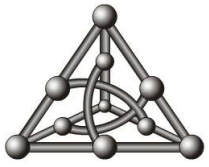
Uma Estratégia para Publicação de um *Linked Open Data* Baseado em *Data Warehouse*

Davison André Zangerolami de Oliveira

Faculdade de Computação da
Universidade Federal de Mato Grosso do Sul

Orientação: Marcelo Augusto Santos Turine



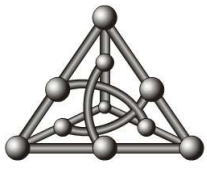
Facom
Agosto de 2012

Davison André Zangerolami de Oliveira

Uma Estratégia para Publicação de um *Linked Open Data* Baseado em *Data Warehouse*

Dissertação apresentada como parte dos requisitos para obtenção do grau de Mestre em Ciência da Computação, Curso de Mestrado em Ciência da Computação, Facom Faculdade de Computação, Fundação Universidade Federal de Mato Grosso do Sul.

Orientação: Marcelo Augusto Santos Turine



Facom
Agosto de 2012

Agradecimentos

Agradeço ao Senhor pela oportunidade que me
foi dada de mais uma realização profissional,
pelas pessoas que cruzaram meu
caminho durante esse percurso,
pela força e paciência, pois sei que nada
se concretizaria se não fosse através dele.
Agradeço a minha esposa pelo apoio e carinho.
Agradeço ao meu pai, a minha irmã, aos amigos,
ao meu orientador, enfim, a todas as pessoas
que me ajudaram nessa conquista.
Agradeço a minha mãe pelo grande exemplo de
que na vida devemos estar sempre
estudando e não parar no tempo!

Uma Estratégia para Publicação de um *Linked Open Data* Baseado em *Data Warehouse*

Resumo

O Ministério da Educação (MEC) e o Instituto de Estudos e Pesquisas em Educação Anísio Teixeira (INEP) vêm aplicando censos e avaliações educacionais a fim de auxiliar a tomada de decisão em relação às políticas educacionais no Brasil. Com isso, é preciso que os dados gerados sejam transformados em dados estatístico-educacionais de forma ágil, auxiliando a visualização da realidade socioeducacional do país. Nesse contexto, este projeto visa trabalhar com os microdados do ENEM (Exame Nacional do Ensino Médio), no período de 1998 até 2008, disponibilizados no formato ASCII, e transformá-los em um Data Warehouse (DW), o qual deve dar suporte a consultas analíticas que subsidiarão a tomada de decisão. Após a finalização do DW, é realizada a criação de um *Linked Open Data* (LOD) para o ENEM, que se destina a dar suporte à pesquisa de informações de forma automatizada e também oferecer suporte à descoberta de novas ligações/associações entre os dados. Além disso, foram selecionados e analisados três processos *open sources* para serem analisados: o primeiro utiliza as ferramentas *Stdtrip*, *Triplify* e Virtuoso; o segundo utiliza *Ontowiki*, *Olap2DataCube* e Virtuoso; e o terceiro processo utiliza as ferramentas Babel e Virtuoso, sendo que a ferramenta *Olap2DataCube* foi a escolhida para uso no projeto, pois foi a que apresentou melhor desempenho no processo de triplificação. Dessa forma, a contribuição prática deste projeto é o desenvolvimento de um modelo de processo de publicação de *Linked Data*, consistindo na triplificação dos dados do DW-ENEM e sua liberação para consulta via SPARQL no banco Virtuoso.

Palavras-chave: Avaliações educacionais, *Open source*, *Data Warehouse*, *Linked Data* e Triplificação

Uma Estratégia para Publicação de um *Linked Open Data* Baseado em *Data Warehouse*

Abstract

The Ministry of Education (MEC), together with Anísio Teixeira Institute of Educational Studies and Research (INEP), has been making assessments through censuses and educational tests, whose purpose is to assist the decision making process regarding Brazilian education policies. To accomplish that, it is necessary that all the collected data be efficiently turned into statistic-educational data, which will help picturing Brazil's socio-educational reality. That being so, this project is going to work on micro data from ENEM, from 1998 to 2008, which were made available in ASCII format, and turn such micro data into a Data Warehouse (DW) that will assist analytical searches supporting the decision making process. After finishing the DW, a Linked Open Data (LOD) will be developed. The purpose of such technology is to help searching information in an automated way and also to promote the discovery of new connections among data. Furthermore, three processes were chosen to be analyzed, each one with its separate set of open source tools: the first uses Stdtrip, Triplify and Virtuoso; the second uses Ontowiki, Olap2Data Cube and Virtuoso, and the third process uses Babel and Virtuoso. Due to a better performance in the triplifying process, the tool Olap2Data Cube was chosen as the best tool for this project. The final result is the triplification of all the data on the DW and its resulting availability for searches on Virtuoso's database via SPARQL.

Key words: Educational tests, Open source, Data Warehouse, Linked Open Data, Triplification

Sumário

Lista de Tabelas

Lista de Figuras

Capítulo 1

Introdução

1.1 Contextualização

O Ministério da Educação (MEC) disponibiliza acesso a vários bancos de dados e informações referentes às avaliações educacionais por meio do Instituto de Estudos e Pesquisas em Educação Anísio Teixeira (INEP) com o intuito de melhorar a qualidade das políticas educacionais no Brasil. Após a transformação efetivada pela Lei nº 9.448, de 14 de março de 1997, o INEP teve papel estratégico para o fortalecimento da gestão das políticas educacionais e para o desenvolvimento da educação brasileira.

A produção de dados e informações estatístico-educacionais de forma ágil e de qualidade, que retrate a realidade do setor educacional, é o instrumento básico de avaliação, planejamento e auxílio ao processo decisório para o estabelecimento de políticas de melhoria da educação brasileira. É por meio dos censos educacionais que se busca garantir a utilização da informação estatística nesse processo, gerando os indicadores necessários ao acompanhamento do setor educacional [JANNUZZI, 2001].

O INEP é responsável pelo levantamento e produção das estatísticas básicas da educação nacional, por meio da realização de levantamentos periódicos que abrangem os diferentes níveis e modalidades de ensino. A fim de ampliar o conhecimento sobre a realidade do sistema educacional brasileiro, o INEP desenvolve vários estudos sobre as avaliações realizadas para servir de base na elaboração de políticas públicas educacionais.

Neste contexto, a Fundação Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) e o INEP iniciaram em 2007 o Programa Observatório da Educação (http://observatorio.inep.gov.br), que é uma iniciativa para fomentar o desenvolvimento de estudos e pesquisas em educação, com a finalidade de estimular a produção acadêmica e a formação de recursos em nível de mestrado e doutorado, em áreas voltadas à pesquisa da educação, por meio de financiamento específico, para consolidar e ampliar o pensamento

crítico e estratégico para o desenvolvimento sustentável do País.

Em janeiro de 2007 (Edital nº 001/2006/INEP/CAPES) foi aprovado pelo INEP/-CAPES o projeto de pesquisa da Fundação Universidade Federal de Mato Grosso do Sul (UFMS) em parceria com a Universidade Federal de São Carlos (UFScar) intitulado "Web-PIDE - Uma Plataforma Aberta para Integração e Avaliação de Dados Educacionais na Web" (http://webpide.ledes.net), como parte do Programa Observatório de Educação que objetiva ser um sistema computacional para integrar e disponibilizar os dados educacionais do INEP por meio de uma linguagem comum e padronizada. O projeto visa retratar a realidade da situação do setor educacional por meio da produção de informações estatístico-educacionais. Assim, a realidade educacional pode ser utilizada como base nas discussões que levam a mudanças na educação brasileira. [TURINE et al., 2006]

O desenvolvimento do projeto Web-PIDE tem como sua principal motivação o fato de que o INEP possui poucas ferramentas de fácil utilização para exibir e visualizar as estatísticas educacionais provenientes do grande volume de dados disponíveis nas avaliações educacionais.

As avaliações aplicadas pelo INEP podem ser mapeadas quanto ao domínio (Educação Básica ou Superior) ou sistema de avaliação. Deste modo, pode-se observar na Figura 1.2 uma estrutura das avaliações mapeadas juntamente com outros dados importantes: nome de divulgação, ano de aplicação e situação.

Atualmente, é por meio dos Sistemas Nacionais de Avaliação e de Informação (SAEB, ENEM, ENC, ENCCEJA, SINAES, ENADE etc.), todos instituídos pelo INEP, que tal realidade é estabelecida. De fato, o INEP vem coletando dados e realizando censos e avaliações em instituições de ensino há 18 anos, e o enorme volume de dados coletados não está padronizado, mas sim distribuído em bases de dados distintas. Os sistemas de avaliação são constituídos por bases de dados distintas coletadas por meio de censos e pesquisas. Essas bases não estão uniformizadas e são difíceis de serem integradas aos sistemas existentes.

Dessa forma, torna-se complexa a reutilização das informações contidas nos vários bancos de dados e, consequentemente, dificulta-se a tomada de decisões. Com a intenção de minimizar esse problema, o INEP e a Capes vêm apoiando vários projetos, entre eles a plataforma Web-PIDE.

Foi com o intuito de modificar o quadro atual da educação brasileira que se propôs a implementação de uma plataforma aberta, a Web-PIDE, que objetiva integrar os dados educacionais e utilizar uma linguagem comum e padronizada, a qual foi dado o nome "Linguagem de Marcação de Dados Educacionais" (LIDE). A partir da construção do DW em cada sistema de avaliação, será possível extrair dados e fazer inferências utilizando

softwares que auxiliarão na visualização dessas realidades. O desenho geral da arquitetura inicial do projeto está evidenciado na Figura 1.1.

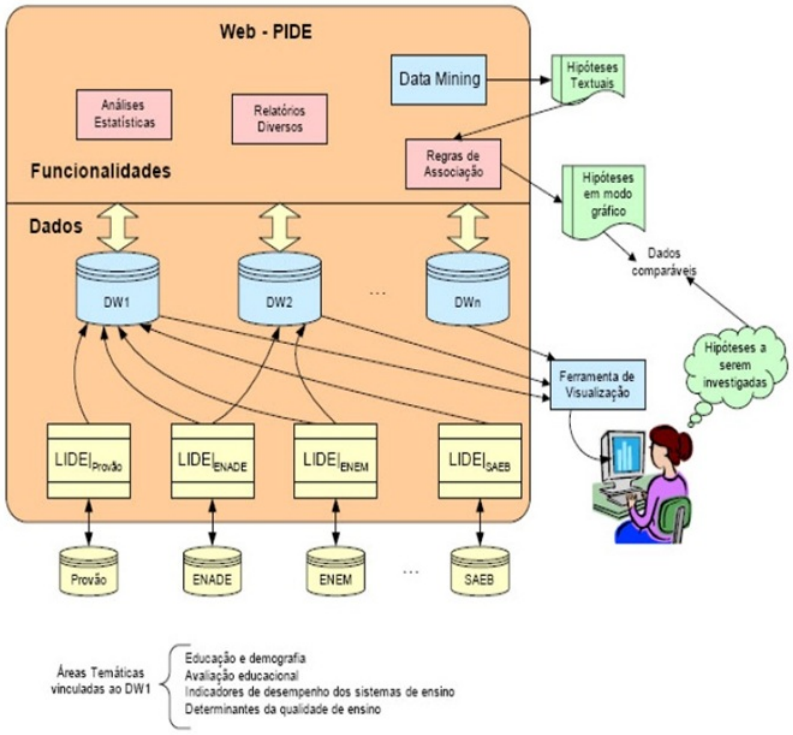


Figura 1.1: Arquitetura da plataforma Web-PIDE.[TURINE et al., 2006]

Na Figura 1.2 é apresentada a Estrutura de Avaliação do INEP entre os anos de 1995 e 2011 e engloba o Ensino Regular, Médio e Superior. Em seguida, tem-se uma breve descrição das avaliações aplicadas pelo INEP entre 1995 e 2011:

1. **SINAES** (Sistema Nacional de Avaliação da Educação Superior (SINAES): avaliar o Ensino Superior no Brasil traçando seu perfil, e com isso promover melhoria na sua qualidade.

2. **Provão ou ENC** (Exame Nacional de Cursos): analisar os cursos de graduação do Ensino Superior no Brasil por meio do desempenho do aluno.

3. **ENCCEJA** (Exame Nacional para Certificação de Competências de Jovens e Adultos): avaliar as pessoas que não tiveram a oportunidade de concluir os estudos na idade adequada.

4. **PNERA** (Pesquisa Nacional de Educação na Reforma Agrária): conhecer e identificar os problemas da educação nos assentamentos da reforma agrária.

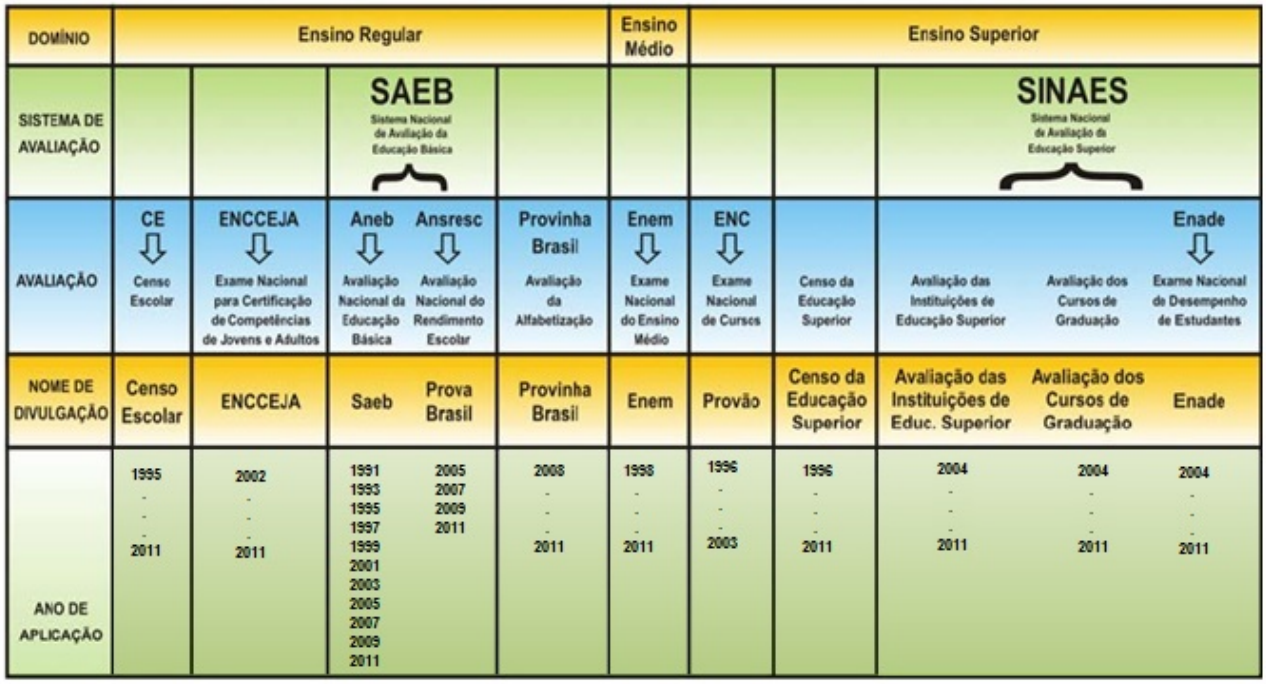


Figura 1.2: Avaliações diponiveis no portal do INEP (03/2012). [Hernandes, 2010]

5. **Provinha Brasil**: avaliar o rendimento escolar aos alunos do ensino regular, proporcionando um instrumento de diagnóstico dos níveis de alfabetização de crianças entre seis e oito anos de idade.

6. **Censo Escolar**: pesquisa realizada anualmente com a finalidade de fazer um levantamento sobre as escolas de educação básica no país. Servir de base para a elaboração de análises, diagnósticos e para o monitoramento das políticas públicas educacionais.

7. **SAEB** (Sistema de Avaliação da Educação Básica): avaliar a qualidade do ensino básico oferecido no Brasil.

8. **ENEM** (Exame Nacional do Ensino Médio): exame lançado em 1998, de caráter individual e voluntário, disponibilizado anualmente, ao fim do ciclo básico e visa avaliar habilidades gerais dos estudantes e proporcionar uma avaliação padronizada dos mesmos.

1.2 Motivações

Como tornar os dados dos Sistemas de Avaliação do Brasil (INEP/MEC) mais acessíveis, com maior semântica e abertos aos gestores de educação, bem como aos educa-

dores e pesquisadores, a fim de melhorar a Educação Brasileira, facilitando a tomada de decisão na gestão pública a partir da recuperação e reutilização dos dados das avaliações educacionais, é a grande questão tecnológica atual.

Tal desafio é atual, observador pela publicação em 18 de novembro de 2011 da Lei de Acesso a Informações Públicas (Lei nº 12.527/2011). A busca de ferramentas e soluções está sendo pesquisada por muitos órgãos da União, Estados, Distrito Federal e Municípios e na comunidade científica.

Com a produção de uma enorme quantidade de dados a partir dos Sistemas de Avaliação da Educação no Brasil e com a necessidade de tornar tais dados acessíveis, deve-se melhorar o processo de extração de informações das bases de dados a fim de gerar dados para guiar o futuro da educação brasileira. O fato de as bases serem demasiadamente grandes e não terem um padrão pré-definido faz com que qualquer forma de extração e manutenção dos dados seja custosa, dificultando o processo de tomada de decisão pelo gestor público.

Nesse contexto, tais problemas motivam a pesquisa de uso da tecnologia Web Semântica, a qual fornece um ambiente em que a aplicação pode consultar os dados e fazer inferências por meio de vocabulários. "Para tornar a Web Semântica uma realidade, é importante ter uma enorme quantidade de dados disponível na Web em um formato padrão, alcançável e gerenciável por ferramentas de Web Semântica."[W3C, 2012]

Segundo Baldus [Baldus, 2011], "Com o uso de Web Semântica é possível criar novas informações por meio das interligações de dados governamentais abertos com outras fontes de dados, o que possibilita para qualquer interessado o desenvolvimento de programas e aplicações de interesse público ou privado. Assim, o interessado utiliza os dados da forma que desejar, somente respeitando o formato previsto pelo fornecedor dos dados. Portanto, neste formato a utilidade desses dados é incalculável."

Neste contexto, o *Linked Data* é um formato de Web Semântica baseado em tecnologias padrões, tais como HTTP, URIs (*Uniform Resource Identifier*) e RDF (*Resource Description Framework*), que descrevem um método de publicação de dados estruturados de forma que possam ser interligados. A sua utilidade se estende à partilha de informações de maneira que estas possam ser lidas automaticamente pelos computadores. Dessa forma, dados de fontes diferentes podem ser conectados e *queries* podem ser aplicadas. O objetivo do *Linked Data* é permitir que os usuários possam compartilhar dados estruturados na Web tão facilmente quanto podem compartilhar documentos.

As primeiras iniciativas de dados abertos governamentais conhecidas foram financiadas pelos governos da Inglaterra e dos Estados Unidos, que criaram os portais de dados abertos "data.gov" e "data.gov.uk". Outras iniciativas incluem o projeto *Linked Open*

Data (LOD) do W3C e o projeto europeu LATC(*Linked Open Data Around The Clock*).

Devido à grande quantidade de bases dos Sistemas de Avaliação, há muito esforço para criar um único LOD com todas as bases juntas. Também seria necessário muito esforço para criar um LOD para cada ano de uma determinada base. Logo, objetiva-se neste projeto utilizar os modelos de DW como entrada para criar LODs específicos. [Siqueira et al., 2008] e [Leite and Rossi, 2010] O Sistema de Avaliação do ENEM foi escolhido como referência neste trabalho, propondo a implementação do DW desde o começo. Isso resulta na obtenção de um novo DW para o projeto Web-PIDE, sendo que já foram criados os DWs do SAEB e do CEB, e também na validação das formas de geração dos modelos de DW propostos em projetos anteriores.

Assim, os conceitos de *Linked Data* e *Open Government Data* são pesquisados neste trabalho e servirão como base estrutural para o desenvolvimento de um modelo de processo de publicação de *Linked Data* que pode ser reutilizado tanto nos DWs existentes quanto nos que poderão vir a serem criados, possibilitando buscas automatizadas, criação de softwares para visualização e interação, entre outros recursos disponíveis.

1.3 Objetivos

O presente projeto tem como objetivo criar um modelo de processo ou estratégia para publicação de dados abertos no padrão LOD, iniciando com a criação de um DW utilizando a abordagem estrela e os dados do ENEM dos anos de 1998 até 2008. A partir do DW triplifica-se os dados do ENEM criando o LOD-ENEM, para tal foi definido um processo que será utilizado como padrão para o projeto Web-PIDE.

Por conseguinte, para alcançar o objetivo proposto, as seguintes atividades devem ser executadas:

1. Importação das bases de dados do ENEM entre os anos 1998-2008 para um banco de dados Postgres;
2. Estudo dos conceitos de DW, principalmente, do modelo estrela;
3. Estudo das bases de DW já existentes no projeto Web-PIDE a fim de padronizar o novo DW;
4. Estudo de todos os conceitos envolvidos para implementação de um LOD, os quais são: Web Semântica, Ontologias, *Ontology Web Language* OWL, Dados abertos governamentais, SPARQL e RDF;
5. Pesquisa e avaliação de processos de triplificação LOD;

6. Implementação, avaliação e validação do DW-ENEM; e
7. Implementação, avaliação e validação do LOD-ENEM.

1.4 Organização do Texto

O presente texto está organizado em seis capítulos descritos a seguir:

1. **Capítulo 1 - Introdução**: contextualizar e caracterizar o problema a ser investigado, as motivações e os objetivos principais do trabalho.
2. **Capítulo 2 - Projeto Web-PIDE e o Sistema de Avaliação ENEM**: descrever a plataforma Web-PIDE, as pesquisas realizadas na plataforma, e o processo de avaliação do Sistema ENEM, destacando as estruturas e os dados da base definidos pelo INEP.
3. **Capítulo 3 - Web Semântica e *Open Linked Data***: apresentar os principais conceitos sobre *Web Semântica*, *Linked Data*, RDF, LOD, OWL, Dados abertos governamentais e *Data Warehouse*.
4. **Capítulo 4 - Processos de geração de LOD**: apresentar os principais processos de triplificação LOD existentes na literatura e que foram pesquisados, demonstrando os passos de cada processo para ser possível uma análise posterior.
5. **Capítulo 5 - Implementação do DW-ENEM e LOD-ENEM**: descrever a proposta no contexto do processo de definição e implementação do DW-ENEM e do LOD-ENEM, destacando suas vantagens e comparações entre as soluções utilizadas.
6. **Capítulo 6 - Conclusão e Trabalhos Futuros**: destacar as principais contribuições da pesquisa, dificuldades e trabalhos futuros.

Capítulo 2

Projeto Web-PIDE e o Sistema de Avaliação ENEM

Neste capítulo é apresentada a situação atual da plataforma Web-PIDE (http://webpide.ledes.net), destacando os trabalhos realizados e as ferramentas geradas. Atenção especial é dada ao Exame Nacional do Ensino Médio (ENEM), pois é o sistema de avaliação utilizado nesta pesquisa. Uma visão geral da estrutura da base de dados do ENEM e de seus microdados é apresentada neste capítulo.

2.1 Projeto Web-PIDE

Como tornar os dados dos Sistemas de Avaliação do Brasil (INEP/MEC) mais acessíveis aos gestores de educação, bem como aos educadores e pesquisadores, a fim de melhorar a Educação Brasileira, facilitando a tomada de decisão na gestão pública a partir da recuperação e reutilização dos dados das avaliações educacionais, é a grande questão tecnológica atual.

O fato de as bases de dados mantidas pelo INEP armazenarem um grande volume de informações e não terem uma estrutura padrão faz com que as consultas computacionais tornem-se custosas e dificulta sua análise, impossibilitando uma tomada de decisão eficiente baseada nos dados das avaliações educacionais.

Nesse contexto, o Programa Observatório da Educação (OBEDUC - http://www.capes.gov.br/educacao-basica/observatorio-da-educacao) é relevante para fomentar o desenvolvimento de estudos e pesquisas em educação com a finalidade de estimular a produção acadêmica e a formação de recursos em nível de mestrado e doutorado, em áreas voltadas à pesquisa da educação.

Em janeiro de 2007, foi aprovado no contexto do Observatório da Educação o projeto de pesquisa da Fundação Universidade Federal de Mato Grosso do Sul (UFMS) em parceria com a Universidade Federal de São Carlos (UFScar) intitulado "Web-PIDE - Uma Plataforma Aberta para Integração e Avaliação de Dados Educacionais na Web" (http://webpide.ledes.net) [TURINE et al., 2006], como parte do Programa Observatório de Educação que objetiva ser um sistema computacional para integrar e disponibilizar os dados educacionais do INEP por meio de uma linguagem comum e padronizada de marcação.

Segundo Turine [TURINE et al., 2004], as transformações que vêm ocorrendo no mundo inteiro e que, mesmo tardiamente, chegam à administração pública exigem das organizações maior agilidade e flexibilidade no atendimento das novas e crescentes demandas da sociedade. Esse quadro, associado ao elevado grau de competição e um crescente desenvolvimento tecnológico, torna possível e necessária à utilização efetiva das tecnologias de informação e comunicação como ferramentas de produtividade do serviço público, de qualidade dos serviços prestados ao cidadão e de democratização do acesso às informações.

Vários trabalhos foram pesquisados no contexto da plataforma Web-PIDE e são descritos resumidamente neste capítulo a fim de fundamentar a importância da presente proposta de pesquisa.

2.2 Pesquisas Realizadas

2.2.1 Um Processo Automatizado para Tratamento de Dados e Conceitualização de Ontologias com Apoio de Visualização

Autor(a): Elis Cristina Montoro Hernandes

Nível: Mestrado/UFSCAR

No projeto de Hernandes [Hernandes, 2010], foram desenvolvidas ferramentas de suporte para apoio da criação da Linguagem para Integração de Dados Educacionais (LIDE). A criação de tal ferramenta envolveu a linguagem de marcação *Extensible Markup Language* (XML). Além da acessibilidade das informações, a LIDE deve ser caracterizada por um formato padrão, pois é imprescindível que os dados educacionais analisados no projeto Web-PIDE possam ser interpretados e armazenados em uma base de dados de forma homogênea, dando suporte a um ambiente em que serão implementadas inúmeras funcionalidades. A pesquisa de Hernandes objetivou a criação de mecanismos de análise de dados relacionados às avaliações do INEP e de identificação das questões usadas em cada ano e em cada questionário. Segundo [Hernandes, 2010], " esse desafio salientou o problema da falta de padronização dos termos usados para compor as questões, indicando

a necessidade de uma ontologia para o domínio, provendo a padronização de termos e também a representação formal do domínio de avaliações educacionais do INEP."

Dessa forma, Hernandes apresentou um Processo para definição da Linguagem para Integração de Dados Educacionais (P-LIDE), utilizando como apoio a *Search and Edition in Visualization Tool* (SEV-Tool), que permitiu que os dados das avaliações do INEP fossem padronizados sintaticamente. Foi por meio do P-LIDE que se solucionou um dos problemas das avaliações do INEP: a identificação do ano em que cada questão contida na avaliação foi utilizada, tarefa que seria praticamente impossível de ser executada manualmente. Identificado o problema, a autora pôs-se à procura de uma ferramenta livre que pudesse solucionar o impasse. Primeiramente, foi considerada a TreeMap [Johnson and Shneiderman, 1991], mas essa não se adequava inteiramente, pois não abordava problemas de busca, edição, visualização e representação de dados em uma região retangular. Conforme a posição hierárquica, os níveis eram desenhados como retângulos dentro da região maior. Depois de descartar o uso da técnica TreeMap, Hernandes decidiu implementar a ferramenta SEV-Tool cujas funcionalidades foram desenvolvidas para a execução do P-LIDE.

Além disso, houve a necessidade do uso de ontologia para promover a padronização dos termos e da representação formal. Assim, foi implementado *Ontology Process* (ONTOP) como forma de suprir a necessidade de padronização de dados educacionais.

Para dar apoio à execução do processo, foi utilizada por Hernandes a ferramenta ONTOP-Tool, que assim como a SEV-Tool, utiliza-se de recursos de visualização de informação. Como descrito pela autora, " (...) o ONTOP dá suporte à fase de conceituação de ontologias, partindo de um glossário até chegar à geração de um arquivo do tipo OWL."[Hernandes, 2010]

Várias ferramentas foram utilizadas no trabalho de Hernandes, tais como, ONTOP-Tool, SEV-Tool, Protegé-2000 e o ambiente Web Moodle, de onde foi extraído o recurso de compartilhamento e controle de glossários.

2.2.2 Uma Abordagem de Integração e Exploração Visual de Dados Educacionais na Plataforma Web-PIDE

Autor(a): Jackson Dias Savitraz

Nível: Mestrado/UFMS

O projeto de Savitraz [Savitraz, 2010] teve como objetivo a integração de diferentes WebApps utilizadas na plataforma Web-PIDE e a visualização de indicadores e de dados estatísticos gerados pelas avaliações do INEP em um único banco de dados.

O INEP disponibiliza diferentes sistemas de consulta em seu portal (http://portal.inep.gov.br/) utilizando técnicas de representação tabular e gráfica dos indicadores educacionais, tais como, Consulta IDEB, EDUDATABRASIL, Data-EscolaBrasil. Além disso, no portal também estão disponibilizados os resultados das avaliações em formato de planilhas para download ou em sinopses em formato PDF. Porém, como as bases de dados do INEP não estão integradas, esses sistemas de consulta não permitem visualizações comparativas.

Para melhor atender as necessidades de visualização dos dados contidos na plataforma Web-PIDE, Savitraz propôs uma abordagem de integração das diversas WebApps. Para gerir a plataforma, criou-se o Portal Web-PIDE (http://webpide.ledes.net/) por meio da utilização do Titan [Carromeu and Turine, 2005] como framework para os Sistemas de Gerenciamento de Conteúdo (CMS). Porém, em primeiro lugar, o projeto criou um mecanismo que permitisse que os dados das avaliações e os seus metadados fossem armazenados de forma consistente e que permitissem reuso, o que não acontece no formato textual ASCII, utilizado atualmente pelo INEP para disponibilizar tais dados. Assim, foi utilizada a ferramenta *Data Extractor ASCII to Relational* (DEAR [Siqueira, 2009]), que carregou os dados em formato ASCII para o banco de dados relacional. A extração de dados foi executada e implementou um DW a partir da base gerada.

Uma vez que os dados foram carregados para o DW, foi necessário o uso de um componente para especificação de consultas e uma interface com o usuário em que este pode acessar os elementos das estruturas dos dados armazenados. Para isso, o autor desenvolveu uma camada de integração entre as WebApps utilizando o conceito *Web Services*.

A proposta de abordagem para integração das WebApps na plataforma Web-PIDE e as técnicas de visualização dos dados desenvolvidas no projeto de Savitraz são inovadoras e trata-se de alternativas significativa. Dessa forma, o ambiente projetado tem uma infraestrutura visual para acesso de dados que forneçam novos conhecimentos sobre os indicadores dos dados das avaliações educacionais do INEP.

Além da proposta de integração e das técnicas de visualização, Savitraz implementou o "Ambiente de Integração e Visualização de Dados" (Amb-PIDE), utilizado para especificação de consultas e também para a visualização de dados provenientes do *Web Services*, conforme arquitetura ilustrada na 2.1.

A interface do Amb-PIDE pode ser acessada a partir de qualquer navegador e dispõe de recursos para especificar novas consultas à base de dados e visualizar de forma gráfica e/ou tabular. Para utilizar a interface, o usuário seleciona uma das bases de dados disponível. Após a seleção, o Amb-PIDE exibirá uma lista de indicadores disponíveis e o usuário poderá consultar os dados desejados conforme especificou em uma estrutura de tabela,

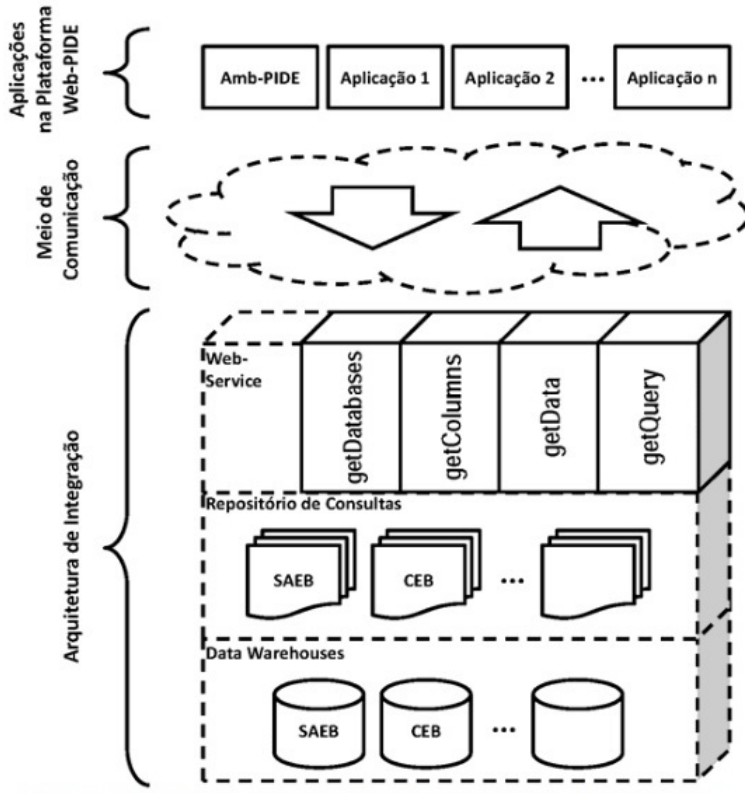


Figura 2.1: Modelo em camadas da arquitetura de integração [Savitraz, 2010]

formada em duas partes: indicador e ano de aplicação.

