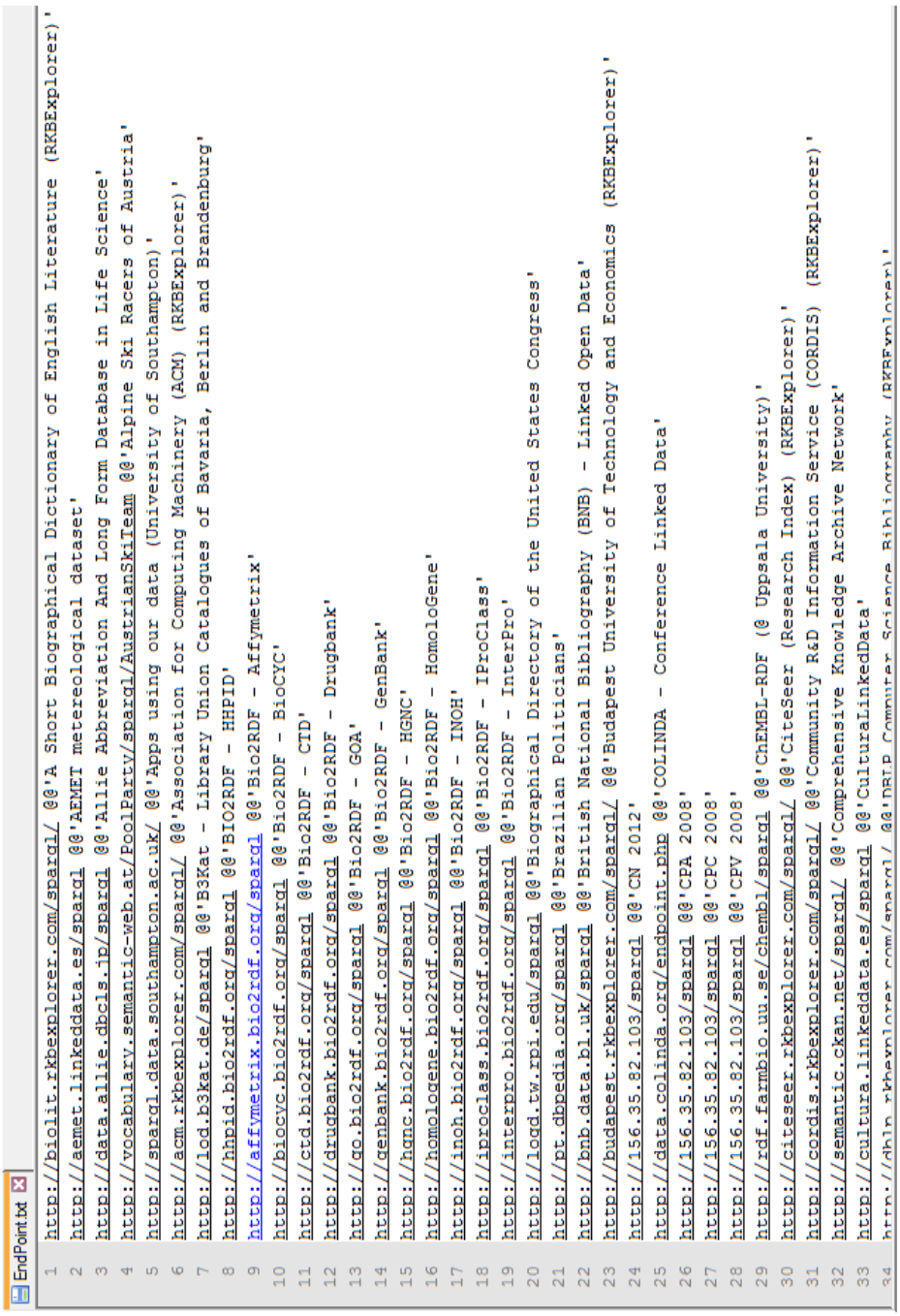
Arquivo enriquecido com a propriedade da língua inglesa. O arquivo faz parte da informação enviada para o Módulo de Recomendação.



PUC-Rio - Certificação Digital Nº 1121837/CA

Apêndice 2 - Arquivo gerado pelo módulo de Recomendação de Fontes de dados.

Modelo da lista emitida pelo Módulo de Recomendação com as possíveis fontes recomendadas.



