UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

RECOMENDAÇÃO DE VOCABULÁRIOS PARA MAPEAMENTO DE DADOS CONECTADOS

Wagner Gomes do Amaral

Orientadores
Bernardo Pereira Nunes
Sean Wolfgand Matsui Siqueira

RIO DE JANEIRO, RJ - BRASIL
JULHO DE 2017

RECOMENDAÇÃO DE VOCABULÁRIOS PARA MAPEAMENTO DE DADOS CONECTADOS

Dissertação apresentada ao Programa de Pós-Graduação em Informática da Universidade Federal do Estado do Rio de Janeiro, como pré-requisito para a obtenção do grau de Mestre em Curso de Informática.

Orientação: Prof. Dr. Bernardo Pereira Nunes, e Prof. Dr. Sean Wolfgand Matsui Siqueira.

Rio de Janeiro

2017

Amaral, Wagner Gomes do
A485 Recomendação de Vocabulários para Mapeamento de Dados Conectados / Wagner Gomes do Amaral. -- Rio de Janeiro, 2017.
82 f.

Orientador: Bernardo Pereira Nunes.
Orientador: Sean Wolfgand Matsui Siqueira.
Dissertação (Mestrado) – Universidade Federal do Estado do Rio de Janeiro, Programa de Pós-Graduação em Informática, 2017.

1. Sistemas de Informação 2. Dados Conectados. 3. Recomendação de Vocabulários. 4. Publicação de Dados. 5. Semântica. I. Pereira Nunes, Bernardo, orient. II. Siqueira, Sean Wolfgand Matsui, orient. III. Recomendação de Vocabulários para Mapeamento de Dados Conectados.

RECOMENDAÇÃO DE VOCABULÁRIOS PARA MAPEAMENTO DE DADOS CONECTADOS

Wagner Gomes do Amaral

DISSERTAÇÃO APRESENTADA COMO REQUISITO PARCIAL PARA OBTENÇÃO DO TÍTULO DE MESTRE PELO PROGRAMA DE PÓSGRADUAÇÃO EM INFORMÁTICA DA UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO (UNIRIO). APROVADA PELA COMISSÃO EXAMINADORA ABAIXO ASSINADA.

Aprovada por:

Bernardo Pereira Nunes, D.Sc (Orientador) – UNIRIO

Sean Wolfgand Matsui Siqueira, D.Sc (Orientador) – UNIRIO

Simone Bacellar Leal Ferreira, D.Sc – UNIRIO

Gilda Helena Bernardino de Campos, D.Sc – PUC-Rio

RIO DE JANEIRO, RJ - BRASIL

JULHO DE 2017

“Para conseguir grandes coisas, é necessário não apenas planejar, mas também acreditar; não apenas agir, mas também sonhar.”
Anatole France

Às mulheres da minha vida, minha esposa Viviane e minha filha Carol.

AGRADECIMENTOS

Agradeço a DEUS.

Aos meus pais Eliezer e Estela que com muito esforço e amor transmitiram a base que abriram portas para novas conquistas.

À minha esposa Viviane por me apoiar sempre em meus sonhos.

Aos meus irmãos por torcerem por mim e pelas palavras de incentivo.

Ao professor Sean primeiramente pela oportunidade de conhecer o curso, foi um passo importante desse processo. Também pela dedicação, incentivo e por compartilhar seu conhecimento e a crença que podemos através da pesquisa melhorar a educação. Foram muitas as discussões durante essa jornada.

Ao professor Bernardo, pela dedicação, por dividir seu conhecimento, pela motivação nos momentos difíceis e por caminhar junto nas pesquisas.

Ao professor Ismael, pela oportunidade de trabalhar e conhecer a pesquisa acadêmica. Foi uma experiência importante.

Aos colegas que de alguma forma contribuíram e participaram dessa jornada, Fritzen, Fabiano, Antônio, Vinicius, Igor, Vanessa, Thiago, Fernando, André.

Aos colegas do grupo Semantics & Learning que durantes esses 2 anos contribuíram para a minha formação através de discussões, dicas e opiniões.

Aos professores do PPGI que contribuíram para a minha formação e aprendizado.

AMARAL, Wagner Gomes do. **Recomendação de Vocabulários para Mapeamento de Dados Conectados.** UNIRIO, 2017. 82 páginas. Dissertação de Mestrado. Programa de Pós-Graduação em Informática, UNIRIO.

RESUMO

Com a crescente adoção de dados conectados (DC) como padrão para publicação e conexão de dados estruturados na Web, cria-se um espaço de dados global que abrange uma grande variedade de domínios (por exemplo, Educação, Governo, Ciências, Linguística etc.), como pode ser visto na nuvem de dados abertos conectados (LOD). Além disso, DC têm fomentado a criação de uma série de aplicações baseadas na interligação de aplicações heterogêneas no nível de dados. Apesar dos muitos benefícios de usar dados conectados, uma série de desafios emerge ao publicá-los seguindo padrões específicos. Um problema comum enfrentado pelos publicadores de dados é como representa-los semanticamente – sendo esta uma das primeiras etapas ao publicar dados conectados. Para esta tarefa, um publicador de dados precisa criar um vocabulário ou reutilizar um ou mais vocabulários existentes e publicados na Web. O último caso leva os publicadores de dados a um problema anterior, ou seja, como encontrar vocabulários que melhor representem seus dados? Este é o problema abordado nesta dissertação. Para resolver os problemas mencionados, esta dissertação apresenta o RVMDC, que consiste de um processo, uma arquitetura e uma ferramenta desenvolvidos para recomendar vocabulários para representar dados que serão publicados como dados conectados na Web. Para avaliar o RVMDC, utilizamos bancos de dados relacionais educacionais e de gerenciamento de conteúdo. Os resultados, em termos de precisão, *recall* e *f-measure*, bem como sua capacidade de auxiliar os

publicadores de dados na tarefa de publicação de conjunto de dados confirmam o potencial da ferramenta para recomendação de vocabulários.

Palavras-chave: Recomendação de Vocabulários, Dados Conectados, Publicação de Dados, Dados Educacionais Conectados, Web Semântica.

ABSTRACT

With the increasing adoption of Linked Data (LD) standards for publishing and connecting structured data on the Web, a global data space has been created covering a large variety of domains (e.g. Education, Government, Life Sciences, Linguistics, etc.) as shown in the LOD cloud. Additionally, LD has also led to the creation of a number of cross-domain and domain-specific applications, allowing the interlinking of heterogeneous applications at the data level. Despite the many benefits of using LD, a number of challenges arises when one wants to publish data following the LD standards. A common problem faced by data publishers is on how to semantically representing data – one of the first steps when publishing linked data. For this task, a data publisher needs to either create his own vocabulary or to reuse one or more of the existing vocabularies published on the Web. The latter takes data publishers to a prior problem, that is, how to find vocabularies that best represent their data? This is the problem addressed in this dissertation. To address these problems, this dissertation introduces RVMDC, a process, an architecture and a tool that can be used for recommending vocabularies to data that will be published as LD in the Web. Evaluation using educational and content management databases shows good results in terms of precision, recall and f-measure as well as its ability on assisting data publishers in the task of dataset publication.

Keywords: Vocabulary Recommendation, Linked Data, Data Publishing, Linked Educational Data, Semantic Web.

Sumário

ÍNDICE DE FIGURAS

ÍNDICE DE TABELAS

LISTA DE NOMENCLATURAS

API	Application Programming Interface
CSV	Comma-Separated Values
DWBP	Data on the Web Best Practices
FOAF	Friend of a Friend
LD-BP	Best Practices for Publishing Linked Data
LMS	Learning Management System
LOD	Linking Open Data
LOV	Linked Open Vocabularies
OWL	Ontology Web Language
RDF	Resource Description Framework
RDFa	Resource Description Framework in Attributes
RDFS	Resource Description Framework - Schema
SGBD	Sistema de Gerenciamento de Banco de Dados
SIOC	Semantically-Interlinked Online Communities
SKOS	Simple Knowledge Organization System
SPARQL	Protocol and RDF Query Language
TSIOC	SIOC Types Ontology Module
URL	Uniform Resource Locator
XML	Extensible Markup Language
W3C	World Wide Web Consortium

1. Introdução

1.1. Motivação

A partir das novas tecnologias, há um crescimento exponencial da criação e compartilhamento dos dados não só por humanos, mas também por máquinas, como indicado pelo relatório da IDC Digital Universe[1]. Para o ano de 2020 são previstos 44 Zettabytes de dados (GANTZ; REINSEL, 2012), sendo que esse grande volume de dados consiste de dados complexos e heterogêneos (SIVARAJAH et al., 2017).

Por outro lado, a Web Semântica (BERNERS-LEE et al., 2001) fornece uma estrutura comum que permite que os dados sejam compartilhados, conectados e reutilizados além dos limites das aplicações, das organizações e da comunidade. É um esforço colaborativo liderado pelo W3C[2], para promover a interoperabilidade e interligação dos dados através de tecnologias como *Resource Description Framework* (RDF), *Protocol and RDF Query Language* (SPARQL), *Ontology Web Language* (OWL) e *Simple Knowledge Organization System* (SKOS).

Entretanto, a criação e publicação de dados conectados são processos complexos que envolvem diversas técnicas, tecnologias e ferramentas que podem variar conforme o domínio (BAUER; KALTENBÖCK, 2011; BIZER, 2009; D'AQUIN, 2012; HALAÇ et al., 2013; HYLAND; WOOD, 2011; VAN NUFFELEN et al., 2014; VILLAZÓN-TERRAZAS et al., 2011). Um dos principais passos para a publicação de dados

[1] https://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm

[2] http://www.w3.org/2001/sw/

conectados é a conversão dos dados atuais (ex.: de bancos de dados relacionais, de documentos e bancos de dados legados) para o formato RDF e de preferência reutilizando um vocabulário padrão. Segundo o W3C[3], a reutilização de vocabulários padrão traz diversos benefícios para os seus conjuntos de dados, entre os quais podemos citar compreensão, processabilidade, reutilização, confiança e interoperabilidade:

- Compreensão: permite uma melhor compreensão por humanos da estrutura de dados, o significado dos dados, os metadados e a natureza do conjunto de dados.
- Processabilidade: permite as máquinas processar e manipular automaticamente os dados dentro de um conjunto de dados.
- Reuso: permite chances maiores de reutilização de conjuntos de dados por diferentes grupos de consumidores de dados.
- Confiança: permite uma maior confiança no conjunto de dados pelos consumidores.
- Interoperabilidade: permite de forma mais fácil chegar a um consenso entre publicadores de dados e consumidores.

Um conjunto de melhores práticas proposto pelo W3C[4] indica como analisar um vocabulário para reuso:

- Foram publicados por um grupo ou organização confiável;
- Têm URIs[5] permanentes;
- Possuem uma política de controle de versão;

[3] https://www.w3.org/TR/dwbp/
[4] https://www.w3.org/TR/ld-bp/
[5] Um URI (ou identificador universal de recurso, do inglês *Universal Resource Identifier*) é definido como uma cadeia de caracteres ASCII usada para identificar coisas na Web Semântica. Fonte: https://www.w3.org/wiki/URI

- São documentados, o que inclui o uso literal de rótulos e comentários, bem como marcadores (*tags*) de idioma apropriado;
- Fornecem páginas legíveis que descrevam o vocabulário, junto a suas classes e propriedades. De preferência, casos de uso devem ser definidos e documentados;
- São auto-descritivos, ou seja, cada propriedade ou termo em um vocabulário deve ter um rótulo, definição e comentário definido;
- São descritos em mais de um idioma, todos os elementos do vocabulário devem ter rótulos, definições e comentários disponíveis na língua oficial e, pelo menos, em inglês;
- São usados por outros conjuntos de dados, o que mostra que já estão estabelecidos na comunidade de *Linked Open Data* (LOD), e, portanto, melhores candidatos para uma maior adoção e reutilização;
- São acessíveis por um longo período.

Assim, considerar reuso de vocabulários na publicação de dados conectados implica em analisar principalmente se deve ou não combinar um conjunto de vocabulários, bem como considerar fatores como popularidade do vocabulário e alinhamento com o domínio modelado.

Embora as fontes de dados para a criação e publicação de dados abertos conectados sejam diversas (por exemplo, bancos de dados relacionais, bancos de dados XML, bancos de dados NoSQL, diretórios LDAP, bancos de dados legados e documentos diversos), esta dissertação se restringiu ao escopo de dados de bancos de dados relacionais.

O mapeamento de dados de banco de dados relacionais em grafos RDF (que

consiste em um dos passos da criação e publicação de dados conectados) é chamada de triplificação ou RDB-to-RDF. Segundo Michel et al. (2014), podemos implementar o mapeamento de forma automática (mapeamento direto ou mapeamento automático), em que um RDF ou um conjunto de Grafos RDF é gerado automaticamente a partir do esquema relacional. Neste caso, as possíveis inconsistências geradas devem ser revistas pela equipe responsável pelo mapeamento, bem como pode haver a necessidade de tratar mapeamentos complexos que as ferramentas utilizadas podem não considerar. Por outro lado, também há o mapeamento dirigido pelo domínio semântico ou mapeamento manual, em que a equipe responsável pelo mapeamento deve considerar a reutilização de vocabulários com base em mapeamentos complexos, obedecendo à semântica do domínio. Lidar com esses mapeamentos complexos para possibilitar a reutilização de vocabulários é o desafio que motiva esse trabalho. Assim, esta dissertação visa prover recomendações de vocabulários existentes para os usuários com base nos dados relacionais, apoiando assim a publicação de dados conectados.

1.2. Problema de Pesquisa

A tarefa de conversão de dados de bancos relacionais em dados conectados não é trivial (SCHAIBLE et al., 2014) e envolve um conjunto de passos a serem seguidos pelo engenheiro de dados conectados, dentre os quais está a seleção de vocabulários a serem reutilizados. Apesar de existirem um conjunto de ferramentas desenvolvidas para apoiar os engenheiros de dados, como: mecanismos de busca de vocabulários e sistemas de recomendação de vocabulários, grande parte das atividades desse processo ainda é desenvolvida manualmente (SCHAIBLE et al., 2013).

Outro fator importante no processo de modelagem de dados conectados é avaliar a combinação ou não de vocabulários. Deste modo, a decisão de qual vocabulário

melhor expressa seus dados pode implicar em uma combinação de vocabulários. O reuso de apenas um vocabulário por domínio pode favorecer a uma estrutura de dados mais clara, em contrapartida reusar vocabulários mais populares pode tornar seus dados mais facilmente consumidos. O uso de vocabulários padronizados implica em tornar os dados conectados mais legíveis para humanos e melhor processáveis por máquinas.

Considerando-se o alto volume de dados de sistemas e complexidade de seus modelos, a publicação de seus dados como dados conectados demanda um profundo conhecimento dos mesmos, bem como de vocabulários existentes.

Um cenário de exemplo seria a publicação dos dados de um ambiente educacional, Moodle. Seu esquema de banco de dados relacional é composto de 314 tabelas e 2.840 atributos. Essas tabelas representam diversas funcionalidades como: organização de conteúdo, calendários, quizzes, conferências de texto assíncronas, conversas de texto em tempo real, espaço de grupo para trabalho colaborativo, avaliação, atribuição de tarefas, blogs, wikis etc. Além disto, diversas tabelas podem estar relacionadas a um mesmo conceito, por exemplo, para a Enquete do Moodle, algumas tabelas relacionadas são:

- mdl_questionnaire_survey,
- mdl_questionnaire,
- mdl_questionnaire_question_type,
- mdl_questionnaire_question,
- mdl_questionnaire_quest_choice,
- mdl_questionnaire_attempts,
- mdl_questionnaire_response,
- mdl_questionnaire_response_bool,

- mdl_questionnaire_response_date,
- mdl_questionnaire_response_other,
- mdl_questionnaire_response_rank,
- mdl_questionnaire_response_text,
- mdl_questionnaire_resp_multiple, e
- mdl_questionnaire_resp_single.

Para publicar seus dados como dados conectados é necessário conhecer todo esquema, bem como combinar diferentes tabelas e seus atributos semanticamente relacionados e identificar os vocabulários existentes que são adequados a serem reutilizados neste domínio. Ainda considerando este cenário, utilizando um ambiente de consultas de vocabulários de dados conectados (LOV[6]), observam-se 1.674 resultados para conteúdo, 594 resultados para usuário, 204 para cursos e assim por diante. Assim, para cada tabela do Moodle seria necessário verificar as possíveis combinações de tabelas, bem como considerar seus atributos para identificar os vocabulários possíveis de serem reutilizados e selecionar o mais adequado. Em outras palavras, isto implica em analisar o esquema dos dados a serem publicados como dados abertos com os resultados das consultas em vocabulários de dados conectados verificando as possíveis correspondências e assegurando a melhor cobertura. Para esta tarefa é necessário o envolvimento de especialistas tanto em tecnologias de dados conectados como no domínio em questão.