"A abordagem proposta foi validada com a especificação do *Web Service* de integração, que disponibiliza recursos sobre os indicadores das bases de dados educacionais. Foi também testada e validada a partir da ferramenta intitulada Amb-PIDE, que utiliza as informações obtidas no *Web Service* em diversos contextos: visualização de dados, *data warehouse* e consultas avançadas."[Savitraz, 2010]

2.2.3 SB-INDEX: Um Índice Espacial Baseado em Bitmap para *Data Warehouse* Geográfico

Autor(a): Thiago Luís Lopes Siqueira

Nível: Mestrado/UFSCAR

Siqueira [Siqueira, 2009] propôs o *Spatial Bitmap Index* (SB-index), uma nova estrutura de indexação de DW geográfico em que se pretende diminuir o tempo de resposta de consultas analíticas multidimensionais. Os diferenciais do SB-index consistem da introdução do Índice Bitmap no contexto de DW geográfico, do tratamento de hierarquias de

atributos espaciais predefinidas e da realização de consultas analíticas multidimensionais.

O autor também definiu estruturas de indexação a fim de aumentar o desempenho da recuperação de registros. Tal estrutura de indexação proporciona formas alternativas em que se podem acessar registros sem que haja alteração da organização física dos dados, além de permitir a seleção e a recuperação dos registros que satisfazem às condições de consulta conforme a chave de busca pré-definida (um conjunto de atributos arbitrariamente escolhidos).

A integração de SIG e seus dados espaciais sob o contexto multidimensional e relacional tem sido motivação para vários autores desenvolverem diferentes abordagens e trabalhos, pois não existe um consenso sobre o que é uma medida espacial. Assim, o autor direcionou o foco da pesquisa à indexação e aos índices *B-tree* e *hashing*, que são comprovadamente eficientes quando empregados pelos Sistemas Gerenciadores de Banco de Dados (SGDB).

Depois da explanação sobre DW geográfico, estruturas de indexação e índices Bitmap, o autor parte para a apresentação do SB-Index, em que define a proposta dessa estrutura de indexação desenvolvida para DW geográfico. Assim, o SB-Index introduz o Índice Bitmap em DW geográfico, reutiliza técnicas como o *binning*, a compressão e a codificação (inicialmente desenvolvidas para o Índice Bitmap), estende as vantagens tradicionais proporcionadas pelo Índice Bitmap ao DW para o DW geográfico.

Conforme afirma o autor, "O SB-Index consiste em uma adaptação do Índice de Projeção, incidindo sobre o atributo que compõe a chave primária da tabela de dimensão espacial." [Siqueira, 2009] Siqueira também implementou em seu trabalho a ferramenta *Data Extractor ASCII to Relational* (DEAR), que visa extrair os dados em formato ASCII e armazenar em um banco de dados relacional.

Além da implementação da ferramenta DEAR, foi criado o DW da base de dados do SAEB. Para a construção do DW do SAEB foram utilizados somente softwares livres ou gratuitos, tais como, PostgreSQL, servidor OLAP Mondrian e a ferramenta DEAR. Na Figura 2.2 é ilustrado o DER do DW.

Como conclusão do processo de criação do DW, [Siqueira et al., 2008] afirmou que "A análise dos dados demonstrou que o formato de armazenamento dos dados adotados pelo INEP possui sérias inconsistências, prejudicando a extração de informações estratégicas. Ainda, as inconsistências encontradas nos metadados (dicionários de variáveis armazenados nos formatos SAS e PDF) determinam a necessidade de mecanismos computacionais que auxiliem corrigi-los."

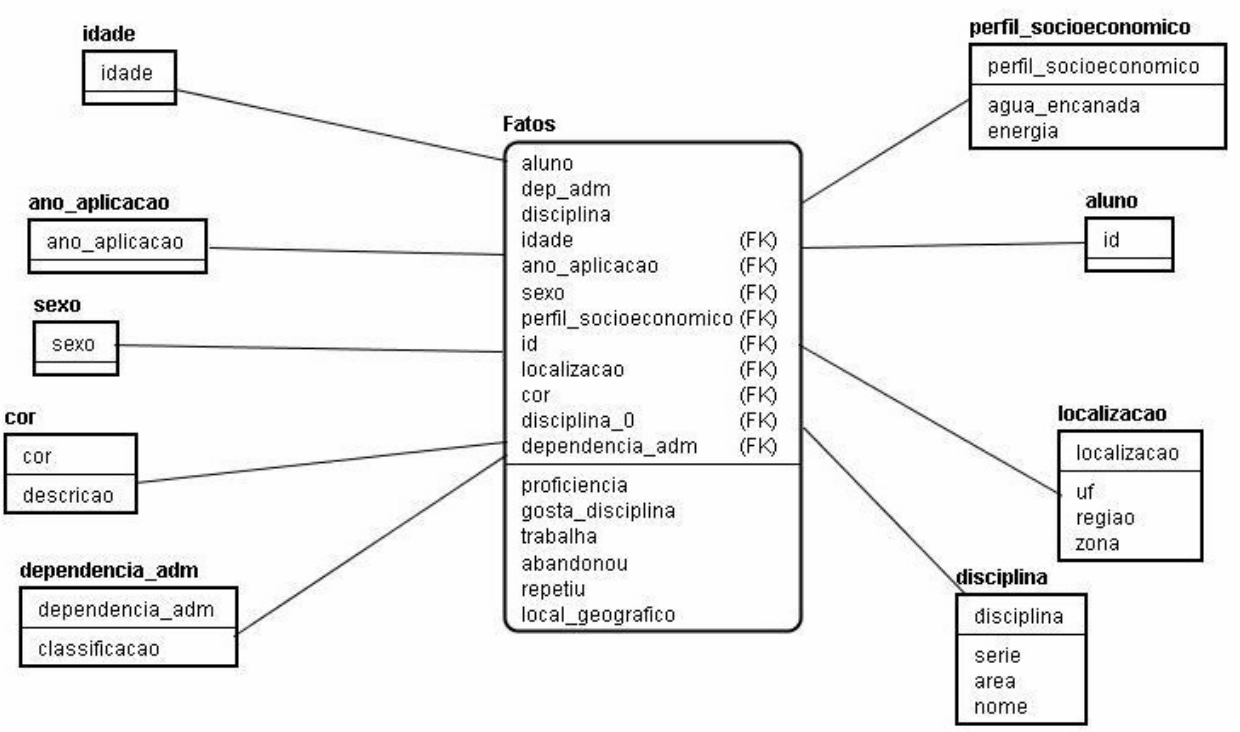


Figura 2.2: Modelo simplificado do DER do SAEB [Siqueira et al., 2008]

2.2.4 Modelo de Processo para Integração de Serviços de Avaliação Educacional na Plataforma Web-PIDE

Autor(a): Maxwell Sampaio dos Santos

Nível: Mestrado/UFMS

Em seu projeto, Santos [Santos, 2011] evidenciou que as ferramentas desenvolvidas ou em desenvolvimento para a plataforma Web-PIDE não estão inteiramente integradas na plataforma e foram frequentemente desenvolvidas sem necessariamente estarem orientadas a serviço.

Santos propôs a construção de um ambiente em que se possam integrar tais ferramentas na plataforma Web-PIDE, permitindo que os usuários do portal possam integrar de forma clara tais ferramentas. Para isso, utilizou o Titan Framework como gerenciador de conteúdo (CMSs) para sistemas Web. Segundo o autor, o Titan Framework proporciona uma solução simples, rápida e completa para a instanciação de gerenciadores de conteúdo.

Na Figura 2.3 é apresentado o modelo da arquitetura do Ambiente Integrador de Serviços na plataforma Web-PIDE.

O autor também explica que "O processo de construção e integração do ambiente proposto foi guiado pelo processo de desenvolvimento de *serviços meet-in-the-middle* que

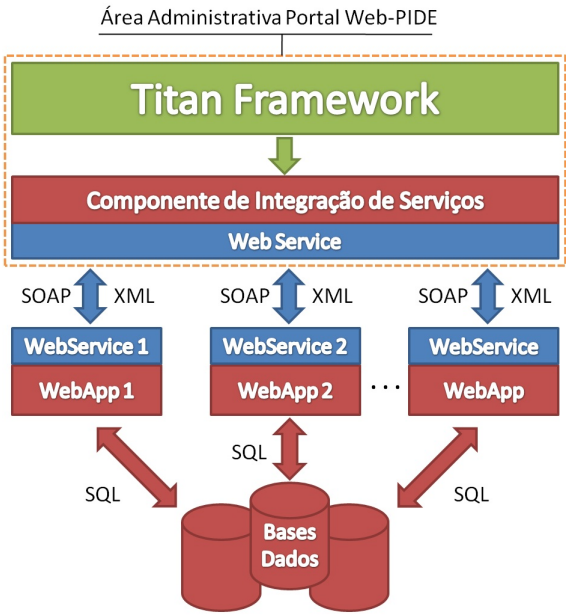


Figura 2.3: Arquitetura do Ambiente Integrador de Serviços na plataforma Web-PIDE [Santos, 2011]

define uma série de procedimentos para a construção eficiente de aplicações orientadas a serviço." [Santos, 2011]

2.2.5 Identificação e Determinação de Serviços para Compor a Plataforma Web-PIDE

Autor(a): Fernanda Aparecida Rocha da Silva

Nível: Mestrado/UFSCAR

No projeto desenvolvido por Silva [Silva, 2011] é abordada uma estratégia de identificação de *Web Services*, composta da fase de engenharia de domínio e da fase de engenharia de aplicação. Essas duas fases se completam para, a partir dos objetivos e dos processos de negócio, estabelecer um processo para que se identifiquem os *Web Services*. Para atingir seu objetivo, a autora utilizou o modelo *AWARE* na fase inicial da estratégia de identificação de *Web Services*.

Conforme [Silva, 2011], "A finalidade da fase de engenharia de domínio é entender o processo de negócio e identificar serviços". Já a engenharia de aplicação "(...) está relacionada ao uso dos serviços identificados e às customizações necessárias para o desenvolvimento de uma aplicação específica".

No caso da engenharia de domínio, o responsável é o engenheiro de domínio que executa etapas, tais como, recuperação de informações sobre o processo de negócio, ela-

boração de GTR (*Goal, Task and Requirements*), elaboração de BPM (*Business Process Management*) e projeto de serviços, conforme ilustradas na Figura 2.4.

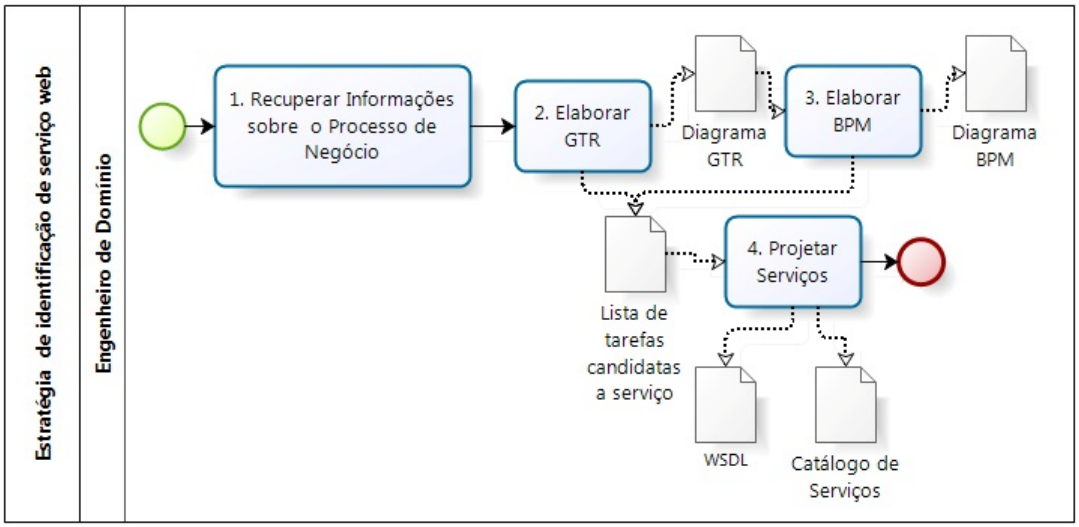


Figura 2.4: Estratégia de Identificação de Serviços - Engenharia de Domínio [Silva, 2011]

Já na engenharia de aplicação, as etapas envolvidas são: elaborar GTR, elaborar BPM, refinar GTR e BPM e selecionar serviços, etapas as quais podem vir a incluir tanto o *stakeholder* quanto o analista ou ambos, conforme pode ser visto na Figura 2.5.

Uma vez aplicada a estratégia de identificação de *Web Services* à plataforma Web-PIDE, ficou definido pela autora que a fase de engenharia de domínio seria de responsabilidade do técnico do INEP, que deverá produzir um catálogo com a descrição dos serviços e informações necessários. A engenharia de aplicação, no entanto, é de responsabilidade dos desenvolvedores da plataforma Web-PIDE.

2.2.6 Simbolização de Mapas Temáticos Utilizando Uma Ontologia Cartográfica

Autor(a): Vinícus Ramos Toledo Ferraz

Nível: Mestrado/UFSCAR

O autor propõe em seu trabalho a criação de Mapas Temáticos (ou Mapas Estatísticos), dentro de um Projeto Cartográfico Temático (PCT), por meio do desenvolvimento de uma Ontologia Cartográfica apropriada que eleve o grau de automatização do PCT. Dessa forma, até mesmo pessoas leigas poderão interpretar um PCT de forma adequada, evitando-se erros no processo de tomada de decisão.

Em princípio, para que seja possível desenvolver os Mapas Temáticos, é preciso abordar os sistemas computacionais destinados ao geoprocessamento, os Sistemas de Informação

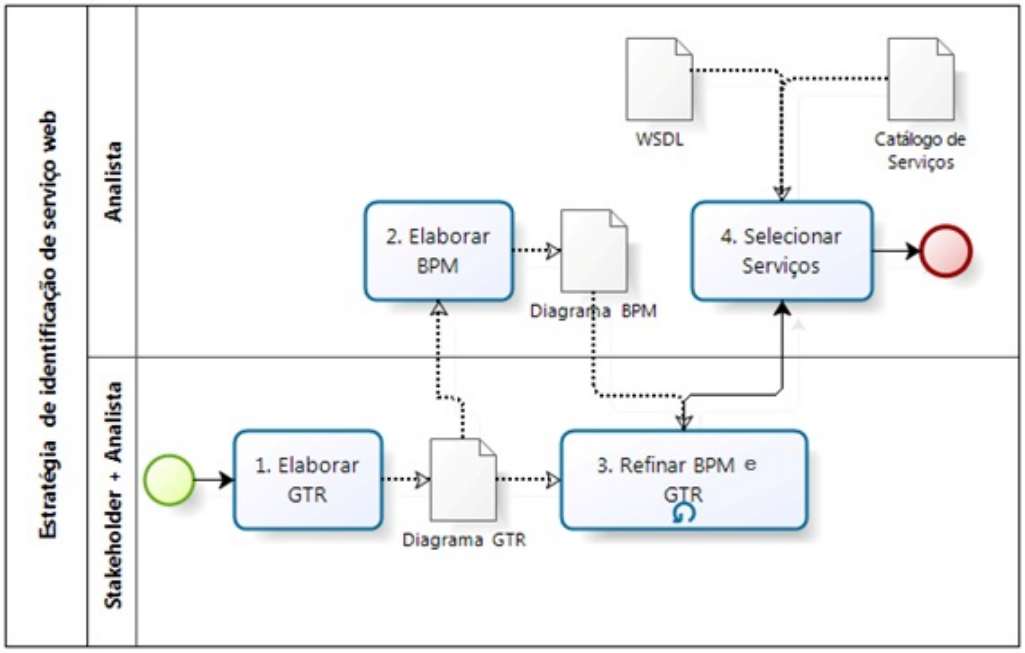


Figura 2.5: Processo de Engenharia de Aplicação [Silva, 2011]

Geográfica (SIG). Estes possuem ferramentas que permitem a produção, o armazenamento, a recuperação, a exportação, a visualização e a análise dos dados geoespaciais coletados.

Em seguida, o autor propõe a criação da metodologia *Cartographer*, cujos principais componentes são uma Ontologia Cartográfica (que codifica o conhecimento cartográfico), uma heurística (cujo objetivo é auxiliar na realização de inferências cartográficas) e um serviço Web geográfico (para que seja possível o reúso online de ambas a heurística e a Ontologia Cartográfica).

Para Ferraz, uma Ontologia Cartográfica apropriada é de suma importância para o desenvolvimento da Metodologia *Cartographer*: "Uma Ontologia Cartográfica descreve formalmente os axiomas que representam o conhecimento estruturado e compartilhado da cartografia. Desses axiomas é possível realizar inferências para certas asserções, ou seja, um motor de inferência utiliza esta representação do conhecimento cartográfico para inferir decisões de PCT automaticamente (por exemplo, definindo uma linguagem cartográfica adequada à natureza do fenômeno mapeado), restringindo possíveis decisões equivocadas."[Ferraz, 2011]

Uma vez estabelecida a metodologia *Cartographer*, Ferraz, então, descreve uma Ontologia Cartográfica denominada *OWL-Cartography*, cujo o objetivo é a representação formal de conhecimento cartográfico suficiente o qual permita, de acordo com informações específicas previamente selecionadas, inferir a simbologia temática adequada. Nas palavras

do autor, a *OWL-Cartography* "(...)foi construída com apoio da metodologia UPON e codificada na linguagem OWL com auxílio da ferramenta Protégé (uma ferramenta gratuita e de código aberto de apoio à construção de ontologias)."[Ferraz, 2011]

Por fim, o autor descreve a criação de um Serviço Web Geográfico que foi denominado *WPS-Cartographer*. O objetivo de tal serviço foi "prover a Metodologia Cartographer com um serviço Web de seleção semiautomática da simbologia cartográfica."[Ferraz, 2011]

Assim, por meio da *OWL-Cartography*, de uma heurística própria e do *WPS-Cartographer*, Ferraz constitui a metodologia *Cartographer* como alternativa que supre as deficiências presentes em trabalhos anteriores e minimizando a possibilidade de equívocos na produção de Mapas Temáticos.

2.2.7 Uma Estratégia para Publicação dos Dados da Base do CEB-INEP/MEC no Padrão*Linked Open Data*

Autor(a): Fernando Maia da Mota

Nível: Graduação/UFMS

Mota [Mota and Rossi, 2010]implementou uma estratégia de publicação das informações contidas na base de dados do CEB-INEP/MEC do ano de 1995 no padrão *Linked Open Data*. Ele propõe a concepção da estratégia com base nas necessidades do INEP em relação aos dados adquiridos ao longo dos anos e que são parte agora do projeto Web-PIDE. O padrão *Linked Open Data* foi escolhido pois apresenta uma possibilidade eficiente para a solução do problema da falta de padronização dos dados acumulados pelo INEP e da consequente dificuldade na reutilização de tais dados.

A Figura 2.6 ilustra o processo executado por Mota, e as etapas são descritas abaixo.

Dessa forma, o primeiro passo no desenvolvimento da estratégia de publicação dos dados foi avaliar as tabelas de bancos de dados do INEP para que fosse feita a criação de um modelo lógico relacional normalizado. O segundo passo consistiu do carregamento dos dados para o SGBD PostgreSQL por meio da utilização da ferramenta DEAR. Na terceira etapa, um modelo de banco de dados normalizado tendo como base a descrição das questões de pesquisa foi criado. Nessa etapa, o modelo foi traduzido para o inglês, pois tanto a normalização quanto a tradução para o Inglês são pré-requisitos para a utilização da ferramenta STDTRIP.

A quarta etapa foi caracterizada pela extração e pelo carregamento dos dados para o modelo normalizado. Para se alcançar esse objetivo, foi criada uma ferramenta para a recuperação dos dados armazenados no banco PostgreSQL e para a inserção dos dados normalizados no SGBD. No quinto passo do processo de publicação, o objetivo é a

criação da ontologia. Nessa etapa, utilizou-se a ferramenta STDTRIP no banco de dados armazenado no MySQL. As saídas desse passo são as ontologias em formato OWL (além do arquivo de configuração da ferramenta TRIPLIFY). A sexta etapa da publicação consistiu na triplificação por meio da ferramenta TRIPLIFY. Por fim, a última etapa foi caracterizada pelo armazenamento das triplas geradas no servidor de triplas OPENLINK VIRTUOSO, o qual também fornece uma interface Web para consultas em SPARQL.

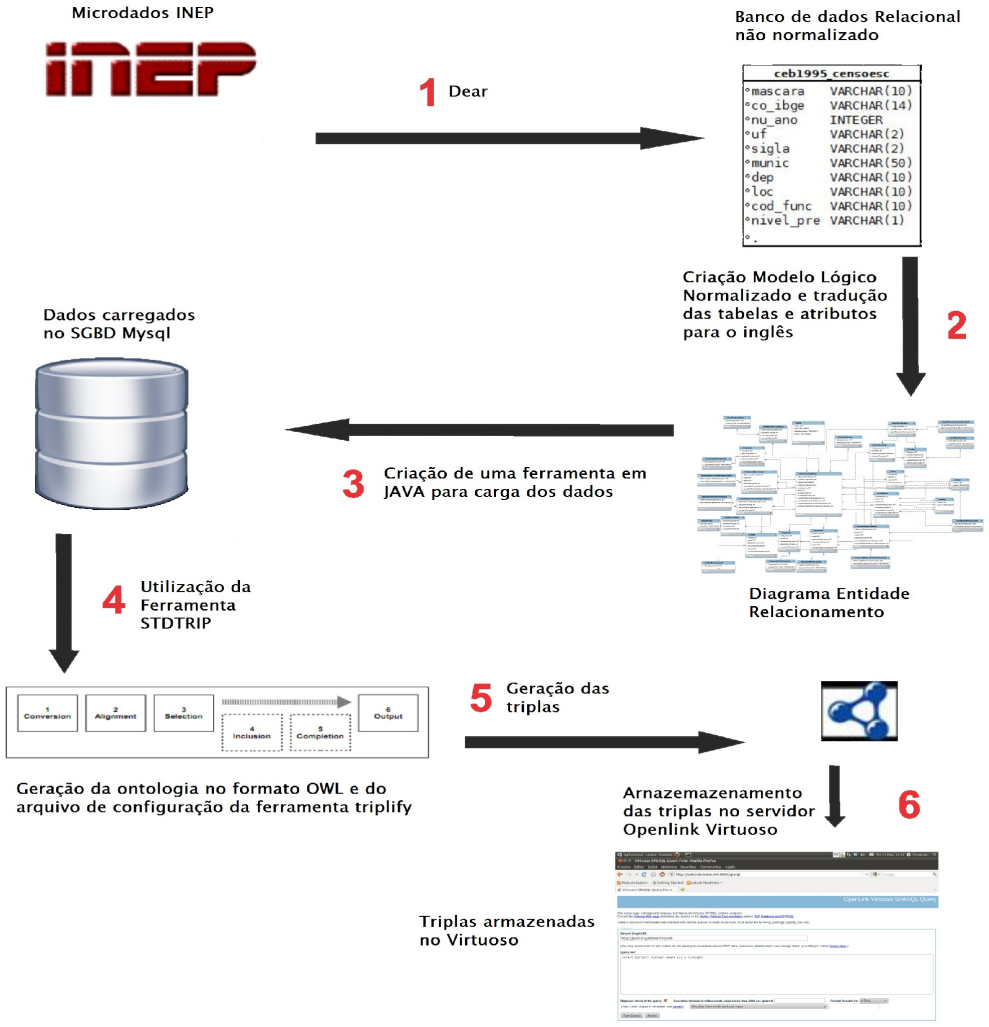


Figura 2.6: Etapas para publicação dos dados no padrão LOD [Mota, 2011]

2.2.8 Estudo e Avaliação das Técnicas e das Ferramentas para Projeto de *Data Warehouse*

Autor(a): Dionísio de Machado Leite, Fernando Maia Mota

Nível: Iniciação Científica/UFMS

Nesse projeto foi implementado o DW do CEB [Leite and Rossi, 2010]. A implementação do projeto deu-se a partir do carregamento dos microdados das bases utilizando a ferramenta DEAR. Foram encontradas várias inconsistências em tais bases, as quais foram sendo sanadas por meio da padronização dos dados.

O DW gerado foi criado utilizando o modelo estrela e a partir deste foi feito o cubo e as consultas *On-Line Analytical Processing* (OLAP). As tecnologias empregadas para a adoção do DW proposto foram: Java, Modrian e JPivot. Na Figura 2.7 é apresentado o modelo simplificado do DER do DW.

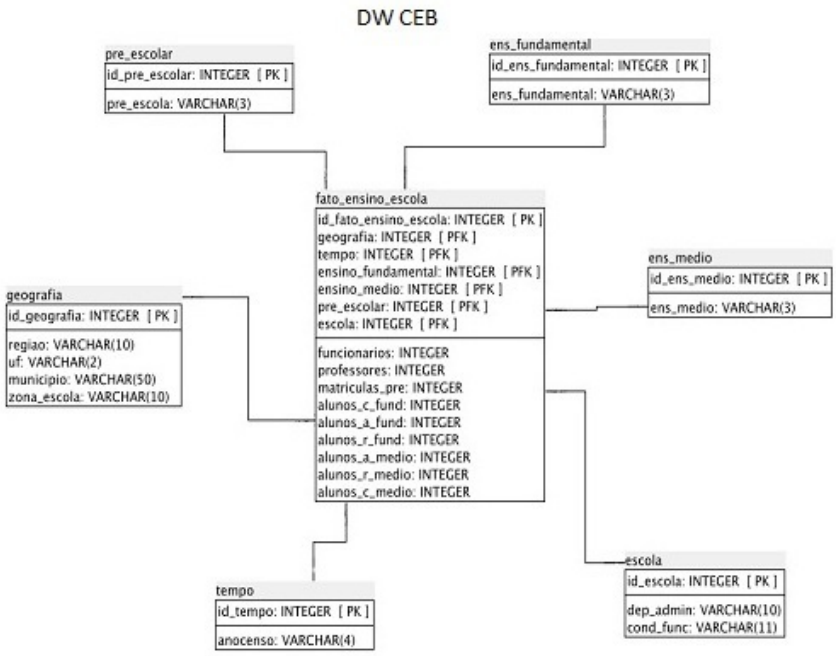


Figura 2.7: Modelo simplificado do DER do CEB [Mota and Rossi, 2010]

O processo de construção dos DWs das bases de dados do SAEB e do CEB foi importante para a construção de um novo DW da base do ENEM, um dos objetivos do projeto em questão. Dentre os facilitadores que ajudaram a construir o DW ENEM, encontram-se as experiências anteriores e o uso das ferramentas: DEAR, Jude, Mondrian e *Pentaho Business Inteligence Server*, além do processo de criação do DW que já estava definido e as experiências com as inconsistências encontradas e relatadas nos relatórios técnicos.

2.3 ENEM e sua Base de Dados

O Exame Nacional do Ensino Médio (ENEM) foi lançado em 1998 e é um exame de caráter individual e voluntário, disponibilizado anualmente, ao fim do ciclo básico. Em

sua primeira edição, em 1998, teve poucos participantes, apenas 115,6 mil. Porém, a partir de 2004, tornou-se o sistema mais popular juntamente com o lançamento do Programa Universidade para Todos (ProUni), que vinculou a concessão de bolsas em Instituições de Ensino Superior ao desempenho do estudante no ENEM.

O objetivo do ENEM é avaliar as habilidades gerais dos estudantes, analisar a capacidade de leitura, interpretação de texto e aplicação de conceitos dos estudantes ao invés de cobrar conteúdos específicos como no tradicional vestibular.

Até o ano de 2008, a prova era composta de um questionário socioeconômico, 63 perguntas interdisciplinares e uma redação. Em 2011, passou a apresentar 180 questões divididas em quatro áreas do conhecimento (linguagem e códigos - português - ciências humanas - geografia e história - ciências da natureza - biologia, física e química - e matemática), além da redação e do questionário socioeconômico.

O exame foi totalmente reformulado em 2009 pelo MEC, pois este pretendia mudar o currículo do Ensino Médio e com isso unificar o processo seletivo das universidades federais. A reformulação continuou com a premissa de avaliar a capacidade de raciocínio do aluno ao invés da capacidade de decorar, mas o faz de uma maneira mais complexa e abrangente. O novo exame foi criado tendo como motivação o SAT (exame da admissão das universidades americanas) e o PISA (teste internacional que afere a qualidade do ensino).

A partir da redefinição do exame, as universidades passaram a ter três opções de utilização do ENEM: podem utilizá-lo como primeira fase da seleção, podem usar a nota do ENEM como percentual na avaliação final, ou podem usá-lo como forma de preencher vagas remanescentes. Nesse novo formato, o ENEM continuou a servir como critério de seleção para o ProUni, substituiu o Encceja e serviu para participação no Sistema de Seleção Unifica (Sisu).

Durante os anos de aplicação do ENEM, foram recolhidos os dados tanto das provas quanto do questionário socioeconômico, e por isso o INEP tem disponível o acervo entre os anos de 1998 e 2010 no formato de microdados. Esses microdados estão publicados no *website* do INEP, em arquivos compactados e categorizados por anos. É também importante saber que esses arquivos estão disponíveis para download para o público em geral.

Cada arquivo contém a seguinte estrutura de pastas: Dados, INPUT_SAS_SPSS e LEIA-ME - conforme pode ser visto na Figura 2.8.

Dentro da pasta Dados há um arquivo de microdados no formato ASCII contendo todos os dados relativos ao ENEM do ano em questão.

Dentro da pasta INPUT_SAS_SPSS, há dois arquivos e ambos servem para auxiliar

Figura 2.8: Estrutura dos arquivos do ENEM

os programas de leitura dos dados (contidos na pasta Dados) de forma programática referente ao ano em questão. A diferença entre eles é que um está em formato SAS e o outro em formato SPSS.

O diretório "Leia-me" possui um arquivo de orientação no formato pdf, ou mais precisamente, um Manual do Usuário dos microdados do ENEM do ano em questão, o qual contém:

- Diretórios (Explicação sobre os diretórios)
- Dicionário das Variáveis
- Input de Leitura do Arquivo - SAS
- Input de Leitura do Arquivo - SPSS
- Questionário Socioeconômico do ENEM
- Prova Objetiva do ENEM
- Gabarito do ENEM

Em anexo, no apêndice A, está exposto parte do dicionário das variáveis do ano de 2007, que exemplifica de forma mais detalhada os dados contidos no arquivo de microdados.

Para dar uma noção melhor da dimensão do ENEM, a Tabela 2.1 contém o ano do exame do ENEM, no período de 1998 a 2008, seguido da quantidade de pessoas que realizaram o ENEM no determinado ano, assim como a quantidade de perguntas do questionário socioeconômico, e por fim a quantidade de dados obtidos no ano em questão.

Tabela 2.1: Esta tabela demonstra a dimensão do exame do ENEM.

Ano	**Inscritos**	**Sócio-econômico**	**Colunas dos Microdados**
1998	156.844	137	161

1999	347.258	129	154
2000	390.089	127	152
2001	1.625.540	75	275
2002	1.836.410	73	251
2003	1.885.230	188	221
2004	1.545.560	205	238
2005	3.020.860	223	256
2006	3.737.870	223	257
2007	3.595.490	223	257
2008	4.029.360	223	261

Observando a tabela, é possível notar que com o passar dos anos o ENEM está se popularizando, pois o número de participantes tem aumentado significamente a cada ano. Além disso, também é possível notar evolução do número de questões, o que acarretou inconsistências nas bases de dados e falta de padronização do questionário, dificultando análises comparativas entre os anos de aplicação.

2.4 Considerações Finais

O projeto Web-Pide teve um avanço considerável abordando temas que foram se complementando, mas ainda não abordou os dados abertos governamentais e web semântica, ideias que vem sendo exploradas na atualidade. Portanto decidiu-se explorar esses conceitos no intuito de dar maior qualidade as funcionalidades do projeto. Para tal, utilizou-se a base do ENEM devido a fato de ela não ter sido explorada, tal qual foram exploradas as bases do SAEB e do CEB.

Capítulo 3

Web Semântica e *Open Linked Data*

Neste capítulo é apresentada uma visão geral da teoria sobre Web Semântica, destacando os principais componentes de sua arquitetura. Conceitos de *Linked Data* e de *Linked Open Data* são apresentados a fim de destacar sua importância para as diretrizes dos portais de dados abertos "data.gov". Além disso, os dois principais modelos de *Data Warehouse* são apresentados, destacando-se o modelo estrela utilizado neste trabalho.