PUC-Rio - Certificação Digital Nº 1121837/CA

Apêndice 3 - Arquivo de configuração do serviço LIMES

Arquivo gerado pela ferramenta Enriquecedor Automático, do componente *Country*, o algoritmo *Levenshtein* e os parâmetros de aceitação de 1 e revisão de 0.95.

```
1.  <?xml version='1.0' encoding='UTF-8'?>
2.  <!--Sample XML file generated by XMLSpy v2010 rel. 3 sp1 (http://www.altova.com)-->
3.  <!DOCTYPE LIMES SYSTEM 'limes.dtd'>
4.  <LIMES>
5.  <PREFIX>
6.  <NAMESPACE>http://www.w3.org/1999/02/22-rdf-syntax-ns#</NAMESPACE>
7.  <LABEL>rdf</LABEL>
8.  </PREFIX>
9.  <PREFIX>
10. <NAMESPACE>http://www.w3.org/2000/01/rdf-schema#</NAMESPACE>
11. <LABEL>rdfs</LABEL>
12. </PREFIX>
13. <PREFIX>
14. <NAMESPACE>http://xmlns.com/foaf/0.1/</NAMESPACE>
15. <LABEL>foaf</LABEL>
16. </PREFIX>
17. <PREFIX>
18. <NAMESPACE>http://www.w3.org/2002/07/owl#</NAMESPACE>
19. <LABEL>owl</LABEL>
20. </PREFIX>
21. <PREFIX>
22. <NAMESPACE>http://dbpedia.org/ontology/</NAMESPACE>
23. <LABEL>dbpedia-o</LABEL>
24. </PREFIX>
25. <PREFIX>
26. <NAMESPACE>http://purl.org/GovDataCube/classes/</NAMESPACE>
27. <LABEL>ex-class</LABEL>
28. </PREFIX>
29. <SOURCE>
30. <ID>http://dbpedia.org/sparql"DBpedia"</ID>
31. <ENDPOINT>http://dbpedia.org/sparql</ENDPOINT>
32. <VAR>?a</VAR>
33. <PAGESIZE>1000</PAGESIZE>
34. <RESTRICTION>?a rdf:typedbpedia-o:Country</RESTRICTION>
35. <PROPERTY>rdfs:label</PROPERTY>
36. </SOURCE>
37. <TARGET>
38. <ID>LinkedDataCubes</ID>
39. <ENDPOINT>http://localhost:8890/sparql</ENDPOINT>
40. <VAR>?b</VAR>
41. <PAGESIZE>1000</PAGESIZE>
42. <RESTRICTION>?b rdf:typeex-class:Country</RESTRICTION>
43. <PROPERTY>rdfs:label</PROPERTY>
```

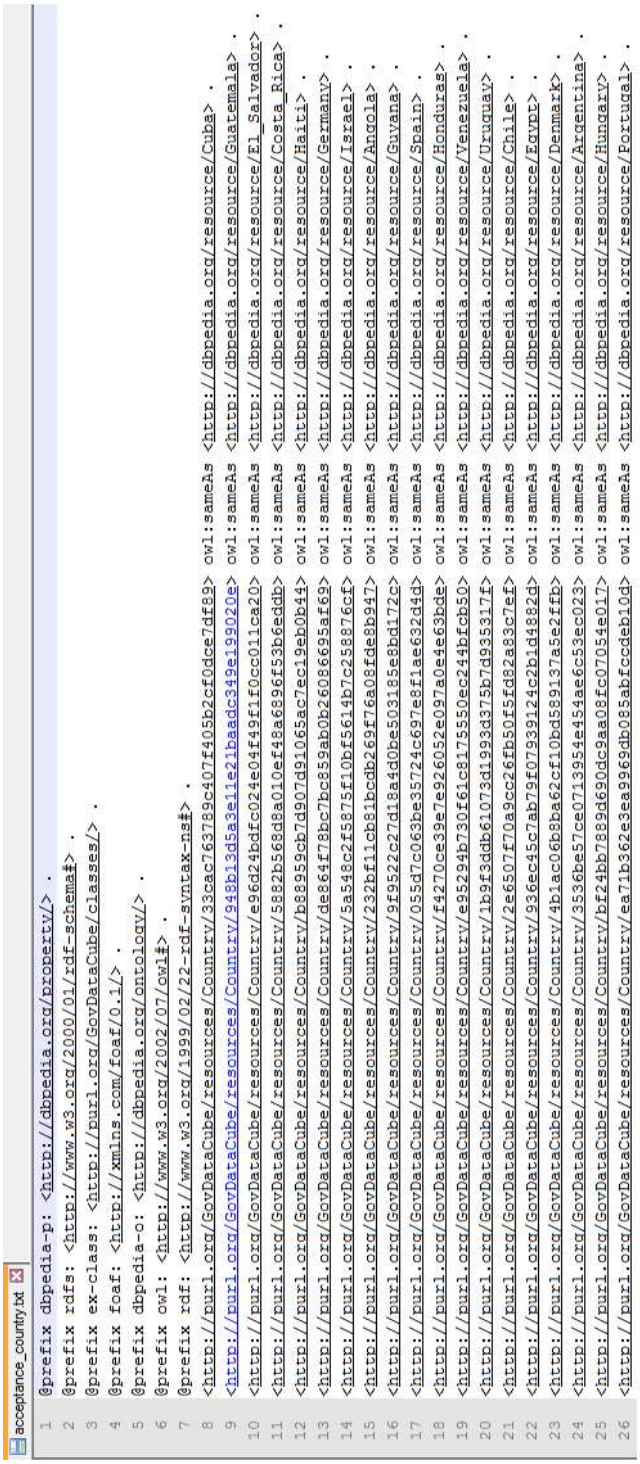
PUC-Rio - Certificação Digital Nº 1121837/CA

```
44. </TARGET>
45. <METRIC>levenshtein(a.rdfs:label, b.rdfs:label)</METRIC>
46. <ACCEPTANCE>
47. <THRESHOLD>1</THRESHOLD>
48. <FILE>acceptance_Country.txt </FILE>
49. <RELATION>owl:sameAs</RELATION>
50. </ACCEPTANCE>
51. <REVIEW>
52. <THRESHOLD>0.95</THRESHOLD>
53. <FILE>review_Country.txt</FILE>
54. <RELATION>owl:sameAs</RELATION>
55. </REVIEW>
56. </LIMES>
```