1.3. Objetivo de Pesquisa

Este trabalho visa apoiar a identificação de vocabulários existentes de dados conectados que possam ser reutilizados na publicação de dados conectados a partir de

6 http://lov.okfn.org/dataset/lov/

dados relacionados. Assim, o objetivo de pesquisa é analisar diferentes abordagens para recomendação de vocabulários para prover uma estrutura clara de dados, fazer os dados serem consumidos facilmente e prover maior cobertura dos dados usando vocabulários existentes. A partir desta análise, espera-se entender as melhores estratégias para os diferentes conjuntos de dados.

1.4. Metodologia de Pesquisa

Nesta dissertação seguimos uma abordagem exploratória com base em diferentes abordagens (ou estratégias) utilizadas para a recomendação de vocabulários: (i) minimizar o número total de vocabulários (recomendar um vocabulário que consiga contemplar o maior número de termos das tabelas e seus atributos); (ii) maximizar o número de vocabulários (utilizar combinações de vocabulários); e (iii) reusar vocabulários mais populares.

Para avaliação dos resultados foram realizados dois experimentos. O primeiro experimento no contexto de sistemas de gerência de aprendizagem (LMS – *Learning Management Systems*), onde quatro LMS serviram de base para o processo de recomendação. Ao longo dos últimos anos temos visto uma enorme quantidade de dados, incluindo dados de cunho educacional, sendo publicadas na Web. Movimentos como: software livre, abertura da literatura de pesquisa e recursos educacionais abertos foram importantes para educação (D'AQUIN, 2012). Instituições de ensino superior público e privado, organizações e companhias estão adotando como prática comum a publicação de seus dados e estão interessadas em soluções eficientes para o compartilhamento e o reuso dos dados publicados (D'AQUIN, 2016). Embora, os dados publicados sejam de extrema relevância, grande parte destes não seguem padrões para sua publicação dificultando o seu reuso e favorecendo a criação de silos de dados (i.e.

fontes de dados heterogêneas e isoladas) (DIETZE et al., 2013). Desafios sobre a integração de dados e a interoperabilidade dos dados dispersos em repositórios heterogêneos colocam em riscos os processos de cooperação de instituições de Ensino Superior (VEGA-GORGOJO et al. 2016; TIROPANIS et al. 2009).

O segundo experimento utilizou esquemas de diversos tamanhos e domínios que foram levantados nos trabalhos relacionados, com base em quatro sistemas comerciais.

Para análise dos resultados foram utilizadas as medidas tradicionais da área de sistemas de recomendação e recuperação de informação: precisão, cobertura e medida-f. Estas medidas foram observadas para o processamento sintático realizado (em que as tabelas foram agrupadas com base em seu nome), bem como para o processamento semântico (em que houve enriquecimento das consultas a vocabulários com base na WordNet[7]).

1.5. Organização da Dissertação

Esta dissertação está organizada em mais cinco capítulos, além desta introdução. No segundo capítulo são apresentados conceitos relacionados a dados conectados, uma vez que a motivação para este trabalho é a publicação de dados conectados. Com base nos conceitos apresentados da área, propomos um processo e uma arquitetura para a recomendação de vocabulários para o mapeamento para dados conectados, cuja descrição é apresentada no Capítulo 3. De modo a explorar as diferentes estratégias de recomendação, são realizados dois experimentos que são descritos no Capítulo 4 e cujos resultados são analisados no Capítulo 5. Finalmente, o Capítulo 6 apresenta as considerações finais, bem como sugestões para trabalhos futuros.

[7] https://wordnet.princeton.edu/

2. Dados Conectados

Neste capítulo são apresentados os principais conceitos relacionados à publicação de dados abertos conectados, bem como os trabalhos relacionados a reuso de vocabulários de dados conectados.

2.1. Definição e Características Fundamentais de Dados Abertos

Segundo a Open Definition (2014), dados abertos são dados que podem ser livremente utilizados, reusados e redistribuídos por qualquer pessoa, sujeitos, no máximo, à exigência de atribuição da fonte original e de compartilhamento pelas mesmas licenças em que as informações foram apresentadas. Sob esta perspectiva, a definição do termo dados abertos deve contemplar três características fundamentais (Open Knowledge Foundation, 2010):

- Disponibilidade e acesso: os dados devem estar disponíveis, preferencialmente possíveis de serem baixados pela Internet. O custo de reprodução deve ser razoável. Os dados devem também estar disponíveis de uma forma conveniente e modificável;
- Reuso e redistribuição: os dados devem ser fornecidos sob termos que permitam a reutilização e a redistribuição, inclusive a combinação com outros conjuntos de dados;
- Participação universal: todos devem ser capazes de usar, reutilizar e redistribuir. Não deve haver discriminação contra áreas de atuação ou contra pessoas ou grupos.

2.2. De Dados Abertos a Dados Conectados

Segundo Isotani e Bittencourt (2015), "é importante frisar que a geração de novos dados baseada em dados anteriormente consumidos é inerente à sociedade e

mostra a importância que a estruturação e a conexão de dados possuem para simplificar e facilitar a recuperação da informação e a produção de novos conhecimentos". Assim, observa-se a motivação e relevância para dados conectados.

O conceito de dados conectados (ou *linked data*) está relacionado a um conjunto de boas práticas para publicar e conectar conjuntos de dados estruturados na Web, com o intuito de criar uma "Web de Dados" (Bizer et al., 2006).

A relação entre dados abertos e dados conectados fica mais clara com base na proposta do sistema de avaliação de 5-estrelas (*5-Star rating system*)[8], de Tim Berners-Lee, que avalia através de estrelas o grau de abertura dos dados. Quanto mais abertos os dados, maior o número de estrelas e mais fácil é enriquecê-los (conectar). As 5 estrelas para dados abertos são:

* Disponível na Web (em qualquer formato), com licença aberta, para ser considerado dado aberto;

** Disponível na Web como dado estruturado possível de ser lido por máquina (por exemplo, em um arquivo Excel ao invés de uma imagem escaneada de uma tabela);

*** Como o anterior, mas em formato não-proprietário (CSV em vez de Excel);

**** Como no anterior, mas usando os padrões estabelecidos pelo W3C (RDF e SPARQL) para identificar coisas, de modo que as pessoas possam apontar para suas coisas;

***** Todas as anteriores, mais: conectar seus dados a dados de outras pessoas para prover contexto.

[8] http://5stardata.info/en/

2.3. Melhores Práticas de Dados na Web

A Web de Dados cria diversas oportunidades para integração semântica do próprio dado, motivando o desenvolvimento de novas aplicações e ferramentas, tais como navegadores (*browsers*) e ferramentas de busca. Deste modo, os padrões, recomendações, guias e boas práticas de dados conectados permitem que qualquer pessoa publique os dados de modo que possam ser lidos por pessoas e processados por máquinas. Isto é possível porque dados que antes estavam "escondidos" na "Web de Documentos" está se tornando agora acessível graças ao uso de padrões para conectar dados. Deste modo, todos (humanos e máquinas) podem trabalhar mais eficientemente juntos.

O guia de melhores práticas de dados na Web[9] visa apoiar a publicação e o uso de dados na Web, sendo projetado para auxiliar em um ecossistema autossustentável: à medida que os dados devem ser descobertos e entendidos por humanos e máquinas, as melhores práticas facilitam a interação entre publicadores e consumidores. De acordo com este guia, os principais benefícios de se aplicar as melhores práticas são:

- Compreensão: humanos terão um melhor entendimento sobre a estrutura dos dados, o significado dos dados, os metadados e a natureza do conjunto de dados;
- Processabilidade: máquinas serão capazes de processar e manipular automaticamente os dados dentro de um conjunto de dados;
- Descoberta: máquinas serão capazes de automaticamente descobrir um conjunto de dados ou dados dentro de um conjunto;

[9] http://www.w3.org/TR/dwbp/

- Reuso: as chances de reuso de um conjunto de dados por diferentes grupos de consumidores de dados aumentará;
- Confiança: a confiança que os consumidores têm em um conjunto de dados melhorará;
- Conectividade: será possível criar links entre recursos de dados (conjuntos e itens de dados);
- Acesso: humanos e máquinas serão capazes de acessar dados atualizados em uma variedade de formas;
- Interoperabilidade: será mais fácil chegar a um consenso entre publicadores e consumidores de dados.

Assim, o uso de Dados Conectados tem trazido inúmeros benefícios, como transparência, reutilização, descoberta de novos conhecimentos, interoperabilidade, entre outros, para diferentes áreas de aplicação.

2.4. Publicação de dados conectados

A iniciativa de abertura de dados, em um formato estruturado, usando uma representação semântica, ainda é recente. Grande parte dos trabalhos ainda está na etapa inicial de todo o processo de publicação e consumo (uso) de dados conectados. Muitos repositórios de dados, de diferentes domínios, precisam passar por uma série de padronizações, análise e adequações para então serem disponibilizados seguindo os princípios de dados conectados.

Para apoiar a etapa de publicação de dados muitas ferramentas são disponibilizadas. Elas auxiliam as tarefas de transformação de dados em diferentes formatos para padrões adotados em dados conectados, como o RDF, por exemplo. Basicamente, essas ferramentas ajudam na tarefa de disponibilização de dados

armazenados em banco de dados relacionais, planilhas, e-mails, entre outros, representando-se através de vocabulários específicos e conectando-os a outros dados em diferentes conjuntos de dados (*datasets*).

2.5. Trabalhos Relacionados

Conforme apresentado por Schaible et al. (2016), podemos classificar os serviços que suportam os engenheiros de dados a reusar vocabulários no processo de modelagem de dados conectados, como descrito nas próximas subseções.

2.5.1. Mecanismo de Busca de Vocabulários

A principal característica desse serviço é prover a busca de vocabulários através de palavras-chave. Mecanismos de busca como Swoogle (DING et al., 2004), vocab.cc (STADTMÜLLER et al., 2013) e LOV (VANDENBUSSCHE et al., 2017) permitem encontrar vocabulários através da similaridade dos termos com o conjunto de tipos (classe e propriedades) RDF. Outros mecanismos como Watson (D'AQUIN et al., 2007), Falcon's (CHENG et al., 2008) e Object Search (CHENG; QU, 2009) utilizam na busca valores específicos de uma entidade retornando vocabulários de fontes de dados encontradas na nuvem de dados conectados que correspondem com a pesquisa.

2.5.2. Sistema de Recomendação de Vocabulários

A principal característica desse serviço é prover a recomendação dos vocabulários através das medidas de similaridade sintáticas e semânticas. Aqui podemos incluir as ferramentas apresentadas por Salas et al. (2011) e mais recentemente atualizado em (Michel et al., 2014), bem como as ferramentas RDB-to-RDF onde uma das características levada em consideração é o reuso de vocabulário. O reuso de vocabulários é a capacidade de mapear entidades relacionais para instâncias de vocabulários e ontologias existentes. Esta é a principal diferença entre o mapeamento

baseado em semântica de domínio e as abordagens de mapeamento direto.

2.5.2.1. D2RQ

D2RQ[10] é uma plataforma de código aberto que permite acessar banco de dados relacional como grafos RDF. É composto de três componentes principais: (i) a linguagem de mapeamento D2RQ - uma linguagem de mapeamento declarativa para descrever a relação entre uma ontologia e um modelo de dados relacional; (ii) o mecanismo D2RQ, um plug-in para reescrever chamadas a API Jena em consultas SQL; e (iii) o servidor D2R, um servidor HTTP que fornece uma visão do dados conectados, uma visão HTML para depuração e um endpoint SPARQL. O reuso de vocabulário é feito manualmente através da edição do arquivo de mapeamento D2RQ, conforme exemplo da Figura 5.11. Esse exemplo faz o relacionamento da base de dados relacional com tabelas relacionadas da ISWC 2002 conference e a ontologia ISWC Ontology[11].

[10] http://d2rq.org/
[11] http://d2rq.org/example/iswc.daml

```
# D2RQ Namespace
@prefix d2rq:        <http://www.wiwiss.fu-berlin.de/suhl/bizer/D2RQ/0.1#> .
# Namespace of the ontology
@prefix : <http://annotation.semanticweb.org/iswc/iswc.daml#> .

# Namespace of the mapping file; does not appear in mapped data
@prefix map: <file:///Users/d2r/example.ttl#> .

# Other namespaces
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

map:Database1 a d2rq:Database;
    d2rq:jdbcDSN "jdbc:mysql://localhost/iswc";
    d2rq:jdbcDriver "com.mysql.jdbc.Driver";
    d2rq:username "user";
    d2rq:password "password";
    .
# ------------------------------------------------
# CREATE TABLE Conferences (ConfID int, Name text, Location text);

map:Conference a d2rq:ClassMap;
    d2rq:dataStorage map:Database1;
    d2rq:class :Conference;
    d2rq:uriPattern "http://conferences.org/comp/confno@@Conferences.ConfID@@";
    .
map:eventTitle a d2rq:PropertyBridge;
    d2rq:belongsToClassMap map:Conference;
    d2rq:property :eventTitle;
    d2rq:column "Conferences.Name";
    d2rq:datatype xsd:string;
    .
map:location a d2rq:PropertyBridge;
    d2rq:belongsToClassMap map:Conference;
    d2rq:property :location;
    d2rq:column "Conferences.Location";
    d2rq:datatype xsd:string;
    .
```

Figura 1.1. Recorte do mapeamento das tabelas da ISWC 2012 com a ontologia ISWC

2.5.2.2. Datalift

Datalift[12] é uma plataforma que tem como objetivo prover aos publicadores de dados todos os passos para publicar e conectar seus conjuntos de dados na web de dados, como dados conectados (SCHARFFE et al., 2012). Cobre os principais desafios da publicação de dados, como: publicar conjuntos de dados como RDF; conectar seus conjuntos de dados a outros dados da web dos dados; descrever seus conjuntos de dados através de uma ontologia.

Através de uma interface gráfica permite aos usuários selecionar vocabulários para reuso. Esse processo consiste em selecionar manualmente os vocabulários no

[12] http://www.datalift.org/

catálogo de vocabulários LOV e através de um conjunto de passos auxiliar o alinhamento entre o RDF já convertido e o vocabulário escolhido. O Datalift segue uma abordagem híbrida, onde é feita inicialmente a conversão seguindo a abordagem de mapeamento direto e depois é aplicada a abordagem de mapeamento dirigido por semântica de domínio.

Segundo Scharffe et al. (2012), a plataforma exige um significativo conhecimento de tecnologias da Web Semântica e não provê um mecanismo automático para seleção de vocabulários para mapeamento de termos entre esquemas de fonte de dados e os vocabulários escolhidos.

2.5.2.3. METAmorphoses

METAmorphoses[13] é uma ferramenta para conversão de dados de banco de dados relacional para RDF (SVIHLA; JELINEK, 2004). Implementado na linguagem Java e disponibilizado sob a licença de código aberto LGPL, tem como proposta ser uma ferramenta de fácil implantação e uso. Propõem uma arquitetura de duas camadas: camada de mapeamento e camada de *template*. A camada de mapeamento é responsável por alinhar o esquema do banco de dados relacional com a ontologia. Nesse mapeamento é possível reusar vocabulários que devem ser informados pelo usuário. A camada de modelo é responsável por descrever como gerar as instâncias dos mapeamentos. Ambos os artefatos são desenvolvidos manualmente pelo usuário através de uma linguagem de mapeamento.

Ainda segundo o autor, o processo é facilitado através de uma interface gráfica, porém ainda é complexo escrever construções manualmente para relações complexas.

[13] http://metamorphoses.sourceforge.net/

2.5.2.4. ODEMapster

ODEMapster é um módulo responsável por converter os dados do banco de dados relacional para RDF automaticamente, desde que os mapeamentos sejam definidos na linguagem de mapeamento R2O (BARRASA et al., 2004). Disponível através de plugin[14] que provê uma interface gráfica para a criação, a execução e o mapeamento entre ontologias e bancos de dados. Nesse mapeamento é possível reusar vocabulários que devem ser informados pelo usuário.

2.5.2.5. SquirrelRDF

SquirrelRDF (SEABORNE et al., 2007) é uma ferramenta que permite criar uma visão RDF através de consultas SPARQL sobre banco de dados relacional e diretórios de serviços, através da implementação de adaptadores que estendem o mecanismo de consulta ARQ[15]. Os adaptadores seguem um conjunto de características básicas de conversão entre base de dados relacional e ontologias, como: toda tabela é uma classe; toda linha é uma instância de classe; toda coluna é uma propriedade; e toda célula é um valor da propriedade para uma instância.

Ainda segundo Seaborne et al. (2007), é possível reusar vocabulários existentes através da personalização manual dos mapeamentos, porém, a personalização é limitada e o processo e essencialmente conduzido por regras de mapeamento direto.

2.5.2.6. Triplify

Triplify é um conversor de banco de dados relacional para vários formatos de serialização de RDF. Disponível a partir de um *plugin* para aplicações Web, oferece um

14 http://neon-toolkit.org/wiki/ODEMapster.html
15 http://jena.apache.org/documentation/query/index.html

conjunto de mapeamentos pré-configurados para aplicações Web populares, como: osCommerce, WordPress, Drupal, Gallery e phpBB. A proposta tem como objetivo ser uma solução focada em fornecer mapeamento RDB-to-RDF adaptada especificamente para aplicações Web, reduzindo a barreira de entrada dos desenvolvedores de aplicações Web (AUER et al., 2009).