3.1 Conceitos e Objetivos

Tim Berners-Lee, criador da *World Wide Web* (WWW), em suas palestras em 1998 criou a ideia da Web Semântica. Em 2001, foi publicado o artigo intitulado "*The Semantic Web*", em cooperação com o Dr. James Hendler (pesquisador na área de Inteligência Artificial no *Rensselaer Polytechnic Institute*) e com Ora Lassila (cientista finlandês que já vem trabalhando com o conceito de Web Semântica desde 1996). "A Web Semântica é uma visão: a ideia de ter dados na Web definidos e conectados de tal forma que possam ser utilizados por máquinas não só para fins de exibição, mas para automatização, integração e reuso dos dados através de várias aplicações". [Berners-Lee, 2000]

De acordo com seus criadores, a Web Semântica é uma extensão da Web atual em que a informação recebe um significado bem definido, o que permite uma cooperação mais avançada entre o computador e seus usuários. No entanto, apesar de a ideia ter surgido em meados de 1998, existem várias pesquisas para melhorar, expandir e padronizar o sistema, e com isso, muitas publicações e ferramentas já foram desenvolvidas.

A Web Semântica é definida por uma rede de dados a qual pode ser processada por máquinas de forma direta e indireta. A ideia principal é baseada no princípio do compartilhamento de dados. Quando um dado em particular é definido, pode-se conectá-lo a outros pedaços de informação estabelecendo categorias relacionadas.[W3C, 2012] Por exemplo,

uma pessoa vai viajar e decide reservar sua passagem na internet, uma aplicação que utiliza a Web Semântica reconheceria a palavra viagem como o tema e forneceria outras opções por associação, tais como, aluguel de carro, hotéis, restaurantes, promoções em viagens, entre outras. Além disso, com os metadados adicionados a seu perfil, a aplicação poderia indicar atividades como museus, concertos, eventos, esporte, de acordo com os interesses da pessoa em questão.

Esse novo conceito aplica métodos que vão além da apresentação linear da informação (que seria a Web 1.0) e da apresentação multilinear (Web 2.0, em que são introduzidas as aplicações web de compartilhamento de informação, colaboração e interação entre usuários) chegando a fazer uso de *hyper-structures* que levam a entidades de *hypertext*. Seu conceito implica em estender as redes de páginas com *hyperlinks* para uso humano por meio da inserção de metadados para uso do computador, além de definir como as páginas podem estar relacionadas entre si, o que permite que agentes automatizados possam acessar os dados de forma mais inteligente no lugar de seus usuários. Assim, o conceito original da Web Semântica é de um sistema que permite que as máquinas, com base no significado, compreendam e respondam às complexas solicitações por parte dos usuários.[Berners-Lee et al., 2001]

No entanto, para que haja "compreensão" por parte das máquinas da informação disponível, é necessária uma estruturação sistemática e prévia dos conteúdos, além de ferramentas próprias numa abordagem caracterizada por camadas. Sem a Web Semântica, as aplicações que possibilitam a integração dos dados oferecem pouco potencial para conectar fontes diversas, além de exigir mapeamento item por item entre os elementos de cada depósito de dados. Entretanto, por meio da Web Semântica, é possível que uma máquina se conecte a qualquer outra máquina e execute a troca e o processamento de dados eficientemente ao fazer uso da informação semântica inserida disponível universalmente e a qual descreva cada recurso. De fato, a Web Semântica irá tornar possível o acesso a toda informação em um único banco de dados gigantesco.

A principal meta da Web Semântica é direcionar a evolução da Web para que seja possível que usuários procurem, partilhem e combinem informações mais facilmente. Atualmente, os usuários são capazes de utilizar a Web para desempenhar todo tipo de tarefa, desde a reserva de um bilhete aéreo até a compra de uma mercadoria. Porém, as máquinas não são capazes de desempenhar tais tarefas sem que haja uma direção humana prévia. Na ideia vislumbrada para a Web Semântica, a informação poderá ser interpretada por máquinas de forma que estas possam desempenhar o trabalho repetitivo e chato de procurar e combinar informações e agir sobre elas.[Breitman, 2006]

Para que se torne possível implementar a Web Semântica, é necessário que metadados semânticos (os quais são dados que descrevem outros dados) sejam adicionados

aos recursos de informação. Assim, será possível que as máquinas processem eficientemente os dados baseadas na informação semântica que os descrevem. Quando houver informação semântica suficiente associada aos dados, os computadores serão capazes de fazer inferências sobre os dados, ou seja, serão capazes de compreender em que consiste um recurso de dados e como ele se relaciona a outros dados.

3.2 Arquitetura da Web Semântica

Atualmente, a Web Semântica está sendo muito utilizada para se referir a formatos e tecnologias que a tornam possível, principalmente quando há coleta, estruturação e recuperação de *collected linked data*. Na Figura 3.1 são ilustradas as camadas da arquitetura da Web Semântica, que incluem a camada Unicode/URI, a XML/Namespace/XML Schema, a RDF/ RDF *Schema*, a Ontologia, a Lógica, a Prova e por último a camada Confiança. Este trabalho explora a base da arquitetura da Web Semântica, trabalhando os conceitos desde Unicode/Uri até Sparql/OWL.

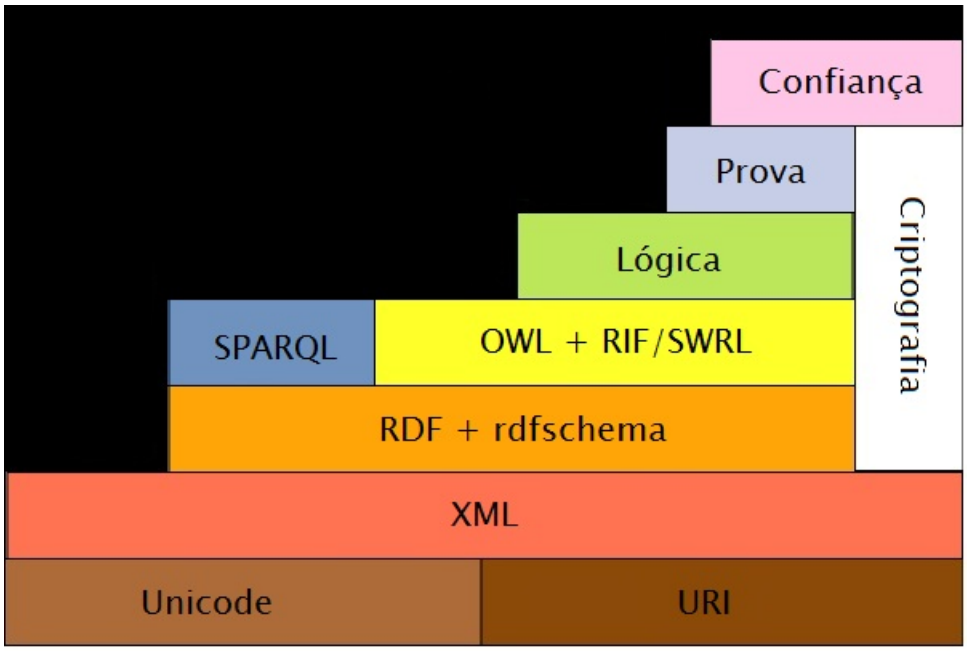


Figura 3.1: Arquitetura da Web Semântica [Berners-Lee, 2005]

A seguir são apresentados, sucintamente, os conceitos e tecnologias utilizados na arquitetura, segundo as W3C *Recommendations*. [W3C, 2012]

URI (*Uniform Resource Identifier*)

O URI, conhecido como identificador da Web, fornece os meios para a identificação de recursos de Web Semântica, pois define uma identificação única para manipulação dos

recursos nas diferentes camadas da arquitetura.

Na computação, o URI é caracterizado por uma cadeia de caracteres utilizada para identificar um nome ou um recurso na Internet. Por meio dessa identificação, é possível que haja uma interação com representações do recurso em uma rede (geralmente a *World Wide Web*) por meio de protocolos específicos. O que define um URI é o esquema que especifica uma sintaxe concreta e seus protocolos associados. "Uniform Resource Identifier(URI) é uma cadeia compacta de caracteres para identificar uma fonte abstrata ou física"[Berners-Lee et al., 1998].

A sintaxe do URI é formada pelo nome do URI *Schema* (por exemplo, "http", "ftp", "mailto", ou "file") seguido do caractere ":" e, depois, de uma parte de *schema* específico, cujas sintaxe e semântica são determinadas pelas especificações que regem o *schema*. O IETF (*Internet Engineering Task Force*) controla cuidadosamente a sintaxe dos URIs. "A missão do IETF é fazer a Internet trabalhar melhor, produzindo documentos técnicos relevantes e de alta qualidade que influenciem a maneira como as pessoas utilizam, definem e gerenciam a Internet." [IETF, 2011]

UNICODE

O Unicode tem como função representar e manipular o texto em várias línguas. "A Web Semântica também deve auxiliar na criação de uma ponte entre os documentos em diversas línguas e, portanto, deve ser capaz de representá-las". [UNICODE, 2012]

O Unicode é um padrão da indústria da computação desenvolvido visando à codificação, à representação e à manipulação consistentes de texto escrito na maioria dos sistemas de escrita existentes no mundo. A última versão ou revisão é o Unicode 6.2, publicado em 2012. [UNICODE, 2012]

XML (*Extensible Markup Language*)

"O XML é o formato universal para documentos e dados estruturados na web".[W3C, 2012] Ele permite a criação de documentos formados por dados estruturados, pois a Web Semântica dá significado (semântica) a tais dados. Essa linguagem consiste de um conjunto de regras para a codificação de documentos no formato de leitura para máquinas. As principais metas do XML é enfatizar a simplicidade, a generalização e a aplicação da Internet. Para isso, suas características incluem um formato de dados textual com uma base forte proporcionada pelo Unicode para as línguas existentes, e, embora seu enfoque seja documentos, é amplamente utilizada para a representação de estruturas de dados arbitrárias como em *Web Services*.

Para processar dados XML, muitas *application programming interfaces* (APIs) têm sido desenvolvidas por desenvolvedores de *software*, e até agora, centenas de linguagens baseadas em XML (tais como SOAP e XHTML) foram desenvolvidas. Além disso, os for-

matos baseados em XML tornaram-se o padrão da maioria das ferramentas para escritório, como o Microsoft Office, o OpenOffice e o iWork da Apple.[W3C, 2012]

XML NAMESPACES

O XML *Namespaces* proporciona um modo de utilização de *markups* provenientes de outras fontes, um método simples para qualificar nomes de elementos e atributos utilizados em documentos XML ao associá-los a *namespaces*. [W3C, 2012]

A utilização do XML *Namespaces* proporciona elementos e atributos com nomes únicos em documentos XML. É importante destacar que um documento XML pode conter nomes de elementos ou atributos provenientes de mais de um vocabulário XML. Porém, se for atribuído um *namespace* a cada vocabulário, põe-se fim à ambiguidade entre elementos ou atributos com nomes idênticos.

Para se atribuir um *namespace* ao documento XML, tal *namespace* necessita de um nome, que vai ser um URI. Normalmente, o URI que for escolhido para o *namespace* de qualquer vocabulário XML irá descrever um recurso controlado pelo autor ou organização que definiu o vocabulário.

RDF (*Resource Description Framework*)

Em sua definição, o RDF é um framework com o propósito de criar demonstrações em forma de triplas, permitindo que as informações sobre recursos sejam representadas na forma de grafos (a Web Semântica é as vezes referida como um Gigante Grafo Global). "O modelo de dados RDF pode ser utilizado para descrever qualquer tipo de fonte que possa ser identificada por um URI. (...) Um documento RDF é uma sequência de afirmações chamada triplas RDF."[Sikos, 2011] Uma vez que o RDF é visto como uma linguagem declarativa, ele proporciona uma forma padronizada para a utilização do XML como modo de representação de metadados. Estes são representados no formato de sentenças sobre as propriedades dos recursos na Web e seus relacionamentos, e desde que possuam um endereço Web, esses recursos podem ser praticamente qualquer objeto. [Breitman, 2006]

A idéia por trás das triplas baseia-se no fato de que as elas criam conexões entre os recursos e os dados relacionados a eles por meio de sua estrutura sujeito-predicado-objeto. Nessa estrutura, o sujeito se refere ao recurso, o objeto é a característica, o valor atribuído e o predicado define a relação entre o sujeito (recurso) e o objeto (valor). No entanto, as triplas são mais do que conexão entre palavras específica, suas partes podem ser substituídas por URIs, que são únicos a um conceito particular. Com tal substituição, possíveis ambiguidades podem ser completamente evitadas.

Caracterizado pelo W3C como uma família de especificações utilizadas para descrever objetos e a relação entre eles por meio de expressões e originalmente desenvolvido como modelo de dados, o RDF tem sido usado como um método para descrição conceitual ou

modelagem de informação. Ele é implementado em recursos de Web por meio da utilização de vários formatos sintáticos. O RDF é bastante similar a abordagens de modelagem conceitual tais como Entidade-Relacionamento ou Diagrama de Classes, já que estes se baseiam na prática de se fazer declarações sobre os recursos na forma de expressões sujeito-predicado-objeto, ou seja, triplas (que permitem a codificação de uma Web Semântica que pode ser lida tanto por computadores como pelos usuários).[W3C, 2012]

Assim, o RDF e seu mecanismo de descrição de recursos é considerado um dos componentes principais da Web Semântica. O modelo de dados simples do RDF e sua capacidade de modelar conceitos abstratos e diferentes fazem com que os usuários utilizem-no cada vez mais em aplicações de gerenciamento de conhecimento que não estão diretamente ligadas à atividade da Web Semântica. Um conjunto de declarações RDF representa de maneira intrínseca um grafo múltiplo, direcionado e rotulado. "Armazenar e utilizar query em dados RDF é uma das tarefas básicas dentro de qualquer aplicação de web semântica".[Harth and Decker, 2005]

RDFS (*RDF Sschema*)

De acordo com o W3C [W3C, 2012], o RDF *Schema* (também abreviado como RDFS, RDF-S ou RDF/S) proporciona um vocabulário básico ao RDF, tornando possível a criação de hierarquias de classes e propriedades. Definido como um conjunto de classes com propriedades específicas que utiliza a representação de conhecimento RDF, o RDFS fornece elementos básicos para a descrição das ontologias (também chamadas vocabulários RDF), as quais visam estruturar os recursos RDF.

Sua primeira versão foi publicada em 1998 pela W3C e a recomendação final em 2004. Segundo Breitman [Breitman, 2006], as principais concepções RDFS são as classes, as propriedades e as propriedades de utilidades RDFS construídas no vocabulário RDF limitado.

OWL (*Web Ontology Language*)

A OWL é uma linguagem de representação de conhecimento utilizada para a criação de ontologias. Elas são caracterizadas por semântica formal e serializações baseadas em RDF/XML para a Web Semântica. "Uma ontologia é uma especificação explícita de uma conceituação. (...) Para todos sistemas, o que "existe" é o que pode ser representado. Quando o conhecimento de um domínio é representado em um formalismo declarativo, o conjunto de objetos que podem ser representados é chamado de universo de discurso. Este conjunto de objetos, e as relações entre eles descritíveis, são refletidos no vocabulário de representação com que um programa de conhecimento representa o conhecimento.". [Gruber, 1993b]

A OWL pode estender as RDFS por meio da adição de conceitos mais avançados os

quais descrevem a semântica das declarações RDF. Ela permite a declaração de conceitos adicionais como por exemplo cardinalidade, restrições de valores e até mesmo características de propriedades tais como transitividade. Essa linguagem é baseada na lógica descritiva e, portanto, acrescenta o poder de racionalização à Web Semântica.

Para compreensão das OWLs, deve-se falar primeiramente sobre a ontologia no ramo da ciência da computação. "Ontologias podem melhorar o funcionamento da Web de muitas formas. Elas podem ser usadas de modo simples para melhorar a precisão das buscas na Web - programas de busca podem procurar apenas aquelas páginas que se referem a um conceito específico em vez de todos os conceitos utilizando palavras-chave ambíguas".[Berners-Lee et al., 2001] Assim, tem-se que uma ontologia representa formalmente o conhecimento como um conjunto de conceitos contidos em um domínio, assim como a relação entre tais conceitos. Uma ontologia interpreta o vocabulário compartilhado e a taxonomia, e estes modelam um domínio com a definição de objetos e/ou conceitos e suas propriedades e relações.

Como a ontologia é composta por conceitos e seus relacionamentos que organizam a informação, ela é utilizada em áreas como a Web Semântica, a inteligência artificial, a engenharia de sistemas, a engenharia de *software*, a informática biomédica, a arquitetura de informação, entre outras, como forma de representação de conhecimento.[Breitman, 2006]

Nas palavras de Gruber, "Uma ontologia especifica um vocabulário com o qual se pode fazer assertivas, que podem ser *inputs* ou *outputs* de agentes de conhecimento (por exemplo um programa de software). Como uma especificação de interface, a ontologia proporciona uma linguagem para comunicação com o agente." [Gruber, 1993a]

A estrutura das ontologias contemporâneas compartilham várias similaridades independentemente da linguagem em que são expressas. Assim, os componentes mais comuns das ontologias incluem indivíduos (objetos ou instâncias), classes (conjuntos e conceitos), atributos (aspectos, parâmetros e propriedades), relações (como relacionam entre si), termos de função, restrições, regras (antecedente-consequência), axiomas (afirmativas em forma lógica) e eventos (alterações de atributos e relações). Além desses componentes estruturais, a ontologia é geralmente codificada por meio da utilização de OWLs.

A família de OWLs possui muitas espécies, serializações, sintaxes e especificações com nomes semelhantes. Com relação às espécies de OWLs, existem três variantes aprovadas pelo W3C [W3C, 2012], são elas: OWL *Lite*, OWL DL e OWL *Full*, que nessa ordem representam um aumento de expressividade entre uma espécie e outra. Cada uma dessas sublinguagens consiste de uma extensão sintática de sua predecessora, a qual é mais simples.

Já em relação à sintaxe, pode-se afirmar que a família de linguagens OWL permite

uma variedade de sintaxes. Existem duas categorias, sintaxe de alto nível (*high level syntax*), que é usada para especificar a estrutura e a semântica da ontologia, e sintaxes de troca (*exchange syntaxes*).[W3C, 2012]

SPARQL (*SPARQL Protocol and RDF Query Language*)

SPARQL é uma linguagem de consulta RDF que pode ser utilizada para fazer consultas em quaisquer dados baseados em RDF, o que inclui declarações envolvendo RDFS e OWL. A linguagem de consulta é necessária na recuperação de informação em aplicações Web Semântica. Portanto, o SPARQL é considerado uma peça chave em tais aplicações, permitindo que uma *query* seja composta de padrões triplos, conjunções, disjunções e padrões opcionais.

"SPARQL pode ser utilizado para expressar *queries* através de diversas fontes de dados, não importando se os dados estão armazenados nativamente como RDF ou visualizados como RDF via *middleware*."[W3C, 2012]

A linguagem SPARQL define quatro variações diferentes de *query* com diferentes propósitos: *Select query*, que tem como propósito extrair valores brutos e apresentar o resultado em formato de tabela; *Construct query*, que visa extrair informações e transformar os resultados em RDF válido; *Ask query*, que proporciona um resultado Verdadeiro/Falso simples; e *Describe query*, cujo objetivo é extrair um grafo RDF que permite que o usuário determine o que é informação útil.[W3C, 2012]

Lógica, Prova e Confiança

A camada de lógica permite que sejam definidas regras lógicas que possibilitam a inferencia automática de novas informações a partir das relações pré-existentes.

Na camada de prova espera-se que seja possível a verificação/comprovação da coerência lógica dos recursos, de modo que os aspectos semânticos das informações estejam descritos de maneira consideravelmente adequada, atendendo a todos os requisitos das camadas inferiores. Espera-se que a a camada confiança possa garantir que as informações estejam representadas de modo correto, possibilitando um certo grau de confiabilidade.[S., 2006]

Criptografia

A Criptografia quando aplicada à ciência da computação se trata da construção e análise de protocolos que superam a influência de adversários e que estão relacionados a vários aspectos da segurança da informação, como por exemplo, confidencialidade e integridade dos dados e autenticação.

3.3 Linked Data

Assim como a Web tradicional o acesso às páginas é por meio de navegadores HTML e a rede de dados criada pelo *Linked Data* pode ser acessada por navegadores de *Linked Data*. No entanto, ao invés de seguir apenas *links* entre páginas HTML, os navegadores de *Linked Data* permitem que os usuários naveguem entre fontes de dados diferentes por meio de RDF. Com isso, um usuário pode começar com uma fonte de dados e mover-se por meio de uma infinidade de fontes de dados conectados por *links* RDF.

Os quatro princípios básicos da tecnologia *Linked Data* são: a utilização de URIs na identificação, uma vez que se não foi usado o grupo de símbolos URI, não é considerado Web Semântica; em segundo está o uso de HTTP URIs para que se possa referenciar as coisas e consultá-las (usuários e máquinas); o terceiro princípio é o fornecimento de informação útil por meio do URI, podendo-se, em geral, consultar as propriedades e as classes dos dados e assim obter informação a partir de RDF, RDFS e ontologias OWL (inclusive a relação entre os termos da ontologia); e, por último, a inclusão de *links* em outros URIs relacionados aos dados expostos para que se possa otimizar a descoberta de outras informações relacionadas na Web. citeBeners06

Devido à sua importância na Web Semântica, Berners-Lee reforçou as três regras extremamente simples do *Linked Data* em sua apresentação sobre o assunto na conferência TED-2009 (*Technology Entertainment and Design*): primeiramente, todos os recursos conceituais devem ter seus nomes começando com HTTP; em segundo lugar, além de obter informação importante sobre o evento, também se deve obter alguns dados em formato padrão que não são dados evidentemente importantes, mas que podem ser úteis para alguém; por fim, juntamente com a informação, deve-se obter as relações, e sempre que uma relação inclua outra coisa, esta tem que ter seu nome começando com HTTP.

A relação entre o *Linked Data* e a Web Semântica é bastante discutida, mas um dos conceitos amplamente aceitos é que a Web Semântica é o todo e as partes são *Linked Data*. Berners-Lee mencionou várias vezes que a forma correta de se fazer a Web Semântica é por meio de *Linked Data*, o que é facilmente aplicado, desde que se utilize o modelo de *Linked Data* para publicar os dados estruturados na Web e se utilize as ligações para conectar dados de fontes diferentes entre si. [Berners-Lee et al., 2009]

3.4 *Linked Open Data*

Linked Open Data (LOD) pode ser definido como *Linked Data* de uso livre. Sua principal meta é promover ampla utilização e acesso a bibliotecas, arquivos, museus, dados governamentais, entre outros. Como a Web ainda não está totalmente preparada

e estruturada para que seus dados sejam conectados de acordo com a tecnologia *Linked Data*, existem muitas iniciativas que propõem essa estruturação contínua.

Entre essas iniciativas, tem-se o projeto LOD do W3C (projeto originalmente proposto por Chris Bizer e Richard Cyganiak), uma iniciativa comunitária cujo principal objetivo é tornar os dados livremente disponíveis a todos. Esse projeto pretende estender a Web com um *data commons* por meio da publicação de vários grupos de *open data* em formato RDF e do estabelecimento de *links* RDF entre os dados de diferentes fontes. Ou seja, resume-se a um esforço comunitário para converter as fontes existentes de *open data* para RDF e tornar tais fontes parte da Web Semântica.

Outra iniciativa importante é o *Linked Open Data Around the Clock* (LATC)[LATC, 2012], um projeto europeu patrocinado pela Comissão Européia como parte do 7º Programa de *Framework* (7th *Framework Programme*) cuja meta é dar apoio à publicação e utilização de LOD. Os objetivos dessa iniciativa incluem melhorar uma infraestrutura contínua que monitora a utilização de LOD assim como a qualidade do LOD; dar suporte à comunidade por meio de tutoriais e guias sobre as melhores práticas; manter uma área de teste para o processamento de *Linked Data* juntamente com dados da União Européia; desenvolver uma biblioteca de ferramentas de processamento de dados *open source*, entre outros.

Na Figura 3.2 é exemplificada a Nuvem LOD do projeto LOD do W3C, que ilustra tanto os grupos de dados numa nuvem de LOD quanto a conexão entre esses dados. Cada nódulo no diagrama representa um conjunto distinto de dados publicado como *Linked Data*, e cada arco indica os *links* que existem entre os dois conjuntos de dados conectados. Arcos mais escuros significam um número maior de *links* entre os dois grupos de dados.

3.5 *Open Government Data*

As primeiras iniciativas foram financiadas pelos governos da Inglaterra e dos Estados Unidos, que criaram os portais de dados abertos "data.gov" (portal americano, lançado em maio de 2009) e "data.gov.uk" (portal britânico, lançado em setembro de 2009). Seguindo as tendências desses dois países, o Canadá, a Austrália e a Nova Zelândia começaram a publicar dados em portais oficiais seguindo os mesmos modelos dos Estados Unidos e do Reino Unido e das diretrizes do *Open Government Data*. Tais diretrizes, ou princípios, foram definidas pelo W3C e totalizam oito regras que guiam as práticas de *Open Government Data*: [W3C, 2009]

- Completos - todos os dados públicos devem estar disponíveis.
- Primários - os dados são apresentados conforme coletados na fonte, sem agregação

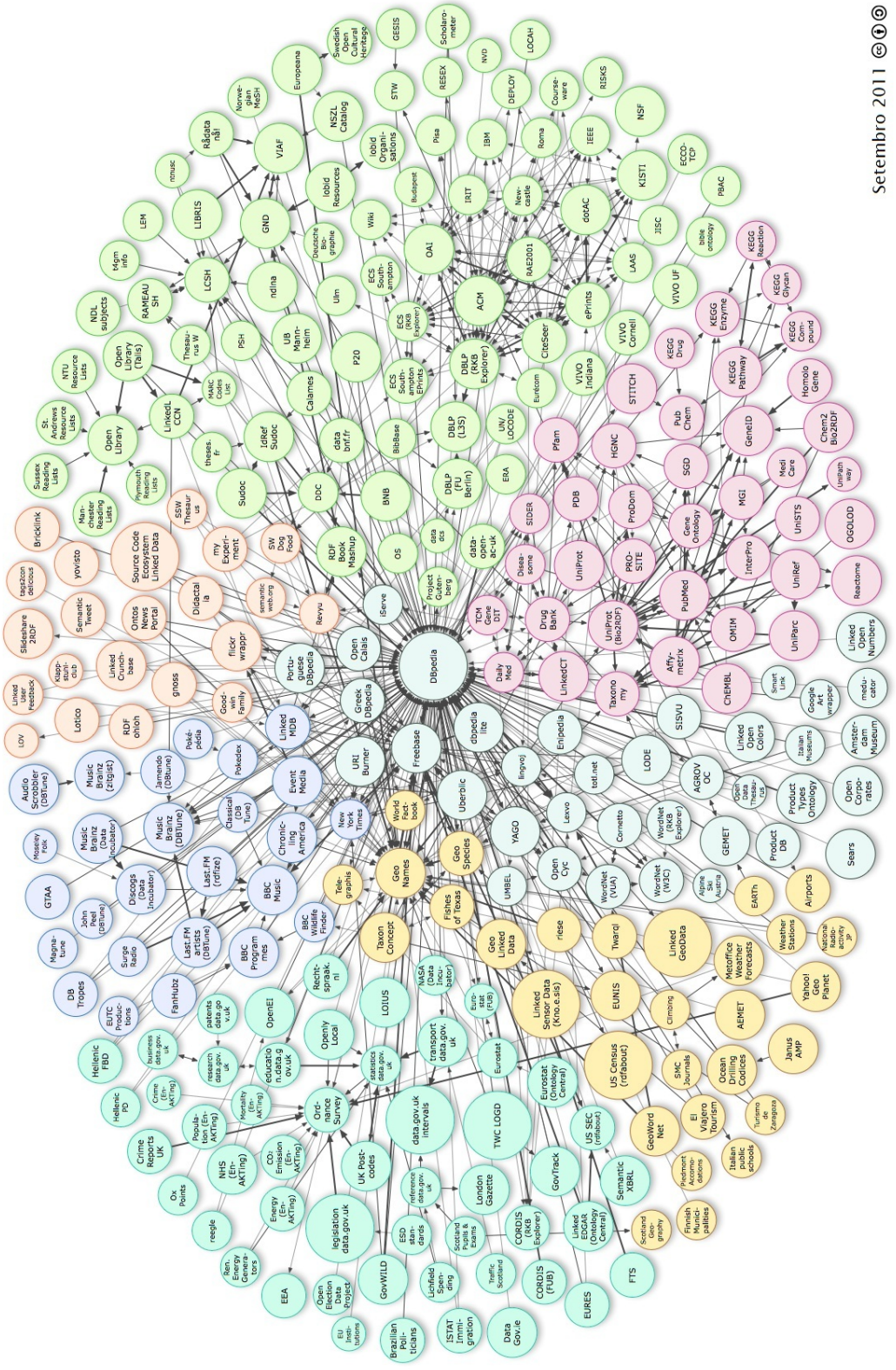


Figura 3.2: Nuvem *Linked Open Data* [Cyganiak and A., 2011]

ou modificação e com o maior nível possível de granularidade.

- Atuais - os dados devem ser disponibilizados tão rapidamente quanto necessários à preservação do seu valor.
- Acessíveis - os dados devem proporcionar o maior alcance possível de usuários e para o maior número possível de finalidades.
- Compreensível por máquinas - os dados devem estar razoavelmente estruturados para que possibilitem o processamento automatizado.
- Não discriminatório - os dados devem estar disponíveis a todos, sem exigências de cadastro ou requerimento.
- Não proprietários - os dados devem estar disponíveis em formatos sobre os quais nenhuma entidade detenha controle exclusivo.
- Livre de licenças - os dados não devem estar sujeitos a restrição de direito autoral, patente, propriedade intelectual ou segredo industrial.

Com a tendência de tornar os dados governamentais abertos ao público, várias iniciativas em todo o mundo estão tornando disponíveis grandes quantidades de dados governamentais brutos na Web. Esse acesso livre aos cidadãos permite transparência e disponibiliza mais serviços públicos, encorajando o uso e reuso de informação governamental por um público cada vez maior. Países como os Estados Unidos e o Reino Unido até mesmo criaram catálogos e portais de forma a tornar mais fácil para o público encontrar e utilizar os dados governamentais. Tais dados estão disponíveis em vários formatos (como *spreadsheets*, banco de dados relacionais e RDF) e podem ser encontrados em diferentes domínios (como geoespacial, estatística, transporte, etc.).

A aplicação dos princípios do *Linked Data* aos bancos de dados governamentais tem grande potencial. No entanto, devido à falta de recursos necessários para transformar dados brutos em *Linked Data* de qualidade em ampla escala, esse potencial não tem sido atualmente aproveitado. David Eaves [Eaves, 2009], canadense e conselheiro sobre as práticas de *Open Government Data* em vários países, definiu e publicou as Três Leis dos Dados Abertos Governamentais, consideradas importantes pelo W3C e também um reflexo dos oito princípios:

- Se o dado não pode ser encontrado e indexado na web, ele não existe.
- Se não estiver aberto e disponível em formato compreensível por máquina, ele não pode ser reaproveitado.

- Se algum dispositivo legal não permitir sua reaplicação, ele não é útil.

O Brasil é o país pioneiro na América Latina quando se trata de dados abertos (*open data*). Em 2006, foi criado o Comitê de Organização de Informações da Presidência da República (COI - PR), que objetiva desenvolver ferramentas que facilitem a coleta, o armazenamento, a validação e a utilização das informações sobre a ação governamental. Em 2009, deu-se início ao projeto DadosGov (http://dados.gov.br/), cujo objetivo principal é criar e disponibilizar um Catálogo Aberto de Informações para a melhoria na gestão pública além de facilitar o acompanhamento das ações governamentais pela sociedade. Hoje, o DadosGov COI-PR é um projeto cuja página web foi desenvolvida pelo Serpro, em linguagem PHP e banco de dados PostGreSQL, de acordo com os critérios definidos pela Presidência da República. Os dados publicados abrangem informações detalhadas de mais de 37 órgãos da administração pública federal. Além do Serpro, outros parceiros do projeto DadosGov são o DATAPREV, o W3C e as Universidades COPPE/UFRJ e PUC/RJ. Todas essas instituições estão trabalhando em parceria para o crescimento do projeto DadosGov e para o fornecimento de dados em formato aberto.

Em regra, os dados abertos governamentais estão fundamentados em três pilares, os quais são a transparência, a participação e a colaboração. De acordo com o W3C, "Dados Abertos Governamentais são a publicação e a disseminação das informações do setor público na Web, compartilhadas em formato bruto e aberto, compreensíveis logicamente, de modo a permitir sua reutilização em aplicações digitais desenvolvidas pela sociedade."[W3C, 2009]

Enfim, para destacar a importância dessa temática de dados abertos no Brasil, em 18 de novembro de 2011, a presidente Dilma Rousseff sancionou a lei de acesso a informações públicas (Lei nº 12.527/2011), que dispõe sobre os procedimentos a serem observados pela União, Estados, Distrito Federal e Municípios, com o fim de garantir o acesso à informação. Muitos órgãos de governo nas diferentes esferas estão à procura de soluções, ferramentas e tecnologias para contribuir com a transparência das informações públicas.