PUC-Rio - Certificação Digital Nº 1121837/CA

Apêndice 4 - Arquivo gerado pela execução do serviço da ferramenta LIMES.

Arquivo *Acceptance* do componente *Country* gerado pela ferramenta Enriquecedor Automático.

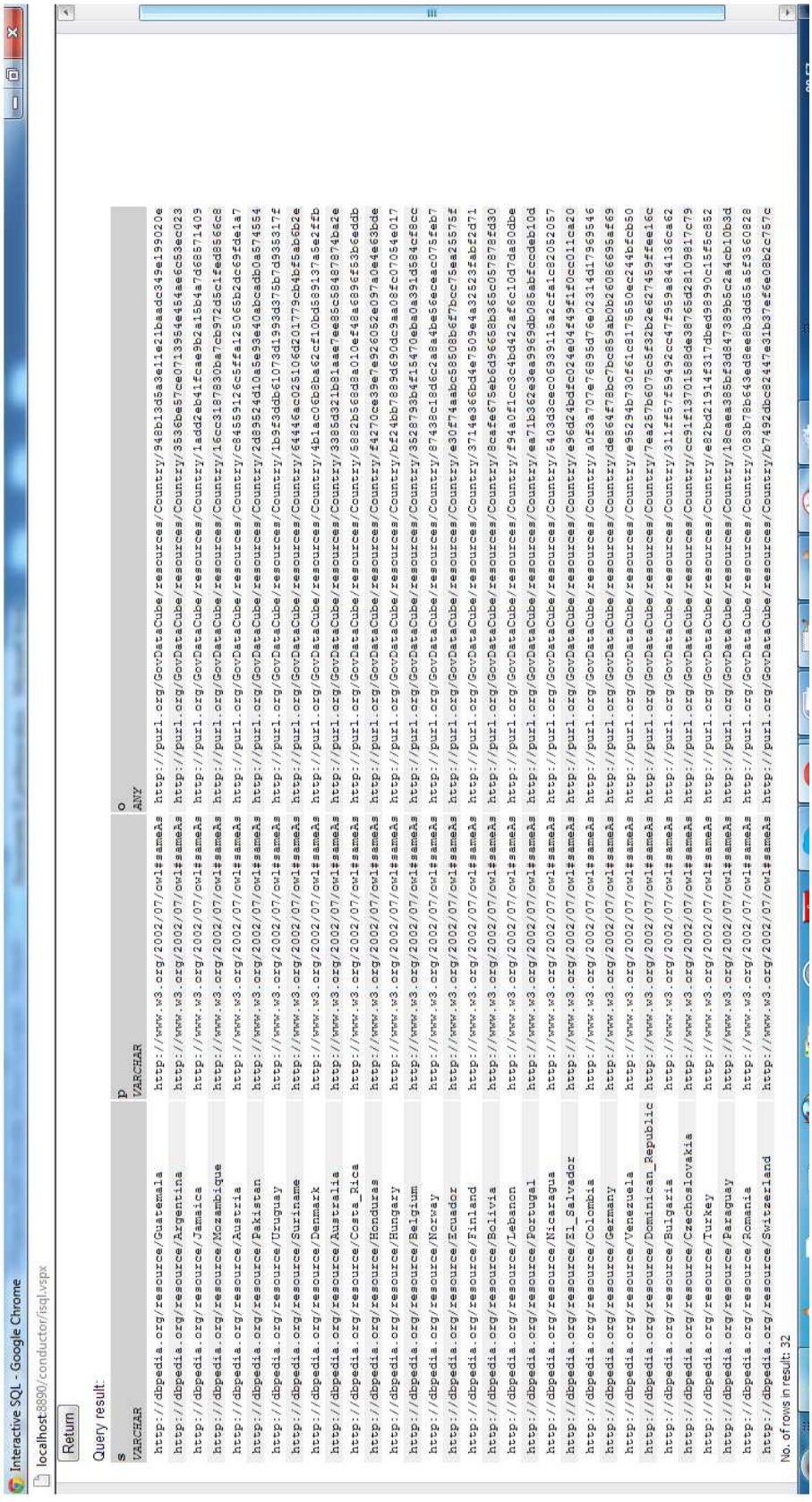


acceptance_country.txt

```
@prefix dbpedia-p: <http://dbpedia.org/property/> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix ex-class: <http://purl.org/GovDataCube/classes/> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix dbpedia-o: <http://dbpedia.org/ontology/> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
<http://purl.org/GovDataCube/resources/Country/33cac763789c407f405b2cf0dce7df89> owl:sameAs <http://dbpedia.org/resource/Cuba> .
<http://purl.org/GovDataCube/resources/Country/948b13d5a3e11e21baadc349e199020e> owl:sameAs <http://dbpedia.org/resource/Guatemala> .
<http://purl.org/GovDataCube/resources/Country/e96d24bdfc024e04f49f1f0cc011ca20> owl:sameAs <http://dbpedia.org/resource/El_Salvador> .
<http://purl.org/GovDataCube/resources/Country/5882b568d8a010ef48a6896f53b6eddb> owl:sameAs <http://dbpedia.org/resource/Costa_Rica> .
<http://purl.org/GovDataCube/resources/Country/b88959cb7d907d91065ac7ec19eb0b41> owl:sameAs <http://dbpedia.org/resource/Haiti> .
<http://purl.org/GovDataCube/resources/Country/de864f78bc7bc859ab0b26086695af69> owl:sameAs <http://dbpedia.org/resource/Germany> .
<http://purl.org/GovDataCube/resources/Country/5a548c2f5875f10bf5614b7c258876cf> owl:sameAs <http://dbpedia.org/resource/Israel> .
<http://purl.org/GovDataCube/resources/Country/232bf11cb81bcdb269f76a08fde8b947> owl:sameAs <http://dbpedia.org/resource/Angola> .