Os mapeamentos são gerados a partir de consultas SQL. Esse mapeamento é feito manualmente utilizando a abordagem tabela para classe e coluna para propriedade. Permite o reuso de vocabulários através de algumas extensões no comando SQL, que são transparentes para o processador SQL, mas alteram o resultado da consulta, em particular os nomes das colunas das relações retornadas (Ex: SELECT id, name AS ’foaf:name’ FROM users).

Pode ser realizado sob demanda, através de solicitação HTTP-URI ou através de processo ETL (Extract-Transform-Load).

2.5.2.7. Morph-RDB

Morph-RDB[16] é um motor de processamento para conversão de banco de dados relacional para RDF. Implementado em Java e Scala[17] e disponibilizado em código aberto sob a licença Apache 2.0. Suporta duas abordagens, na primeira faz a conversão e geração de dados em RDF a partir de um banco de dados relacional utilizando as descrições do mapeamento. Na segunda, faz a conversão através da reescrita de consultas SPARQL em consultas SQL e de acordo com as descrições do mapeamento.

Nessa abordagem (PRIYATNA et al., 2014) foi desenvolvido um método para

16 http://www.datalift.org/https://github.com/oeg-upm/morph-rdb
17 http://www.scala-lang.org/

habilitar a utilização da linguagem de mapeamento R2RML, através da extensão de um algoritmo (CHEBOTKO et al., 2009) para a tradução do SPARQL para SQL.

2.5.2.8. Virtuoso RDF views

Virtuoso's RDF Views é um recurso do Virtuoso Universal Server que implementa a funcionalidade de conversão de banco de dados relacional para RDF. Virtuoso Universal Server[18] é uma suíte de desenvolvimento para soluções corporativas modernas de acesso a dados, integração e gerenciamento de banco relacional (SQL e/ou Grafo RDF). É disponibilizada em versão comercial e de código aberto.

Virtuoso's RDF View (ERLING; MIKHAILOV, 2009) segue a abordagem de criar visões RDF a partir de dados de banco de dados relacional. Difere de outras propostas similares (D2RQ, SquirrelRDF) porque combina o mapeamento com armazenamento nativo de triplas e pode oferecer melhor otimização de consultas SQL distribuídas através das decisões de compilação feitas com o conhecimento dos dados e sua localização. Isto é especialmente importante ao misturar dados de triplas, relacionais ou dados relacionais distribuídos em muitos bancos de dados externos.

Os arquivos de mapeamento são formados por um conjunto básico de elementos e permitem o reuso de vocabulários já existentes através de declarações denominadas "quad map patterns", que especificam como os valores das colunas das tabelas são mapeados para triplas RDF.

2.5.2.9. Ultrawrap

Ultrawrap é um sistema que provê serviço de consultas SPARQL em base de

[18] https://virtuoso.openlinksw.com/

dados relacional, através da virtualização de triplas em tempo real. A proposta apresenta os seguintes benefícios: consistência em tempo real entre o modelo relacional e o modelo semântico e permite utilizar técnicas de otimização em SQL através de visões SQL das representações das triplas (SEQUEDA et al., 2009).

A conversão do esquema relacional para RDF é feita através da criação do chamado "putative ontology", definido como: "qualquer transformação sintática de um esquema de fonte de dados para uma ontologia" (SEQUEDA et al., 2009). Esse processo incorpora a proposta definida por (TIRMIZI et al., 2008), que define um conjunto de regras de transformação utilizando lógica de primeira ordem.

Recentemente uma nova ferramenta foi proposta (SEQUEDA; MIRANKER, 2015), para dar suporte a linguagem de mapeamento R2RML. Ultrawrap Mapper é uma ferramenta gráfica que permite aos usuários criarem mapeamentos entre um banco de dados relacional e uma ontologia alvo, escondendo a complexidade do mapeamento R2RML. Provê técnicas de alinhamento de ontologia de forma semiautomática, permitindo aos usuários escolherem entre uma lista de sugestões. Ultrawrap Mapper está disponível em versão comercial.

2.5.2.10. StdTrip

StdTrip é uma ferramenta de conversão de banco de dados relacional para triplas RDF, também chamado de triplificação. Sua abordagem é baseada no design a priori (SALAS, 2011), que promove a reutilização de Vocabulários padrões já adotados por outras fontes de dados em RDF, com objetivo de prover maior interoperabilidade e facilidade na integração com a nuvem de dados conectados.

O mapeamento entre o esquema de banco de dados relacional e o modelo RDF é

auxiliado por um processo cujas etapas são resumidas a seguir:

- Conversão - Essa etapa consiste em transformar o esquema de banco de dados relacional para uma ontologia RDF, seguindo um conjunto de conversões para ER (Entidade-Relacionamento) para OWL. Possui duas operações principais: primeiro transformar o esquema relacional em um modelo ER. Segundo, converter o ER em uma ontologia RDF. Essa etapa ainda não associa o esquema relacional a uma ontologia padrão. Converter o modelo de banco de dados relacional diretamente para OWL, não mapearia adequadamente alguns dos atributos.
- Alinhamento - Essa etapa tem como objetivo o alinhamento da ontologia gerada com uma correspondente no conjunto de vocabulários padrões definidos a priori. O alinhamento é implementado utilizando a ferramenta K-match.
- Seleção - Nessa etapa o usuário deve selecionar os resultados do alinhamento entre os termos das ontologias que melhor represente o conceito, através de uma lista ordenada por valor de similaridade em ordem decrescente. Esse processo demanda do usuário conhecimento do domínio.
- Inclusão - Essa etapa permite ao usuário incluir novas ontologias utilizadas por outras fontes de dados conectados, quando não são encontrados elementos nos vocabulários que correspondam entre si. Esse processo utiliza um motor de busca de ontologias, Watson[19], que é baseado em busca por palavras chaves.

[19] http://watson.kmi.open.ac.uk/REST_API.html

- Conclusão - Essa etapa auxilia o usuário a criar uma ontologia, baseada nas recomendações e melhores práticas se nenhuma das etapas anteriores prover uma ontologia adequada.

2.5.2.11. Aureli

AuReLi é um software para conversão de dados de banco de dados relacional para RDF. A proposta é uma extensão do D2R Server e D2RQ map que adiciona as seguintes funcionalidades: geração automática do mapeamento entre o esquema do banco relacional e a ontologia alvo, e busca automática de interligação dos dados convertidos com instâncias de entidades da DBPedia[20] através de um sistema de consultas pré-definidas (POLFLIET; ICHISE, 2010).

O mapeamento automático utiliza diversas medidas de similaridade para encontrar correspondência entre o esquema banco de dados relacional e a ontologia especificada a priori, buscando o alinhamento entre os nomes dos atributos do esquema com as propriedades de uma ontologia.

5.6 Limitações das Propostas Existentes

Foram levantados trabalhos e ferramentas de conversão de banco de dados relacional para RDF que suportam o reuso de vocabulários. Na grande parte dos trabalhos pesquisados, o processo de geração e mapeamento dos atributos do esquema para uma ontologia é feito manualmente pelo usuário com suporte de uma interface gráfica para minimizar a complexidade da geração dos arquivos de mapeamento. O usuário deve informar o vocabulário que melhor represente os dados que serão

[20] A DBpedia é um esforço colaborativo para extrair informações estruturadas da Wikipedia, tornando estas informações disponíveis na Web. http://pt.dbpedia.org/

convertidos. Na proposta desse trabalho esse processo é semiautomático.

Os trabalhos propostos por (POLFLIET; ICHISE, 2010) e (SALAS, 2011) apresentam um processo automático de geração de mapeamento entre o banco relacional e uma ontologia utilizando um conjunto de ontologias definidas a priori. Nossa proposta estende a implementação da proposta (POLFLIET; ICHISE, 2010) permitindo a descoberta de novos conjuntos de vocabulários dinamicamente.

Nossa proposta também inclui a possibilidade do usuário escolher estratégias de mapeamento que serão usadas no processo de reuso dos vocabulários, que não são disponibilizadas em nenhum dos trabalhos pesquisados.

3. Recomendação de Vocabulários para Mapeamento de Dados Conectados

Nesse capítulo apresentamos a solução para a Recomendação de Vocabulários para Mapeamento de Dados Conectados (RVMDC). O objetivo geral é recomendar vocabulários de dados conectados para apoiar o mapeamento em larga escala a partir de dados relacionais. Deste modo, busca-se auxiliar os provedores de dados a publicarem seus conjuntos de dados provenientes de banco de dados relacionais para o formato RDF, que é o formato recomendado no esquema de implementação das 5 estrelas, abordado com Capítulo 2.

3.1. O Processo da RVMDC

Para a recomendação de vocabulários para o mapeamento de dados relacionais para dados conectados é proposto um processo (processo da RVMDC) (Figura 3.1), que possui seis tarefas principais, detalhadas nas próximas subseções.

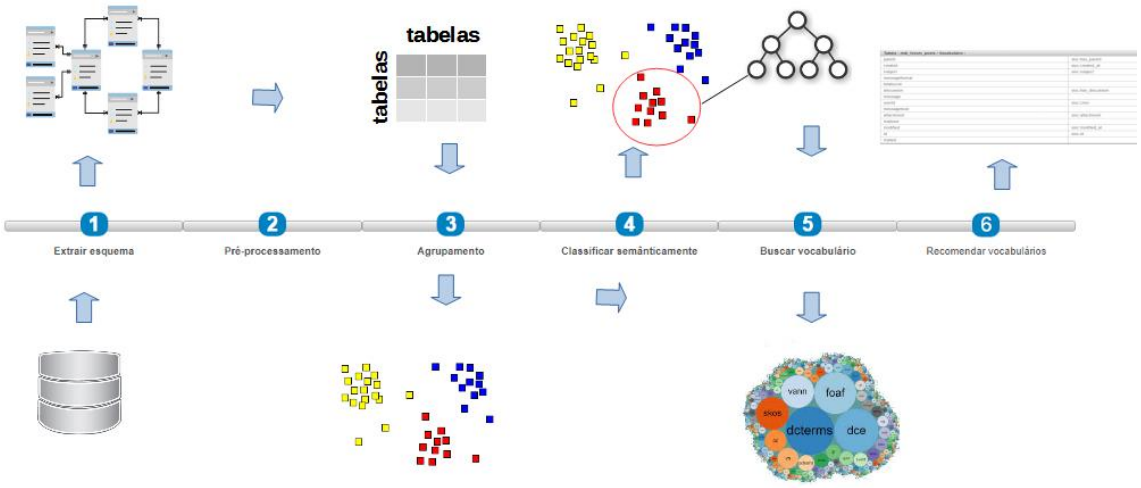


Figura 3.1. Processo de Recomendação de Vocabulários para Mapeamento de Dados Conectados.

Vale observar que as três primeiras tarefas do processo definido nessa abordagem estão baseadas nas pesquisas de alinhamento de esquemas (RAHM; BERNSTEIN, 2001). A Figura 3.2 mostra a classificação das propostas de alinhamento de esquemas e as que foram utilizadas nessa pesquisa.

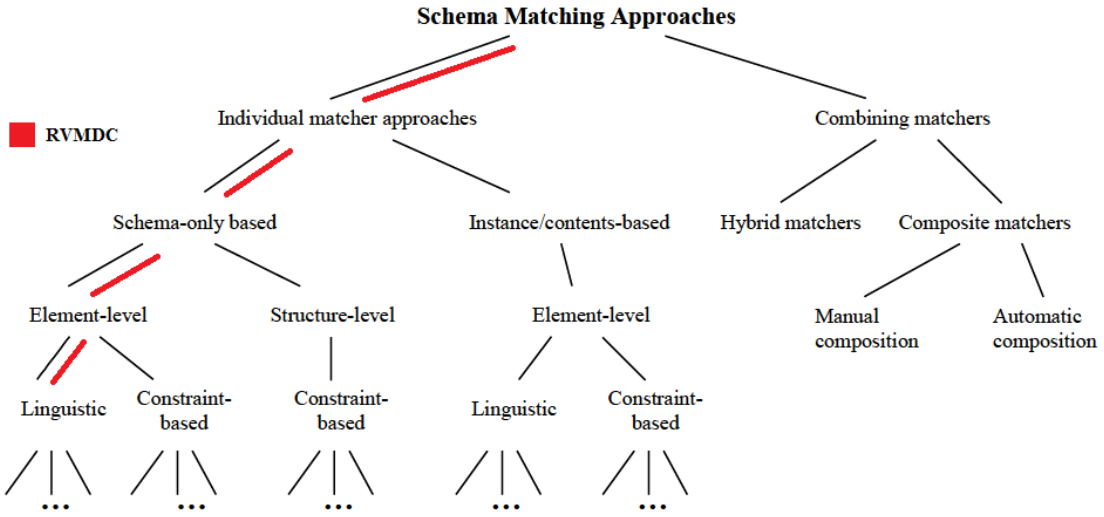


Figura 3.2. Classificação de propostas de alinhamento de esquemas, adaptado de (RAHM; BERNSTEIN, 2001).

Segundo esta classificação (RAHM; BERNSTEIN, 2001), o alinhamento de nível-esquema (*schema-only based*) considera apenas informações relacionadas à estrutura, de modo que nenhuma informação de instância de dados é usada. O alinhamento de nível-elemento (*element-level*) pode ser aplicado usando uma granularidade alta, onde o elemento de alto nível é alinhando com outro elemento, ignorando sua subestrutura e componentes. Finalmente, o alinhamento linguístico ou baseado na linguagem pode ser usado para verificar a terminologia utilizada. Em nossa abordagem utilizaremos o alinhamento baseado em nomes, onde o alinhamento de esquemas é feito através da similaridade dos nomes dos seus elementos. Um exemplo será detalhado na tarefa de aplicar agrupamento (clusterização) das tabelas (Seção 3.1.3).

3.1.1. Extrair esquema do banco de dados relacional

A tarefa de extrair o esquema do banco de dados relacional é responsável por coletar as informações de como os dados são originalmente armazenados. Assim, estamos interessados em coletar informações das estruturas de dados de bancos de dados relacionais para que seja possível entender as fontes de dados que serão publicadas como dados conectados.

O modelo de banco de dados relacional pode ser descrito de forma simplificada através dos conceitos básicos (Tabela 3.1): relação, tupla e atributo (ELMASRI; NAVATHE, 2005), também descritos informalmente como: tabelas, linhas e colunas.

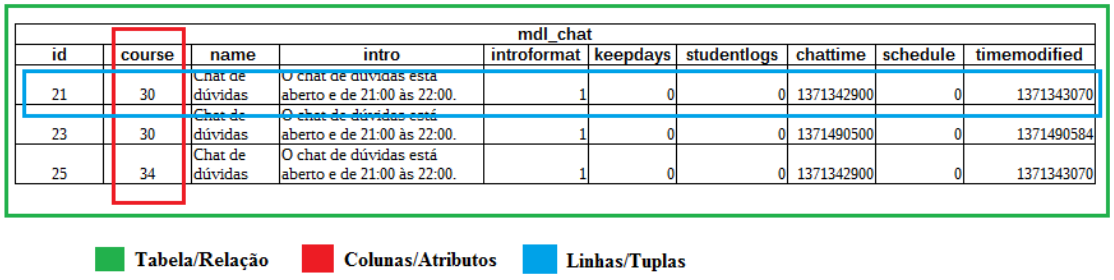


Tabela 3.1. Modelo Relacional

3.1.2. Executar pré-processamento

Uma vez capturado o esquema do banco de dados relacional, é executada a tarefa de alinhamento linguístico, que faz uso de técnicas de processamento de linguagem natural para melhorar o alinhamento de nomes (SHVAIKO; EUZENAT, 2005). Busca-se através dessas técnicas, automatizar a compreensão das linguagens naturais (ex.: nomes dados em língua inglesa às tabelas pelo administrador de dados) para processamento por máquinas. Em muitos casos são necessários limpar caracteres indesejados e remover palavras que não agregam valor semântico ao alinhamento

(stopwords), identificar padrões como, por exemplo, camelcase[21], palavras separadas por caracteres especiais ou simplesmente concatenadas (tokenização) e removendo flexões de gênero, número e grau (lemantização).

3.1.3. Aplicar agrupamento (clusterização) das tabelas

Após o pré-processamento é possível agrupar as tabelas que estão semanticamente relacionadas. As motivações para usar o processo de agrupamento de tabelas de um esquema nessa abordagem estão relacionadas às diversas alterações resultantes da transformação de um esquema conceitual para o esquema lógico e posterior projeto físico, que levam em consideração o modelo (na nossa proposta modelo relacional) e aspectos relacionados ao SGBD. Um exemplo é o processo de normalização, que visa decompor as relações de um esquema de banco de dados relacional com base em suas dependências funcionais e chaves primárias para minimizar a redundância de dados e minimizar anomalias de inserção, exclusão e atualização (ELMASRI; NAVATHE, 2005). A normalização dos dados normalmente traz a divisão da tabela em duas ou mais tabelas.