3.6 *Data Warehouse*

O DW (*Data Warehouse*) é essencialmente um banco de dados central de larga escala carregado de informações provenientes de vários bancos de dados operacionais. Ele é um tipo especial de *database* utilizado para armazenar grandes quantias de dados, incluindo dados analíticos, históricos e/ou sobre o cliente, e consolidar fontes de dados diferentes e tornar os recursos disponíveis para apoiar a geração de informação crítica à tomada de decisões estratégicas. [Inmon, 2005]

O DW é bastante diferente dos sistemas de banco de dados tradicionais. Os dados provenientes de diversos sistemas operacionais são trazidos para um DW. Uma vez nesse novo ambiente, formatos que anteriormente deixavam a base inconsistente são alterados, deixando um formato de dados uniforme, de fácil utilização e ilimitado. Ou seja, o DW prepara dados diferentes e inacessíveis para transformação em informação utilizável. Só essa característica já demonstra a superioridade dos benefícios do DW.

Além disso, o DW mantém suas funções em três fases: a fase de obtenção de dados na qual é usado para obter e armazenar grandes quantidades de informações; a fase de integração em que é usado para tornar os dados armazenados uniformes e consistentes; e a fase de acesso que se resume na facilitação da utilização da informação pelo usuário final. Ao fim do processo, a principal fonte de dados é limpa, transformada, catalogada e disponibilizada aos usuários para aplicação de técnicas de *data mining* e *online analytical processing* (OLAP).

São considerados componentes essenciais do sistema de DW os meios utilizados para obter e analisar os dados (isto é, a extração, a transformação e o carregamento dos dados) e para gerenciar o dicionário de dados. Em relação à variação de tempo: "Os dados de um *Data Warehouse* são precisos em relação ao tempo, representam resultados operacionais em determinado momento de tempo, o momento em que foram capturados. Os dados de um DW são um *snapshot* (...)."[Machado, 2010]

Os vários benefícios de um DW são: oportunidade de manter um histórico de dados (mesmo se os sistemas-fonte não o fizerem), integrar dados de múltiplos sistemas-fonte e assim permitir uma visualização central, melhorar os dados ao proporcionar códigos e descrições consistentes, apresentar a informação de maneira consistente, proporcionar um único modelo de dados comum para todos os dados relevantes independentemente da fonte e reestruturar os dados de forma que façam sentido para o usuário final e que permitam um excelente desempenho mesmo em consultas analíticas complexas.

Segundo Inmon [Inmon, 2005], existem duas abordagens principais no armazenamento de dados em DW: a abordagem dimensional ou esquema Estrela e a abordagem normalizada ou esquema Floco de Neve.

A abordagem dimensional é "uma metodologia para modelar dados lógicos com o objetivo de melhorar o desempenho das *queries* e a facilidade de utilização, que começa com um grupo de eventos básicos de medida". [Kimball and Ross, 2002]. O processo inicia-se com uma tabela de fatos, que é geralmente construída em um ambiente SGBD relacional com um registro para cada medida discreta. Uma vez que essa tabela de fatos é construída, ela é envolta em um grupo de tabelas dimensionais que descrevem com acurácia aquilo que se sabe sobre o contexto de cada registro de medida.

O nome Estrela foi criado devido à estrutura característica de uma abordagem dimensional, que se assemelha a uma estrela. Esse modelo tem sido amplamente utilizado, pois provou apresentar enormes vantagens em áreas como compreensão, predicabilidade, extensão etc. "Os modelos dimensionais são também a fundação lógica para todos os sistemas OLAP" [Kimball and Ross, 2002]. Além disso, o modelo dimensional tem como características a simplicidade e a simetria, que fazem com que os otimizadores de banco de dados processem mais eficientemente e com menos junções. Outro benefício é sua capacidade de se expandir para acomodar mudanças, pois todas as dimensões são equivalentes e possuem pontos de entrada simetricamente iguais na tabela de fato.

Na abordagem dimensional, os dados transacionais são repartidos em "fatos" (geralmente dados transacionais numéricos) ou "dimensões" (que são um conjunto de atributos que contextualizam uma informação e que estão muito relacionados). Por exemplo, uma venda pode ser dividida em fatos, tais como número de produtos e preço pago por tais produtos, e as dimensões podem ser a data do pedido, o nome do cliente, o número de referência do produto e o responsável pela venda. Na Figura 3.3 é ilustrado um exemplo simples de modelo estrela, onde a tabela fato é "Data" e as dimensões são "Tempo", "Produto" e "Local".

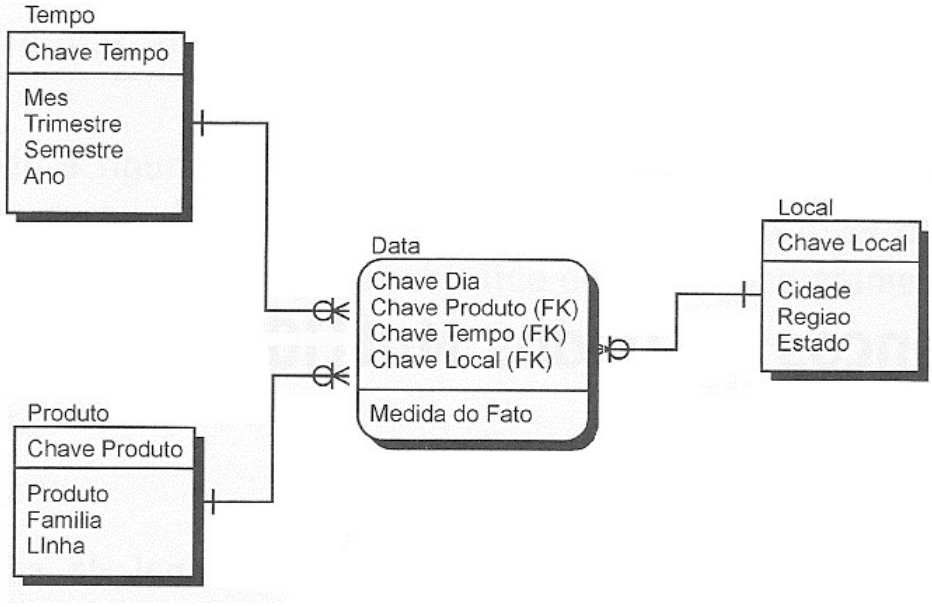


Figura 3.3: Exemplo de modelo estrela[Machado, 2010]

A principal vantagem da abordagem dimensional é que o DW se torna mais fácil de ser compreendido e consequentemente utilizado pelo usuário. "As estruturas dimensionais são mais fáceis de serem compreendidas pelos usuários na empresa, pois é dividida em medidas/fatos e contexto/dimensões. Os fatos estão relacionados aos processos empresariais da organização e aos sistemas operacionais enquanto as dimensões contêm o contexto da

medida". [Kimball et al., 2008]

As principais desvantagens do modelo estrela são: manter a integridade dos fatos e dimensões, o carregamento de dados provenientes de sistemas operacionais variados no DW é um processo complicado e é difícil modificar a estrutura do DW se a organização que adotou a abordagem dimensional mudar a forma como faz negócio.

Na abordagem normalizada ou esquema Floco de Neve, os dados armazenados no DW seguem, de certa forma, regras de normalização de base de dados. As tabelas são agrupadas de acordo com o assunto que reflete a categoria dos dados em geral. Essa estrutura divide os dados em entidades, o que cria várias tabelas no banco de dados, conforme pode ser visto na Figura 3.4. Quando utilizado em grandes empresas, resulta em dezenas de tabelas ligadas em uma rede. Além disso, cada entidade criada é convertida em uma tabela separada quando o banco de dados é implementado.

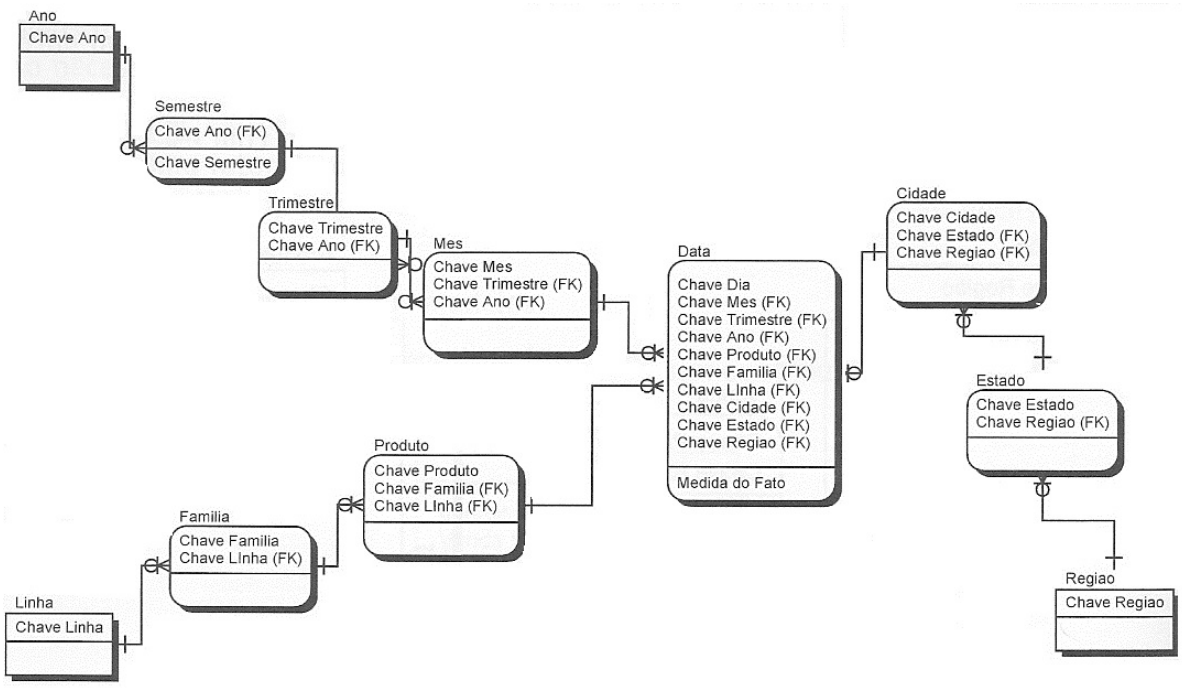


Figura 3.4: Exemplo de modelo normalizado ou *Snow Flake*[Machado, 2010]

A principal vantagem do modelo Floco de Neves é a forma direta e clara usada para adicionar informação ao DW. Sua desvantagem é que, com o grande número de tabelas envolvidas, pode se tornar difícil para os usuários agrupar dados de fontes diferentes e transformá-los em informação coerente e, então, acessar tais informações sem uma compreensão precisa das fontes de dados e da estrutura do DW.

3.6.1 Sistemas OLAP e OLTP

Em geral, para dar suporte aos recursos de consulta, extração e manipulação em um DW são utilizados os sistemas OLAP (*Online Analytical Processing*) e OLTP (*Online Transaction Processing*). O OLAP é uma abordagem que rapidamente responde *multi-dimensional analytical* (MDA) *queries* - *queries* analíticas multidimensionais. Esse sistema é caracterizado por um volume relativamente pequeno de transações. As ferramentas OLAP permitem que o usuário analise interativamente os dados multidimensionais a partir de várias perspectivas. O núcleo de qualquer sistema OLAP é um cubo multidimensional que consiste de fatos numéricos (medidas) categorizados por dimensões. "Combinando tais dimensões, o usuário tem uma visão dos dados de um DW, podendo efetuar operações ditas básicas como *slice and dice*, que é uma forma de mudança das dimensões a serem visualizadas, *drill down* e *roll up*, como se denomina a navegação entre os níveis de detalhamento dos dados de DW".[Machado, 2010]

O OLTP é caracterizado por um enorme número de pequenas transações *online* (tais como Inserir, Atualizar, Apagar, Recuperar etc.). Sua principal característica é o processamento extremamente rápido de *queries*, mantendo a integridade dos dados em ambientes de multi-acesso. Esse sistema tem como objetivo facilitar e gerenciar aplicações orientadas para transação e normalmente é representado por modelos normalizados.

3.7 Considerações Finais

Este capítulo discorreu sobre o embasamento necessário em relação à Web semântica e ao LOD. Nele foi possível entender a arquitetura e todo arcabouço necessário para construção de um LOD utilizando os conceitos apresentados anteriormente. Além disso, foi apresentado o conceito de DW e também de modelo estrela, conceitos estes fundamentais na construção do DW-ENEM.

Capítulo 4

Processos de geração de LOD

Neste capítulo serão apresentados os processos que fazem a triplificação de um DW, apontando os passos de cada processo. Para implantar um LOD é necessário um arcabouço de ferramentas que triplifiquem os dados, ou seja, faça uso dos dados que estão no DW e produza triplas no formato RDF - representações semânticas. As ferramentas que poderão desempenhar tal papel são: a *Stdtrip* e a *Triplify*, a qual foi indicada respectivamente nos trabalhos de [Auer et al., 2009] e [Salas et al., 2010]; a Babel, a qual é demonstrada no trabalho [Marx, 2012]; ou ainda a ferramenta *OLAP2DataCube*, a qual é referenciada no trabalho [Salas et al., 2011].

4.1 *Stdtrip, Triplify* e Virtuoso

O primeiro processo analisado foi a utilização das ferramentas Stdtrip, *Triplify* e Virtuoso. A análise do processo utilizou como base o trabalho de [Mota, 2011], que implementou a triplificação da base do CEB 1995. A primeira ferramenta a ser utilizada nessa etapa é a Stdtrip. Os requisitos iniciais para utilizá-la incluem a execução de três passos:

- Carga dos microdados no SGBD MySQL
- Execução da normalização da base de dados
- Tradução dos termos que estão na base de dados para o inglês

Segundo [Mota, 2011], após esses passos já é possível utilizar o Stdtrip, e o resultado atingido ao executá-lo é a obtenção das representações semânticas dos dados em formato OWL, além de um arquivo de configuração da ferramenta *Triplify*. Esse processo de geração das representações semânticas tenta capturar a estrutura do banco de dados e transformá-la em representações semânticas. A partir dessas representações semânticas, é

feita uma comparação com representações semânticas padrões para que o usuário selecione o melhor vocabulário gerado. Caso a ferramenta não consiga satisfazer o usuário, então é possível buscar um vocábulo em uma interface web. Mas se mesmo assim não for encontrado o vocábulo, é possível criá-lo.

A próxima fase tem início a partir dos dois artefatos gerados pelo Stdtrip. Essa fase consiste na utilização da ferramenta *Triplify*, que tem como entrada o arquivo OWL, um arquivo de configuração da ferramenta *Triplify*, e a base de dados a ser triplificada. A partir deles, gera-se o arquivo triplas RDF. Por último, deve-se carregar esse arquivo de triplas em um banco de triplas, que nesse caso foi utilizado o Virtuoso.

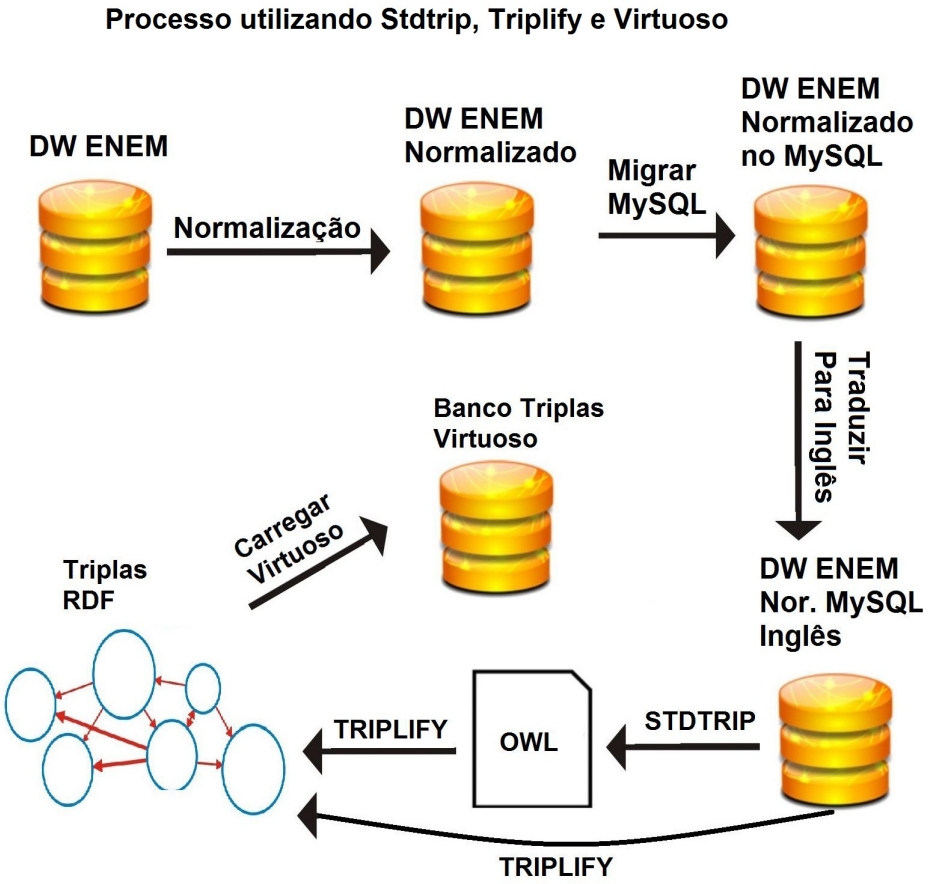


Figura 4.1: Processo utilizando Stdtrip, *Triplify* e Virtuoso no DW ENEM.

A Figura 4.1 exemplifica como seria o processo aplicado ao DW do ENEM, conforme descrito anteriormente. De acordo com a figura, foram executados seis passos até chegar ao resultado final: triplas RDF carregadas no banco de triplas. Além disso, fica evidente no processo a etapa de normalização, o que no caso dos modelos de DW utilizados para o projeto seria um contra-senso, pois o esquema estrela não é normalizado e caso se fosse normalizar a base, estar-se-ia mudando do esquema Estrela para o esquema Floco de Neve. Portanto, decidiu-se buscar outras ferramentas que facilitassem o processo de triplificação.

4.2 Ontowiki, *OLAP2DataCube* e Virtuoso

O próximo processo analisado foi o uso da ferramenta *OLAP2DataCube*, desenvolvida em conjunto por equipes das Universidade de Leipzig e Pontifícia Universidade Católica (PUC) e consiste de um *plug-in* para a aplicação Ontowiki. Portanto, para utilizar *OLAP2DataCube* é necessária a configuração da ferramenta Ontowiki [Auer et al., 2006]. Assim como a Ontowiki, a *OLAP2DataCube* foi desenvolvida em PHP e ambas são *open source*.

A Ontowiki foi desenvolvida no intuito de suportar criação, manutenção e publicação de bases RDF. Seus criadores e mantenedores fazem parte do grupo *Agile Knowledge Engineering and Semantic Web* (AKSW) da Universidade de Leipzig. São também conhecidos pelo projeto DBpedia.

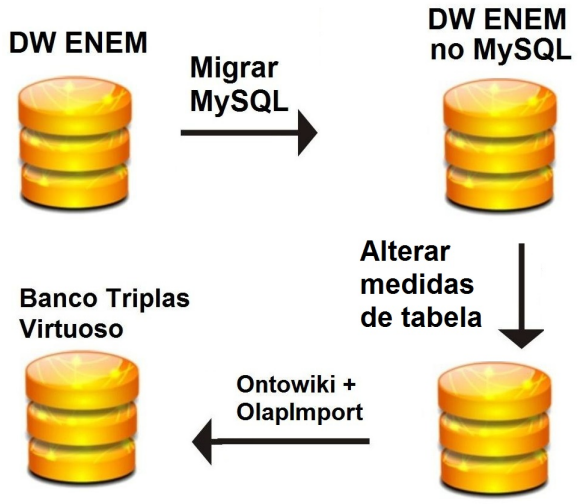


Figura 4.2: Processo utilizando Ontowiki, *OLAP2DataCube* e Virtuoso no DW ENEM.

Antes de iniciar o estudo da *OLAP2DataCube*, é preciso entender que o software foi projetado para ser o mais simples possível, e por isso ele foi criado para utilizar o relacionamento entre a tabela de fato e as de dimensão. Para isso, ele se utiliza das *primary keys* e *foreign keys* para definir quais sãos as tabelas de fato e as tabelas de dimensão. Outro fator importante é que as medidas presentes na tabela de fato devem estar em tabelas à parte, ou seja, em tabelas como se fossem uma dimensão, o que requer, portanto, uma readequação do DW.

A entrada para o *OLAP2DataCube* é um DW em esquema estrela carregado no SGBD MySQL e a saída do mesmo são as triplas RDF já inseridas no Virtuoso. Para uso nos DW do projeto Web-PIDE é necessário a migração do SGBD do Postgres para o MySQL e alteração das medidas da tabela de fato para uma dimensão. Esse processo para triplificação pode ser visto na Figura 4.2.

4.3 Babel e Virtuoso

O último processo analisado foi o da ferramenta Babel. Essa ferramenta foi desenvolvida em um projeto da PUC e tem como objetivo um framework de publicação RDF que utiliza diversas fontes de dados e cuja abordagem é baseada em *template*.

Para utilizar a ferramenta não é necessário nenhum ajuste na base do DW. Além disso, a ferramenta faz uso do *Java Database Connectivity*(JDBC). Portanto, também suporta todos os SGBD que tem versão de JDBC, que é uma API Java que encapsula comandos SQL para um SGBD específico.

O processo utilizando o Babel tem quatro atividades bem definidas que podem ser visualizadas na Figura 4.3.

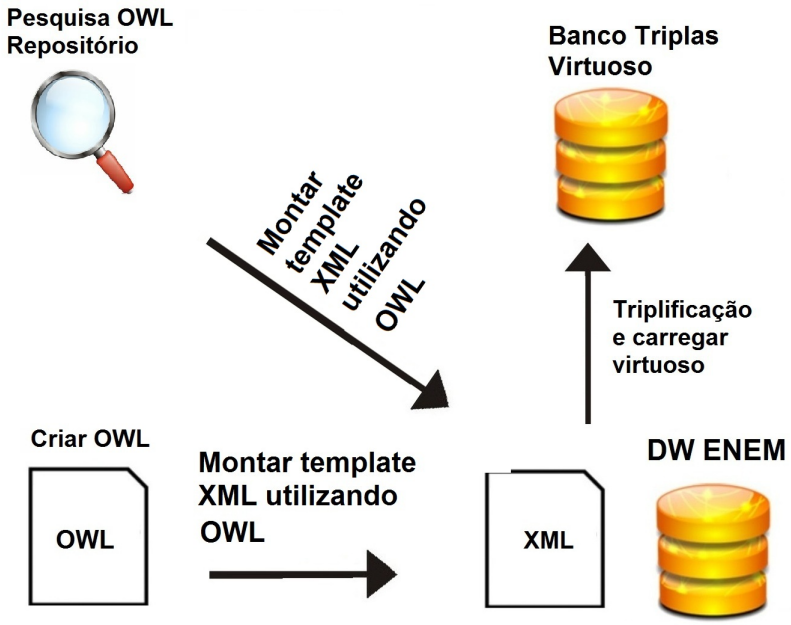


Figura 4.3: Processo utilizando Babel e Virtuoso no DW ENEM.

A primeira atividade é a de pesquisa de representações semânticas. Essa atividade é de muita importância, pois o reúso de representações semânticas facilita a interligação entre bases de triplas RDF e também dá qualidade à tripla RDF, uma vez que uma representação semântica muito utilizada provavelmente já foi aperfeiçoada. A próxima atividade é a de construção das representações semânticas para serem utilizadas nas triplas RDF. A terceira atividade é a construção do *template* XML, que requer, como entrada, as representações semânticas pesquisadas e as construídas. Por fim, na última atividade, é somente necessário utilizar a interface do Babel.

4.4 Considerações Finais

Neste capítulo foi descrito o funcionamento de três processos que podem triplificar dados de DW. Visto que o reúso do processo é grande e o modelo de dados do *Stdtrip, Triplify* e do Virtuoso necessita da normalização, este processo foi descartado. Os outros dois processos são compreendidos, mais facilmente, após execução dos mesmos utilizando os dados do DW ENEM no próximo capítulo.

Capítulo 5

Implementação do DW-ENEM e do LOD-ENEM

Neste capítulo será apresentado como foi implementado o DW-ENEM, evidenciando os problemas que foram encontrados nas bases. Além disso, será demonstrado como foi executada a triplificação dos dados do DW-ENEM.

Trabalhos anteriores relacionados ao projeto Web-PIDE já resolveram com eficiência o problema de manter uma base unificada e homogênea a partir do DW para cada censo existente. A inserção de novos microdados pertencentes a cada DW desenvolvido é resolvido com as tecnologias citadas. No entanto, agora se tem essa grande massa de dados armazenados em DWs variados, e é preciso prover formas de acesso a esses dados garantindo que todos os interessados possam ter acesso à informação. Por essa razão, este trabalho tem como objetivo construir um LOD para uso posterior em aplicações de Web Semântica.

No trabalho de [Mota, 2011], foi proposta uma estratégia de publicação para o CEB do ano de 1995 em um LOD. Conforme analisado no trabalho citado e visto na Figura 5.1, a base do CEB foi sendo alterada durante os anos, e consequentemente, para se publicar todos os dados categorizados em anos, seria preciso refazer os passos de publicação para cada ano, o que acabaria por ser muito dispendioso e consumiria muito tempo. No modelo de publicação a partir de um DW, ao se inserir novos dados, a estrutura do DW se mantém. Logo, não é preciso refazer todos os passos de publicação, somente alimentar o banco que contém os dados de publicação acrescentando os novos dados.

Pela nova abordagem proposta neste trabalho, seria necessário realizar os passos de publicação para cada DW ao invés de realizá-los uma vez para cada base, como pode ser visto na Figura 5.2. Além disso, no caso do ENEM, em que existem atualmente onze bases de microdados (1998-2008), seria possível economizar muito tempo e trabalho, pois

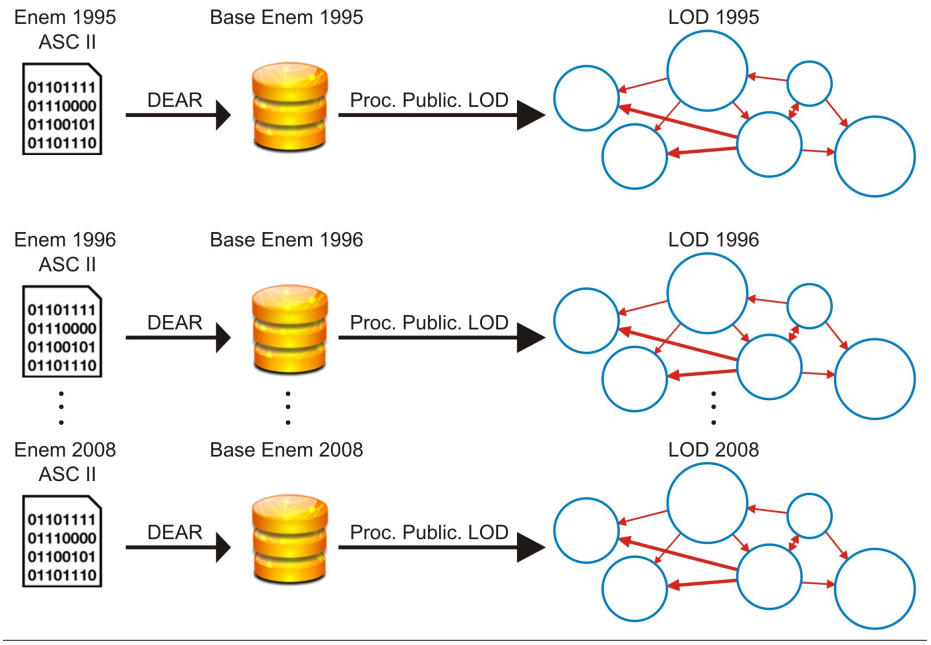


Figura 5.1: Abordagem de publicação LOD proposta por [Mota, 2011]

evitar-se-ia fazer doze processos independentes de publicação.

Outro ponto que deve ser abordado é que com o uso da tecnologia LOD e com a reutilização das representações semânticas, tornou-se possível a integração entre as várias bases provenientes do INEP. Isso seria feito ao se criar o primeiro LOD utilizando o DW do ENEM e, após esse passo, ir incrementando o modelo criado conforme os processos fossem sendo refeitos em cima dos outros DWs e reutilizando assim as representações semânticas. O resultado de todo esse longo processo seria um LOD integrado, em que todos os interessados poderiam interagir de forma fácil e eficaz uma vez que existirá um modelo de acesso a todas as informações em um único repositório, o que ainda não foi obtido pelo INEP. A integração entre os repositórios do LOD não é objetivo deste trabalho, mas poderia agregar valor à Plataforma Web-PIDE e consequentemente beneficiar o INEP.

A seguir, será definido o processo de carga dos microdados do ENEM e também a criação de um DW. Em seguida, a arquitetura e os passos para a implementação do LOD serão apresentados.

5.1 Carga de Dados e Construção do DW

Inicialmente, para que se possa começar a popular o banco de dados, é preciso que seja decido com qual base de dados dever-se-á trabalhar. Devido a sua importância e dimensão, escolheu-se a base do ENEM, pois ela contém dados desde 1998 até 2008, além

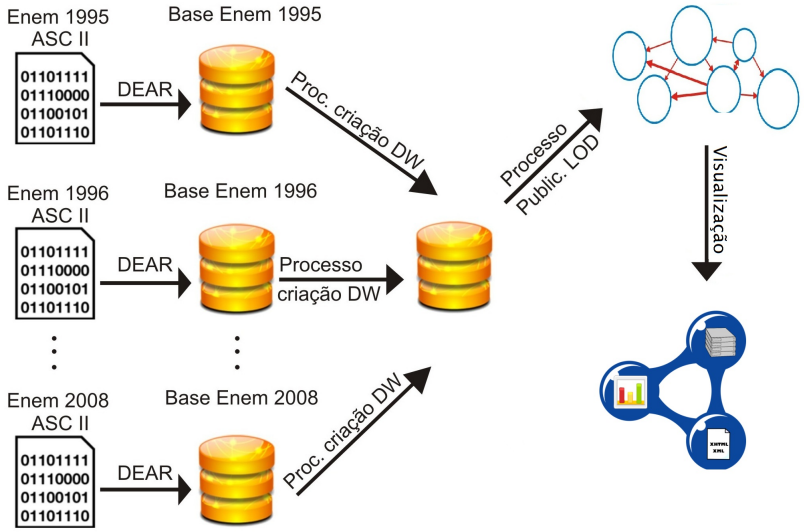


Figura 5.2: Abordagem de publicação LOD proposta

de conter em torno de 22 milhões de tuplas, o que irá gerar uma grande massa de dados a ser explorada como objeto de pesquisa. Para a carga dos microdados do ENEM, foi utilizada a ferramenta DEAR implementada em trabalho anterior.

Na fase de implementação do DW serão utilizadas as ferramentas do projeto *open source* Pentaho, que auxilia a extração, a transformação e o carregamento (ETL), os relatórios, o processamento analítico online (OLAP) e a mineração de dados. O projeto consiste nas seguintes ferramentas: *Pentaho Data Integration* também conhecido como *Kettle*, *Pentaho Analysis* também conhecido como *Mondrian*, *Pentaho Reporting*, *Pentaho Data Mining*, *Pentaho BI Suite*. Em especial, o interesse foi na ferramenta *Kettle*, na qual é executada a ETL, e na ferramenta *Mondrian*, por meio da qual se pode fazer o OLAP.

Após o término da construção do DW e para validação do mesmo, o DW será instanciado no aplicativo *Data Webhouse* presente no portal Web-PIDE. Uma vez instanciado, será possível fazer consultas e visualizar os dados. Abaixo segue os passos que foram seguidos para construção do DW.

5.1.1 Fase 1: A Ferramenta DEAR

Como a proposta deste projeto é a manipulação dos microdados do ENEM desde 1998 até 2008, o primeiro passo foi carregar os dados no SGBD Postgres por meio da utilização da ferramenta DEAR. Essa ferramenta foi desenvolvida por integrantes do projeto Web-PIDE já com vistas a carregar os microdados dos arquivos disponibilizados pelo INEP, tais como SAEB, ENEM, SEB, Provinha Brasil, entre outros.

A escolha da ferramenta DEAR e do SGDB Postgres se deu devido ao fato de que ambos já são utilizados com sucesso em outras etapas do projeto Web-PIDE.

Os microdados do ENEM utilizados neste projeto foram obtidos no site do INEP, que disponibiliza para todos as informações pertinentes a todas as avaliações aplicadas pelo MEC.