<http://purl.org/GovDataCube/resources/Country/9f9522c27d184d0be503185e8bd172c> owl:sameAs <http://dbpedia.org/resource/Guyana> .
<http://purl.org/GovDataCube/resources/Country/055d7c063be35724c697e8f1ae632d4d> owl:sameAs <http://dbpedia.org/resource/Spain> .
<http://purl.org/GovDataCube/resources/Country/f4270ce39e7e926052e097a0e4e63bde> owl:sameAs <http://dbpedia.org/resource/Honduras> .
<http://purl.org/GovDataCube/resources/Country/e95294b730f61c8175550ec244bfcb50> owl:sameAs <http://dbpedia.org/resource/Venezuela> .
<http://purl.org/GovDataCube/resources/Country/1b9f3ddb61073d1993d375b7d935317f> owl:sameAs <http://dbpedia.org/resource/Uruguay> .
<http://purl.org/GovDataCube/resources/Country/2e6507f70a9cc26fb50f5fd82a83c7ef> owl:sameAs <http://dbpedia.org/resource/Chile> .
<http://purl.org/GovDataCube/resources/Country/936ec45c7ab79f07939124c2b1d4882d> owl:sameAs <http://dbpedia.org/resource/Egypt> .
<http://purl.org/GovDataCube/resources/Country/4b1ac06b8ba62cf10bd589137a5e2ffb> owl:sameAs <http://dbpedia.org/resource/Denmark> .
<http://purl.org/GovDataCube/resources/Country/3536be57ce0713954e454ae6c53ec023> owl:sameAs <http://dbpedia.org/resource/Argentina> .
<http://purl.org/GovDataCube/resources/Country/bf24bb7889d690dc9aa08fc07054e017> owl:sameAs <http://dbpedia.org/resource/Hungary> .
<http://purl.org/GovDataCube/resources/Country/ea71b362e3ea9969db085abfccdeb10d> owl:sameAs <http://dbpedia.org/resource/Portugal> .
```

PUC-Rio - Certificação Digital Nº 1121837/CA

Grafo do componente *Country* inserido no Virtuoso através da ferramenta Enriquecedor Automático.