Deste modo, com esta tarefa, pretende-se minimizar o problema de engenharia reversa do banco de dados (SPANOS et al., 2012). Conceitos como generalização, especialização e agregação também podem ser agrupados nesse processo.

A tarefa de agrupamento das tabelas divide-se em duas etapas. Na primeira etapa, define-se a matriz quadrada de similaridade semântica dos nomes das tabelas. A matriz de similaridade é formada por n x n elementos, onde cada elemento da matriz representa o resultado da medida de similaridade do par de nomes das tabelas que são

[21] CamelCase é a denominação em inglês para a prática de escrever palavras compostas ou frases, onde cada palavra é iniciada com maiúsculas e unidas sem espaços.

comparados através de alinhamento baseados em nomes. De acordo com Shvaiko e Euzanat (2005), podemos identificar a similaridade de várias formas:

- Igualdade de nomes
- Igualdade de representações de nomes canônicos após a radicalização e outros pré-processamentos.
- Igualdade de Sinônimos.
- Igualdade de tipo de ou subtipo.
- Similaridade de nomes com base em substrings comuns, de edição de distância, pronúncia e soundex.
- Correspondências de nome fornecidas pelo usuário.

Na segunda etapa é aplicado o processo de agrupamento, que, a partir da matriz de similaridade, tem objetivo de identificar grupos que representam conceitos semanticamente equivalentes. Para cada grupo ou cluster gerado, o resultado deve ser de alta similaridade intercluster e baixa similaridade extracluster com base nos elementos (HAN et al., 2011).

3.1.4. Classificar os grupos semanticamente

Após a etapa de agrupamento é necessário, para cada instância, definir os termos de pesquisa que serão utilizados na próxima etapa de buscar vocabulários e seus metadados. Os termos de pesquisa são definidos pelas palavras-chaves que caracterizam os grupos, adicionadas de termos relacionados que são extraídos de estruturas de representação do conhecimento, utilizando técnicas de expansão de consultas. Essas técnicas de expansão de consultas foram adotadas com intuito de melhorar a precisão dos resultados das consultadas, como realizado em (PRATES et al., 2013). A expansão de consultas é uma técnica onde se acrescenta novos termos à consulta original com

objetivo de minimizar a ambiguação e melhorar os resultados da pesquisa em motores de busca.

A abordagem proposta usa o WordNet (MILLER, 1995) para expandir o conjunto inicial de termos com termos das relações semânticas fornecidas (como realizado em (FERNÁNDEZ et al., 2006)) e executa uma pesquisa para cada uma das palavras-chave definidas em um índice de ontologias. A abordagem utilizada difere de (FERNÁNDEZ et al., 2006), uma vez que em nosso trabalho as relações semânticas são escolhidas automaticamente. No APÊNDICE A, apresenta-se uma descrição da base de dados WordNet utilizada por essa proposta.

3.1.5. Buscar vocabulário e seus metadados

Para que a proposta de reuso de vocabulários seja efetiva é preciso tornar o acesso e a descoberta de vocabulários facilitada. Vários são os desafios para encontrar vocabulários apropriados para o reuso (D'AQUIN; NOY, 2012), como: encontrar vocabulários que representem o domínio tratado; determinar se o vocabulário cobre o domínio suficientemente e se é possível avaliar a sua qualidade; e se está disponível em formato acessível. Bibliotecas de ontologias (D'AQUIN; NOY, 2012) (ou de vocabulários para o contexto de dados conectados) são endereçadas para tratar esses desafios, e são definidas como "um sistema baseado na Web que fornece acesso a uma coleção extensível de ontologias com o objetivo principal de permitir que os usuários encontrem e utilizem uma ou várias ontologias dessa coleção".

As principais bibliotecas de ontologias normalmente possuem as seguintes características.

- Quanto à proposta e cobertura – se atendem a um domínio específico, tipos de ontologias ou são gerais.

- Quanto ao conteúdo – se fornecem conteúdos adicionais às ontologias como metadados das ontologias e mapeamentos.
- Quanto a funções chaves – se consideram mecanismos facilitadores para encontrar ontologias como mecanismos de busca e navegador.

Utilizamos o catálogo de vocabulários LOV (*Linked Open Vocabularies*) (VANDENBUSSCHE et al., 2017), onde é possível extrair indicadores como: interconexões entre vocabulários, histórico de versões, editor (indivíduo ou organização), *link* para o site do vocabulário e estatísticas de uso na rede de dados abertos conectados (LOD – *linked open data*).

3.1.6. Recomendar Vocabulários

A Recomendação de Vocabulários é feita nessa tarefa, a partir de um conjunto de estratégias que foram derivadas da proposta de Schaible et al. (2014). Foram identificadas as estratégias comuns utilizadas no reuso de vocabulários por especialistas na modelagem de dados conectados quem cobrem as seguintes questões: (i) prover uma clara estrutura de dados; (ii) fazer os dados serem consumidos facilmente; e (iii) prover maior cobertura dos dados. As estratégias são:

3.1.6.1. Minimizar o número total de vocabulários

A estratégia de minimizar o número total de vocabulários consiste em utilizar um único vocabulário que apresentou maior pontuação no retorno da busca dos termos elencados na tarefa de buscar vocabulários no catálogo de vocabulários LOV. Essa estratégia normalmente está associada à questão (i), que é prover uma clara estrutura de dados.

3.1.6.2. Maximizar o número de vocabulários

A estratégia de maximizar o número total de vocabulários consiste em utilizar um conjunto de combinações de vocabulários. Essa estratégia normalmente está associada à questão (iii), que é prover maior cobertura dos dados.

3.1.6.3. Reusar vocabulários mais populares

A estratégia de reusar vocabulários mais populares consiste em utilizar um único vocabulário tido como popular na comunidade de dados conectados. Essa informação é extraída das métricas fornecidas pelo catálogo de vocabulários LOV, que incluem o número de conjuntos de dados que usam esse vocabulário e o total de ocorrência do termo do vocabulário. Essa estratégia normalmente está associada à questão (ii) que é permitir que os dados sejam consumidos facilmente.

3.2. Arquitetura do Sistema de RVMDC

A arquitetura do sistema de RVMDC está organizada em módulos. A arquitetura foi pensada como uma API para permitir a flexibilidade nas escolhas de implementação de cada módulo através das interfaces comuns.

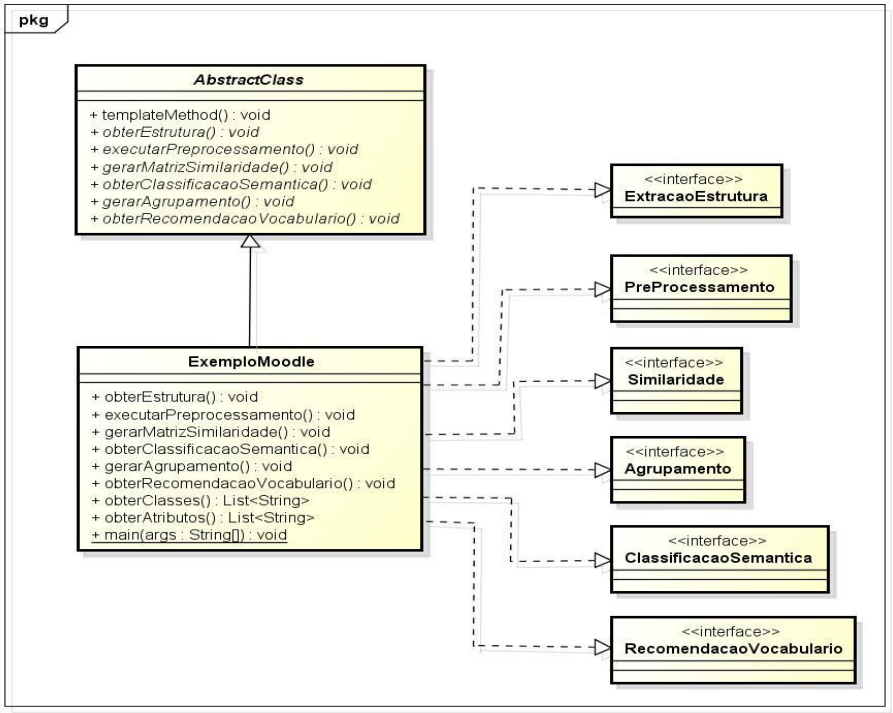


Figura 3.3. Diagrama de classe simplificado com as principais interfaces.

3.2.1. Módulo de Extração de Estrutura

O módulo de extração de estrutura é responsável por coletar as informações de como os dados são armazenados. Estamos trabalhando com a coleta da estrutura de dados de bancos de dados relacional nessa proposta. Um exemplo de informações relevantes extraídas da execução desse módulo é apresentado para a tabela do Moodle mdl_course_categories (Tabela 3.2).

Nome da tabela	Colunas
mdl_course_categories	id, name, idnumber, description, descriptionformat, parent, sortorder, coursecount, visible, visibleold, timemodified, depth, path, theme

Tabela 3.2. Exemplo de resultado da estrutura de dados de bancos de dados relacional

3.2.2. Módulo Pré-processamento

O módulo pré-processamento é responsável por padronizar os termos. Na implementação desse trabalho estamos trabalhando com a biblioteca Apache OpenNLP,

que é um *toolkit* baseado na aprendizagem de máquinas para o processamento de texto em linguagem natural (PLN), implementado na linguagem Java e disponibilizado sobre a licença Apache License, version 2.0. Suporta as tarefas mais comuns de PNL, tais como tokenização, segmentação de sentenças, tagging, extração de entidade nomeada, fragmentação, análise e resolução de correferência.

Após a execução desse módulo temos os termos mais limpos como no exemplo demonstrado na Tabela 3.3. O prefixo mdl é removido assim como as palavras que foram submetidas ao processo de tokenização; os termos foram convertidos para o formato de treinamento OpenNLP Tokenizer; e os tokens foram separados por um espaço em branco.

Nome tabelas originais	**Nomes tratados**
mdl_user	User
mdl_course_modules	course module
mdl_forum_posts	forum post

Tabela 3.3. Exemplo de resultado do pré-processamento

3.2.3. Módulo de Similaridade

O módulo de Similaridade é responsável por gerar a matriz de similaridade. A matriz de similaridade possui a pontuação, ou a medida de similaridade, de cada par de termos resultante do processo de pré-processamento dos nomes das tabelas.

Nesse trabalho foi proposto um algoritmo para cálculo de similaridade dos nomes das tabelas, que foi adaptado de (FENG et al., 2008). Para cada termo complexo (que contem mais de um token no nome tratado da tabela) é feita a separação de dois conjuntos de classes de palavras: substantivo e verbo. Para cada termo é aplicado o processo de lemantização e anotação (passo 2 e passo 3, respectivamente). Nessa implementação foi utilizada a biblioteca Stanford CoreNLP que fornece um conjunto de ferramentas de análise de linguagem natural (MANNING et al., 2014). No passo 4, para

cada par de termos simples e de mesma classe é feito o cálculo de similaridade utilizando as medidas de similaridade disponíveis na API WS4J (WordNet Similarity for Java)[22] e SimMetrics (a Java library of similarity and distance metrics)[23]. No passo 5 é escolhido o valor máximo do cálculo de similaridade e no passo 6 é apresentado o cálculo final após a aplicação da regra de integração onde é dado peso 0.80 para substantivo e 0.20 para verbo (devido ao fato de conceitos serem representadas principalmente através de substantivos) quando há medidas de similaridade em ambos os grupos.

Um exemplo da aplicação do algoritmo considerando duas tabelas, uma relacionada a categorias de cursos e outra relacionada a módulos de cursos é apresentado a seguir.

Passo 1 - Dados de entrada
Sentença 1 = course categories
Sentença 2 = course modules
Passo 2 - Vetores de palavras
Sena [course, categories]
Senb [course, modules]
Passo 3 - Vetores de substantivos(nSena) e verbos(vSena)
nSena [course, category]
vSena []
nSenb [course, module]
vSenb []
Passo 4 - Correlação similaridade WuPalmer
word1=course
termo [course, module] medidas [1.0, 0.5]
word2=category
termo [course, module] medidas [0.4, 0.46153846153846156]
Passo 5 - Valor máximo da correlação
similarityWuPalmer = [1.0]
Passo 6 - Resultado final
The final similarity of sentences = 1.0

22 https://github.com/Sciss/ws4j
23 https://github.com/Simmetrics/simmetrics

3.2.4. Módulo de Agrupamento

Uma vez computada a similaridade entre termos, é possível realizar o agrupamento de tabelas relacionadas. O módulo de agrupamento é responsável por processar a matriz de similaridade e gerar grupos que representam conceitos de um domínio. A RVMDC faz uso da implementação do algoritmo de clustering K-Means, disponível em WEKA (*Waikato Environment for Knowledge Analysis*) (HALL et al., 2009). O WEKA é disponibilizado em código aberto sobre a licença GNU (General Public License), possui um conjunto de ferramentas para tarefas de mineração de dados, como: pré-processamento de dados, classificação, regressão, agrupamento, regras de associação e visualização.

3.2.5. Módulo de Classificação Semântica

O módulo de classificação semântica é responsável por implementar a comunicação com a base de dados da WordNet através da API JWI (*Java Wordnet Interface*)[24] e realizar a expansão de termos para possibilitar a consulta e recomendação. Deste modo, são selecionados os termos que aparecem com maior frequência nos grupos resultantes do processo de agrupamento. As palavras chaves que caracterizam os grupos são enriquecidas semanticamente com as categorias propostas na WordNet. Neste trabalho abordamos relações semânticas entre synsets[25], visando extrair, automaticamente, a relação semântica hiperonímia. Ilustramos na Figura 3.4 um exemplo da palavra *user*.

[24] https://projects.csail.mit.edu/jwi/
[25] Grupo de sinônimos para as palavras (ou termos) utilizados

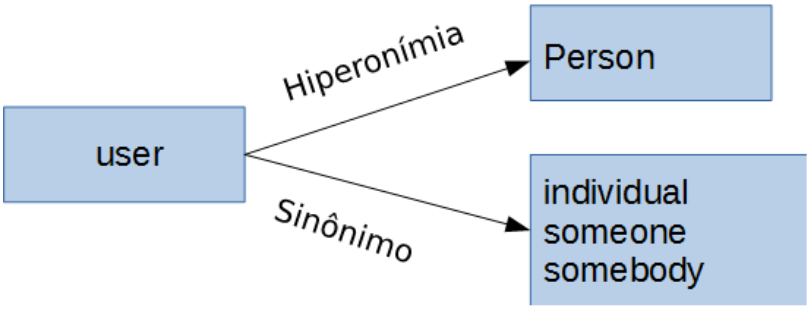


Figura 3.4. Relação semântica de hiperonímia e sinônimo da palavra user.

3.2.6. Módulo de Recomendação de Vocabulários

O módulo de recomendação de vocabulários tem como entrada o resultado do agrupamento e sua classificação e é responsável por implementar a comunicação com o catálogo de vocabulários LOV. Esse acesso é feito através da API do LOV que permite acesso a chamadas HTTP GET e resposta no formato JSON, como pode ser vista no exemplo da Figura 3.5. Nesse exemplo, pesquisamos todos os Vocabulários que possuem a classe com o termo "Person" e do tipo "Classe", http://lov.okfn.org/dataset/lov/api/v2/term/search?q=Person&type=class.

Esse módulo estende as funcionalidades da ferramenta Aureli (*Automatic Relational Database to Linked Data Converter*) (POLFLIET; ICHISE, 2010), que converte de base de dados relacional para RDF, mapeando atributos e relações com classes e propriedades. O Aureli fornece alguns vocabulários genéricos mais comuns, porém é necessário que o usuário adicione novos vocabulários manualmente. Neste trabalho o processo é automatizado através da recomendação de vocabulários presentes nesse módulo, baseado nas estratégias (conforme discutido anteriormente na seção 3.1.6): (a) reutilizar vocabulários mais populares, (b) minimizar o número de vocabulários e (c) maximizar o número de vocabulários.

```
... {
      "prefixedName": [
        "bbccore:Person"
      ],
      "metrics.reusedByDatasets": [
        0
      ],
      "vocabulary.prefix": [
        "bbccore"
      ],
      "metrics.occurrencesInDatasets": [
        0
      ],
      "uri": [
        "http://www.bbc.co.uk/ontologies/coreconcepts/Person"
      ],
      "type": "class",
      "score": 0.511834,
      "highlight": {
        "http://www.w3.org/2000/01/rdf-schema#label@en-gb": [
          "<b>Person</b>"
        ],
        "localName.ngram": [
          "<b>Person</b>"
        ]
      }
    } ...
```


Figura 1.5. Informações no formato JSON da consulta ao catálogo de Vocabulários LOV.

No APÊNDICE B contém as telas da ferramenta RVMDC que suportam o processo de recomendação de vocabulários para mapeamento de dados conectados.