Uma vez instalado o SGDB Postgres, utilizou-se a ferramenta DEAR para importar todos os microdados necessários para a realização do projeto, ano a ano. Cada ano gerou uma tabela, totalizando onze tabelas ao fim do processo. Na Figura 5.3 é exemplificada essa etapa do processo, onde constam três janelas, uma para cada um dos passos necessários: o primeiro mostra como é feita a escolha da base de microdados a ser importada; o segundo passo demonstra a importação do dicionário de dados, em que é possível a intervenção do usuário de forma a resolver possíveis problemas; e o terceiro passo resume-se na escolha do arquivo de microdados de texto e sua execução, inserindo-o no SGBD Postgres.

Para que fosse possível utilizar a ferramenta DEAR, foi necessário criar um banco de dados auxiliar que contivesse uma tabela com o nome microdados_inep. Além disso, foi preciso inserir registros nessa tabela para cada ano em que se importou microdados do site do INEP. A própria ferramenta DEAR dispõe de um arquivo exemplo sobre como montar essa tabela, bastando que se faça a troca de valores dos registros inseridos (relativos aos microdados em uso). Esse passo é imprescindível, pois a ferramenta utiliza essa tabela auxiliar como facilitador no momento de escolher qual base de microdados será importada.

5.1.2 Fase 2: Análise dos itens gerados em cada tabela

Assim que todas as tabelas foram carregadas, tornou-se necessário fazer um estudo de cada tabela para analisar todas as informações contidas em cada ano em que o ENEM foi realizado. Isso significa que cada coluna relativa à avaliação socioeconômica, a redação e a prova objetiva teve que ser analisada para que fosse possível homogeneizar as informações. Visto que as tabelas de cada ano diferiam enormemente entre si, optou-se por selecionar as perguntas e itens que estivessem contidos em todas as tabelas, desconsiderando-se as colunas que não estivessem presentes em todos os anos. São exemplos de tais itens as perguntas sobre sexo, raça, estado civil, tamanho da família, frequência em curso de língua, de informática, notas das competências em redação, entre muitos outros, totalizando quarenta e oito colunas. Tudo isso permitiu o estabelecimento de uma linha de tempo. Mantendo-se os dados que estão sempre presentes, tornou-se possível aos interessados o estudo da evolução dos estudantes e de seu estilo de vida, além disso, o desenvolvimento de novas estratégias na educação.

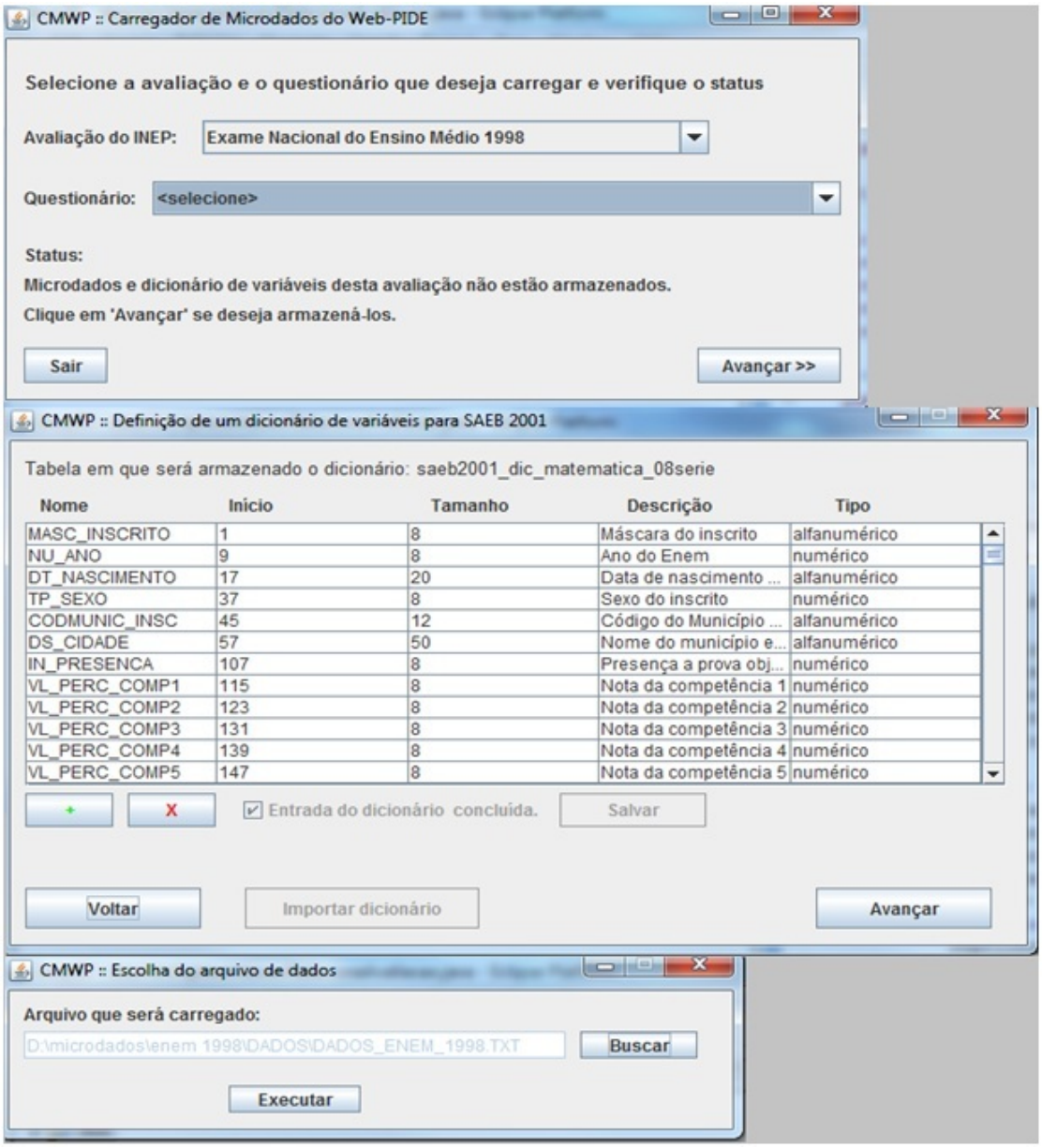


Figura 5.3: Três passos necessários para importar os microdados a serem trabalhados no DW.

Em anexo, no apêndice B, pode-se visualizar uma tabela exemplificando o resultado dessa etapa, com todas as colunas selecionadas: nome da coluna selecionada para inserir no DW, uma descrição da coluna, ou seja, o que ela significa, a dimensão que esse dado se tornou no DW, os dados que estão contidos nas dimensões e por fim a descrição de tais dados. Foram catalogados 48 itens. Na tabela 5.1 são demonstrados os campos utilizados na construção do DW-ENEM.

Tabela 5.1: Tabela demonstrativa dos campos utilizados na construção do DW-ENEM

Nome	**Descrição**	**Dados**	**Descrição Dados**
1-nu_ano	Ano do Enem	1998,1999...2008	-
4-cod_municipio	Código do município do participante	Vários códigos correspondente a região tais como: região, uf, mesoregião, microregião, município, dv	-
7-nu_nota _objetiva	Nota da prova objetiva	Valor Ex.: 10,00, 55,00	-
14-nu_nota_global _redacao	Nota da prova de redação	Valor Ex.: 10,00, 55,00	-
21-q1	Sexo	Múltipla escolha: "1-M, 2-F"	-

Além da análise coluna a coluna, foi preciso analisar os dados contidos nos itens, pois a quantidade de opções de resposta para cada item do questionário socioeconômico variou de acordo com o ano do ENEM em análise. Por exemplo, a questão tipo de prova continham os seguintes valores:

- **1998:** A - Amarela; B - Branca; G - Grafite; Z - Azul.
- **1999 - 2004:** A - Amarela; B - Branca; R - Rosa; V - Verde.
- **2005 - 2007:** 1 - Amarela; 2 - Azul; 3 - Branca; 4 - Rosa.
- **2008:** Amarela; Azul; Branca; Rosa.

Para fazer o ajuste necessário, houve a união dos diferentes valores de dados em uma única estrutura de valores, ou seja, os atributos que significavam letras ou número iguais foram ajustados para somente um caractere. Por exemplo, os caracteres "A" em 1998 e "1" em 2005 significavam a cor de prova amarela, contudo foram ajustados para o caractere "3". Essas adequações foram feitas com os outros atributos conforme os valores abaixo:

- **1998 - 2008:** 1 - Grafite; 2 - Rosa; 3 - Amarela; 4 - Branca; 5 - Verde; 6 - Azul

5.1.3 Fase 3: Construção do DW

Uma vez completada a seleção de colunas (questões) a serem mantidas e a homogeneização dos dados presentes em tais colunas, partiu-se para a última etapa, a montagem do DW.

Na escolha do modelo de DW mais adequado, definiu-se a adoção do modelo estrela, implementado por meio da ferramenta *Kettle*. Na Figura 5.4, é ilustrada a representação do DER do DW ENEM, demonstrando como ficaram dispostos os dados do ENEM depois de finalizado o DW. Nela, pode-se visualizar como as dimensões e a tabela de fato foram montadas, além da demonstração modelo Estrela, em que a tabela de fato se relaciona às dimensões. Assim, para a montagem do DW, foram selecionadas as seguintes colunas:

- q1 (sexo)
- nu_ano
- cod_municipio
- nu_nota_objetiva
- nu_nota_global_redacao

As colunas nu_nota_objetiva e nu_nota_global_redacao se mostraram como medidas na tabela de fato e se tornaram os campos respectivamente media_objetiva e media_redacao. Além disso, nu_ano se tornou a dimensão dim_ano e nela estará contido os anos de aplicação do ENEM. O cod_municipio se tornou a dimensão dim_regiao que conterá o cod_municipio decomposto nos seguintes campos: regiao/descr_regiao(onde o primeiro campo contém um código e o segundo uma descrição que representa este código, e assim sucessivamente para os próximos campos), uf/descr_uf, mesoregiao/descr_mesoregiao, microregiao/descr_microregiao, municipio/descr_municipio e dv(dígito verificador). Por fim, a coluna q1 tornou-se a dimensão dim_sexo e esta contém a designação quanto ao gênero do candidato. Além disso, as dimensões contêm um coluna identificadora (id) que se tornou *foreign key* na tabela de fato. Ao final do processo a tabela de fato continha em torno de vinte e dois milhões de tuplas.

Esse DW é um modelo inicial para suprir respostas a questões específicas, mas pode ser acrescido de novas dimensões e medidas utilizando-se das colunas que já foram previamente selecionadas, bastando determinar a que novos tipos de questões pretende-se responder. Além desse acréscimo, também podem ser construídos novos DWs do ENEM

seguindo a mesma ideia, ou seja, selecionando os tipos de questões que ele deve responder. Pode-se inclusive reutilizar as dimensões já criadas, chegando a um modelo de constelações de fatos.

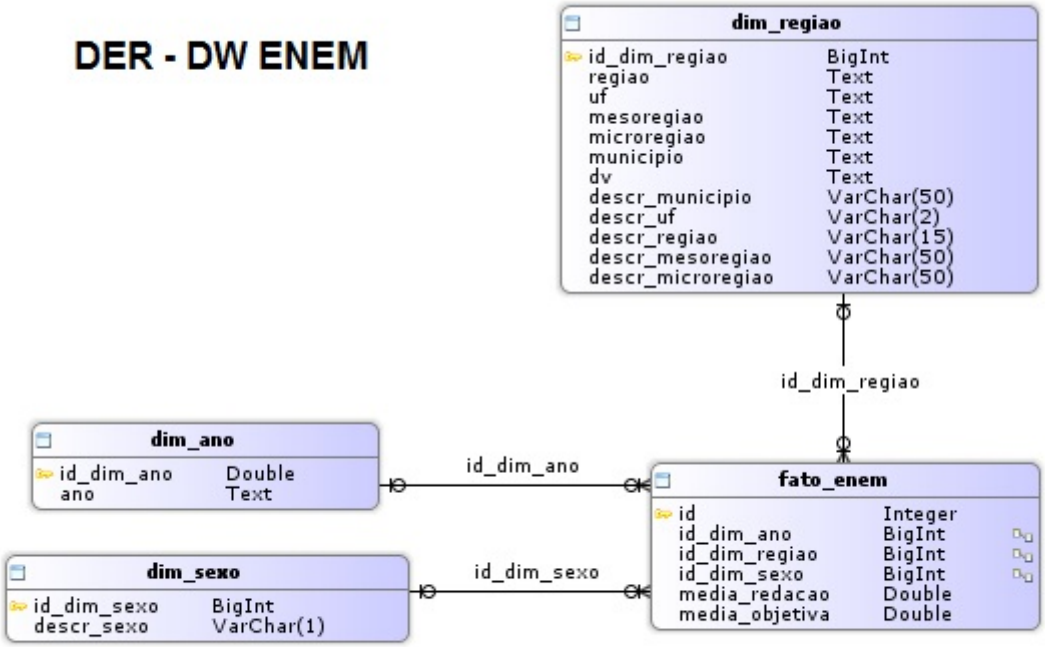


Figura 5.4: DER do DW Enem, que representa o resultado alcançado.

5.1.4 Fase 4: Validação do DW

A etapa de validação consiste em verificar se o DW é funcional, apresenta erros e inconsistências e permite análise para verificar se há possíveis melhorias. Nessa etapa, também foi utilizado o projeto Pentaho, com a ferramenta *Pentaho Analysis (Mondrian)*, a qual realiza consultas no DW e permite ao usuário sua visualização, além de se basear em uma interface amigável.

O aplicativo *Data Web House* é uma instância do *Mondrian* que está localizada no portal WebPide e foi customizada para o projeto com intuito de conter os DWs criados e disponibilizá-lo para a avaliação/acompanhamento dos dados por parte dos integrantes do projeto, dos avaliadores e dos educadores.

Ao se fazer a validação do DW proposto, foi reutilizado o aplicativo mencionado acima. Para isso, foi criado um arquivo de configuração xml que descreve o DW ENEM de forma que a ferramenta pudesse disponibilizar as consultas. Além disso, foi necessário inserir o

DW ENEM no SGBD Postgres do servidor do Web PIDE.

Com isso, tornou-se possível a qualquer indivíduo interessado no projeto navegar no *Data Web House* (uma vez que é redirecionado a partir do portal WebPIDE) e consultar os dados disponibilizados. A ferramenta *Mondrian* permite gerar arquivos em .xls, gráficos, impressão de arquivos .pdf, filtrar determinadas colunas durante a pesquisa, ou seja, possui todo um arcabouço para auxiliar o usuário.

Além disso, foram formuladas duas questões, que demonstram que tipo de resultados esse DW pretende responder. São elas:

- Consulta 1: Quais são as médias da prova objetiva e da redação, agrupadas por ano e região. Com essa pergunta é possível verificar se no ano de 1999 os candidatos da região Sul obtiveram nota maior do que os candidatos das regiões Norte, Nordeste, Centro-oeste e Sudeste.
- Consulta 2: Qual a quantidade de alunos que tiraram nota zero, agrupados por ano e região. Já nesta pergunta é possível responder, por exemplo, se a quantidade de alunos que tiraram nota zero foi crescendo ou diminuindo e ver essa variação por região.

A partir dessas questões, foram implementadas quatro consultas SQL. As duas primeiras consultam o DW e estão em anexo, no apêndice C.1 e C.2. As outras duas consultam as tabelas carregadas com os microdados e estão disponíveis no apêndice C.3 e C.4.

O diferencial das consultas feitas no DW fica evidente, pois as mesmas ficaram com tamanho menor do que 15 linhas e não são complexas. Além disso, o tempo de execução foi menor do que 30 segundos. Já as consultas feitas nas tabelas com os microdados ficaram com tamanho aproximado de 164 linhas, ficando evidente nessas consultas o maior grau de complexidade em relação às feitas no DW, pois elas utilizam vários SELECT e UNION para unir as tabela, além de utilizarem CASE e SUBSTRING para tratar os diferentes tipos de dados de cada tabela. Por fim, o tempo de execução das consultas realizadas nas tabelas dos microdados ficou em torno de 100 segundos.

Visto que tanto a complexidade de elaboração quanto o desempenho das consultas sobre o DW foram melhores, fica comprovada a eficiência do DW ENEM, atestando a necessidade de construção do mesmo.

5.2 *Linked Open Data*

Antes de iniciar essa etapa, é necessário que se termine a implementação do DW, pois este será o provedor de dados. Depois da implementação do DW, vem a adoção da tecno-

logia LOD, e para tal, a primeira atividade a ser desempenhada é a escolha do processo de triplificação. As abordagens de triplificação já foram demonstradas anteriormente, a partir de agora será feito uso destas abordagens para ser possível melhor análise de cada uma.

Nesse momento, pode ser necessário um projeto de representações semânticas, em que será preciso aplicar o reúso de representações semânticas, pesquisando em outros repositórios como, por exemplo, o [data.gov.uk, 2011]. Inicialmente, esse reúso visa garantir a integração a outros repositórios. Além disso, para fazer a reutilização dos termos encontrados no DW, deve-se traduzir todos esses termos para o inglês, já que a maioria das bases de LOD no mundo estão nessa língua. Caso tais termos não sejam encontrados em nenhum repositório, será necessária a formulação de novos termos. Todas as representações semânticas encontradas ou formuladas serão representadas por RDF.

É necessário que as triplas RDF geradas durante a triplificação sejam carregado em um banco de dados para que o acesso seja facilitado. Assim, as triplas são importadas para o Virtuoso [Virtuoso, 2011], que é um motor de base de dados que armazena as triplas RDF e também outros formatos como o XML. Além de armazenar os dados, ele também pode servir como um servidor de aplicação web. Como último passo, pode-se então validar as triplas geradas por meio da extração de dados através de consultas utilizando a característica de servidor do Virtuoso. Deve-se ainda ressaltar que a arquitetura RDF e o Virtuoso são recomendados pelo W3C, que é a união das organizações interessadas em definir novos conceitos, protocolos e padrões de arquiteturas Web, fortalecendo o uso dessas tecnologias.

Após a triplificação dos dados, as triplas RDF ficam disponíveis para consulta via SPARQL no Virtuoso. Essa triplificação é o passo inicial para utilização de um *LOD*. A partir dela, já é possível fazer consultas semânticas, mas para aprimorar a qualidade do LOD, poderiam ser feitas várias melhorias, tais como: estudo e implementação de novos LODs a partir de outros DWs do Web-PIDE, interligação dos LODs implementados e pesquisa de formas de visualização para esconder do usuário as consultas SPARQL e facilitar a pesquisa semântica. As melhorias citadas não são objetivos deste trabalho. Quanto mais dados e ligações vão sendo inseridos no LOD, maior o valor da base triplificada, sendo que no fim de todo o processo, seria possível pesquisar dados na Web, assim como é feito no Google, com a vatangem de se poder pesquisar dados utilizando semântica.

5.2.1 Criação de representações semânticas para uso no RDF

Para se ter uma idéia de como começar a montar as anotações ontológicas, será utilizado como base um campo da dimensão região do DW ENEM como estudo de caso. Nele

existem vários campos que são candidatos a serem utilizados na nova base triplificada. Essa escolha de campos é subjetiva, e deve-se levar em conta campos relevantes para a futura base triplificada. Dentre os campos dessa dimensão, pode-se citar como fortes candidatos para utilização os seguintes campos: descr_uf, descr_municipio e descr_regiao.

No próximo passo, utilizar-se-á o campo descr_uf. Como a dimensão conta com vários campos passíveis de serem utilizados, pode-se criar uma classe chamada localidade e um atributo chamado uf. Essas definições são expressas pela seguinte representação:

XML 5.1: Exemplo representações semânticas

```
<owl:Class rdf:about=http://meuEndereco/ontologia/localidade">
 <rdfs:label>Localidade</rdfs:label>
 <rdfs:comment>Contém todas as localidades do Enem</rdfs:comment>
 <rdfs:isDefinedBy>http://meuEndereco/ontologia</rdfs:isDefinedBy>
</owl:Class>

<owl:ObjectProperty rdf:about="http://meuEndereco/ontologia/localidade#uf ">
 <rdfs:label>Unidade Federativa</rdfs:label>
 <rdfs:comment>Cada unidade federativa da união</rdfs:comment>
 <rdfs:domain>http://meuEndereco/ontologia/localidade</rdfs:domain>
 <rdfs:isDefinedBy>http://meuEndereco/ontologia</rdfs:isDefinedBy>
</owl:Class>
```

Cada atributo do XML tem uma função bem definida pelo RDFS. O mas importante a se saber nessa representação é que existe uma classe e essa classe tem um atributo. Poder-se-ia inserir novos atributos nessa classe e ainda serem mapeados os atributos em classes diferentes.

A partir do arquivo criado, deve-se montar o RDF, que conterá um cabeçalho dizendo quais as representações serão utilizadas e as instâncias de cada classe definida no arquivo das anotações semânticas. Para o exemplo anterior, foi definido um RDF representando o identificador da localidade e o uf pertencente a essa localidade pré-definida. Essas definições estão descritas a seguir:

XML 5.2: Exemplo RDF

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
  xmlns:owl="http://www.w3.org/2002/07/owl#"
  xmlns:novaOntologia="http://meuEndereco/ontologia#">
 <novaOntologia:localidade rdf:about="http://meuEndereco/enem/localidade/1">