PUC-Rio - Certificação Digital Nº 1121837/CA

Apêndice 5 - Arquivo de Proveniência

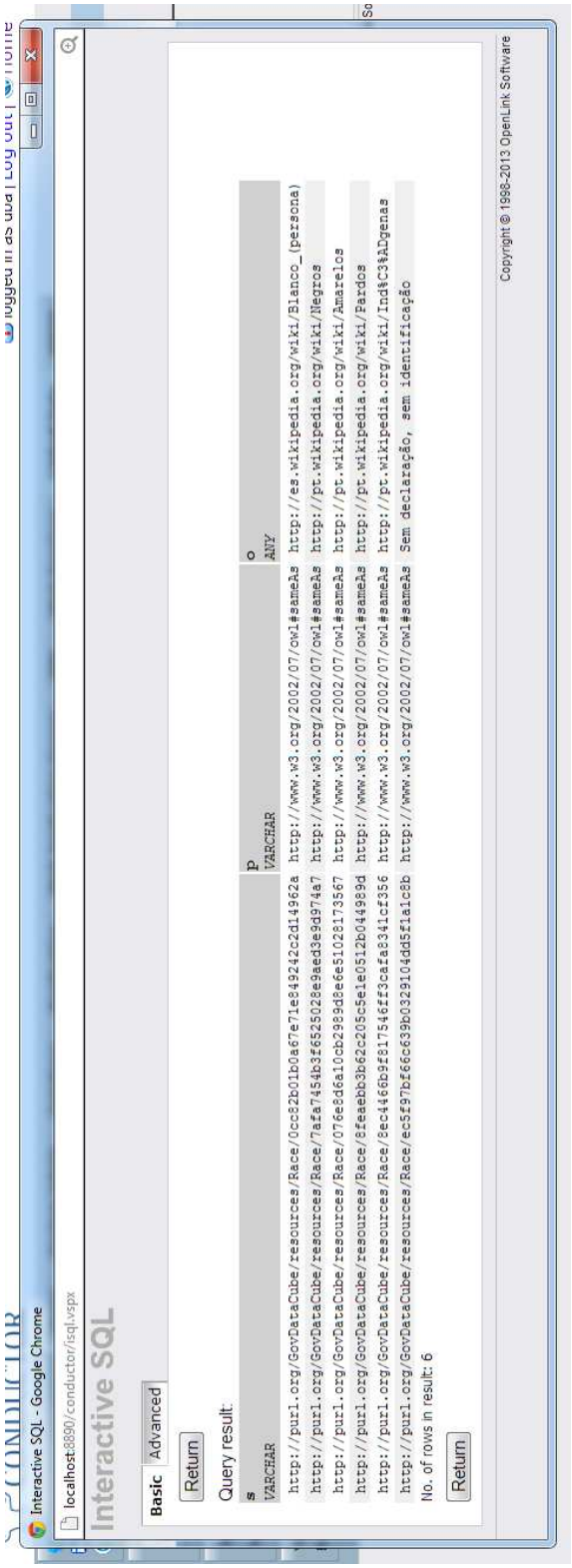
Arquivo gerado pela ferramenta Enriquecedor Automático com os dados utilizados para configurar o serviço LIMES.

```
@prefix rdf:    <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
@prefix dc:     <http://purl.org/dc/elements/1.1/>
@prefix rdfs:   <http://www.w3.org/2000/01/rdf-schema#>.
@prefix dct:    <http://purl.org/dc/terms/>

<http://acceptance>     rdf:type  <http://xmlns.com/foaf/0.1/Project> .
<http://acceptance>     dc:creator      "Limes" ;
                        dc:date         "2013-06-16T06:00TZD" ;
                        dc:description  "Calculo de similaridade, entre dois bancos de dados".

<http://dbpedia.org/sparql>rdf:type<http://vocab.deri.ie/cogs#Source> .
<http://dbpedia.org/sparql/Country>rdf:type<http://vocab.deri.ie/cogs#Input> .
<http://purl.org/GovDataCube/listaSourceI1I2>rdf:type<http://vocab.deri.ie/cogs#File> .
<http://purl.org/GovDataCube/listaSourceI1I2>rdfs:label "Null" .

<http://localhost:8890/sparql>rdf:type<http://vocab.deri.ie/cogs#Endpoint> .

<http://localhost:8890/sparql/Country>rdf:type<http://vocab.deri.ie/cogs#Input> .

<http://purl.org/GovDataCube/metricaI1I2>rdf:type<http://vocab.deri.ie/cogs#Formula> .
<http://purl.org/GovDataCube/metricaI1I2>rdfs:label "levenshtein" .

<http://purl.org/GovDataCube/thresholdAI1I2>rdf:type<http://vocab.deri.ie/cogs#Filter> .
<http://purl.org/GovDataCube/thresholdAI1I2>rdfs:label "1" .

<http://purl.org/GovDataCube/thresoldRI1I2>rdf:type<http://vocab.deri.ie/cogs#Filter>  .
<http://purl.org/GovDataCube/thresoldRI1I2>rdfs:label "0.95" .

<http://acceptance>     dct:relation    <http://dbpedia.org/sparql>;
                        dct:relation    <http://dbpedia.org/sparql/Country>;
                        dct:relation    <http://purl.org/GovDataCube/listaSourceI1I2>;
                        dct:relation    <http://localhost:8890/sparql>;
                        dct:relation    <http://localhost:8890/sparql/Country>;
                        dct:relation    <http://purl.org/GovDataCube/metricaI1I2>;
                        dct:relation    <http://purl.org/GovDataCube/thresholdAI1I2>;
                        dct:relation    <http://purl.org/GovDataCube/thresholdRI1I2> .
```