4. Experimentos

Para avaliar a implementação da proposta foram realizados dois experimentos com a RVMDC. Nós adotamos as métricas padrão de precisão, cobertura e medida-F para medir a qualidade do alinhamento entre modelos conceituais/dados (SHVAIKO; EUZENAT, 2005). A precisão (Eq1) é calculada como a razão entre as correspondências positivas encontradas e o número total de correspondências geradas. A cobertura (Eq2) é calculada como a razão entre o número de correspondências positivas encontradas e o número total de correspondências. A medida-F (Eq3) é a média harmônica entre a precisão e a cobertura.

$$\frac{vp}{vp+fp}, onde vp = verdadeiros positivos e fp = falsos positivos \quad \text{(Eq1)}$$

$$\frac{vp}{vp+fn}, onde vp = verdadeiros positivos e fn = falsos negativos \quad \text{(Eq2)}$$

$$\frac{2*(Eq1)*(Eq2)}{(Eq1)+(Eq2)}, onde (Eq1) e (Eq2) são ponderados uniformemente \quad \text{(Eq3)}$$

4.1. Experimento 1

Pesquisas apontam para tecnologias semânticas como forma de enfrentar os desafios da integração e interoperabilidade dos dados educacionais. Nesse experimento foram utilizados os esquemas de quatro ambientes educacionais (Moodle, Sakai, Atutor e Forma.lms), que tem alta penetração no mercado de LMS (*Learning Management Systems*) (DRON; ANDERSON, 2014) e normalmente possuem componentes chaves como: organização conteúdo, calendários, quizzes, conferências de texto assíncronas, conversas de texto em tempo real, espaço de grupo para trabalho colaborativo, avaliação, atribuição de tarefas, blogs e wikis.

4.1.1. Esquemas de banco de dados

Nessa seção serão descritos os ambientes educacionais e seus respectivos esquemas de banco de dados utilizados no experimento.

4.1.1.1. Moodle

Moodle[26] é uma plataforma de código aberto que vem sendo desenvolvida desde 2001 para fornecer aos educadores, administradores e estudantes um sistema robusto, seguro e integrado, criando assim um ambiente de aprendizagem personalizado. Atualmente existem 81.299 sites ativos registados em 236 países. Esse grande volume de dados com registros detalhados das atividades desenvolvidas dentro da plataforma tem um grande potencial para ser publicado como dados abertos conectados. Seu esquema possui 314 tabelas e 2.840 atributos.

4.1.1.2. Sakai

Sakai[27] oferece um ambiente flexível e rico em recursos para o ensino, a aprendizagem, a pesquisa e outras formas de colaboração. Desenvolvido sobre a plataforma Java e distribuída 100% código aberto, é atualmente utilizada por mais de 350 instituições em todo o mundo com mais de 4 milhões de estudantes. Seu esquema possui 286 tabelas e 2.364 atributos.

4.1.1.3. ATutor

ATutor[28] é um Sistema de Gerenciamento de Aprendizagem baseado na Web distribuído sob os termos da GNU General Public License (GPL). Criado a partir de estudos sobre acessibilidade de Sistemas de Gestão de Aprendizagem, proporcionando

[26] https://moodle.org/
[27] https://sakaiproject.org/
[28] http://www.atutor.ca/

aos alunos aprenderem em um ambiente de aprendizagem social acessível e adaptativo. Seu esquema possui entorno de 120 tabelas e 745 atributos.

4.1.1.4. Forma.lms

Forma.lms[29] é uma plataforma de gestão de aprendizagem de código aberto, desenvolvida com foco em treinamento corporativo e usada para gerenciar e oferecer cursos de treinamento on-line. Seu esquema possui entorno de 242 tabelas e 1.615 atributos.

4.1.1.5. Amostra utilizada

Para o experimento 1 foram considerados subconjuntos de tabelas de cada um dos LMS, conforme demostrado na Tabela 4.1. Esses subconjuntos de dados representam os elementos chave de LMS, segundo Dron e Anderson (2014) e são formados por 180 tabelas e 1.458 atributos.

Esquema	Tabelas	Atributos
Moodle	91	835
Sakai	25	213
ATutor	28	185
Forma.lms	36	225

Tabela 4.1. Subconjunto de dados do experimento 1

4.1.2. Execução do Experimento

Para execução do experimento foi necessário percorrer os passos do processo de recomendação de vocabulários para mapeamento de dados conectados definidos no capítulo 3. Para cada iteração, um esquema precisou ser definido e seus respectivos parâmetros informados. Os valores foram definidos após observações empíricas conforma Tabela 4.2.

[29] https://www.formalms.org/

Iteração	Esquema	Algoritmo de similaridade	Número de agrupamento	Qtde vocabulário na busca	Alinhamento semântico	Valor de corte
I	Moodle	Lin	7	5	Não	0.9
II	Sakai	Jaccard	5	5	Não	0.9
III	ATutor	WuPalmer	6	5	Não	0.9
IV	Forma.lms	Jaccard	8	5	Não	0.9

Tabela 4.2. Valores empíricos definidos no experimento 1

4.1.2.1 Coleta de Dados

O processo de coleta de dados é feito após cada iteração através dos artefatos gerados pelo protótipo RVMDC. Como resultado da execução é gerado um arquivo com o alinhamento entre tabelas/atributos com vocabulários/propriedades. Para cada linha é necessária a avaliação manual por um especialista do domínio, como pode ser vista no fragmento apresentado na Tabela 4.3.

Tabela/Atributo	Vocabulário/Propriedade	Sintático	Semântico
mdl_forum_posts			
parent	theatre:parent_venue	Não	Não
created	sioc:created_at	Sim	Sim
subject	sioc:subject	Sim	Sim
messageformat			
totalscore			
discussion	sioc:has_discussion	Sim	Sim
message			
userid	sioc:User	Sim	Sim
messagetrust			
attachment	sioc:attachment	Sim	Sim
mailnow			
modified	sioc:modified_at	Sim	Sim
id	sioc:id	Sim	Sim
mailed			

Tabela 4.3. Fragmento da avaliação do resultado do alinhamento tabela/atributo com vocabulários/propriedades.

O especialista deve levar em consideração a associação da tabela/atributo com a classe/propriedade da possível correspondência, nomeado nesse trabalho como associação sintática. Outra avaliação é feita levando em consideração o domínio do vocabulário além da associação sintática, chamada de associação semântica. No

exemplo apresentado na Tabela 4.4 é feita a associação entre a tabela/atributo mdl_question_match[30]/question e o vocabulário/atributo disco[31]/Question que é valida sintaticamente. Ambas representam informações que tratam um questionamento, porém a tabela faz referência a uma questão no domínio educacional e o vocabulário representa uma questão no domínio de pesquisas científicas.

Tabela/Atributo	Vocabulário/Propriedade	Sintático	Semântico
mdl_question_match	Disco		
question	disco:Question	Sim	Não

Tabela 4.4. Fragmento da avaliação do resultado do alinhamento mdl_question_match/question com disco/Question.

4.2 Experimento 2

Nesse segundo experimento foram utilizados esquemas de diversos tamanhos e domínios que também foram utilizados na validação de propostas levantadas nos trabalhos relacionados (AUER et al., 2009; POLFLIET; ICHISE, 2010).

4.2.1 Esquemas de banco de dados

Nessa seção são descritos os ambientes e os esquemas de banco de dados utilizados no experimento.

4.2.1.1 World

World[32] é um banco de dados de exemplo que contém informações de países e cidades ao redor do mundo, que utiliza dados extraídos de Statistics Finland[33]. Seu esquema possui 3 tabelas e 24 atributos.

30 http://www.examulator.com/er/
31 http://rdf-vocabulary.ddialliance.org/discovery.html
32 https://dev.mysql.com/doc/world-setup/en/
33 http://www.stat.fi/worldinfigures

4.2.1.2 Sakila

Sakila[34] é um banco de dados de exemplo projetado para representar uma loja de aluguel de DVD on-line. Seu esquema possui 16 tabelas e 132 atributos.

4.2.1.3 Wordpress

Wordpress[35] é uma plataforma de gerenciamento de conteúdo distribuída sob a licença GPL. Possui uma grande quantidade de funcionalidades que estendem a plataforma e disponibiliza novas funcionalidades através de plugins construídos por uma comunidade de desenvolvedores. Seu esquema básico possui 12 tabelas e 94 atributos.

4.2.1.4 OSCommerce

O OsCommerce Online Merchant[36] é uma solução completa para disponibilização de lojas virtuais. Possui tanto um frontend de catálogo de produtos como um backend de ferramentas de administração baseado em tecnologias Web. É distribuído sob a licença GPL. Possui uma comunidade de desenvolvedores ativas que estendem a plataforma e disponibilizam um conjunto de recursos adicionais. Seu esquema básico possui 49 tabelas e 343 atributos.

Para este segundo experimento foram consideradas todas as tabelas, conforme demostrado na Tabela 4.5. Esse conjunto de dados é formado por 80 tabelas e 593 atributos.

34 https://dev.mysql.com/doc/sakila/en/
35 https://wordpress.org/
36 https://www.oscommerce.com/

Esquema	Tabelas	Atributos
World	3	24
Sakila	16	132
Wordpress	12	94
OSCommerce	49	343

Tabela 4.5. Conjunto de dados do experimento 2

4.2.2 Execução do Experimento

Para execução do experimento é necessário percorrer os passos do processo de recomendação de vocabulários para mapeamento de dados conectados definidos no capítulo 3. Para cada iteração, um esquema deve ser definido e seus respectivos parâmetros informados. Os valores foram definidos após observações empíricas, conforme Tabela 4.6.

Iteração	Esquema	Algoritmo de similaridade	Número de agrupamento	Qtde vocabulário na busca	Alinhamento semântico	Valor de corte
I	World	Wupalmer	1	5	não	0.9
II	Sakila	Wupalmer	15	5	não	0.9
III	Wordpress	WuPalmer	5	5	não	0.9
IV	OSCommerce	Wulpalmer	10	5	não	0.9

Tabela 4.6. Valores empíricos definidos no experimento 2

4.2.3 Coleta de dados

O processo de coleta de dados foi feito após cada iteração através dos artefatos gerados pelo protótipo RVMDC. Como resultado da execução foi gerado um arquivo com o alinhamento entre tabelas/atributos com vocabulários/propriedades.

5. Análise dos Resultados

Neste capítulo são apresentados os resultados relacionados aos dois experimentos realizados para validar a abordagem proposta. No primeiro experimento realizado no domínio educacional foram agrupados os resultados em relação aos esquemas e as estratégias de mapeamento entre as tabelas dos esquemas e os vocabulários disponíveis no repositório LOV.

5.1. Análise das Métricas do Primeiro Experimento

A Figura 5.1 mostra os resultados da precisão sintática dos alinhamentos, onde a estratégia de minimizar a quantidade de vocabulários apresentou os melhores resultados em 3 dos 4 esquemas.

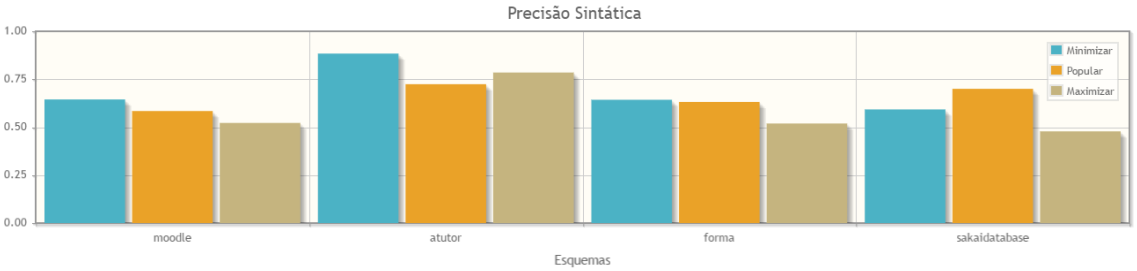


Figura 5.1. Relação da precisão sintática entre as estratégias de mapeamento e a recomendação de vocabulários

Analisando os resultados da precisão semântica do gráfico da Figura 5.2, a estratégia de reutilização de vocabulários mais populares apresentou os melhores resultados em 3 dos 4 esquemas.

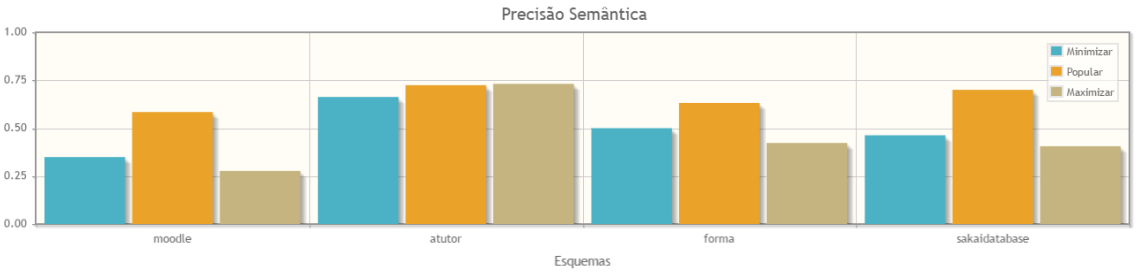


Figura 5.2. Relação da precisão semântica entre as estratégias de mapeamento e a recomendação de vocabulários

Os resultados observados que levam em consideração apenas a precisão sintática apresentam melhor performance com relação aos resultados da precisão semântica.

Porém nessa avaliação não é levado em consideração o domínio do vocabulário, o que implica em perda semântica. A proposta aqui foi mostrar que há um conjunto de classes e propriedades que poderiam ser reusadas se forem criados vocabulários mais genéricos ou que possam ser estendidos por vocabulários mais especializados.

Observando o gráfico da Figura 5.3, onde em vez de agrupar por esquemas foi feito o agrupamento por conceitos que representam os principais componentes de um LMS. Os maiores valores de precisão semântica foram alcançados pela estratégia de reutilizar vocabulários mais populares. Os vocabulários recomendados foram os vocabulários genéricos como FOAF[37], SIOC[38] e TSIOC[39], que são amplamente usados na comunidade de dados conectados.

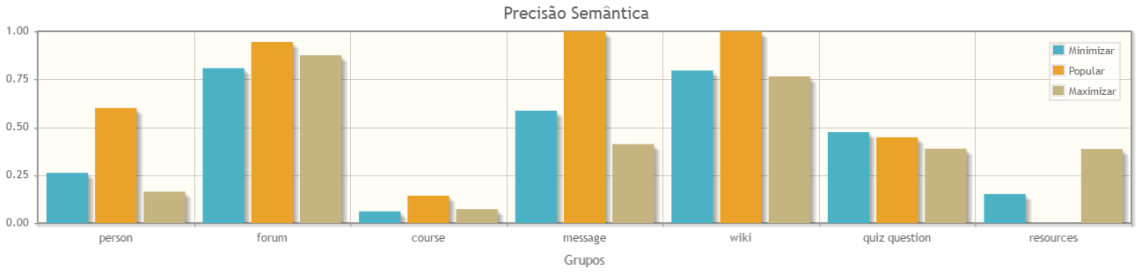


Figura 5.3. Relação da precisão sintática entre as estratégias de mapeamento e a recomendação de vocabulários por conceitos

Os resultados que medem a cobertura do mapeamento do alinhamento tabela/atributo com o vocabulário/propriedade apresentaram resultados muito baixos em ambas as estratégias, com uma leve vantagem para a estratégia de maximizar a quantidade de vocabulários e minimizar a quantidade vocabulários com relação à estratégia de reutilizar vocabulários mais populares na recomendação, conforme podemos observar nas Figuras 5.4 e 5.5. É importante ressaltar que todos os atributos das tabelas foram levados em consideração e muitos tratam de questões específicas da

37 http://xmlns.com/foaf/spec/
38 http://rdfs.org/sioc/spec/
39 Estende a Ontologia Core SIOC (*Semantically-Interlinked Online Communities*) definindo subclasses e subpropriedades de termos SIOC.

aplicação, podendo ser interessante ou não, dependendo da natureza da disponibilização dos dados.

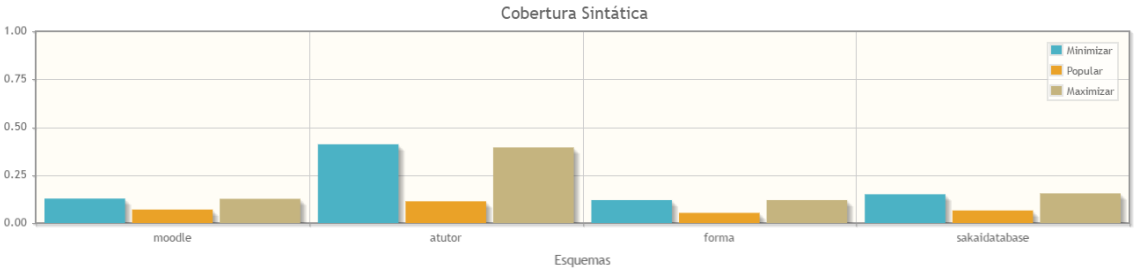


Figura 5.4. Relação da cobertura sintática entre as estratégias de mapeamento e a recomendação de vocabulários

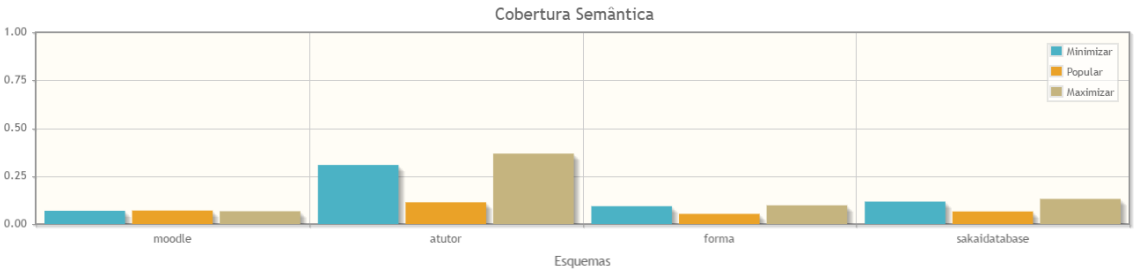


Figura 5.5. Relação da cobertura semântica entre as estratégias de mapeamento e a recomendação de vocabulários.