  <novaOntologia:uf>MS</novaOntologia:uf>
 </novaOntologia:localidade>
</rdf:RDF>
```

Agora, para cada uf contido no DW ENEM, deve-se criar uma instância do RDF, como

descrito acima, e depois inseri-los no banco de triplas. Ao fim do procedimento, está feita a triplificação do atributo uf.

5.2.2 Execução Ontowiki, *OLAP2DataCube* e Virtuoso

A configuração da ferramenta Ontowiki para trabalhar juntamente com a *OLAP2DataCube* e o Virtuoso foi muito difícil de ser executada, devido ao alto grau de atividade de desenvolvimento da Ontowiki. A versão atual desta ferramenta não suportava o *plug-in OLAP2DataCube*, e em trabalho conjunto com os desenvolvedores de ambas, chegou-se a uma solução funcional, sendo assim possível utilizá-las. Além disso, são pré-requisitos para a *OLAP2DataCube* a instalação no sistema operacional Linux e ainda o uso de uma base de dados MySQL como fonte de dados para triplificação.

O primeiro passo para tornar possível a utilização da *OLAP2DataCube* foi a migração do DW ENEM do Postgres para o MySQL. Depois, foi preciso adequar a estrutura das medidas da tabela de fato, trocando-as para tabelas com relacionamento com a tabela de fato. Com o tratamento da base de dados terminado, já se pode aplicar a triplificação feita pelo *OLAP2DataCube*, e assim inicia-se a utilização por meio da interface da Ontowiki. Nessa interface, é necessário que se especifique a configuração da base de conhecimento que é o grafo que representa as triplas no banco Virtuoso. Após essa instanciação da base de conhecimento, deve-se fazer a configuração da conexão ao banco relacional para acessar a estrutura do DW. Esses dois primeiros passos podem ser observados nas Figuras 5.5 e 5.6.

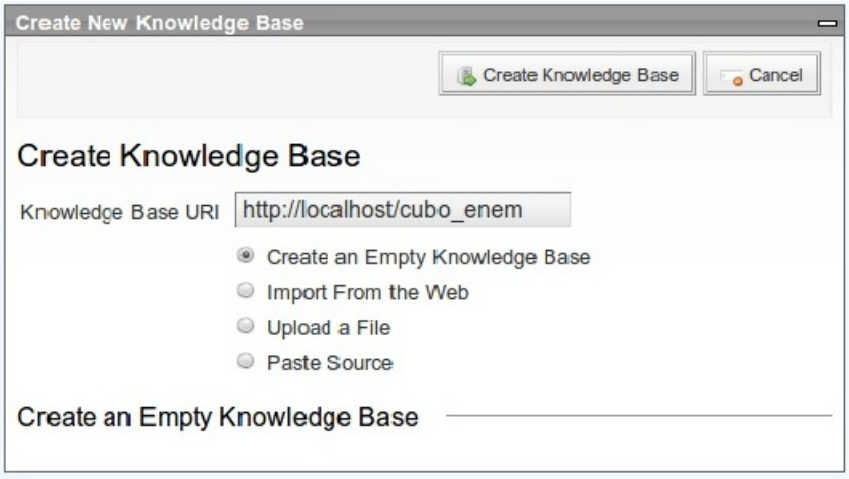


Figura 5.5: Demonstração de instanciação de uma base de conhecimento pela Ontowiki

A partir desse momento, entra em ação o facilitador da ferramenta, que extrai os metadados do banco e monta a próxima tela da ferramenta com as tabelas de fato e de dimensões encontradas. Na Figura 5.7, a *OLAP2DataCube* sugere a estrutura do DW

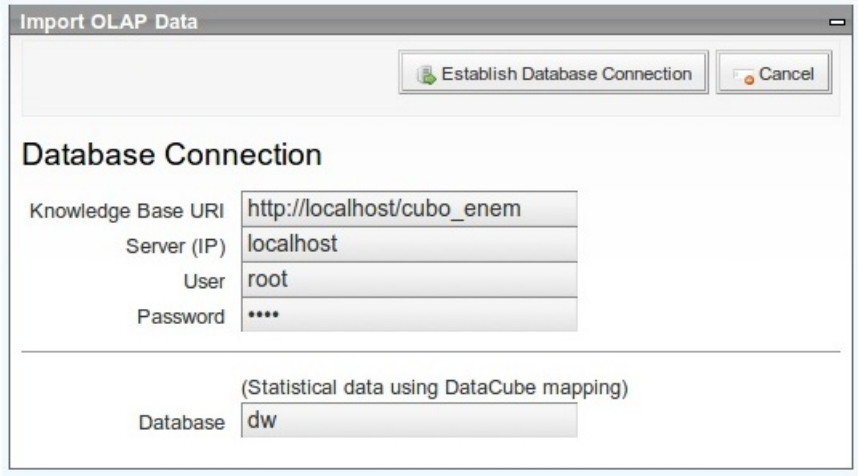


Figura 5.6: Configuração da conexão com o Mysql para acessar o DW ENEM pela *OLAP2DataCube*.

ENEM.

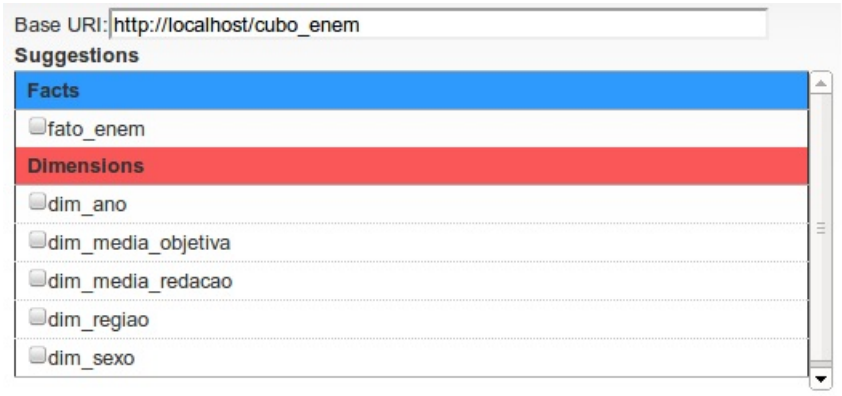


Figura 5.7: Sugestão de esquema feito pela *OLAP2DataCube*.

Ao selecionar a tabela de fato, surgem novos campos para configuração do mapeamento das dimensões e medidas. Ao mapear uma dimensão, deve-se configurar o nome que esta receberá na triplificação, a propriedade *concept* que define a qual conceito essa propriedade pertence, a propriedade *subPropertyOf* que define a qual propriedade superior essa propriedade pertence e, por último, a seleção dos campos que serão triplificados. Caso seja mapeada uma medida, deve-se selecionar somente os campos que serão triplificados. Na Figura 5.8, pode-se observar a configuração da dimensão dim_ano e a medida dim_media_redacao.

Com todas as dimensões e medidas escolhidas devidamente configuradas, é possível iniciar o processo de extração das triplas, que serão inseridas automaticamente no Virtuoso. Após o término da triplificação, já é possível utilizar consultas SPARQL para acessar a base de dados.

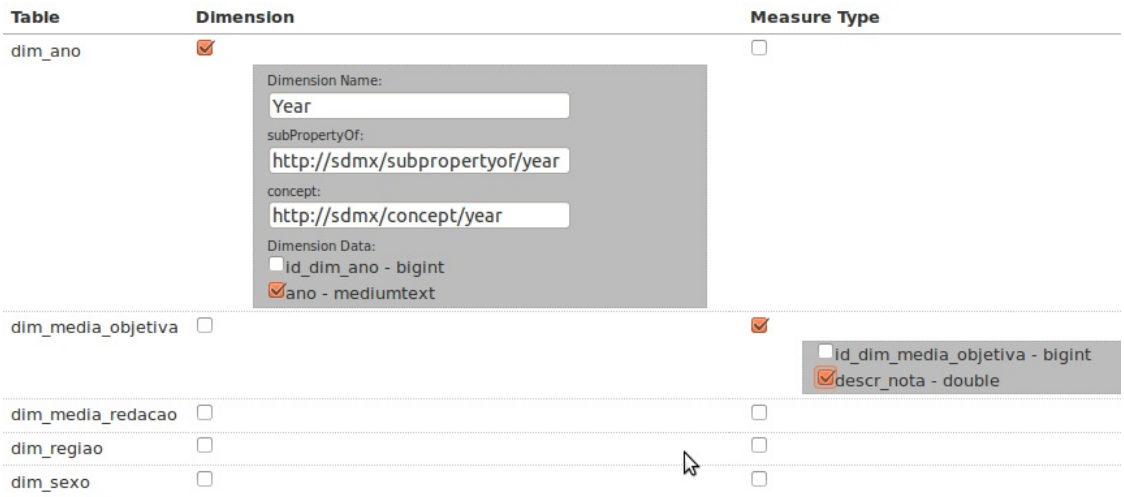


Figura 5.8: Configuração das dimensões e medidas a partir da *OLAP2DataCube*.

Um conceito interessante utilizado na *OLAP2DataCube* é que ela se baseia no RDF *Data Cube Vocabulary* [J., 2012], que é uma ontologia recomendada pelo W3C e tem foco apenas na publicação de dados multi-dimensionais na web. A adoção dessa ontologia tem como vantagem o fato de que já existem trabalhos de visualização que se baseiam na estrutura dessa ontologia, sendo possível a utilização de ferramentas como o CubeViz, que é uma ferramenta para facilitar a navegação e a exploração de cubos RDF que utilizam a *Data Cube Vocabulary*.

Para demonstrar o resultado da triplificação foram feitas duas consultas no módulo SPARQL do Virtuoso. Na primeira consulta, são recuperadas todas as dimensões presente no grafo de triplas do ENEM. Essa consulta pode ser visualizada na Figura 5.9.

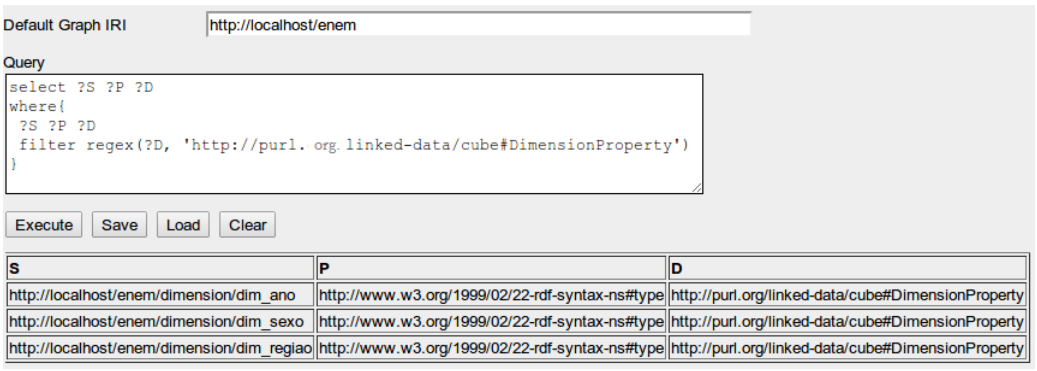


Figura 5.9: Consulta SPARQL demonstrando as dimensões no grafo do ENEM no Virtuoso.

A próxima consulta demonstra como fica a triplificação da dimensão ano. A Figura 5.10 demonstra o resultado obtido a partir da consulta.

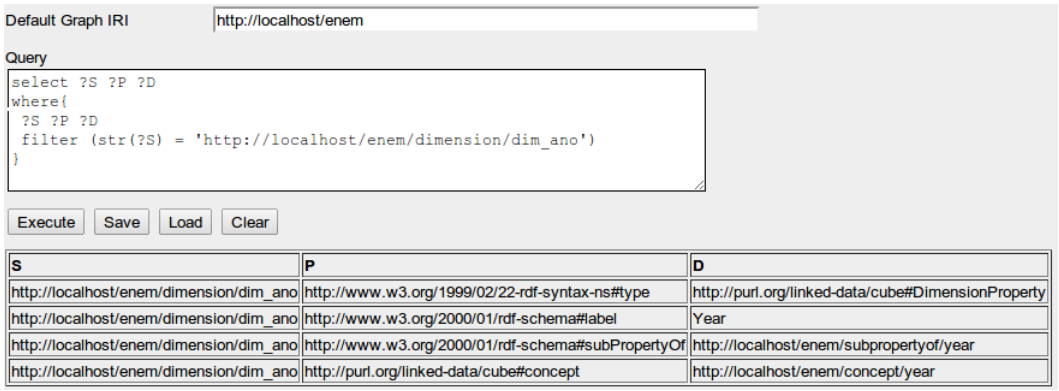


Figura 5.10: Consulta SPARQL demonstrando a dimensão ano.

5.2.3 Execução Babel e Virtuoso

Durante o processo de triplificação utilizando o Babel, foi considerado o uso da ontologia *Data Cube Vocabulary*, mas devido ao fato de que a ferramenta *OLAP2DataCube* faz uso da mesma, e de a *OLAP2DataCube* ter um processo mais intuitivo e rápido, foi decidido pesquisar e criar anotações semânticas para destacar a dificuldade de se trabalhar com as mesmas.

A ferramenta não fornece nenhuma forma de facilitador para a atividade de pesquisa de representações semânticas, o que faz com que seja muito difícil de ser executada. Existem muitas bases de triplas RDF, e foi feita uma pesquisa manual nessas bases procurando algo que se encaixasse ao dados do DW ENEM. Foi encontrada uma representação semântica para os dados de região na base GeoName. Nela foram encontradas as definições para: *Place*, *City*, *State*, *Region* e *Country*. Não foram encontradas outras representações semânticas que combinassem com a base do DW ENEM.

Após a atividade de pesquisa, foi necessário a definição de novas representações semânticas que pode ser visualizada no anexo, no apêndice D. Essa atividade é de elevado grau de complexidade. Segundo [Hernandes, 2010], "(...) o processo de construção de ontologias é considerado uma das principais dificuldades relacionadas ao tema".

A próxima fase é bem simples de ser executada, deve-se contruir o arquivo de template XML que é dividido em duas partes. A primeira é a de conexão com o banco de dados e extração dos dados mediante consultas SQL que, no caso do DW ENEM, para cada dimensão ou medida foi feita uma consulta SQL extraindo os dados que serão necessários. Na segunda parte do template, fez-se o mapeamento campo/ontologia no RDF. As configurações do template XML utilizado na triplificação do DW ENEM podem ser visualizadas no anexo, no apêndice E.

Por fim, deve-se selecionar o arquivo com o *template* que será triplificado e configurando

a forma de saída, que no caso é para o banco de triplas Virtuoso e seus dados de conexão e executar a triplificação. Os dois últimos passos estão demonstrados na Figura 5.11.

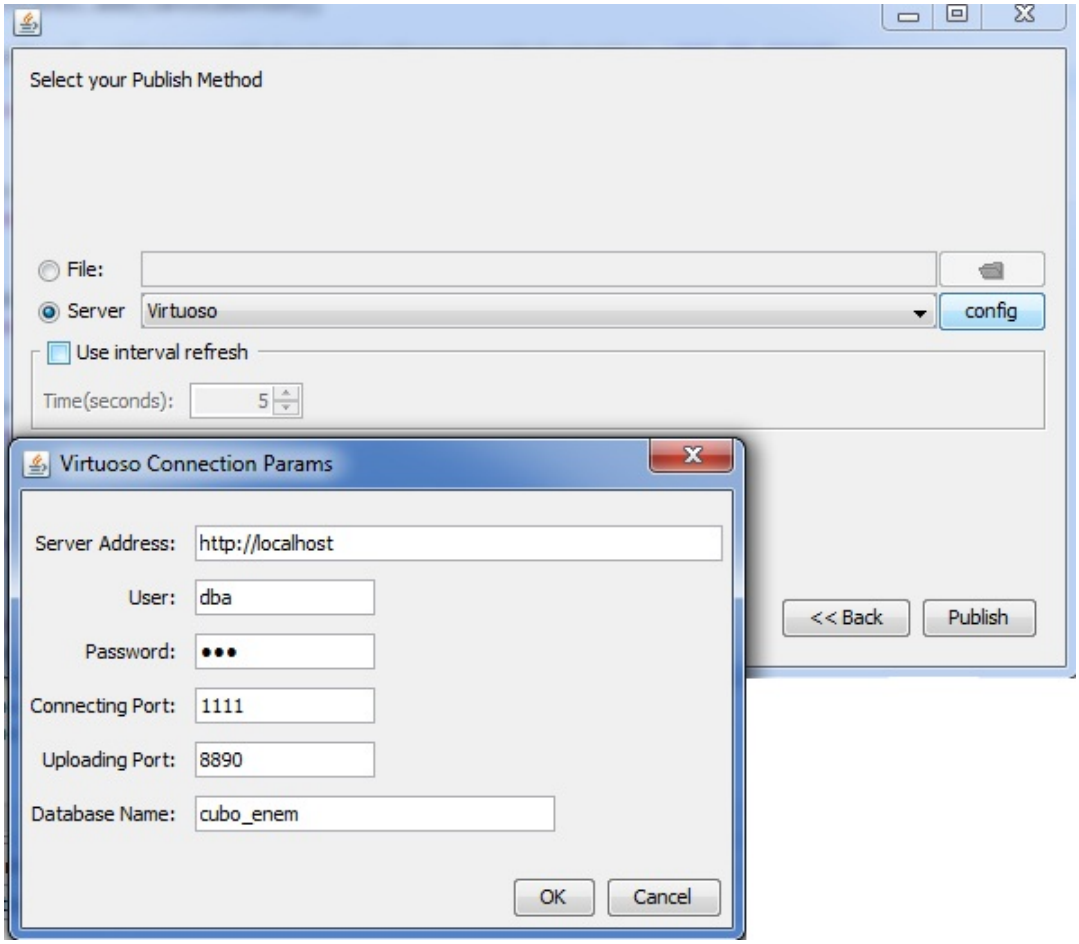


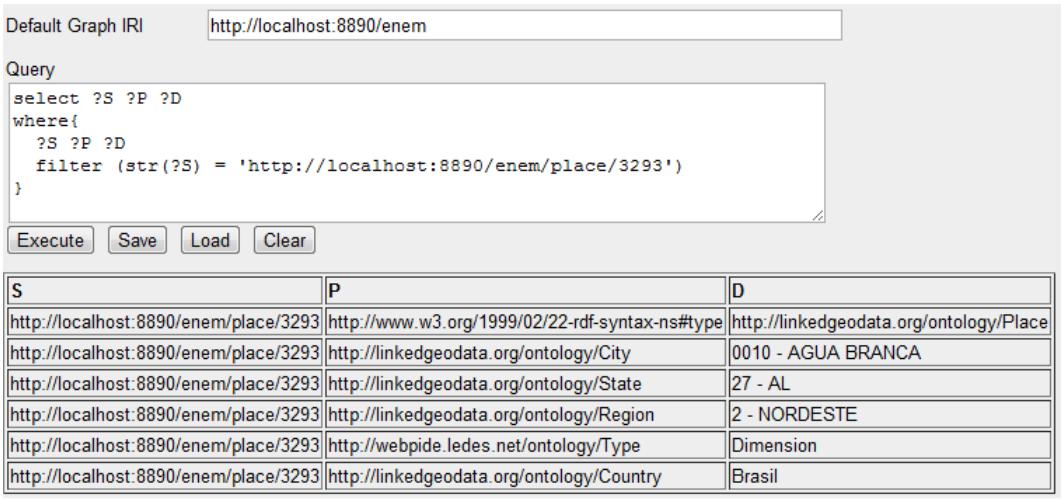
Figura 5.11: Interface do Babel para configuração do Virtuoso e execução da triplificação.

Nesse último passo, o carregamento das triplas no banco Virtuoso estava sendo feito de forma manual, acrescentando mais uma atividade ao processo. Porém, em parceria com o desenvolvedor do Babel, foi acrescentado um módulo que carrega os dados automaticamente, bastando a configuração correta dos parâmetros do Virtuoso.

Ao final do processo, a ferramenta Babel concluiu com êxito o processo de triplificação, ficando disponível para consulta no Virtuoso. A Babel se mostrou uma ferramenta eficaz na triplificação, mas não foi inteiramente satisfatória no quesito construção de representações semânticas. Ela poderia prover de forma automática o mapeamento para um DW utilizando a *Data Cube Vocabulary* assim como a *OLAP2DataCube* o faz, aproveitando assim a estrutura de um DW que já é pré-definida.

A partir do RDF gerado pela Babel foram feitas duas consultas SPARQL para demonstrar o resultado no Virtuoso. A primeira consulta demonstra o uso das representações

semânticas pesquisadas. Essa consulta é demonstrada na Figura 5.12. A partir da Figura, nota-se que pesquisando o *place* 3293 obtém-se como resposta as definições do que é esse *place*.



S	P	D
http://localhost:8890/enem/place/3293	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://linkedgeodata.org/ontology/Place
http://localhost:8890/enem/place/3293	http://linkedgeodata.org/ontology/City	0010 - AGUA BRANCA
http://localhost:8890/enem/place/3293	http://linkedgeodata.org/ontology/State	27 - AL
http://localhost:8890/enem/place/3293	http://linkedgeodata.org/ontology/Region	2 - NORDESTE
http://localhost:8890/enem/place/3293	http://webpide.ledes.net/ontology/Type	Dimension
http://localhost:8890/enem/place/3293	http://linkedgeodata.org/ontology/Country	Brasil

Figura 5.12: Consulta SPARQL demonstrando a consulta das representações semânticas encontradas no repositório GEONAME.

A outra consulta feita demonstra todas as representações semânticas utilizadas para definir as triplas RDF. A Figura 5.13 demonstra o resultado obtido a partir da consulta.

A tabela 5.2 demonstra sinteticamente os prós e os contras do processo utilizando o Babel e o *OLAP2DataCube*.

5.3 Considerações Finais

Neste capítulo foi apresentado o processo de criação do DW-ENEM, o qual servirá de apoio à comunidade na tomada de decisão sobre aspectos da educação brasileira. Além disso, o DW é utilizado como fonte de dados para uso na triplificação da base do ENEM. Também foram apresentados os processos para implementação do LOD. Tais processos foram avaliados e executados. A partir dessas avaliações conclui-se que o processo que utiliza a ferramenta *OLAP2DataCube* obteve o melhor desempenho, e por isso deve ser utilizada como padrão no projeto.

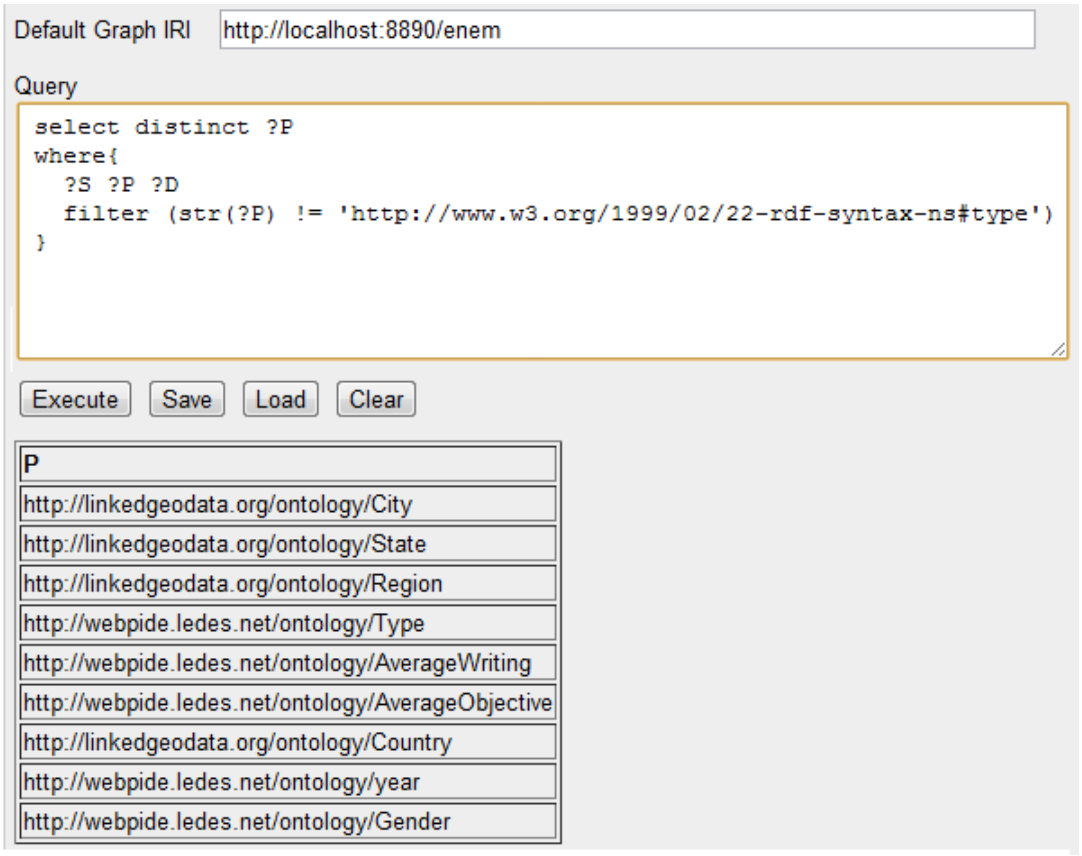


Figura 5.13: Consulta SPARQL demonstrando as representações semânticas utilizadas no processo.

Tabela 5.2: Comparação entre ferramentas Babel e *OLAP2DataCube*

	Babel	*OLAP2DataCube*
Opensource	Sim	Sim
Dependência do modelo banco dados	Para qualquer base de dados	Somente em DW em esquema Estrela
Multiplataforma	Qualquer SO que possível de instalar Java	Somente Linux
SGBD	Qualquer SGBD que tenha JDBC	Somente MySql
Necessidade de adequação da Base antes de iniciar o processo	Nenhuma	Migrar para MySql; Ajuste das medidas em dimensões;
Dependência de outras ferramentas	Não	Sim (Ontowiki)
Passos para executar a triplificação	4	3
Representações semânticas	Pesquisar e criar representações semânticas	Específico (*Data Cube Vocabulary*)
Qualidade representações semânticas	Depende da noção do usuário	Controlada pelo W3C
Interface e usabilidade	Fácil de utilizar, mas depende de um conhecimento prévio	Interface simples de entender e de utilizar
Modelo dados	Banco de dados, arquivos proprietários e planilhas	Banco de dados em esquema Estrela

Capítulo 6

Conclusão e Trabahos Futuros

Neste capítulo são apresentadas as conclusões do projeto, trabalhos futuros e dificuldades encontradas.

6.1 Resultados alcançados

Este trabalho teve dois objetivos principais: a criação e a disponibilização de um DW do ENEM e a proposta de um processo para implementação de um LOD a partir de um DW.

Existe uma grande quantidade de informações que o MEC vem acumulando por meio das avaliações educacionais que são aplicadas pelo INEP. A partir desses dados, foi decidido implementar uma nova base de DW. Para tal, foi utilizado os microdados do ENEM, e após sua realização, o primeiro objetivo deste projeto foi atingido com sucesso, visto ter sido possível a implementação eficiente do DW ENEM. Essa etapa validou toda a estratégia já proposta pelo projeto Web-PIDE, as quais são: o projeto, a construção e a publicação de DW. Além disso, foi disponibilizada mais uma base de DW para utilização pela comunidade interessada.

Com o DW implementado, partiu-se para o desenvolvimento do LOD. Para isso, foram analisados três processos: o processo *Stdtrip, Triplify* e Virtuoso, o processo Ontowiki, *OLAP2DataCube* e Virtuoso e o processo Babel e Virtuoso. O primeiro processo demonstrou ser um tanto quanto custoso, pois ele requer seis atividades até atingir o objetivo almejado, as triplas RDF carregada no Virtuoso. Além disso, ele vai contra a escolha do DW em esquema Estrela, pois utiliza em suas atividades um modelo relacional normalizado. Após análise do processo, decidiu-se descartar todo o processo.

O processo Ontowiki, *OLAP2DataCube* e Virtuoso foi o próximo a ser analisado. Esse

processo requereu uma quantidade total de três atividades para que fosse possível a implementação das triplas RDF no Virtuoso. Outro fator interessante é que o plug-in *OLAP2DataCube*, ferramenta chave para a triplificação, foi desenvolvido para ser de fácil utilização e por isso é guiado por uma interface de fácil manuseio e entendimento. Por fim, a ontologia usada para triplificação nesse processo foi a *Data Cube Vocabulary*, que é uma ontologia para modelos dimensionais referenciada pelo W3C. O uso do vocabulário mencionado é visto como uma vantagem, pois poderia vir a facilitar a integração entre outras bases LOD que utilizam a mesma ontologia e a triplificação dos dados, pois já se tem uma ontologia pré-definida e com isso é possível esconder do usuário o processo de triplificação.

Os fatores negativos no uso dessas ferramentas são a dificuldade de configuração do ambiente, que consumiu muito tempo devido à incompatibilidade entre a ferramenta Ontowiki e *OLAP2DataCube*, e a falta de uma documentação de uso da *OLAP2DataCube*. Porém, superados esses obstáculos, a ferramenta teve bom desempenho no aspecto de triplificação de um DW e foi considerada ideal para o projeto Web-PIDE.

O último processo analisado foi o Babel e Virtuoso, em que foram necessárias quatro atividades para se atingir o objetivo. Esse processo foi de fácil realização devido ao uso de um template XML, que é utilizado para acessar a base a qual será triplificada e para transformá-la em triplas RDF. A possibilidade de utilizar qualquer SGBD como entrada de dados na ferramenta é uma grande vantagem, pois não aumenta a quantidade de atividades no processo de triplificação.

Os fatores negativos são o fato de ser necessário a pesquisa e a construção de uma representação semântica, o que demanda um esforço grande, pois a pesquisa de representações semânticas para reúso é manual, e é muito complexo elaborar uma ontologia, principalmente se for levado em consideração o aspecto do reúso. O que pode ser feito para se contornar o problema é o uso da ontologia *Data Cube Vocabulary*, pois se está utilizando um DW ao se triplificar, mas ainda assim é necessário configurar todo o *template* XML. Outro fator negativo nesse processo foi o carregamento do Virtuoso, que foi feito de forma manual e tornou-se uma complexidade extra. No entanto, em parceria com o desenvolvedor do Babel, foi desenvolvido um módulo para automatizar o carregamento das triplas. O processo Babel foi considerado de bom desempenho, tendo cumprindo seu papel no ato de triplificação do DW ENEM, mas tendo um desempenho inferior ao processo feito pela ferramenta *OLAP2DataCube*.

Por fim, foram utilizados os processos Ontowiki, *OLAP2DataCube* e Virtuoso e Babel e Virtuoso para triplificação do DW ENEM, sendo possível consultas via SPARQL para acesso à base RDF. Dessa forma, o processo utilizando o Ontowiki, *OLAP2DataCube* e Virtuoso ficou estabelecido como padrão do projeto Web-PIDE quando se desejar tripli-

ficar novas DWs do projeto.

O interessante de trabalhar com o projeto LOD é que ele vem de encontro com as políticas que estão se estabelecendo no mundo de dados abertos governamentais, a implementação do LOD ENEM a partir de um DW consegue atingir seis das oitos regras definidas pelo W3C para as práticas de *Open Government Data*. Não atinge as duas que outras regras, pois a triplificação está sendo feita em DWs, que são fontes de dados que não contém todos os dados da base ENEM e nem são dados primários. O que é compreensível devido o excesso de dados que o INEP coletou durantes anos.

6.2 Dificuldades

Durante a execução da construção do DW, ficaram evidentes as inconsistências que a base de microdados continha, além da falta de padronização dos dados em relação a cada ano em que a avaliação foi aplicada, um fator que dificultou muito o processo de seleção de dados para construção do DW. Esses problemas são recorrentes, pois já foram relatados pelos outros integrantes do projeto Web-PIDE que implementaram DWs. Outro fator complicador foi a quantidade de bases, que somaram um total de onze bases, e a quantidade de atributos das bases, que no total foram mais de dois mil e quatrocentos atributos. Por fim, o DW do ENEM ficou com mais de 22 milhões de tuplas, o que é por um lado é excelente, pois o elevado número de tuplas eleva consequentemente o valor do trabalho, mas também diminui a velocidade das consultas e dificultou muito o trabalho de construção do DW, uma vez que consumiu muito hardware e tempo de trabalho.

Outro problema foi a pesquisa de ferramentas e consequentemente de processos de triplificação que fossem fáceis de serem aplicados e que não envolvessem muitos passos para triplificação devido a grande quantidade de bases que podem ser triplificadas.

Também foi considerada uma dificuldade a configuração das ferramentas e o uso das mesmas, visto que a ferramenta Babel ainda estava em desenvolvimento durante o processo de triplificação, e algumas funcionalidades não estavam finalizadas e foram sendo ajustadas durante o processo. Já a *OLAP2DataCube* teve sua versão defasada em relação ao Ontowiki, o que fez com que o uso das duas ferramentas em conjunto tivesse que ser ajustado.

Por fim, a construção das representações semânticas foi de alto grau de complexidade, pois não é fácil construir uma representação semântica que seja eficiente e reutilizável. Além disso, o processo de obtenção de conhecimento no domínio da representação semântica demanda tempo.

6.3 Trabalhos futuros

A partir do trabalho proposto, existem melhorias que podem ser aplicadas as quais tornariam ainda mais abrangentes os resultados obtidos, tais como:

- A partir do estudo comparativo entre consultas feitas neste trabalho, ficou evidente a melhora significativa que um DW proporciona, mas ainda poderia ser melhorado o desempenho do DW. Assim sendo, seria interessante fazer um estudo utilizando visões materializadas. Esse estudo serviria para construir visões que seriam armazenadas, e quando uma determinada consulta fosse executada, ela responderia ao usuário de forma mais rápida do que o DW provê.

- A ferramenta *OLAP2DataCube* escolhida para ser utilizada na triplificação dos DWs do projeto Web-PIDE poderia ter acrescentada em suas funcionalidades a capacidade de triplificação a partir de um DW contido no Postgres, pois este é o SGBD padrão do projeto, e assim diminuiria uma atividade na triplificação de DWs do projeto.

- Utilizando todos os processos analisados neste trabalho, é possível implementar novos DWs e, a partir dos mesmos, utilizar a abordagem de triplificação a fim de obter novas bases triplificadas, validar os processos aqui analisados e finalmente utilizar alguma técnica que interligue o conteúdo das várias bases triplificadas, pois quanto maior a quantidade de dados triplificados, melhor é o resultado de pesquisas semânticas.

- Outra linha de pesquisa de muita relevância na área de triplas RDF e para este trabalho é a interligação entre repositórios de triplas. Atualmente, existem muitos repositórios, como foi possível observar na nuvem LOD na Figura 3.2 citada anteriormente, mas a pesquisa de dados de forma que estes sejam interligados a tais repositórios ainda é muito manual, assim como foi muito manual a pesquisa de representações semânticas para triplificação do DW ENEM. No entanto, com o reúso da representação semântica feito na ferramenta Babel, tornou-se evidente, por exemplo, que é possível interligar dados das triplas do ENEM com as triplas da base GeoName.

- Estudo de técnicas de visualização, pois a forma de consulta do LOD está sendo feita por meio de consultas SPARQL, o que obriga o usuário a ter conhecimento sobre o vocabulário utilizado a fim de conseguir implementar consultas e dessa forma conseguir obter informações sobre a base. Logo, isso acaba por filtrar os usuários por obrigá-los a ter conhecimento na área, e com o estímulo de facilitar a consulta e esconder a complexidade do vocabulário, seria muito interessante ter uma ferramenta de visualização.

- A implementação de *Mashups*, que para a área de LOD, em geral são programas Web que, por meio da comunicação via *Web Services* ou SPARQL, consultam dados de bases distintas e disponibilizam através de uma interface simples. Isso possibilita, portanto, criar dados mais elaborados e completos melhorando a qualidade dos dados disponibilizados.

Referências

[Auer et al., 2009] Auer, S., Dietzold, S., Lehmann, J., and Hellmann, S., A. D. (2009). *Triplify: light-weight linked data publication from relational databases.* In *Proceedings of the 18th international conference on World wide web, Nova York, NY, EUA. ACM*, pages p. 621–630, Observatório da Educação, Brasília.

[Auer et al., 2006] Auer, S., Dietzold, S., and Riechert, T. (2006). OntoWiki - A Tool for Social, Semantic Collaboration.

[Baldus, 2011] Baldus, L. H. S. (2011). Estratégia para publicação de dados governamentais abertos no padrão linked data.

[Berners-Lee et al., 1998] Berners-Lee, T., Irvine, U., Fielding, and R., Masinter, L. (1998). *Uniform Resource Identifiers* (uri): *Generic Syntax. Standard Tracks*, RFC 2396.

[Berners-Lee, 2000] Berners-Lee, T. J. (2000). *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web.* Editora HarperBusiness, 1º edição edition.

[Berners-Lee, 2005] Berners-Lee, T. J. (2005). Semantic web - xml2000. Disponível em: http://www.w3.org/2000/Talks/1206-xml2k-tbl/Overview.html. Acesso 30 em junho 2011.

[Berners-Lee et al., 2009] Berners-Lee, T. J., Bizer, C., and Heath, T. (2009). *Linked Data - The Story So Far. International Journal on Semantic Web and Information Systems (IJSWIS).* Special Issue on Linked Data.

[Berners-Lee and Hendle, 2001] Berners-Lee, T. J. and Hendle, J. A. (2001). *Scientic publishing on the semantic web. Nature 410, 1023 - 1024.* Disponível em: http://www.nature.com/nature/debates/eaccess/Articles/bernerslee.htm. Acesso em: 15 março 2011.

[Berners-Lee and Hendle, 2006] Berners-Lee, T. J. and Hendle, J. A. (2006). *Linked data - design issues. Nature 410, 1023 - 1024.* Disponível em: http://www.w3.org/DesignIssues/LinkedData.html Acesso em: 12 março 2011.

[Berners-Lee et al., 2001] Berners-Lee, T. J., Hendle, J. A., and Lassila, O. (2001). *The Semantic Web. Scientific American.* Exemplar nº 501 - Maio.

[Breitman, 2006] Breitman, K. (2006). *Web Semântica - A Internet do Futuro.* Editora LTC, 4º edição edition.

[Campos, 2010] Campos, M. L. M. (2010). Linkeddatabr - exposição, compartilhamento e conexão de recursos de dados abertos na web (*Linked Open Data*).

[Cantele, 2009] Cantele, R. C. (2009). Construindo ontologias a partir de recursos existentes: uma prova de conceito no domínio da educação. Projeto aprovado no Edital - Tese de Doutorado em Engenharia Elétrica.

[Carromeu and Turine, 2005] Carromeu, C. and Turine, M. A. S. (2005). Um framework de aplicações web de apoio e gestão de fomento. In Workshop de Teses e Dissertações do XI Simpósio Brasileiro de Sistemas Multimidia e Web - WebMedia (Poços de Caldas, MG, Brasil, 2005).

[Cyganiak, 2005] Cyganiak, R. A. (2005). *Relational algebra for sparql*. Disponível em: http://www.hpl.hp.com/techreports/2005/HPL-2005-170.pdf, Acessado em 14 julho 2011.

[Cyganiak and A., 2011] Cyganiak, R. A. and A., J. (2011). *Linking open data cloud diagram*. Disponível em:http://lod-cloud.net, Acessado em 20 outubro 2011.

[data.gov.uk, 2011] data.gov.uk (2011). Disponível em: http://data.gov.uk, acessado em 25 junho 2011.

[Eaves, 2009] Eaves, D. (2009). *The three laws of open government*. Disponível em: http://eaves.ca/2009/09/30/three-law-of-open-government-data/, Acessado em 15 março 2011.

[Ferraz, 2011] Ferraz, V. R. T. (2011). Simbolização de mapas temáticos utilizando uma ontologia cartográfica. UFSCar.

[Gruber, 1993a] Gruber, T. R. (1993a). *A Translation Approach to Portable Ontologies*.

[Gruber, 1993b] Gruber, T. R. (1993b). *Toward Principles for the Design of Ontologies*.

[Harth and Decker, 2005] Harth, A. and Decker, S. (2005). *Optimized Index Structures for Querying RDF from the Web*.

[Hernandes, 2010] Hernandes, E. C. M. (2010). Um processo automatizado para tratamento de dados e conceitualizão de ontologias com apoio de visualização. Dissertação para pós gradução da UFSCar.

[Heuser, 2008] Heuser, C. A. (2008). *Projeto de Banco de Dados*. Editora Bookman, 4º edição edition.

[IETF, 2011] IETF (2011). Disponível em: http://www.ietf.org/. acessado em 25 setembro 2011.

[INEP, 2011] INEP (2011). Disponível em: www.inep.gov.br. acessado em 20 janeiro 2011.

[Inmon, 2005] Inmon, W. H. (2005). *Building the Data Warehouse*. Editora Wiley Publishing, 4º edição edition.

[J., 2012] J., T. (2012). *The RDF Data Cube Vocabulary*. Disponível em: http://www.w3.org/TR/vocab-data-cube/, Acessado em 15 junho 2012.

[JANNUZZI, 2001] JANNUZZI, M. P. (2001). Indicadores sociais no brasil: Conceitos, fontes de dados e aplicações.

[Johnson and Shneiderman, 1991] Johnson, B. and Shneiderman, B. (1991). *Tree-maps: a space-filling approach to the visualization of hierarchical information structures*. Conference on Visualization - VIS,2 San Diego. Proceedings... New York: IEEE Computer Society.

[Kimball and Ross, 2002] Kimball, R. and Ross, M. (2002). *The Data Warehouse Toolkit - The Complete Guide to Dimensional Modeling*. Editora *Wiley Publishing*, 2º edição edition.

[Kimball et al., 2008] Kimball, R., Ross, M., W., T., J., M., and Becker, B. (2008). *The Data Warehouse Lifecycle Toolkit*. Wiley, 2 edition, 4º edição edition.

[LATC, 2012] LATC (2012). Disponível em: http://latc-project.eu/. acessado em 03 março 2012.

[Leite and Rossi, 2010] Leite, D. M. and Rossi, L. L. (2010). Estudo e avaliação das técnicas e das ferramentas para projeto de *Data Warehouse*. Coxim - MS, UFMS-CPCX.

[Machado, 2010] Machado, F. N. R. (2010). *Tecnologia e Projeto de Data Warehouse*. Editora Érica Ltda, 4º edição edition.

[Marx, 2012] Marx, E. L. (2012). Babel. um framework extensível para a publicação de rdf de várias fontes de dados utilizando templates. Pontifícia Universidade Católica do Rio de Janeiro.

[Mota, 2011] Mota, F. M. (2011). Uma estratégia para publicação dos dados da base do ceb-inep/mec no padrão linked open data. Coxim - MS, UFMS-CPCX.

[Mota and Rossi, 2010] Mota, F. M. and Rossi, L. L. (2010). Estudo e avaliação das técnicas e das ferramentas para o projeto de data warehouse. Coxim - MS, UFMS-CPCX.

[Pentaho, 2011] Pentaho (2011). Disponível em: http://community.pentaho.com//, acessado em 9 março 2011.

[Rigotti and Cerqueira, 2001] Rigotti, J. R. and Cerqueira, C. A. (2001). As bases de dados do inep e os indicadores educacionais: conceitos e aplicações. *In: Proceedings of the international union for scientific study of population.*

[S., 2006] S., R. R. A. (2006). Web semântica: aspectos interdisciplinares da gestão de recursos informacionais no âmbito da ciência da informação.

[Salas et al., 2011] Salas, P. E., Breitman, Mota, F. M., M. A., , Casanova, M. A., Auer, S., and M., M. (2011). *Publishing Statistical Data on the Web*. Pontifícia Universidade Católica do Rio de Janeiro e Universidade de Leipzig.

[Salas et al., 2010] Salas, P. E., Breitman, K. K., Casanova, M. A., , and Viterbo, J. (2010). Stdtrip: *An a priori design approach and process for publishing open government data*. Departamento de Informática - PUC-Rio e Departamento de Ciência e Tecnologia - UFF.

[Santos, 2011] Santos, M. S. (2011). Modelo de processo para integração de serviços de avaliação educacional na plataforma web-pide. FACOM-UFMS.

[Savitraz, 2010] Savitraz, J. D. (2010). Uma abordagem de integração e exploração visual de dados educacionais na plataforam web-pide. FACOM-UFMS.

[Sikos, 2011] Sikos, L. (2011). *Web Standards*. Apress, 1º edição edition.

[Silva, 2011] Silva, F. A. R. (2011). Identificação e determinação de serviços para compor a plataforma web-pide. UFSCAR.

[Siqueira, 2009] Siqueira, T. (2009). Sb-index: Um Índice espacial baseado em bitmap para data warehouse geográfico. UFSCar.

[Siqueira et al., 2008] Siqueira, T. L. L., Ciferri, R. R., and Santos, M. T. P. (2008). Projeto, construção e manutenção de data warehouses para auxiliar o planejamento de políticas públicas de educação. UFSCar.

[TURINE et al., 2006] TURINE, M. A. S., CÁCERES, E. N., MONGELLI, H., PAIVA, D. M. B., FABBRI, S., ROCHA, M. G. B., CIFERRI, R. R., and SANTOS, M. T. P. (2006). Web-pide: uma plataforma aberta de integração e avaliação de dados educacionais.

[TURINE et al., 2004] TURINE, M. A. S., OLIVEIRA, M., ISHY, E.and BLANES, D., ACOSTA, A. R., and WANDERLEY, M. B. (2004). Sigs: Uma tecnologia social na web para gestão, monitoramento e avaliação de programas sociais. Conferência IADIS Ibero-Americana WWW/Internet 2004, v. 1, p. 437-440, Madrid.

[UNICODE, 2012] UNICODE (2012). Disponível em: http://www.unicode.org/. acessado em 02 março 2012.

[Vanni, 2009] Vanni, R. M. P. (2009). Integração de serviços em ambientes heterogêneos: uso de semântica para comunicação entre entidades em mudanças de contexto. Tese (Doutorado em Ciências de Computação e Matemática Computacional) - São Carlos - USP.

[Virtuoso, 2011] Virtuoso (2011). Virtuoso, open link. Disponível em: http://virtuoso.openlinksw.com. Acesso 20 em junho 2011.

[W3C, 2009] W3C (2009). Dados abertos governamentais. disponível em: http://www.w3c.br/divulgacao/pdf/dados-abertos-governamentais.pdf, acessado em 25 abril 2011.

[W3C, 2012] W3C (2012). Disponível em: http://www.w3.org/standards/semanticweb/, acessado em 9 março 2012. W3C *Recommendations*.

Apêndice A

Parte do Manual do Usuário (Dicionário de Dados) - Enem 2007

DICIONÁRIO DAS VARIÁVEIS - ENEM 2007

Nome da Variável	Descrição	Variáveis Categóricas		Início	Tamanho	Tipo
		Categoria	Descrição			
VARIÁVEIS DE CONTROLE DO INSCRITO						
MASC_INSCRITO	**Máscara do inscrito:** Número gerado como código de acesso ao participante do exame. A mesma MASCARA_INSCRITO para anos diferentes não identifica o mesmo participante no exame e não permite o acesso aos dados cadastrais como nome, endereço, RG, etc. A mesma MASCARA_INSCRITO não identifica o mesmo participante em microdados de pesquisas diferentes.			1	8	Alfanumérica
NU_ANO	Ano do Enem			9	8	Numérica
DT_NASCIMENTO	Data de nascimento do inscrito			17	20	Alfanumérica
TP_SEXO	Sexo do inscrito	1 2	Masculino Feminino	37	8	Numérica
CODMUNIC_INSC	Código do Município em que o inscrito mora: 1º dígito: região; 1º e 2º dígitos: UF; 3º, 4º, 5º e 6º dígitos: município; 7º dígito: DV;			45	8	Alfanumérica
NO_MUNICIPIO_INSC	Nome do município em que o inscrito mora			53	150	Alfanumérica
UF_INSC	Sigla da Unidade da Federação do inscrito no Enem			203	2	Alfanumérica
IN_CONCLUIU	Situação em relação ao ensino médio	0 1 2	Já concluiu Concluirá em 2007 Concluirá após 2007	205	1	Numérica
IN_SUPLETIVO	Aluno do ensino médio ou da educação de jovens e adultos	1 2	Regular Educação para jovens e adultos(EJA)	206	8	Numérica
VARIÁVEIS DE CONTROLE DA ESCOLA						
PK_COD_ENTIDADE	**Código da Escola:** Número gerado como identificação da escola			214	8	Alfanumérica
COD_MUNICIPIO_ESC	Código do Município da escola em que estudou: 1º dígito: região; 1º e 2º dígitos: UF; 3º, 4º, 5º e 6º dígitos: município; 12º dígito: DV;			222	8	Alfanumérica
NO_MUNICIPIO_ESC	Nome do município da escola			230	150	
UF_ESC	Sigla da Unidade da Federação da escola			380	2	Alfanumérica
ID_DEPENDENCIA_ADM	Dependência administrativa	1 - Federal 2 - Estadual 3 - Municipal 4 - Privada		382	1	Alfanumérica
ID_LOCALIZACAO	Localização/Zona da escola	1 - Urbana 2 - Rural		383	1	Alfanumérica

DICIONÁRIO DAS VARIÁVEIS - ENEM 2007

Nome da Variável	Descrição	Variáveis Categóricas		Início	Tamanho	Tipo
		Categoria	Descrição			
SIT_FUNC	Situação de funcionamento	EM ATIVIDADE PARALIZADA EXTINTA	Em Funcionamento Atividades temporariamente suspensas Atividades definitivamente encerradas	384	15	Alfanumérica
VARIÁVEIS DA PROVA OBJETIVA						
IN_PRESENCA	Presença a prova objetiva	0 1	Faltou à prova Presente à prova	399	8	Numérica
VL_PERC_COMP1	Nota da competência 1			407	8	Numérica
VL_PERC_COMP2	Nota da competência 2			415	8	Numérica
VL_PERC_COMP3	Nota da competência 3			423	8	Numérica
VL_PERC_COMP4	Nota da competência 4			431	8	Numérica
VL_PERC_COMP5	Nota da competência 5			439	8	Numérica
NU_NOTA_OBJETIVA	Nota da prova objetiva			447	8	Numérica
VARIÁVEIS DA PROVA DE REDAÇÃO						
IN_SITUACAO	Presença à redação	B F N P	Entregou a redação em branco Faltou à prova Redação anulada Presente à prova	455	1	Alfanumérica
NU_NOTA_REDACAO_COMP1	Nota da competência 1			456	8	Numérica
NU_NOTA_REDACAO_COMP2	Nota da competência 2			464	8	Numérica
NU_NOTA_REDACAO_COMP3	Nota da competência 3			472	8	Numérica
NU_NOTA_REDACAO_COMP4	Nota da competência 4			480	8	Numérica
NU_NOTA_REDACAO_COMP5	Nota da competência 5			488	8	Numérica
NU_NOTA_GLOBAL_REDACAO	Nota da prova de redação			496	8	Numérica
QUESTIONÁRIO SOCIOECONÔMICO DO ENEM						
VOCÊ E SUA FAMÍLIA						
IN_QSE	Resposta ao Questionário Socioeconômico	1 0	Respondeu o questionário socioecômico Não respondeu o questionário socioecômico	504	1	Alfanumérica
Q1	Sexo	A B	Feminino Masculino	505	1	Alfanumérica
Q2	Ano em que nasceu	A B C D E F G H I	Após 1990 Em 1990 Em 1989 Em 1988 Em 1987 Em 1986 Em 1985 Entre 1981e 1984 Antes de 1981	506	1	Alfanumérica
Q3	Como se considera	A B C D E	Branco(a) Pardo(a) Preto(a) Amarelo(a) Indígena	507	1	Alfanumérica
Q4	Se indicou indígena, qual(is) língua(s) você domina	A B C D E F G	Minha língua materna é o português Falo uma língua indígena e o português Falo mais de uma língua indígena e o português Falo uma língua indígena, o português e o espanhol Falo mais de uma língua indígena, o português e o espanhol Além do português, falo francês e inglês Além do português, falo o creole	508	1	Alfanumérica

DICIONÁRIO DAS VARIÁVEIS - ENEM 2007

Nome da Variável	Descrição	Variáveis Categóricas		Início	Tamanho	Tipo
		Categoria	Descrição			
Q5	Qual o estado civil	A	Solteiro(a)	509	1	Alfanumérica
		B	Casado(a) / mora com um(a) companheiro(a)			
		C	Separado(a) / divorciado(a) / desquitado(a)			
		D	Viúvo(a)			
Q6	Onde e como mora atualmente	A	Em casa ou apartamento, com sua família	510	1	Alfanumérica
		B	Em casa ou apartamento, sozinho(a)			
		C	Em quarto ou cômodo alugado, sozinho(a)			
		D	Em habitação coletiva: hotel, hospedaria, quartel, pensionato, república, etc			
		E	Outra situação			
Q7	Mora sozinho(a)	A	Sim	511	1	Alfanumérica
		B	Não			
Q8	Mora com o pai	A	Sim	512	1	Alfanumérica
		B	Não			
Q9	Mora com a mãe	A	Sim	513	1	Alfanumérica
		B	Não			
Q10	Mora com esposa / marido / companheiro(a)	A	Sim	514	1	Alfanumérica
		B	Não			
Q11	Mora com filho(s)	A	Sim	515	1	Alfanumérica
		B	Não			
Q12	Mora com irmão(s)	A	Sim	516	1	Alfanumérica
		B	Não			
Q13	Mora com outro(s) parente(s)	A	Sim	517	1	Alfanumérica
		B	Não			
Q14	Mora com amigo(s) ou colega(s)	A	Sim	518	1	Alfanumérica
		B	Não			
Q15	Quantidade de pessoas que moram em sua casa	A	Duas pessoas	519	1	Alfanumérica
		B	Três pessoas			
		C	Quatro pessoas			
		D	Cinco pessoas			
		E	Seis pessoas			
		F	Mais de 6 pessoas			
		G	Moro sozinho			
Q16	Quantos filhos tem	A	Um filho	520	1	Alfanumérica
		B	Dois filhos			
		C	Três filhos			
		D	Quatro ou mais filhos			
		E	Não tenho filhos			
Q17	Até quando seu pai estudou	A	Não estudou	521	1	Alfanumérica
		B	Da 1ª a 4ª série do ensino fundamental (antigo primário)			
		C	Da 5ª a 8ª do ensino fundamental (antigo ginásio)			
		D	Ensino Médio (2º grau) incompleto			
		E	Ensino Médio (2º grau) completo			
		F	Ensino Superior incompleto			
		G	Ensino Superior completo			
		H	Pós-graduação			
		I	Não sei			

DICIONÁRIO DAS VARIÁVEIS - ENEM 2007

Nome da Variável	Descrição	Variáveis Categóricas		Início	Tamanho	Tipo
		Categoria	Descrição			
Q18	Até quando sua mãe estudou	A	Não estudou	522	1	Alfanumérica
		B	Da 1ª a 4ª série do ensino fundamental (antigo primário)			
		C	Da 5ª a 8ª do ensino fundamental (antigo ginásio)			
		D	Ensino Médio (2º grau) incompleto			
		E	Ensino Médio (2º grau) completo			
		F	Ensino Superior incompleto			
		G	Ensino Superior completo			
		H	Pós-graduação			
		I	Não sei			
Q19	Área que o pai trabalha ou trabalhou, na maior parte da vida	A	Na agricultura, no campo, em fazenda ou na pesca	523	1	Alfanumérica
		B	Na indústria			
		C	No comércio, banco, transporte ou outros serviços			
		D	Funcionário público do governo federal, estadual, ou do município, ou militar			
		E	Profissional liberal, professor ou técnico de nível superior			
		F	Trabalhador do setor informal (sem carteira assinada)			
		G	Trabalha em casa em serviços (costura, cozinha, aulas particulares, etc)			
		H	No lar			
		I	Não trabalha			
		J	Não sei			
Q20	Qual a posição do seu pai neste trabalho, na maior parte do tempo	A	Gerente, administrador ou diretor de empresa privada	524	1	Alfanumérica
		B	Funcionário público (federal, estadual, ou municipal), com funções de direção			
		C	Militar (guarda-civil, polícia estadual ou Forças Armadas), com posto de comando			
		D	Empregado no setor privado, com carteira assinada			
		E	Funcionário público (federal, estadual, ou municipal), sem função de direção			
		F	Militar (guarda-civil, polícia estadual ou Forças Armadas), sem posto de comando			
		G	Trabalho temporário, informal, sem carteira assinada			
		H	Trabalho por conta própria			
		I	Desempregado			
		J	Aposentado			
		K	Outra situação			

DICIONÁRIO DAS VARIÁVEIS - ENEM 2007

Nome da Variável	Descrição	Variáveis Categóricas		Início	Tamanho	Tipo
		Categoria	Descrição			
Q21	Área que sua mãe trabalha ou trabalhou, na maior parte da vida	A	Na agricultura, no campo, em fazenda ou na pesca	525	1	Alfanumérica
		B	Na indústria			
		C	No comércio, banco, transporte ou outros serviços			
		D	Como trabalhadora doméstica			
		E	Como funcionária do governo federal, do estado ou do município ou militar			
		F	Como profissional liberal, professor ou técnica de nível superior			
		G	No lar			
		H	Trabalha em casa em serviços (costura, cozinha, aulas particulares, etc)			
		I	Não trabalha			
		J	Não sei			
Q22	Qual a posição de sua mãe neste trabalho, na maior parte do tempo	A	Gerente, administrador ou diretor de empresa privada	526	1	Alfanumérica
		B	Funcionário público (federal, estadual, ou municipal), com funções de direção			
		C	Militar (guarda-civil, polícia estadual ou Forças Armadas), com posto de comando			
		D	Empregado no setor privado, com carteira assinada			
		E	Funcionário público (federal, estadual, ou municipal), sem função de direção			
		F	Militar (guarda-civil, polícia estadual ou Forças Armadas), sem posto de comando			
		G	Trabalho temporário, informal, sem carteira assinada			
		H	Trabalho por conta própria			
		I	Desempregado			
		J	Aposentado			
		K	Outra situação			
Q23	Renda familiar (somando a do respondente e com a das pessoas que moram com ele)	A	Até 1 salário mínimo (até R$ 380,00 inclusive)	527	1	Alfanumérica
		B	De 1 a 2 salários mínimos (R$ 380,00 a R$ 760,00 inclusive)			
		C	De 2 a 5 salários mínimos (R$ 760,00 a R$ 1.900,00 inclusive)			
		D	De 5 a 10 salários mínimos (R$ 1.900,00 a R$ 3.800,00 inclusive)			
		E	De 10 a 30 salários mínimos (R$ 3.800,00 a R$ 11.400,00 inclusive)			
		F	De 30 a 50 salários mínimos (R$ 11.400,00 a R$ 19.000,00 inclusive)			
		G	Mais de 50 salários mínimos (mais de R$ 19.000,00)			
		H	Nenhuma renda			
Q24	Tem TV e quantas	A	1	528	1	Alfanumérica
		B	2			
		C	3 ou mais			
		D	Não tem			

DICIONÁRIO DAS VARIÁVEIS - ENEM 2007

Nome da Variável	Descrição	Variáveis Categóricas		Início	Tamanho	Tipo
		Categoria	Descrição			
Q25	Tem Videocassete e/ou DVD e quantos	A B C D	1 2 3 ou mais Não tem	529	1	Alfanumérica
Q26	Tem Rádio e quantos	A B C D	1 2 3 ou mais Não tem	530	1	Alfanumérica
Q27	Tem Microcomputador e quantos	A B C D	1 2 3 ou mais Não tem	531	1	Alfanumérica