PUC-Rio - Certificação Digital Nº 1121837/CA

Apêndice 6 - Exemplo de Enriquecimento Manual

Grafo do componente *Race* inserido no Virtuoso através da ferramenta Enriquecedor Manual.



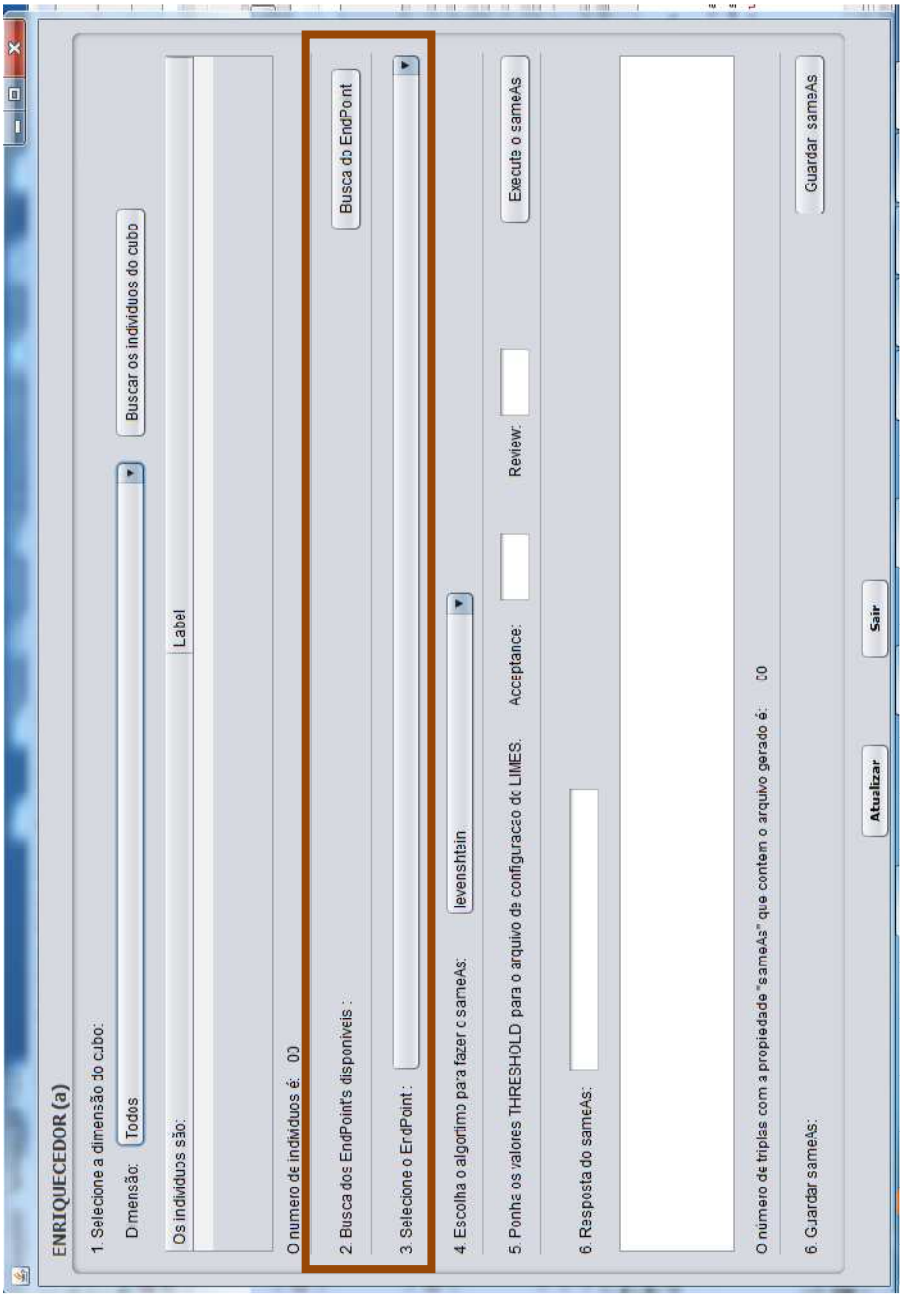
PUC-Rio - Certificação Digital Nº 1121837/CA