A Medida-F é utilizada para medir a qualidade do alinhamento conforme apresentado no Capítulo 4. Os resultados mostram que não houve diferenças significativas entre as estratégias de mapeamento, como pode ser observado nas Figuras 5.6 e 5.7.

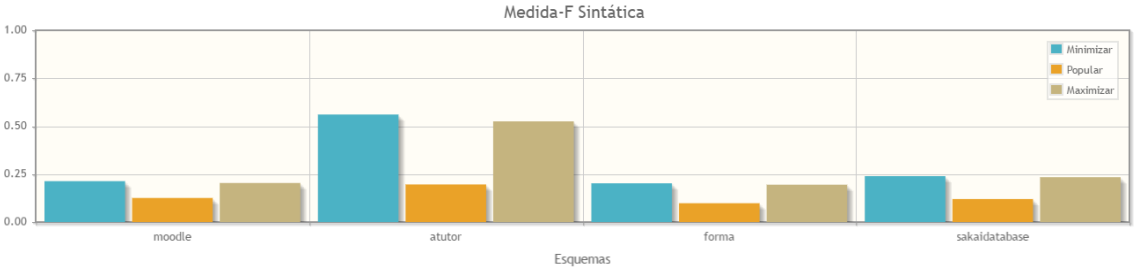


Figura 5.6. Relação da medida-f sintática entre as estratégias de mapeamento e a recomendação de vocabulários

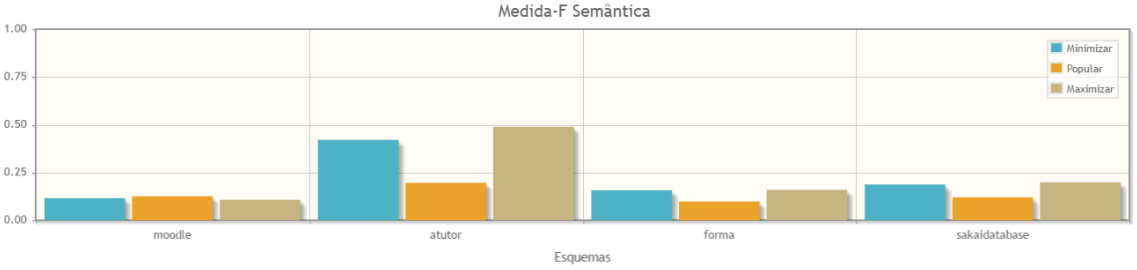


Figura 5.7. Relação da medida-f Semântica entre as estratégias de mapeamento e a recomendação de vocabulários

5.2. Análise das Métricas do Segundo Experimento

O segundo experimento foi realizado em um conjunto de domínios diversos e os resultados foram agrupados em relação aos esquemas e às estratégias de mapeamento entre as tabelas dos esquemas e os vocabulários disponíveis no repositório LOV.

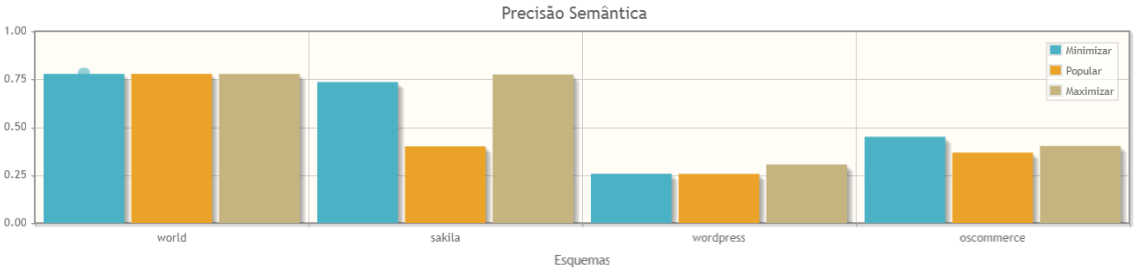


Figura 5.8. Relação da precisão semântica entre as estratégias de mapeamento e a recomendação de vocabulários

A Figura 5.8 mostra os resultados da precisão semântica dos alinhamentos. Não houve diferença significativa entre 3 dos 4 esquemas com relação as estratégias de mapeamento. No esquema Sakila, a precisão usando a estratégia de reuso de vocabulários mais populares apresentou o pior resultado, porém não podemos atribuir o resultado a estratégia e sim a uma limitação da proposta em recuperar vocabulários a partir de termos enriquecidos. Por exemplo, para as tabelas de aluguel de DVD foi recuperado o vocabulário com classes de aluguel de veículos. Os esquemas Wordpress e OSCommerce apresentaram uma precisão abaixo de 50%, que foi devido a alguns atributos possuírem o nome da tabela como prefixo do seu nome.

A Figura 5.9 mostra os resultados da cobertura semântica dos alinhamentos. Como apresentado na análise do experimento 1, os resultados da cobertura apresentaram resultados baixos em ambas as estratégias, com uma leve vantagem para a estratégia de maximizar em dois dos esquemas. A cobertura apresentou um resultado um pouco melhor para os esquemas mais simples e menores, por exemplo o esquema World.

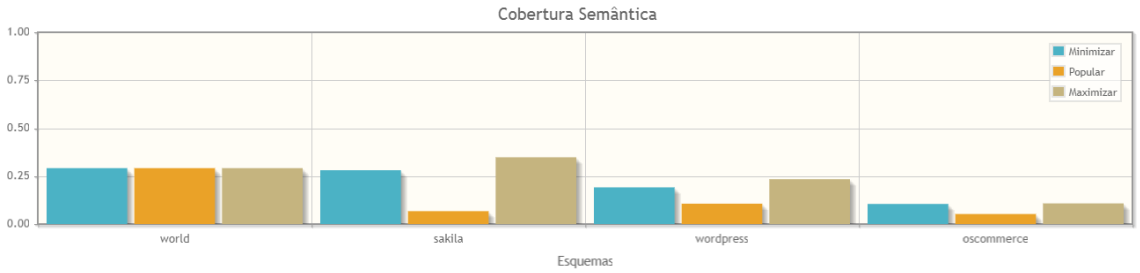


Figura 5.9. Relação da cobertura semântica entre as estratégias de mapeamento e a recomendação de vocabulários

Os resultados do gráfico da medida-F apresentado na Figura 5.10, mostram uma pequena vantagem em dois dos esquemas com relação à estratégia de maximizar vocabulários, porém a diferença dos valores não é significativa. Não é possível concluir qual estratégia é a melhor nessa abordagem.

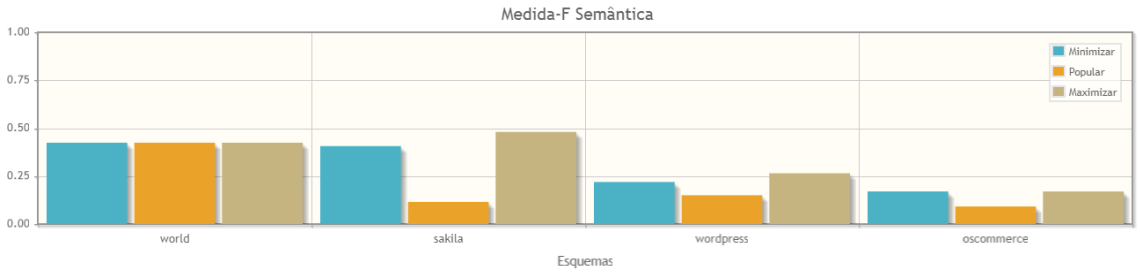


Figura 5.10. Relação da medida-f semântica entre as estratégias de mapeamento e a recomendação de vocabulários

A figura 5.11 apresenta um resumo de todos os resultados dos experimentos. Foi possível perceber que a recomendação (semi-)automática apresenta bons resultados, especialmente em termos de precisão em ambas as estratégias. A cobertura apresentou resultados baixos com leve vantagem para a estratégia de maximizar a quantidade de vocabulários.

Database schema	Popular Vocabularies			Minimize vocabularies			Maximize vocabularies		
	Precision	Recall	F-measure	Precision	Recall	F-measure	Precision	Recall	F-measure
Moodle	0.58	0.07	0.12	0.34	0.06	0.11	0.27	0.06	0.10
Sakai	0.7	0.06	0.12	0.46	0.11	0.18	0.40	0.13	0.19
Atutor	0.72	0.11	0.19	0.66	0.30	0.42	0.78	0.36	0.48
Forma	0.63	0.05	0.09	0.5	0.09	0.15	0.42	0.09	0.15
World	0.77	0.29	0.42	0.77	0.29	0.42	0.77	0.29	0.42
Sakila	0.4	0.06	0.11	0.73	0.28	0.40	0.77	0.34	0.48
Wordpress	0.30	0.23	0.26	0.25	0.19	0.21	0.30	0.23	0.26
Oscommerce	0.36	0.05	0.09	0.45	0.10	0.17	0.40	0.10	0.17

Figura 5.11 - Resumo de todos os resultados dos experimentos.

6. Conclusão

Neste capítulo apresentamos as principais contribuições desta dissertação, bem como sugerimos trabalhos futuros.

6.1. Contribuições da Dissertação

Nesta dissertação diferentes abordagens (estratégias) para recomendação de vocabulários de dados conectados foram analisadas para apoiar a publicação de dados relacionais como dados conectados.

Este trabalho visa promover a reutilização de vocabulários de dados já existentes na publicação de novos conjuntos de dados, o que consiste em uma das melhores práticas de dados na Web. Trata-se de um trabalho relevante para o contexto de Ciência da Web, pois está relacionado à melhor compreensão, processabilidade, reuso, confiança e interoperabilidade dos dados na Web.

Em um contexto de Informática na Educação, este trabalho, ao considerar um experimento utilizando quatro diferentes ambientes de aprendizagem (também chamados de LMS – *Learning Management Systems*), analisou a recomendação de vocabulários relacionados aos conceitos adotados nestes ambientes, provendo mecanismos para a abertura de dados dos ambientes de aprendizagem, bem como sua interoperabilidade e conectividade (através da publicação de dados conectados) entre as diversas plataformas.

Diferentemente das propostas encontradas na literatura, nesta dissertação foi apresentado um processo e uma arquitetura que possibilitam uma recomendação semiautomática de vocabulários de dados conectados com base nos esquemas fontes relacionais dos dados a serem publicados como dados conectados. Deste modo, torna-se possível não apenas recomendar vocabulários que possam ser reutilizados na publicação

de novos dados conectados, mas também analisar diferentes estratégias

O processo adotado se baseia na extração do esquema do banco de dados relacional original (fonte de dados), seu pré-processamento, com posterior agrupamento (*clustering*) das tabelas e classificação dos grupos semanticamente, para então buscar vocabulários e seus metadados e recomendar os vocabulários existentes de dados conectados. A recomendação seguiu três diferentes estratégias: minimizar o número total de vocabulários, maximizar o número de vocabulários e reusar vocabulários mais populares.

A recomendação foi analisada em dois diferentes experimentos, cada um com quatro esquemas relacionais diferentes de sistemas reais. De acordo com as análises dos resultados, se observa que as diferentes estratégias podem ser melhor aplicadas em cenários específicos, dependendo da especificidade do domínio e a complexidade do esquema. Apesar das variações dos resultados em relação às estratégias, também foi possível perceber que a recomendação (semi-)automática apresenta bons resultados, especialmente em termos de precisão, embora diversas melhorias possam ser trabalhadas.

Finalmente, vale ressaltar as contribuições técnicas, através do desenvolvimento da ferramenta de recomendação e dos experimentos realizados.

6.2. Trabalhos Futuros

O principal trabalho futuro refere-se à combinação das diferentes estratégias na recomendação de vocabulários de dados conectados para publicação de dados relacionais como dados abertos conectados. A partir do trabalho realizado nesta dissertação é possível definir heurísticas de uso das estratégias com base nas características do esquema relacional fonte de modo a gerar uma melhor recomendação

de vocabulários.

Também vale ressaltar a importância de um trabalho futuro no sentido de validar a recomendação automática com usuários responsáveis pela publicação de dados conectados, considerando diferentes complexidades dos esquemas fonte.

Outros trabalhos futuros envolvem a melhoria das soluções adotadas no processo de recomendação, em termos de implementação do sistema de recomendação, em especial aqueles relacionados ao tratamento sintático e tratamento semântico. Assim, diversas otimizações podem ser consideradas para melhorar a performance da recomendação e consequentemente os resultados.

Avaliações mais detalhadas em termos de reutilização de vocabulários de dados conectados, bem como outras características relacionadas como interoperabilidade, compreensão, processabilidade, confiança e acessibilidade dos dados poderiam ser consideradas como trabalhos futuros.

Finalmente, alguns trabalhos que podem ser de interesse com base nos resultados deste trabalho, envolvem a proposição de políticas públicas voltadas para a abertura de dados, em especial os de ambientes de aprendizagem de instituições brasileiras. Pode-se trabalhar para que os dados destes ambientes sejam publicados como dados conectados, resultando em um grande conjunto de dados a serem explorados para o entendimento de procedimentos acadêmicos, bem como questões relacionadas a sucessos e fracassos no ensino-aprendizagem. Este conhecimento poderia promover melhorias significativas na Educação Brasileira.

REFERÊNCIAS BIBLIOGRÁFICAS

AUER, Sören; DIETZOLD, S.; LEHMANN, J.; HELLMANN, S.; AUMUELLER, D. Triplify: light-weight linked data publication from relational databases. In: Proceedings of the 18th international conference on World wide web. ACM, 2009. p. 621-630.

BARRASA, Jesús; CORCHO, Óscar; GÓMEZ-PÉREZ, Asunción. R2O, an Extensible and Semantically based Database-to-Ontology Mapping Language. In: in In **Proceedings of the 2nd Workshop on Semantic Web and Databases (SWDB2004)**. 2004.

BAUER, Florian; KALTENBÖCK, Martin. Linked open data: The essentials - A Quick Start Guide for Decision Makers. **Edition mono/monochrom**, Vienna, 2011.

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The semantic web. **Scientific american**, v. 284, n. 5, p. 28-37, 2001.

BIZER, C.; LEHMANN, J.; KOBILAROV, G.; AUER, S.; BECKER, C.; CYGANIAK, R.; HELLMANN, S. DBpedia-A crystallization point for the Web of Data. **Web Semantics: science, services and agents on the world wide web**, v. 7, n. 3, p. 154-165, 2009.

CHEBOTKO, Artem; LU, Shiyong; FOTOUHI, Farshad. Semantics preserving SPARQL-to-SQL translation. **Data & Knowledge Engineering**, v. 68, n. 10, p. 973-1000, 2009.

CHENG, Gong; GE, Weiyi; QU, Yuzhong. Falcons: searching and browsing entities on the semantic web. In: **Proceedings of the 17th international conference on World Wide Web**. ACM, 2008. p. 1101-1102.

CHENG, Gong; QU, Yuzhong. Searching Linked Objects with Falcons: Approach, Implementation and Evaluation. **International Journal on Semantic Web and Information Systems (IJSWIS)**, v. 5, n. 3, p. 49-70, 2009.

D'AQUIN, Mathieu. Linked data for open and distance learning. **Commonwealth of Learning, Vancouver**, 2012.

D'AQUIN, Mathieu. On the use of Linked Open Data in education: Current and future practices. In: **Open data for education**. Springer International Publishing, 2016. p. 3-15.

D'AQUIN, Mathieu; NOY, Natalya F. Where to publish and find ontologies? A survey of ontology libraries. **Web Semantics: Science, Services and Agents on the World Wide Web**, v. 11, p. 96-111, 2012.

D'AQUIN, MATHIEU; BALDASSARRE, C.; GRIDINOC, L.; SABOU, M.; ANGELETOU, S.; MOTTA, E. Watson: Supporting next generation semantic web applications. In: **Proceedings of the IADIS WWW/Internet International Conference**. IADIS, 2007.

DIETZE, Stefan; SANCHEZ-ALONSO, S.; EBNER, H.; QING YU, H.; GIORDANO, D.; MARENZI, I.; PEREIRA NUNES, B. Interlinking educational resources and the web of data: A survey of challenges and approaches. **Program**, v. 47, n. 1, p. 60-91, 2013.