Q28	Tem Automóvel e quantos	A B C D	1 2 3 ou mais Não tem	532	1	Alfanumérica
Q29	Tem Máquina de lavar roupa e quantos	A B C D	1 2 3 ou mais Não tem	533	1	Alfanumérica
Q30	Tem Geladeira e quantas	A B C D	1 2 3 ou mais Não tem	534	1	Alfanumérica
Q31	Tem Telefone fixo e quantos	A B C D	1 2 3 ou mais Não tem	535	1	Alfanumérica
Q32	Tem Telefone celular e quantos	A B C D	1 2 3 ou mais Não tem	536	1	Alfanumérica
Q33	Tem Acesso à Internet e quantos	A B C D	1 2 3 ou mais Não tem	537	1	Alfanumérica
Q34	Tem TV por assinatura e quantas	A B C D	1 2 3 ou mais Não tem	538	1	Alfanumérica
Q35	Tem casa própria	A B	Sim Não	539	1	Alfanumérica
Q36	Se a casa é em rua calçada ou asfaltada	A B	Sim Não	540	1	Alfanumérica
Q37	Se a casa tem água corrente de torneira	A B	Sim Não	541	1	Alfanumérica
Q38	Se a casa tem eletricidade	A B	Sim Não	542	1	Alfanumérica
Q39	Se a casa é situada em comunidade indígena	A B	Sim Não	543	1	Alfanumérica

DICIONÁRIO DAS VARIÁVEIS - ENEM 2007

Nome da Variável	Descrição	Variáveis Categóricas		Início	Tamanho	Tipo
		Categoria	Descrição			
Q40	Motivo para fazer o Enem	A	Para testar seus conhecimentos / capacidade de raciocínio	544	1	Alfanumérica
		B	Para entrar na faculdade / conseguir pontos para o vestibular			
		C	Para ter um bom emprego / saber se está preparado(a) para o futuro profissional			
		D	Não sei			
VOCÊ E O TRABALHO						
Q41	O motivo mais importante para se ter um trabalho?	A	Para ter mais responsabilidade	545	1	Alfanumérica
		B	Independência financeira			
		C	Adquirir experiência			
		D	Crescer profissionalmente			
		E	Sentir-me útil			
		F	Para fazer amigos, conhecer pessoas			
		G	Não acho importante ter um trabalho			
		H	Para ajudar minha comunidade indígena			
Q42	Trabalha, ou já trabalhou, ganhando algum salário ou rendimento	A	Sim	546	1	Alfanumérica
		B	Nunca trabalhei			
		C	Nunca trabalhei, mas estou procurando trabalho			
Q43	Trabalhou ou teve alguma atividade remunerada durante o ensino médio (2º grau)	A	Sim, todo o tempo	547	1	Alfanumérica
		B	Sim, menos de 1 ano			
		C	Sim, de 1 a 2 anos			
		D	Sim, de 2 a 3 anos			
		E	Não			
Q44	Quantas horas trabalhava, durante o ensino médio (2º grau)	A	Sem jornada fixa, até 10 horas semanais	548	1	Alfanumérica
		B	De 11 a 20 horas semanais			
		C	De 21 a 30 horas semanais			
		D	De 31 a 40 horas semanais			
		E	Mais de 40 horas semanais			
Q45	Com que finalidade você trabalhava enquanto estudava o ensino médio (2º grau)	A	Para ajudar meus pais nas despesas com a casa, sustentar a família	549	1	Alfanumérica
		B	Para ser independente (ter meu sustento, ganhar meu próprio dinheiro)			
		C	Para adquirir experiência			
		D	Para ajudar minha comunidade			
		E	Outra finalidade			
Q46	Com que idade você começou a exercer atividade remunerada enquanto estudava o ensino médio (2º grau)	A	Antes dos 14 anos	550	1	Alfanumérica
		B	Entre 14 e 16 anos			
		C	Entre 17 e 18 anos			
		D	Após 18 anos			
Q47	Se você está trabalhando atualmente, qual a sua renda ou seu salário mensal?	A	Até 1 salário mínimo (até R$ 380,00 inclusive)	551	1	Alfanumérica
		B	De 1 a 2 salários mínimos (R$ 380,00 a R$ 760,00 inclusive)			
		C	De 2 a 5 salários mínimos (R$ 760,00 a R$ 1.900,00 inclusive)			
		D	De 5 a 10 salários mínimos (R$ 1.900,00 a R$ 3.800,00 inclusive)			
		E	De 10 a 30 salários mínimos (R$ 3.800,00 a R$ 11.400,00 inclusive)			
		F	De 30 a 50 salários mínimos (R$ 11.400,00 a R$ 19.000,00 inclusive)			
		G	Mais de 50 salários mínimos (mais de R$ 19.000,00)			
		H	Não estou trabalhando			

DICIONÁRIO DAS VARIÁVEIS - ENEM 2007

Nome da Variável	Descrição	Variáveis Categóricas		Início	Tamanho	Tipo
		Categoria	Descrição			
Q48	Você está trabalha atualmente em alguma atividade para o qual se preparou	A	Sim	552	1	Alfanumérica
		B	Não			
Q49	Em que trabalha atualmente	A	Na agricultura, no campo, em fazenda ou na pesca	553	1	Alfanumérica
		B	Na indústria			
		C	No comércio, banco, transporte ou outros serviços			
		D	Como trabalhador(a) doméstico(a)			
		E	Como funcionária do governo federal, do estado ou do município ou militar			
		F	Como profissional liberal, professor ou técnica de nível superior			
		G	No lar			
		H	Trabalho em casa em serviços (costura, cozinha, aulas particulares)			
		I	Não trabalha			
Q50	Qual a sua posição neste trabalho	A	Gerente, administrador(a) ou diretor(a) de empresa privada	554	1	Alfanumérica
		B	Funcionário público (federal, estadual, ou municipal), com funções de direção			
		C	Militar (guarda-civil, polícia estadual ou Forças Armadas), com posto de comando			
		D	Empregado no setor privado, com carteira assinada			
		E	Funcionário público (federal, estadual, ou municipal), sem função de direção			
		F	Trabalho temporário, informal, sem carteira assinada			
		G	Militar (guarda-civil, polícia estadual ou Forças Armadas), com posto de comando			
		H	Trabalho por conta própria			
		I	Trabalhando por conta própria			
		J	Aposentado(a)			
		K	Outra situação			
Q51	Quanto tempo está trabalhando na atividade	A	Menos de 1 ano	555	1	Alfanumérica
		B	Entre 1 e 2 anos			
		C	Entre 2 e 4 anos			
		D	Mais de 4 anos			
Q52	Os conhecimentos no ensino médio foram adequados ao que o mercado de trabalho solicita	A	Sim	556	1	Alfanumérica
		B	Não			
Q53	Os conhecimentos no ensino médio tiveram relação com a profissão que escolheu / que exerce	A	Sim	557	1	Alfanumérica
		B	Não			
Q54	Os conhecimentos no ensino médio foram bem desenvolvidos, com aulas práticas, laboratórios, etc	A	Sim	558	1	Alfanumérica
		B	Não			
Q55	Os conhecimentos no ensino médio proporcionaram cultura e conhecimento	A	Sim	559	1	Alfanumérica
		B	Não			
Q56	Avaliação de ter estudado e trabalhado, simultaneamente, durante o ensino médio	A	Atrapalhou os estudos	560	1	Alfanumérica
		B	Possibilitou crescimento pessoal			
		C	Atrapalhou os estudos, mas possibilitou crescimento pessoal			
		D	Não atrapalhou os estudos			
		E	Não trabalho / não trabalhei			

DICIONÁRIO DAS VARIÁVEIS - ENEM 2007

Nome da Variável	Descrição	Variáveis Categóricas		Início	Tamanho	Tipo
		Categoria	Descrição			
Q57	A escola que freqüenta ou freqüentou durante o ensino médio levou em conta que trabalhava ao mesmo tempo que estudava	A B C	Sim Não Não sei	561	1	Alfanumérica
Q58	Como prova de consideração por parte da escola em consideração pelo aluno trabalhar e estudar ao mesmo tempo, a escola tinha o horário flexível	A B	Sim Não	562	1	Alfanumérica
Q59	Como prova de consideração por parte da escola em consideração pelo aluno trabalhar e estudar ao mesmo tempo, a escola tinha menor carga de trabalho ou de tarefas extraclasse	A B	Sim Não	563	1	Alfanumérica
Q60	Como prova de consideração por parte da escola em consideração pelo aluno trabalhar e estudar ao mesmo tempo, a escola tinha programa de recuperação de notas	A B	Sim Não	564	1	Alfanumérica
Q61	Como prova de consideração por parte da escola em consideração pelo aluno trabalhar e estudar ao mesmo tempo, a escola tinha abono de faltas	A B	Sim Não	565	1	Alfanumérica
Q62	Como prova de consideração por parte da escola em consideração pelo aluno trabalhar e estudar ao mesmo tempo, a escola tinha aulas mais dinâmicas, com didática difenciada	A B	Sim Não	566	1	Alfanumérica
Q63	Como prova de consideração por parte da escola em consideração pelo aluno trabalhar e estudar ao mesmo tempo, a escola tinha aulas de revisão da matéria aos (às) interessados(as)	A B	Sim Não	567	1	Alfanumérica
Q64	Como prova de consideração por parte da escola em consideração pelo aluno trabalhar e estudar ao mesmo tempo, a escola tinha fornecimento de refeição aos (às) alunos(as)	A B	Sim Não	568	1	Alfanumérica
Q65	A escola deve oferecer horário flexível para o aluno que trabalha	A B	Sim Não	569	1	Alfanumérica
Q66	A escola deve oferecer menor carga de trabalho ou de tarefas extraclasse para o aluno que trabalha	A B	Sim Não	570	1	Alfanumérica
Q67	A escola deve oferecer programa de recuperação de notas para o aluno que trabalha	A B	Sim Não	571	1	Alfanumérica
Q68	A escola deve oferecer abono de faltas para o aluno que trabalha	A B	Sim Não	572	1	Alfanumérica
Q69	A escola deve oferecer aulas mais dinâmicas, com didática diferenciada para o aluno que trabalha	A B	Sim Não	573	1	Alfanumérica
Q70	A escola deve oferecer aulas de revisão da matéria aos alunos que interessados que trabalham	A B	Sim Não	574	1	Alfanumérica
Q71	A escola deve oferecer fornecimento de refeição para o aluno que trabalha	A B	Sim Não	575	1	Alfanumérica
VOCÊ E OS ESTUDOS						
Q72	Anos que levou para concluir o ensino fundamental (1° grau)	A B C D E F	Menos de 8 anos 8 anos 9 anos 10 anos 11 anos Mais de 11 anos	576	1	Alfanumérica

DICIONÁRIO DAS VARIÁVEIS - ENEM 2007

Nome da Variável	Descrição	Variáveis Categóricas		Início	Tamanho	Tipo
		Categoria	Descrição			
Q73	Em que tipo de escola cursou o ensino fundamental (1ª grau)	A	Somente em escola pública	577	1	Alfanumérica
		B	Parte em escola pública e parte em escola particular			
		C	Somente em escola particular			
		D	Somente em escola indígena			
		E	Parte na escola indígena e parte em escola não-indígena			
Q74	Em que ano concluiu ou concluirá o ensino médio (2º grau)	A	Vai concluir após 2007	578	1	Alfanumérica
		B	Vai concluir no segundo semestre de 2007			
		C	No primeiro semestre de 2007			
		D	Em 2006			
		E	Em 2005			
		F	Em 2004			
		G	Em 2003			
		H	Em 2002			
		I	Em 2001			
		J	Antes de 2001			
Q75	Quantos anos levou para cursas o ensino médio (2º grau)	A	Menos de 3 anos	579	1	Alfanumérica
		B	3 anos			
		C	4 anos			
		D	5 anos			
		E	6 anos			
		F	Mais de 6 anos			
Q76	Em que turno cursou ou esta cursando o ensino médio	A	Somente no turno diurno	580	1	Alfanumérica
		B	Maior parte no turno diurno			
		C	Somente no turno noturno			
		D	Maior parte no turno noturno			
Q77	Em que escola cursou ou está cursando o ensino médio (2º grau)	A	Somente em escola pública	581	1	Alfanumérica
		B	Maior parte em escola pública			
		C	Somente em escola particular			
		D	Maior parte em escola particular			
		E	Somente em escola indígena			
		F	Maior parte em escola não-ingígena			
Q78	Em que modalidade de ensino concluiu ou vai concluir o ensino médio (2º grau)	A	Ensino regular	582	1	Alfanumérica
		B	Educação para jovens e adultos (antigo supletivo)			
		C	Ensino técnico / ensino profissional			
Q79	Fez curso de língua estrangeira fora da escola durante o ensino médio (2º grau)	A	Sim	583	1	Alfanumérica
		B	Não			
Q80	Fez curso de computação ou informática fora da escola durante o ensino médio (2º grau)	A	Sim	584	1	Alfanumérica
		B	Não			
Q81	Fez curso preparatório para o vestibular (cursinho) fora da escola durante o ensino médio (2º grau)	A	Sim	585	1	Alfanumérica
		B	Não			
Q82	Fez artes plásticas ou atividades artísticas em geral fora da escola durante o ensino médio (2º grau)	A	Sim	586	1	Alfanumérica
		B	Não			
Q83	Fez esportes, atividades físicas fora da escola durante o ensino médio (2º grau)	A	Sim	587	1	Alfanumérica
		B	Não			
Q84	Fez outro curso fora da escola durante o ensino médio (2º grau)	A	Sim	588	1	Alfanumérica
		B	Não			
Q85	Com que freqüência lê jornais	A	Freqüentemente (todo dia ou quase todo dia)	589	1	Alfanumérica
		B	Às vezes			
		C	Nunca			

DICIONÁRIO DAS VARIÁVEIS - ENEM 2007

Nome da Variável	Descrição	Variáveis Categóricas		Início	Tamanho	Tipo
		Categoria	Descrição			
Q86	Com que freqüência lê revistas de informação geral (Veja, Isto é, Época, etc)	A B C	Freqüentemente (todo dia ou quase todo dia) Às vezes Nunca	590	1	Alfanumérica
Q87	Com que freqüência lê revistas de humor / quadrinhos	A B C	Freqüentemente (todo dia ou quase todo dia) Às vezes Nunca	591	1	Alfanumérica
Q88	Com que freqüência lê revistas de divulgação científica (Ciência Hoje, Galileu, etc)	A B C	Freqüentemente (todo dia ou quase todo dia) Às vezes Nunca	592	1	Alfanumérica
Q89	Com que freqüência lê romances, livros de ficção	A B C	Freqüentemente (todo dia ou quase todo dia) Às vezes Nunca	593	1	Alfanumérica
Q90	Avaliação da escola que fez o ensino médio quanto o conhecimento que os(as) professores(as) têm das matérias e a maneira de transmiti-lo	A B C	Insuficiente a regular Regular a bom Bom a excelente	594	1	Alfanumérica
Q91	Avaliação da escola que fez o ensino médio quanto a dedicação dos(as) professores(as) para preparar aulas e atender aos alunos	A B C	Insuficiente a regular Regular a bom Bom a excelente	595	1	Alfanumérica
Q92	Avaliação da escola que fez o ensino médio quanto as iniciativas da escola para realizar excursões, estudos do meio ambiente	A B C	Insuficiente a regular Regular a bom Bom a excelente	596	1	Alfanumérica
Q93	Avaliação da escola que fez o ensino médio quanto a biblioteca	A B C	Insuficiente a regular Regular a bom Bom a excelente	597	1	Alfanumérica
Q94	Avaliação da escola que fez o ensino médio quanto as condições das salas de aula	A B C	Insuficiente a regular Regular a bom Bom a excelente	598	1	Alfanumérica
Q95	Avaliação da escola que fez o ensino médio quanto as condições dos laboratórios	A B C	Insuficiente a regular Regular a bom Bom a excelente	599	1	Alfanumérica
Q96	Avaliação da escola que fez o ensino médio quanto o acesso a computadores e outros recursos de informática	A B C	Insuficiente a regular Regular a bom Bom a excelente	600	1	Alfanumérica
Q97	Avaliação da escola que fez o ensino médio quanto o ensino de língua estrangeira	A B C	Insuficiente a regular Regular a bom Bom a excelente	601	1	Alfanumérica
Q98	Avaliação da escola que fez o ensino médio quanto o interesse dos(as) alunos(as)	A B C	Insuficiente a regular Regular a bom Bom a excelente	602	1	Alfanumérica
Q99	Avaliação da escola que fez o ensino médio quanto o trabalho de grupo	A B C	Insuficiente a regular Regular a bom Bom a excelente	603	1	Alfanumérica
Q100	Avaliação da escola que fez o ensino médio quanto a práticas de esportes	A B C	Insuficiente a regular Regular a bom Bom a excelente	604	1	Alfanumérica
Q101	Avaliação da escola que fez o ensino médio quanto a atenção e o respeito dos(as) funcionários(as)	A B C	Insuficiente a regular Regular a bom Bom a excelente	605	1	Alfanumérica
Q102	Avaliação da escola que fez o ensino médio quanto a direção dela	A B C	Insuficiente a regular Regular a bom Bom a excelente	606	1	Alfanumérica

DICIONÁRIO DAS VARIÁVEIS - ENEM 2007

Nome da Variável	Descrição	Variáveis Categóricas		Início	Tamanho	Tipo
		Categoria	Descrição			
Q103	Avaliação da escola que fez o ensino médio quanto a organização dos horários de aulas	A B C	Insuficiente a regular Regular a bom Bom a excelente	607	1	Alfanumérica
Q104	Avaliação da escola que fez o ensino médio quanto a localização dela	A B C	Insuficiente a regular Regular a bom Bom a excelente	608	1	Alfanumérica
Q105	Avaliação da escola que fez o ensino médio quanto a segurança (iluminação, policiamento, etc)	A B C	Insuficiente a regular Regular a bom Bom a excelente	609	1	Alfanumérica
Q106	Avaliação da escola que fez o ensino médio quanto a atenção à identificação étnica dos(as) alunos(as)	A B C	Insuficiente a regular Regular a bom Bom a excelente	610	1	Alfanumérica
Q107	A escola em que estuda ou estudou realiza palestras / debates	A B	Sim Não	611	1	Alfanumérica
Q108	A escola em que estuda ou estudou realiza jogos / esportes / campeonatos	A B	Sim Não	612	1	Alfanumérica
Q109	A escola em que estuda ou estudou realiza teatro	A B	Sim Não	613	1	Alfanumérica
Q110	A escola em que estuda ou estudou realiza coral	A B	Sim Não	614	1	Alfanumérica
Q111	A escola em que estuda ou estudou realiza dança / música	A B	Sim Não	615	1	Alfanumérica
Q112	A escola em que estuda ou estudou realiza estudos do meio ambiente / passeios	A B	Sim Não	616	1	Alfanumérica
Q113	A escola em que estuda ou estudou realiza feira de ciências / feira cultural	A B	Sim Não	617	1	Alfanumérica
Q114	A escola em que estuda ou estudou realiza festas / gincanas	A B	Sim Não	618	1	Alfanumérica
Q115	De acordo com os ensinamentos no ensino médio, como considera o preparo para conseguir um emprego, exercer alguma atividade	A B C D	Eu me considero preparado(a) para entrar no mercado de trabalho Apesar de ter freqüentado uma boa escola, eu me considero despreparado(a), pois não aprendi o suficiente para conseguir um emprego Eu me considero despreparado(a) devido à baixa qualidade do ensino de minha escola, que não me preparou o suficiente Não sei	619	1	Alfanumérica
Q116	Os(as) professores(as) têm autoridade, firmeza	A B	Sim Não	620	1	Alfanumérica
Q117	Os(as) professores(as) são distantes, têm pouco envolvimento	A B	Sim Não	621	1	Alfanumérica
Q118	Os(as) professores(as) têm respeito	A B	Sim Não	622	1	Alfanumérica
Q119	Os(as) professores(as) são indiferentes, ignoram sua existência	A B	Sim Não	623	1	Alfanumérica
Q120	Os(as) professores(as) são preocupados(as) e dedicados(as)	A B	Sim Não	624	1	Alfanumérica
Q121	Os(as) professores(as) são autoritários(as), rígidos(as), abusam do poder	A B	Sim Não	625	1	Alfanumérica
Q122	Os(as) professores(as) valorizam a identidade étnica dos(as) alunos(as)	A B	Sim Não	626	1	Alfanumérica

DICIONÁRIO DAS VARIÁVEIS - ENEM 2007

Nome da Variável	Descrição	Variáveis Categóricas		Início	Tamanho	Tipo
		Categoria	Descrição			
Q123	Avaliação sobre a escola quanto a liberdade de expressar a idéias	A B C	Insuficiente a regular Regular a bom Bom a excelente	627	1	Alfanumérica
Q124	Avaliação sobre a escola quanto o respeito aos alunos e alunas	A B C	Insuficiente a regular Regular a bom Bom a excelente	628	1	Alfanumérica
Q125	Avaliação sobre a escola quanto a amizade e respeito entre alunos(as) e funcionários(as)	A B C	Insuficiente a regular Regular a bom Bom a excelente	629	1	Alfanumérica
Q126	Avaliação sobre a escola quanto a levar em conta suas opiniões	A B C	Insuficiente a regular Regular a bom Bom a excelente	630	1	Alfanumérica
Q127	Avaliação sobre a escola quanto a discussão dos problemas da atualidade nas aulas	A B C	Insuficiente a regular Regular a bom Bom a excelente	631	1	Alfanumérica
Q128	Avaliação sobre a escola quanto a convivência entre alunos(as)	A B C	Insuficiente a regular Regular a bom Bom a excelente	632	1	Alfanumérica
Q129	Avaliação sobre a escola quanto a organização para apoiar a resolução de problemas de relacionamento entre alunos(as)	A B C	Insuficiente a regular Regular a bom Bom a excelente	633	1	Alfanumérica
Q130	Avaliação sobre a escola quanto a iniciativa para apoiar a resolução de problemas de relacionamento entre alunos(as) e professores(as)	A B C	Insuficiente a regular Regular a bom Bom a excelente	634	1	Alfanumérica
Q131	Avaliação sobre a escola quanto a levar em conta seus problemas pessoais e familiares	A B C	Insuficiente a regular Regular a bom Bom a excelente	635	1	Alfanumérica
Q132	Avaliação sobre a escola quanto a realização de Programas e Palestras contra drogas	A B C	Insuficiente a regular Regular a bom Bom a excelente	636	1	Alfanumérica
Q133	Avaliação sobre a escola quanto a capacidade relacionar os conteúdos das matérias com o cotidiano	A B C	Insuficiente a regular Regular a bom Bom a excelente	637	1	Alfanumérica
Q134	Avaliação sobre a escola quanto a capacidade de a escola avaliar conhecimento, o que aprendeu	A B C	Insuficiente a regular Regular a bom Bom a excelente	638	1	Alfanumérica
Q135	Avaliação sobre a escola quanto o reconhecimento e valorização da identidade étnica dos(as) alunos(as)	A B C	Insuficiente a regular Regular a bom Bom a excelente	639	1	Alfanumérica
Q136	Nota para a formação que obteve no ensino médio	A B C D E F G H I J K L	0 1 2 3 4 5 6 7 8 9 10 Não sei	640	1	Alfanumérica
Q137	Considera-se racista	A B	Sim Não	641	1	Alfanumérica

DICIONÁRIO DAS VARIÁVEIS - ENEM 2007

Nome da Variável	Descrição	Variáveis Categóricas		Início	Tamanho	Tipo
		Categoria	Descrição			
Q138	Parentes racistas	A B	Sim Não	642	1	Alfanumérica
Q139	Amigos(as) racistas	A B	Sim Não	643	1	Alfanumérica
Q140	Colegas de escola e/ou de trabalho racistas	A B	Sim Não	644	1	Alfanumérica
Q141	Vizinhos e/ou conhecidos em geral racistas	A B	Sim Não	645	1	Alfanumérica
Q142	Já sofreu discriminação econômica	A B	Sim Não	646	1	Alfanumérica
Q143	Já sofreu discriminação étnica, racial ou de cor	A B	Sim Não	647	1	Alfanumérica
Q144	Já sofreu discriminação de gênero (ou por ser mulher ou por ser homem)	A B	Sim Não	648	1	Alfanumérica
Q145	Já sofreu discriminação por ser (ou parecer ser) homossexuais	A B	Sim Não	649	1	Alfanumérica
Q146	Já sofreu discriminação religiosa	A B	Sim Não	650	1	Alfanumérica
Q147	Já sofreu discriminação por causa do local de origem	A B	Sim Não	651	1	Alfanumérica
Q148	Já sofreu discriminação por causa da idade	A B	Sim Não	652	1	Alfanumérica
Q149	Já sofreu discriminação por ser portador de necessidades especiais	A B	Sim Não	653	1	Alfanumérica
Q150	Já sofreu discriminação por outro(s) motivo(s)	A B	Sim Não	654	1	Alfanumérica
Q151	Já presenciou discriminação econômica	A B	Sim Não	655	1	Alfanumérica
Q152	Já presenciou discriminação étnica, racial ou de cor	A B	Sim Não	656	1	Alfanumérica
Q153	Já presenciou discriminação de gênero (ou por ser mulher ou por ser homem)	A B	Sim Não	657	1	Alfanumérica
Q154	Já presenciou discriminação contra homossexuais	A B	Sim Não	658	1	Alfanumérica
Q155	Já presenciou discriminação religiosa	A B	Sim Não	659	1	Alfanumérica
Q156	Já presenciou discriminação por causa do local de origem	A B	Sim Não	660	1	Alfanumérica
Q157	Já presenciou discriminação por causa da idade	A B	Sim Não	661	1	Alfanumérica
Q158	Já presenciou discriminação por ser portador de necessidades especiais	A B	Sim Não	662	1	Alfanumérica
Q159	Já presenciou discriminação por outro(s) motivo(s)	A B	Sim Não	663	1	Alfanumérica
Q160	Se incomodaria se tivesse como parente ou colega de escola ou de trabalho uma pessoa de outra classe social	A B	Sim Não	664	1	Alfanumérica
Q161	Se incomodaria se tivesse como parente ou colega de escola ou de trabalho uma pessoa de outra cor ou etnia	A B	Sim Não	665	1	Alfanumérica
Q162	Se incomodaria se tivesse como parente ou colega de escola ou de trabalho uma pessoa de outra religião	A B	Sim Não	666	1	Alfanumérica

DICIONÁRIO DAS VARIÁVEIS - ENEM 2007

Nome da Variável	Descrição	Variáveis Categóricas		Início	Tamanho	Tipo
		Categoria	Descrição			
Q163	Se incomodaria se tivesse como parente ou colega de escola ou de trabalho uma pessoa com posições político-ideológicas diferentes	A B	Sim Não	667	1	Alfanumérica
Q164	Se incomodaria se tivesse como parente ou colega de escola ou de trabalho uma pessoa de outra origem geográfica	A B	Sim Não	668	1	Alfanumérica
Q165	Se incomodaria se tivesse como parente ou colega de escola ou de trabalho uma pessoa homossexual	A B	Sim Não	669	1	Alfanumérica
Q166	Se incomodaria se tivesse como parente ou colega de escola ou de trabalho uma pessoa muito mais nova ou mais velha	A B	Sim Não	670	1	Alfanumérica
Q167	Se incomodaria se tivesse como parente ou colega de escola ou de trabalho uma pessoa com necessidades especiais	A B	Sim Não	671	1	Alfanumérica
Q168	Participa de um Grêmio estudantil	A B	Sim Não	672	1	Alfanumérica
Q169	Participa de um Sindicato ou Associação Profissional	A B	Sim Não	673	1	Alfanumérica
Q170	Participa de um Grupo de bairro ou associação comunitária	A B	Sim Não	674	1	Alfanumérica
Q171	Participa de uma Igreja ou grupo religioso	A B	Sim Não	675	1	Alfanumérica
Q172	Participa de um partido político	A B	Sim Não	676	1	Alfanumérica
Q173	Participa de uma Ong ou movimento social	A B	Sim Não	677	1	Alfanumérica
Q174	Participa de um clube recreativo ou associação esportiva	A B	Sim Não	678	1	Alfanumérica
Q175	O quanto você se interessa pela política nacional, o papel dos(as) deputados(as) e senadores(as), o Presidente da República, etc	A B C	Muito Pouco Não me interesso	679	1	Alfanumérica
Q176	O quanto você se interessa pela política dos outros países	A B C	Muito Pouco Não me interesso	680	1	Alfanumérica
Q177	O quanto você se interessa pela economia nacional, a questão da inflação	A B C	Muito Pouco Não me interesso	681	1	Alfanumérica
Q178	O quanto você se interessa a política da sua cidade, o prefeito, os vereadores	A B C	Muito Pouco Não me interesso	682	1	Alfanumérica
Q179	O quanto você se interessa por esportes	A B C	Muito Pouco Não me interesso	683	1	Alfanumérica
Q180	O quanto você se interessa pelas questões sobre o meio ambiente, poluição, etc	A B C	Muito Pouco Não me interesso	684	1	Alfanumérica
Q181	O quanto você se interessa pelas questões sociais como a desigualdade, a pobreza, o desemprego, a miséria	A B C	Muito Pouco Não me interesso	685	1	Alfanumérica
Q182	O quanto você se interessa pelas questões sobre artes, teatro, cinema	A B C	Muito Pouco Não me interesso	686	1	Alfanumérica

DICIONÁRIO DAS VARIÁVEIS - ENEM 2007

Nome da Variável	Descrição	Variáveis Categóricas		Início	Tamanho	Tipo
		Categoria	Descrição			
Q183	O quanto você se interessa sobre a questão das drogas e suas conseqüências	A	Muito	687	1	Alfanumérica
		B	Pouco			
		C	Não me interesso			
Q184	O quanto você se interessa sobre assuntos sobre seu ídolo (cantor/a, artista, ou conjunto musical)	A	Muito	688	1	Alfanumérica
		B	Pouco			
		C	Não me interesso			
Q185	O quanto você se interessa sobre questões sociais como acesso aos servidores públicos de saúde e educação	A	Muito	689	1	Alfanumérica
		B	Pouco			
		C	Não me interesso			
Q186	O quanto você se interessa sobre sexualidade	A	Muito	690	1	Alfanumérica
		B	Pouco			
		C	Não me interesso			
SEUS VALORES						
Q187	Qual dos pontos te preocupa em 1° lugar no momento	A	O meio ambiente	691	1	Alfanumérica
		B	A aids e as doenças sexualmente transmissíveis			
		C	O racismo e a discriminação étnico-racial			
		D	A discriminação de gênero			
		E	A discriminação contra homossexuais			
		F	A discriminação etária			
		G	A discriminação religiosa e os conflitos religiosos			
		H	A desigualdade social no Brasil			
		I	A pobreza, as favelas, os meninos de rua			
		J	As drogas e a violência			
		K	A situação econômica do país			
		L	A precariedade dos serviços públicos de saúde e de educação			
Q188	Qual dos pontos te preocupa em 2° lugar no momento	A	O meio ambiente	692	1	Alfanumérica
		B	A aids e as doenças sexualmente transmissíveis			
		C	O racismo e a discriminação étnico-racial			
		D	A discriminação de gênero			
		E	A discriminação contra homossexuais			
		F	A discriminação etária			
		G	A discriminação religiosa e os conflitos religiosos			
		H	A desigualdade social no Brasil			
		I	A pobreza, as favelas, os meninos de rua			
		J	As drogas e a violência			
		K	A situação econômica do país			
		L	A precariedade dos serviços públicos de saúde e de educação			

DICIONÁRIO DAS VARIÁVEIS - ENEM 2007

Nome da Variável	Descrição	Variáveis Categóricas		Início	Tamanho	Tipo
		Categoria	Descrição			
Q189	É a mais importante contribuição que obteve ao realizar o ensino médio(2º grau)	A	Obtenção de um certificado de conclusão de curso / obtenção de um diploma	693	1	Alfanumérica
		B	Formação básica necessária para obter um emprego melhor			
		C	Condições de melhorar minha posição no emprego atual			
		D	Obtenção de cultura geral / ampliação de minha formação pessoal			
		E	Formação básica necessária para continuar os estudos em uma universidade / faculdade			
		F	Fazer muitos amigos / conhecer várias pessoas			
		G	Atender à expectativa de meus pais sobre meus estudos			
		H	Ajudar minha comunidade indígena			
Q190	É a segunda mais importante contribuição que obteve ao realizar o ensino médio(2º grau)	A	Obtenção de um certificado de conclusão de curso / obtenção de um diploma	694	1	Alfanumérica
		B	Formação básica necessária para obter um emprego melhor			
		C	Condições de melhorar minha posição no emprego atual			
		D	Obtenção de cultura geral / ampliação de minha formação pessoal			
		E	Formação básica necessária para continuar os estudos em uma universidade / faculdade			
		F	Fazer muitos amigos / conhecer várias pessoas			
		G	Atender à expectativa de meus pais sobre meus estudos			
		H	Ajudar minha comunidade indígena			
Q191	É a terceira mais importante contribuição que obteve ao realizar o ensino médio(2º grau)	A	Obtenção de um certificado de conclusão de curso / obtenção de um diploma	695	1	Alfanumérica
		B	Formação básica necessária para obter um emprego melhor			
		C	Condições de melhorar minha posição no emprego atual			
		D	Obtenção de cultura geral / ampliação de minha formação pessoal			
		E	Formação básica necessária para continuar os estudos em uma universidade / faculdade			
		F	Fazer muitos amigos / conhecer várias pessoas			
		G	Atender à expectativa de meus pais sobre meus estudos			
		H	Ajudar minha comunidade indígena			
Q192	A principal decisão que vai tomar quando concluir o ensino médio (2º grau)	A	Já conclui o ensino médio	696	1	Alfanumérica
		B	Prestar vestibular e continuar os estudos no ensino superior			
		C	Procurar um emprego			
		D	Prestar vestibular e continuar a trabalhar			
		E	Fazer curso(s) profissionalizante(s) e me preparar para o trabalho			
		F	Trabalhar por conta própria / trabalhar em negócio da família			
		G	Trabalhar em atividade ligada à comunidade indígena			
		H	Ainda não decidiu			
		I	Outro plano			

DICIONÁRIO DAS VARIÁVEIS - ENEM 2007

Nome da Variável	Descrição	Variáveis Categóricas		Início	Tamanho	Tipo
		Categoria	Descrição			
Q193	E a médio prazo, daqui a uns 4 ou 5 anos já planejou o que gostaria que acontecesse	A	Gostaria de ter um diploma universitário para conseguir um bom emprego	697	1	Alfanumérica
		B	Gostaria de prestar um concurso e trabalhar no setor público			
		C	Gostaria de ganhar dinheiro em meu próprio negócio			
		D	Gostaria de ter um emprego			
		E	Gostaria de estar envolvido em projeto de desenvolvimento de minha comunidade indígena			
		F	Não planejei			
		G	Outro plano			
Q194	Que profissão escolheu seguir	A	Ainda não escolhi	698	1	Alfanumérica
		B	Profissão ligada às Engenharias / Ciências Tecnológicas			
		C	Profissão ligada às Ciências Humanas			
		D	Profissão ligada às Artes			
		E	Profissão ligadas às Ciências Biológicas e de Saúde			
		F	Professor(a) de Ensino Fundamental e Médio (1° e 2° graus)			
		G	Não vou seguir nenhuma profissão			
Q195	Meus pais ajudaram a tomar minha decisão sobre minha profissão	A	Ajudou muito	699	1	Alfanumérica
		B	Ajudou pouco			
		C	Não ajudou			
Q196	A escola ajudou a tomar minha decisão sobre minha profissão	A	Ajudou muito	700	1	Alfanumérica
		B	Ajudou pouco			
		C	Não ajudou			
Q197	Meus amigos ajudaram a tomar minha decisão sobre minha profissão	A	Ajudou muito	701	1	Alfanumérica
		B	Ajudou pouco			
		C	Não ajudou			
Q198	Informações gerais, revistas, jornais, TV ajudaram a tomar minha decisão sobre minha profissão	A	Ajudou muito	702	1	Alfanumérica
		B	Ajudou pouco			
		C	Não ajudou			
Q199	Meu trabalho ajudou a tomar minha decisão sobre minha profissão	A	Ajudou muito	703	1	Alfanumérica
		B	Ajudou pouco			
		C	Não ajudou			
Q200	Estímulo financeiro ajudou a tomar minha decisão sobre minha profissão	A	Ajudou muito	704	1	Alfanumérica
		B	Ajudou pouco			
		C	Não ajudou			
Q201	Facilidade de obter emprego ajudou a tomar minha decisão sobre minha profissão	A	Ajudou muito	705	1	Alfanumérica
		B	Ajudou pouco			
		C	Não ajudou			
Q202	Minha identificação com a profissão me ajudou a tomar minha decisão sobre minha profissão	A	Ajudou muito	706	1	Alfanumérica
		B	Ajudou pouco			
		C	Não ajudou			
Q203	Querer ajudar minha comunidade indígena ajudou a tomar minha decisão sobre minha profissão	A	Ajudou muito	707	1	Alfanumérica
		B	Ajudou pouco			
		C	Não ajudou			
EGRESSOS						
Q204	Continuou os estudos depois de ter concluído o ensino médio (2° grau)	A	Sim, estou estudando no momento atual	708	1	Alfanumérica
		B	Sim, mas não estou estudando no momento atual			
		C	Não			
Q205	Está freqüentando um curso profissionalizante	A	Sim	709	1	Alfanumérica
		B	Não			

DICIONÁRIO DAS VARIÁVEIS - ENEM 2007

Nome da Variável	Descrição	Variáveis Categóricas		Início	Tamanho	Tipo
		Categoria	Descrição			
Q206	Está freqüentando um curso preparatório para vestibular	A B	Sim Não	710	1	Alfanumérica
Q207	Está freqüentando um curso superior	A B	Sim Não	711	1	Alfanumérica
Q208	Está freqüentando um curso de língua estrangeira	A B	Sim Não	712	1	Alfanumérica
Q209	Está freqüentando um curso de computação ou informática	A B	Sim Não	713	1	Alfanumérica
Q210	Está freqüentando outro curso	A B	Sim Não	714	1	Alfanumérica
Q211	Concluiu curso profissionalizante	A B	Sim Não	715	1	Alfanumérica
Q212	Concluiu curso preparatório para vestibular, mas não ingressei no curso superior	A B	Sim Não	716	1	Alfanumérica
Q213	Concluiu curso superior	A B	Sim Não	717	1	Alfanumérica
Q214	Fiz curso superior mas não me formei	A B	Sim Não	718	1	Alfanumérica
Q215	Concluiu curso de língua estrangeira	A B	Sim Não	719	1	Alfanumérica
Q216	Concluiu curso de computação ou informática	A B	Sim Não	720	1	Alfanumérica
Q217	Concluiu outro curso	A B	Sim Não	721	1	Alfanumérica
Q218	O curso profissionalizante fez mais falta para minha vida, depois que terminei o ensino médio	A B	Sim Não	722	1	Alfanumérica
Q219	O curso preparatório para vestibular fez mais falta para minha vida, depois que terminei o ensino médio	A B	Sim Não	723	1	Alfanumérica
Q220	O curso superior fez mais falta para minha vida, depois que terminei o ensino médio	A B	Sim Não	724	1	Alfanumérica
Q221	O curso de língua estrangeira fez mais falta para minha vida, depois que terminei o ensino médio	A B	Sim Não	725	1	Alfanumérica
Q222	O curso de computaçao ou informática fez mais falta para minha vida, depois que terminei o ensino médio	A B	Sim Não	726	1	Alfanumérica
Q223	Outro curso fez mais falta para minha vida, depois que terminei o ensino médio	A B	Sim Não	727	1	Alfanumérica
VT_RESP_OBJETIVA	Vetor com as respostas da parte objetiva da prova (Obs: as 63 primeiras possições deste campo são refrentes as rescpectivas respostas)			728	100	Alfanumérica
TP_PROVA	Tipo de prova	1 2 3 4	Amarela Azul Branca Rosa	828	1	Alfanumérica
VT_GABARITO_PROVA	Vetor com o gabarito da parte objetiva da prova (Obs: as 63 primeiras possições deste campo são refrentes ao respectivo gabarito)			829	100	Alfanumérica

Apêndice B

Tabela demonstrativa dos campos disponíveis para a criação do DW

Tabela B.1: Esta tabela demonstra os campos que foram selecionados dos microdados, a decrição do campo, os valores que esses campos assumiam e a descrição dos valores que os campos assumiam.

Nome	Descrição	Dados	Descrição Dados
1-nu_ano	Ano do Enem	1998,1999...2008	-
2-dt_nascimento	Data nascimento	Data Ex.: "01/01/1983"	-
3-nu_inscrito	Número gerado como código de acesso do participante	código Ex.: "00000029"	-
4-cod_municipio	Código do município do participante	Vários códigos correspondente a região tais como: região, uf, mesoregião, microregião, município, dv	-
5-tp_prova	Tipo de prova	Múltipla escolha: "1-2,2-5, 3-3, 4-1,5-6, 6-4"	1-Branca, 2-Grafite, 3-Amarela, 4-Azul, 5-Rosa, 6-Verde
6-in_presenca	Presente a prova objetiva	Múltipla escolha: "1-1,2-0"	0-Faltou à prova, 1-Presente à prova
7-nu_nota _objetiva	Nota da prova objetiva	Valor Ex.: 10,00, 55,00	-
8-vl_perc_comp1	Nota da competência 1 - Objetiva	Valor Ex.: 10,00, 55,00	-

Tabela B.1 – Continuação

Nome	Descrição	Dados	Descrição Dados
9-vl_perc_comp2	Nota da compe tência 2 - Objetiva	Valor Ex.: 10,00, 55,00	-
10-vl_perc_comp3	Nota da competência 3 - Objetiva	Valor Ex.: 10,00, 55,00	-
11-vl_perc_comp4	Nota da competência 4 - Objetiva	Valor Ex.: 10,00, 55,00	-
12-vl_perc_comp5	Nota da competência 5 - Objetiva	Valor Ex.: 10,00, 55,00	-
13-in_situacao	Presente à redação	Múltipla escolha: "1-F,2-B, 3-P, 4 -N"	B-Entregou a redação em branco, F-Faltou à prova, N-Redação anulada ,P -Presente à prova
14-nu_nota_global _redacao	Nota da prova de redação	Valor Ex.: 10,00, 55,00	-
15-nu_nota_redacao _comp1	Nota da competência 1 - Redação	Valor Ex.: 10,00, 55,00	-
16-nu_nota_redacao _comp2	Nota da competência 2 - Redação	Valor Ex.: 10,00, 55,00	-
17-nu_nota_redacao _comp3	Nota da competência 3 - Redação	Valor Ex.: 10,00, 55,00	-
18-nu_nota_redacao _comp4	Nota da competência 4 - Redação	Valor Ex.: 10,00, 55,00	-
19-nu_nota_redacao _comp5	Nota da competência 5 - Redação	Valor Ex.: 10,00, 55,00	-
20-in_qse	Resposta ao Questionário Socioeconômico	Múltipla escolha: "1-1, 2-0"	0-Não respondeu o questionário socioecônomico, 1-Respondeu o questionário socioeconômico
21-q1	Sexo	Múltipla escolha: "1-M, 2-F"	-
22-q3	Como se considera Cor	Múltipla escolha: "1-B,2-C,3-A,4-D, 5-E"	A-Branco(a), B-Pardo(a)/ mulato(a), C-Negro(a), D-Amarelo (a)(De origem asiática), E-Indígena
23-q4	Qual o estado civil	Múltipla escolha: "1-B,2-A, 3-D, 4-C"	A-Solteiro(a), B-Casado(a)/mora com um companheiro(a), C-Separado(a)/divorciado(a)/desquitado(a), D-Viúvo(a)

Tabela B.1 – Continuação

Nome	Descrição	Dados	Descrição Dados
24-q15	Quantidade de pessoas que moram em sua casa	Múltipla escolha: "1-B,2-F, 3-G, 4-C, 5-E, 6-D, 7-A"	A-Moro sozinho, B-Duas pessoas, C-Três pessoas, D-Quatro pessoas, E-Cinco pessoas, F-Seis pessoas, G-Mais de seis pessoas
25-q24	Tem TV e quantas	Múltipla escolha: "1-C,2-B,3-A, 4-D"	A-1, B-2, C-3 ou +, D-Não tem
26-q25	Tem Videocassete e/ou DVD e quantos	Múltipla escolha: "1-C,2-B,3-A, 4-D"	A-1, B-2, C-3 ou +, D-Não tem
27-q27	Tem Microcomputador e quantos	Múltipla escolha: "1-C,2-B,3-A, 4-D"	A-1, B-2, C-3 ou +, D-Não tem
28-q28	Tem Automóvel	Múltipla escolha: "1-C,2-B,3-A, 4-D"	A-1, B-2, C-3 ou +, D-Não tem
29-q29	Tem Máquina de lavar roupa e quantos	Múltipla escolha: "1-C,2-B,3-A, 4-D"	A-1, B-2, C-3 ou +, D-Não tem
30-q35	Tem casa própria	Múltipla escolha: "1-B,2-A"	A-Sim,B-Não
31-q43	Trabalha, ou já trabalhou, ganhando algum salário ou rendimento	Múltipla escolha: "1-B,2-C,3-A"	A-Sim,B-Sim, eventualmente,C-Não
32-q44	Quantas horas trabalhava, durante o ensino médio (2º grau)	Múltipla escolha: "1-C,2-B,3-A"	A-Até 20 horas semanais, B-De 21 a 30 horas semanais, C-De 31 a 40 horas semanais
33-q46	Com que idade começou a exercer atividade remunerada	Múltipla escolha: "1-B,2-A, 3-D, 4-C"	A-Antes dos 14 anos, B-Entre 14 e 16 anos, C-Entre 17 e 18 anos, D-mais de 18
34-q47	Se está trabalhando atualmente, qual é a renda ou o	Múltipla escolha: "1-H,2-C, 3-G, 4-E,5-A, 6-D, 7-B, 8-F"	A-Até 1 salário mínimo, B-Entre 1 e 2 salário mínimos, C-De 2 a 5 salários mínimos,

Tabela B.1 – Continuação

Nome	Descrição	Dados	Descrição Dados
	seu salário mensal		D-De 5 a 10 salário mínimos, E-De 10 a 3 0 salários mínimos, F-De 30 a 50 salários mínimos, G-Mais de 5 0 salários mínimos, H-Não estou trabalhando
35-q72	Anos que levou para concluir o ensino fundamental (1º grau)	Múltipla escolha: "1-B,2-F, 3-C, 4-E,5-A, 6-D"	A-Menos de 8 anos, B-8 anos,C-9 anos, D-10 anos, E-11 anos, F-mais de 11
36-q79	Fez curso de língua estrangeira fora da escola durante o ensino médio (2º grau)	Múltipla escolha: "1-B,2-A"	A-Sim,B-Não
37-q80	Fez curso de computação ou informática fora da escola durante o ensino médio	Múltipla escolha: "1-B,2-A"	A-Sim,B-Não
38-q81	Fez curso preparatório para o vestibular (cursinho) fora da escola durante o ensino médio	Múltipla escolha: "1-B,2-A"	A-Sim,B-Não
39-q82	Fez artes plásticas em geral fora da escola durante o ensino médio	Múltipla escolha: "1-B,2-A"	A-Sim,B-Não
40-q85	Com que frequência lê jornais	Múltipla escolha: "1-C,2-B,3-A"	A-Frequentemente, B-Às vezes, C-nunca
41-q86	Com que frequência lê revistas de informação geral	Múltipla escolha: "1-C,2-B,3-A"	A-Frequentemente, B-Às vezes ,C-nunca
42-q87	Com que frequência lê revistas de humor / quadrinhos	Múltipla escolha: "1-C,2-B,3-A"	A-Frequentemente, B-Às vezes, C-nunca
43-q88	Com que frequência lê revistas de divulgação científica (Ciência Hoje, Super Interessante, etc.)	Múltipla escolha: "1-C,2-B,3-A"	A-Frequentemente, B-Às vezes, C-nunca
44-q89	Com que frequência	Múltipla escolha:	A-Frequentemente,

Tabela B.1 – Continuação

Nome	Descrição	Dados	Descrição Dados
	lê romances, livros de ficção	"1-C,2-B,3-A"	B-Às vezes, C-nunca
45-q208	Está frequentando um curso de línguas	Múltipla escolha: "1-B,2-A"	A-Sim,B-Não
46-q209	Está frequentando um curso de computação ou informática	Múltipla escolha: "1-B,2-A"	A-Sim,B-Não
47-vt_resp _objetiva	Vetor com as respostas da parte objetiva da prova	Vetor de respostas Ex.: ("ACEC...")	-
48-vt_gabarito _prova	Vetor com o gababarito da parte objetiva da prova	Vetor de garito Ex.: ("ACEC...")	-

Apêndice C

Consultas implementadas para validação dos DW's

Consulta C.1: Média das notas de redação e prova objetiva agrupados por ano e região para o DW do ENEM