Grafo do componente *Country* inserido no Virtuoso através da ferramenta Enriquecedor Manual.

PUC-Rio - Certificação Digital Nº 1121837/CA

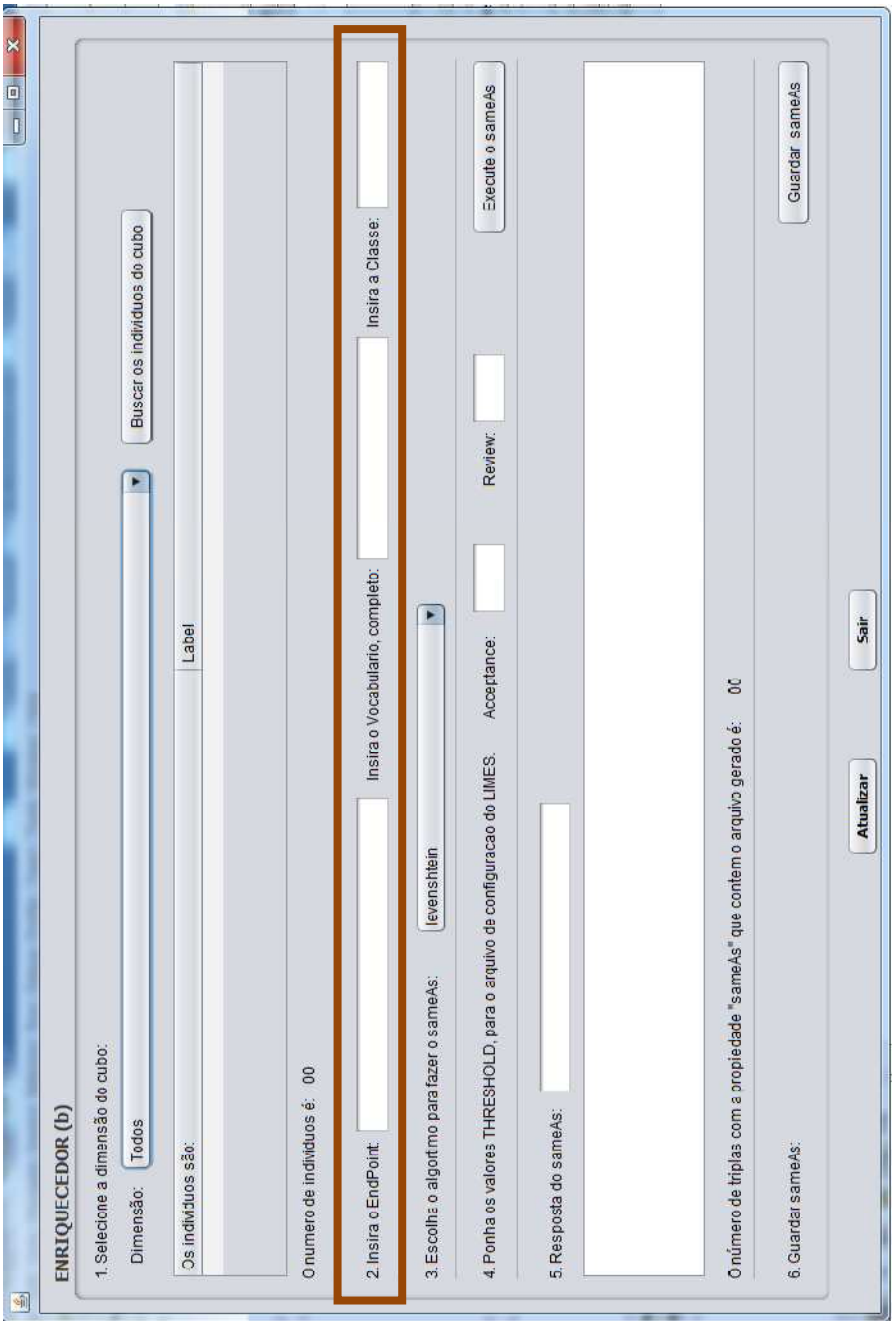
Apêndice 7 - Tela do Enriquecedor Automático

Tela da ferramenta Enriquecedor Automático, com o serviço de recomendação.