DING, L.; FININ, T.; JOSHI, A.; PAN, R.; COST, R. S.; PENG, Y.; REDDIVARI, P.; DOSHI, V.; SACHS, J. Swoogle: a search and metadata engine for the semantic web. In: **Proceedings of the thirteenth ACM international conference on Information and knowledge management**. ACM, 2004. p. 652-659.

DRON, John; ANDERSON, Terry. **Teaching crowds: Learning and social media**. Athabasca University Press, 2014.

ELMASRI, Ramez; NAVATHE, Shamkant B. **Sistemas de banco de dados**. Addison. Ribeirão Preto SP, 2005.

ERLING, Orri; MIKHAILOV, Ivan. RDF Support in the Virtuoso DBMS. In: **Networked Knowledge-Networked Media**. Springer Berlin Heidelberg, 2009. p. 7-24.

FENG, Jin; ZHOU, Yi-Ming; MARTIN, Trevor. Sentence similarity based on relevance. In: **Proceedings of IPMU**. 2008. p. 833.

FERNÁNDEZ, Miriam; CANTADOR, Iván; CASTELLS, Pablo. CORE: A Tool for Collaborative Ontology Reuse and Evaluation. In: **4th International EON (Evaluation of Ontologies for the Web) Workshop @ the 15th International World Wide Web Conference**. 2006.

GANTZ, John; REINSEL, David. The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east. **IDC iView: IDC Analyze the future**, v. 2007, n. 2012, p. 1-16, 2012.

HALAÇ, T. G.; ERDEN, B.; INAN, E.; OGUZ, D.; GOCEBE, P.; DIKENELLI, O. Publishing and linking university data considering the dynamism of datasources. In: **Proceedings of the 9th International Conference on Semantic Systems**. ACM, 2013. p. 140-145.

HALL, Mark; FRANK, E.; HOLMES, G.; PFAHRINGER, B.; REUTEMANN, P.; WITTEN, I. H. The WEKA data mining software: an update. **ACM SIGKDD explorations newsletter**, v. 11, n. 1, p. 10-18, 2009.

HAN, Jiawei; PEI, Jian; KAMBER, Micheline. **Data mining: concepts and techniques**. Elsevier, 2011.

HYLAND, Bernadette; WOOD, David. The joy of data-a cookbook for publishing linked government data on the web. **Linking government data**, p. 3-26, 2011.

ISOTANI, Seiji; BITTENCOURT, Ig Ibert. **Dados Abertos Conectados: Em busca da Web do Conhecimento**. Novatec Editora, 2015.

MANNING, C. D.; SURDEANU, M.; BAUER, J.; FINKEL, J. R.; BETHARD, S.; MCCLOSKY, D. The stanford corenlp natural language processing toolkit. In **Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations**. 2014. p. 55-60.

MICHEL, Franck; MONTAGNAT, Johan; FARON-ZUCKER, Catherine. **A survey of RDB to RDF translation approaches and tools**. 2014. Tese de Doutorado. I3S.

MILLER, George A. WordNet: a lexical database for English. **Communications of the ACM**, v. 38, n. 11, p. 39-41, 1995.

POLFLIET, Simeon; ICHISE, Ryutaro. Automated mapping generation for converting databases into linked data. In: **Proceedings of the 2010 International Conference on Posters & Demonstrations Track-Volume 658**. CEUR-WS. org, 2010. p. 173-176.

PRATES, J. C., FRITZEN, E.; SIQUEIRA, S. W.; BRAZ, M. H. L.; ANDRADE, L. C. Contextual web searches in Facebook using learning materials and discussion messages. **Computers in Human Behavior**, v. 29, n. 2, p. 386-394, 2013.

PRIYATNA, Freddy; CORCHO, Oscar; SEQUEDA, Juan. Formalisation and experiences of R2RML-based SPARQL to SQL query translation using morph. In:

Proceedings of the 23rd international conference on World wide web. ACM, 2014. p. 479-490.

RAHM, Erhard; BERNSTEIN, Philip A. A survey of approaches to automatic schema matching. **the VLDB Journal**, v. 10, n. 4, p. 334-350, 2001.

SALAS, P. E. R. *StdTrip: An a priori design process for publishing Linked Data*. 2011. Dissertação (Mestrado em Informática) - Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro. 2011.

SALAS, Percy; VITERBO, J.; BREITMAN, K.; CASANOVA, M. A.. Stdtrip: Promoting the reuse of standard vocabularies in open government data. **Linking government data**, p. 113-133, 2011.

SCHAIBLE, J.; GOTTRON, T.; SCHEGLMANN, S.; SCHERP, A. Lover: support for modeling data using linked open vocabularies. In: **Proceedings of the Joint EDBT/ICDT 2013 Workshops**. ACM, 2013. p. 89-92.

SCHAIBLE, Johann; GOTTRON, Thomas; SCHERP, Ansgar. Survey on common strategies of vocabulary reuse in linked open data modeling. In: **European Semantic Web Conference**. Springer, Cham, 2014. p. 457-472.

SCHAIBLE, Johann; GOTTRON, Thomas; SCHERP, Ansgar. TermPicker: Enabling the reuse of vocabulary terms by exploiting data from the Linked Open Data cloud. In: **International Semantic Web Conference**. Springer, Cham, 2016. p. 101-117.

SCHARFFE, François; ATEMEZING, G.; TRONCY, R.; GANDON, F.; VILLATA, S.; BUCHER, B.; HAMDI, F.; BIHANIC, L.; KÉPÉKLIAN, G.; COTTON, F.; EUZENAT, J.; FAN, Z.; VANDENBUSSCHE, P-Y.; VATANT, B. Enabling linked-data publication with the datalift platform. In: **Proc. AAAI workshop on semantic cities**. 2012.

SEABORNE, A.; STEER, D.; WILLIAMS, S. SQL-RDF (SquirrelRDF). In: **W3C Workshop on RDF Access to Relational Databases**. Cambridge, USA, 2007.

SEQUEDA, Juan F.; DEPENA, Rudy; MIRANKER, Daniel P. Ultrawrap: Using sql views for rdb2rdf. **Proc. of ISWC2009**, 2009.

SEQUEDA, Juan F.; MIRANKER, Daniel P. Ultrawrap Mapper: A Semi-Automatic Relational Database to RDF (RDB2RDF) Mapping Tool. In: International Semantic Web Conference (Posters & Demos). 2015.

SHVAIKO, Pavel; EUZENAT, Jérôme. A survey of schema-based matching approaches. **Journal on data semantics** IV, p. 146-171, 2005.

SIVARAJAH, U.; KAMAL, M. M.; IRANI, Z.; WEERAKKODY, V. Critical analysis of Big Data challenges and analytical methods. **Journal of Business Research**, v. 70, p. 263-286, 2017.

SPANOS, Dimitrios-Emmanuel; STAVROU, Periklis; MITROU, Nikolas. Bringing relational databases into the semantic web: A survey. **Semantic Web**, v. 3, n. 2, p. 169-209, 2012.

STADTMÜLLER, Steffen; HARTH, Andreas; GROBELNIK, Marko. Accessing information about linked data vocabularies with vocab. cc. In: **Semantic Web and Web Science**. Springer, New York, NY, 2013. p. 391-396.

SVIHLA, Martin; JELINEK, Ivan. Two layer mapping from database to RDF. **Proceedings of Electronic Computers and Informatics (ECI)**, 2004.

TIRMIZI, Syed Hamid; SEQUEDA, Juan; MIRANKER, Daniel. Translating sql applications to the semantic web. In: **International Conference on Database and Expert Systems Applications**. Springer, Berlin, Heidelberg, 2008. p. 450-464.

TIROPANIS, Thanassis et al. Semantic technologies for learning and teaching in the Web 2.0 era. **IEEE Intelligent Systems**, v. 24, n. 6, p. 49-53, 2009.

VAN NUFFELEN, B.; JANEV, V.; MARTIN, M.; MIJOVIC, V.; TRAMP, S. Supporting the linked data life cycle using an integrated tool stack. In: **Linked Open Data--Creating Knowledge Out of Interlinked Data**. Springer International Publishing, 2014. p. 108-129.

VANDENBUSSCHE, P. Y.; Atemezing, G. A.; Poveda-Villalón, M.; Vatant, B. Linked Open Vocabularies (LOV): a gateway to reusable semantic vocabularies on the Web. **Semantic Web**, v. 8, n. 3, p. 437-452, 2017.

VEGA-GORGOJO, Guillermo; TIROPANIS, Thanassis; MILLARD, David E. The opportunity of linked data for the European higher education area. **International Journal of Information and Education Technology**, v. 6, n. 1, p. 58, 2016.

VILLAZÓN-TERRAZAS, B.; VILCHES-BLÁZQUEZ, L. M.; CORCHO, O.; GÓMEZ-PÉREZ, A. Methodological guidelines for publishing government linked data. **Linking government data**, p. 27-49, 2011.

APÊNDICE A

WordNet é um banco de dados léxico da língua inglesa, criado e disponibilizado pela Universidade de Princenton. Nele substantivos, verbos, adjetivos e advérbios são agrupados através das suas relações semânticas, conhecido como synsets (conceitos). Os conceitos ou synsets são conjunto de palavras que podem ser utilizados sem que o contexto seja alterado. Isto é, palavras que pertençam ao mesmo synset são consideradas semanticamente equivalentes.

As principais relações semânticas encontradas na WordNet são:

• Sinônimos: o mesmo que (X significa o mesmo que Y).

• Hiperônimos: termo geral ou supertipo para (X é o termo geral para Y).

• Hipônimos: tipo de ou subtipo (X é um tipo de Y).

• Merônimos: parte de (X é parte de Y).

• Holônimos: tem parte (X tem a parte de Y).

• Antônimos: contrário de (X é o contrário de Y).

Os substantivos e os verbos são hierarquicamente organizados por suas relações de Hiperonímia e Hiponímia. Isso permite criar conceitos genéricos que formam a raiz de múltiplas hierarquias que representam campos semânticos distintos com seu próprio vocabulário. Essa organização permite criar uma classificação utilizando as 25 entidades para os substantivos, conforme Tabela 1 e 15 para os verbos, apresentado na Tabela 2.

Atualmente a WordNet está disnponível na versão 3.0 e disponibiliza um total de 117659 synsets, sendo 82115 substantivos, 13767 verbos, 18156 adjetivos e 3621

advérbios.

Atributos semânticos	Descrição
noun.tops	topo exclusivo para substantivo.
noun.act	substantivos que denotam atos ou ações.
noun.animal	substantivos animais que denotam animais.
noun.artifact	substantivos que denotam objetos sintéticos.
noun.attribute	substantivos que denotam atributos de pessoas e objetos.
noun.body	substantivos que indicam partes do corpo.
noun.cognition	substantivos que denotam processos e conteúdos cognitivos.
noun.communication	substantivos denotando processos e conteúdos comunicativos.
noun.event	substantivos que denotam eventos naturais.
noun.feeling	substantivos que denotam sentimentos e emoções.
noun.food	substantivos que denotam alimentos e bebidas.
noun.group	substantivos que denotam agrupamentos de pessoas ou objetos.
noun.location	substantivos que denotam posição espacial.
noun.motive	substantivos que denotam objetivos.
noun.object	substantivos que denotam objetos naturais.
noun.person	substantivos que denotam pessoas.
noun.phenomenon	substantivos que denotam fenômenos naturais.
noun.plant	substantivos que denotam plantas.
noun.possession	substantivos denotam posse e transferência de posse.
noun.process	substantivos de processo denotando processos naturais.
noun.quantity	substantivos que indicam quantidades e unidades de medida.
noun.relation	substantivos que denotam relações entre pessoas ou coisas ou idéias.
noun.shape	substantivos que denotam formas bidimensionais e tridimensionais.
noun.state	substantivos que denotam estados de coisas estáveis.
noun.substance	substantivos que denotam substâncias.
noun.time	substantivos que denotam tempo e relações temporais.

Tabela 1 - Lista dos atributos semânticos para substântivos.

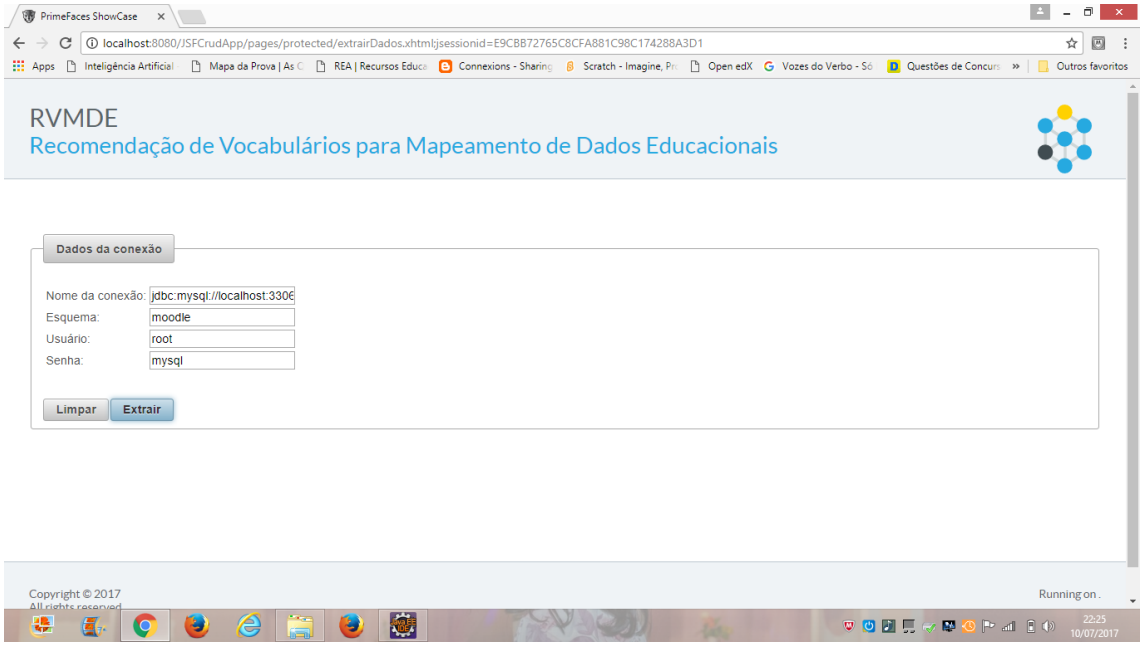
Atributos semânticos	Descrição
verb.body	verbos de aparência, de vestimentas e

	cuidados corporais.
verb.change	verbos de mudança de tamanho, temperatura, intensidade, etc.
verb.cognition	verbos de pensar, julgar, analisar, duvidar, etc.
verb.communication	verbos de contar, perguntar, pedir, cantar, etc.
verb.competition	verbos de luta, atividades atléticas, etc.
verb.consumption	verbos de comer e beber.
verb.contact	verbos de tocar, bater, amarrar, cavar, etc.
verb.creation	verbos de costura, cozimento, pintura, performance, etc.
verb.emotion	verbos de sentimentos.
verb.motion	verbos de andar, voar, nadar, etc.
verb.perception	verbos de ver, ouvir, sentir, etc.
verb.possession	verbos de compra, venda, propriedade e transferência.
verb.social	verbos de atividades sociais e eventos políticos e sociais.
verb.stative	verbos do ser, ter, relações espaciais.
verb.weather	verbos relacionados ao clima, chover, nevar, descongelar, trovejar, etc.

Tabela 2 - Lista dos atributos semânticos para verbos.

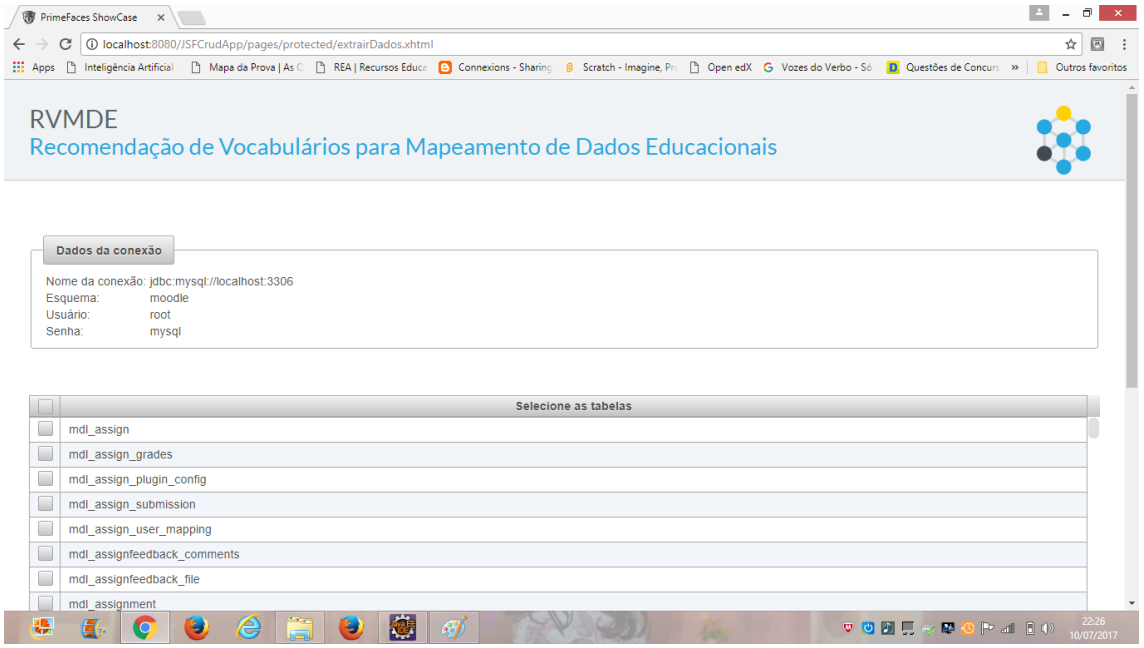
APÊNDICE B

A tela 1 permite ao publicador de dados informar os parâmetros da conexão do banco de dados relacional para extração das informações da estrutura de dados.



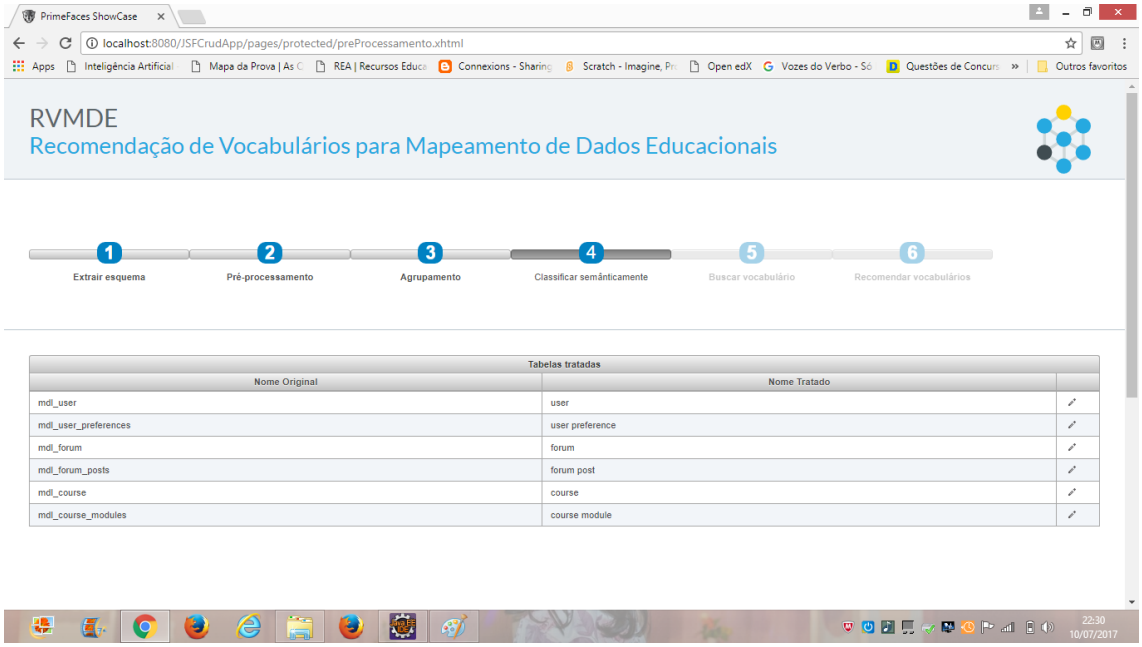
Tela 1 - Parâmetros da conexão do banco de dados relacional.

A tela 2 permite ao publicador de dados escolher as tabelas do esquema que serão passadas para as próximas etapas.



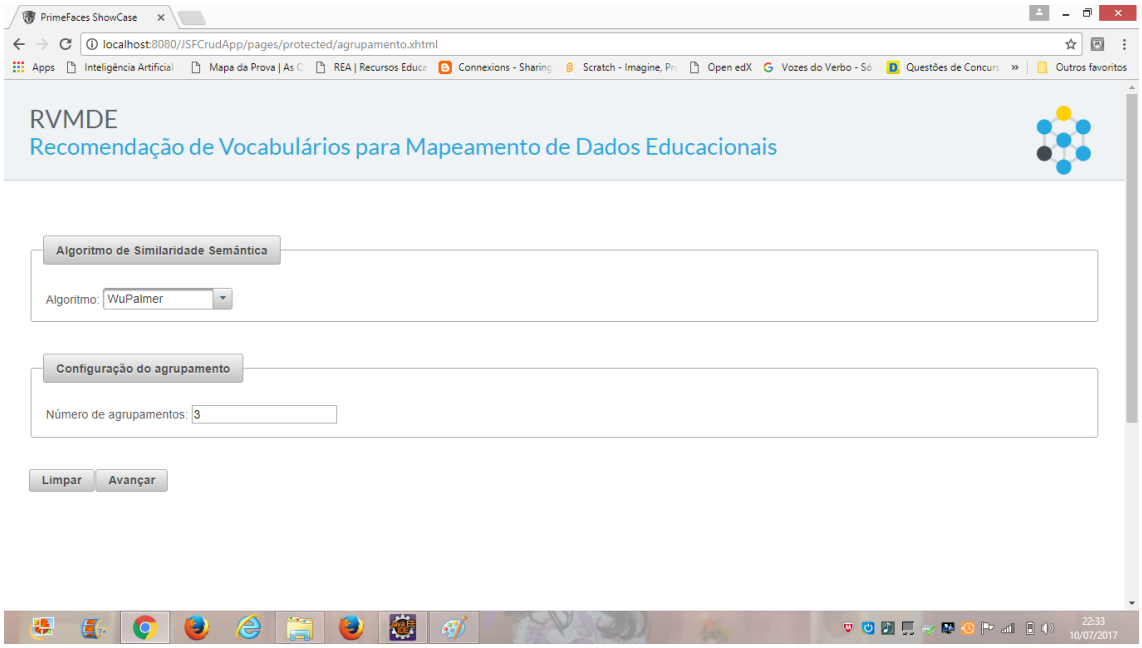
Tela 2 - Escolher tabelas do esquema

A tela 3 mostra o resultado do pré-processamento nos nomes das tabelas selecionadas.



Tela 3 - Resultado do pré-processamento.

A tela 4 permite ao publicador de dados escolher o algoritmo de similaridade e informar os parâmetros do agrupamento.

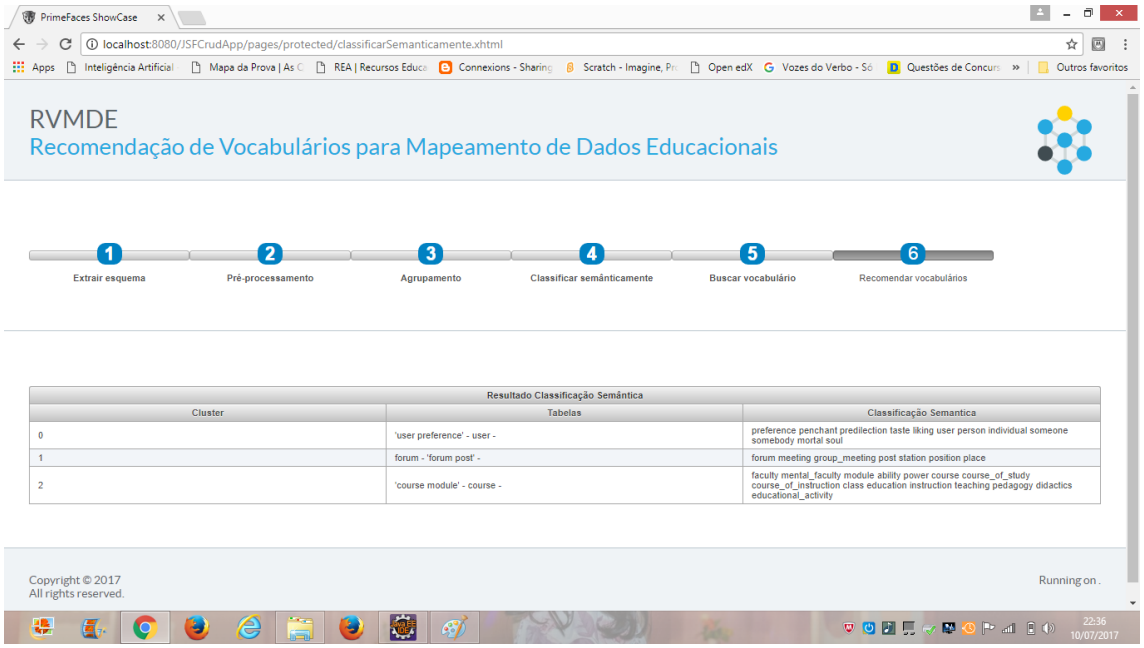


Tela 4 - Escolher algoritmo de similaridade e parâmetros do agrupamento.

A tela 5 mostra o resultado do agrupamento, permite ao publicador de dados editar os grupos criados.

Tela 5 - Resultado do agrupamento.

A tela 6 mostra o resultado da classificação semântica bem como as palavras chaves que caracterizam os grupos mais os termos relacionados extraídos das relações semânticas da WordNet.

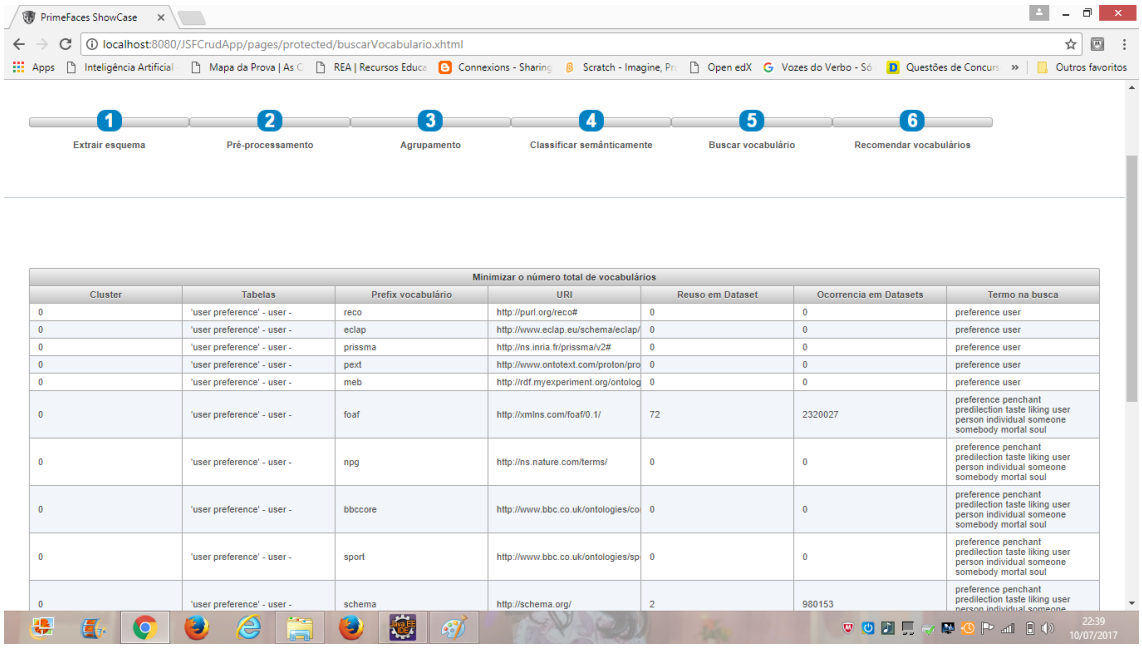


Tela 6 - Resultado da classificação semântica.

A tela 7 permite ao publicador de dados informar os parâmetros de busca de vocabulários e escolher as estratégias de mapeamento.

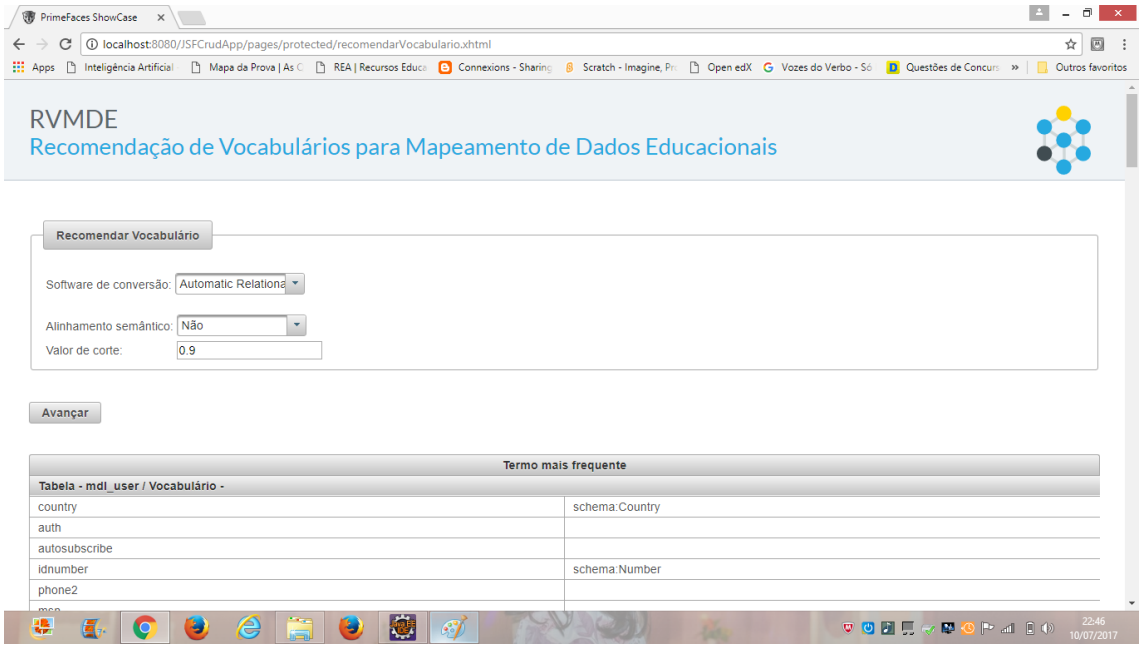
Tela 7 - Parâmetros de busca de vocabulários.

A tela 8 apresenta o resultado da busca de vocabulários.



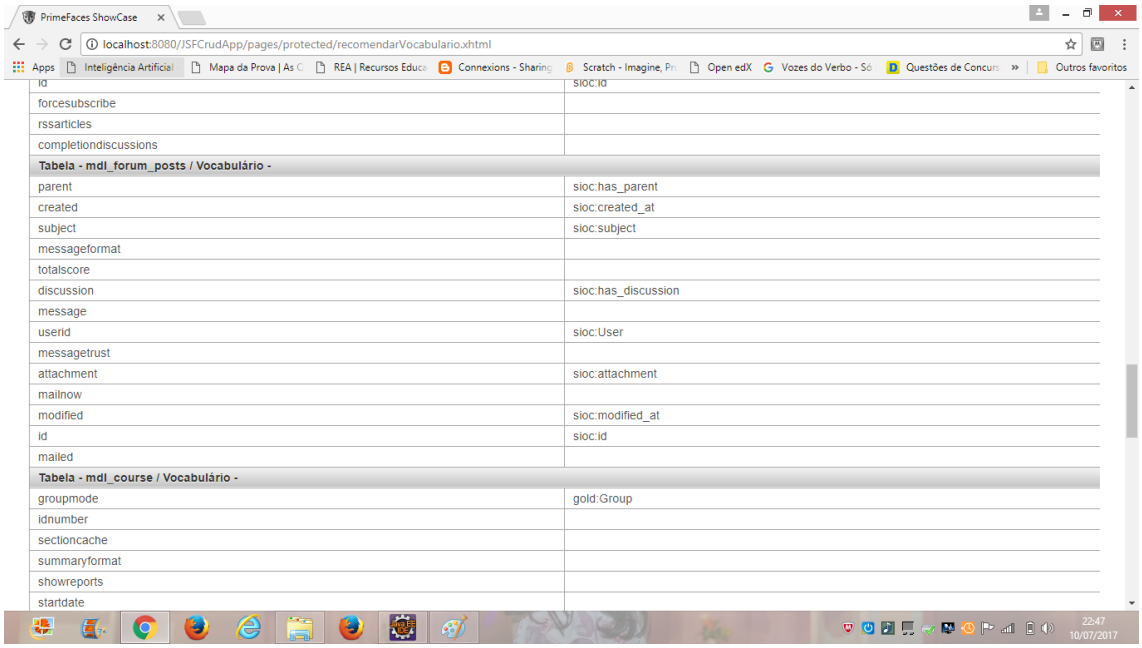
Tela 8 - Resultado da busca de vocabulários.

A tela 9 permite ao publicador de dados definir os parâmetros que serão usados na recomendação de vocabulários.



Tela 9 - Parâmetros para recomendação de vocabulários.

Na tela 10 mostra o resultado da recomendação de vocabulários com alinhamento entre os atributos das tabelas e as propriedades dos vocabulários.



Tela 10 - Resultado da recomendação de vocabulários.