PUC-Rio - Certificação Digital Nº 1121837/CA

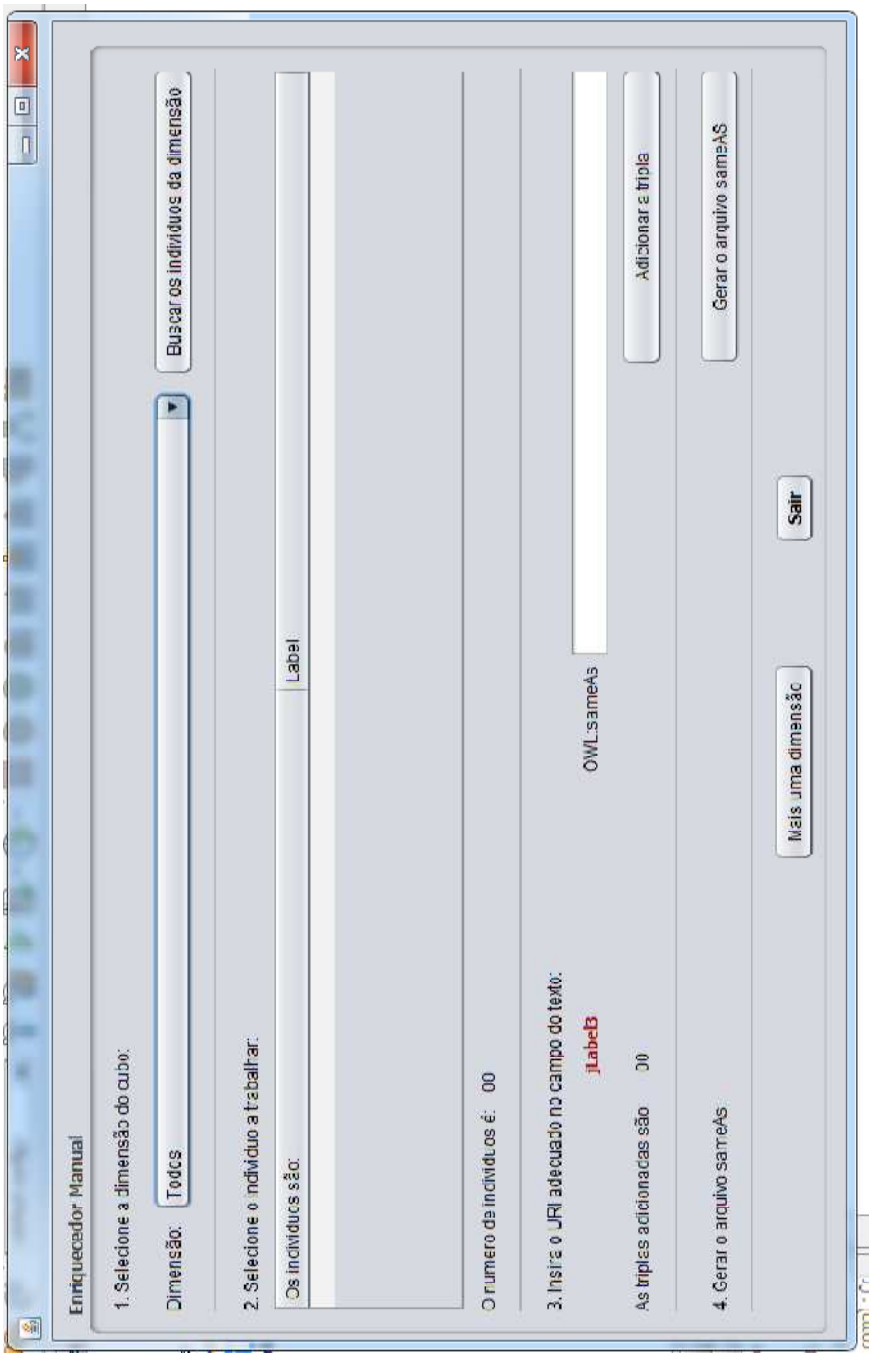
Tela da ferramenta Enriquecedor Automático, sem o serviço de recomendação.



PUC-Rio - Certificação Digital Nº 1121837/CA

Apêndice 8 - Tela do Enriquecedor Manual

Tela da ferramenta Enriquecedor Manual.



PUC-Rio - Certificação Digital Nº 1121837/CA