```
select
 round(CAST (avg(f.media_redacao) AS NUMERIC), 2)  redacao,
 round(CAST (avg(f.media_objetiva) AS NUMERIC), 2) objetiva,
 reg.descr_regiao regiao,
 ano.ano ano
from dw.fato_enem f
 inner join dw.dim_regiao reg on f.id_dim_regiao
  = reg.id_dim_regiao
 inner join dw.dim_ano    ano on f.id_dim_ano
  = ano.id_dim_ano
where
 (media_redacao <> 0 or media_objetiva <> 0)
 and reg.descr_regiao is not null
group by reg.descr_regiao, ano.ano
order by ano.ano, reg.descr_regiao;
```

Consulta C.2: Quantidade de alunos que tiraram zero na redação agrupados por ano e região para o DW do ENEM

```
select
    count(f.media_redacao)  qtdRedacaoZeros,
    reg.descr_regiao regiao,
    ano.ano ano
from dw.fato_enem f
    inner join dw.dim_regiao reg on f.id_dim_regiao = reg.id_dim_regiao
    inner join dw.dim_ano    ano on f.id_dim_ano = ano.id_dim_ano
where
    (media_redacao = 0 or media_objetiva <> 0)
    and reg.descr_regiao is not null
group by reg.descr_regiao, ano.ano
order by ano.ano, reg.descr_regiao;
```

Consulta C.3: Média das notas de redação e prova objetiva agrupados por ano e regiao para as tabelas do ENEM que compreendem os anos de 1998-2008

```
select
 round(CAST (avg(x.mediaObjetiva) AS NUMERIC), 2)  mediaRedacao,
 round(CAST (avg(x.mediaRedacao) AS NUMERIC), 2) mediaObjetiva,
 case
  when x.regiao = '1' then 'NORTE'
  when x.regiao = '2' then 'NORDESTE'
  when x.regiao = '3' then 'SUDESTE'
  when x.regiao = '4' then 'SUL'
  when x.regiao = '5' then 'CENTRO-OESTE'
 end regiao,
 x.ano ano
from (
 (select
  round(CAST (avg(e1998.nu_nota_objetiva) AS NUMERIC), 2)  mediaRedacao,
  round(CAST (avg(e1998.nu_nota_global_redacao) AS NUMERIC), 2) mediaObjetiva,
  substring(e1998.codmunic_insc from 1 for 1) regiao,
  '1998' as ano
 from enem1998_conceitos_ies e1998
 where  (e1998.nu_nota_objetiva <> 0 or e1998.nu_nota_global_redacao <> 0)
  and substring(e1998.codmunic_insc from 1 for 1) is not null
 group by substring(e1998.codmunic_insc from 1 for 1), ano)

 union

 (select
  round(CAST (avg(e1999.nu_nota_objetiva) AS NUMERIC), 2)  mediaRedacao,
  round(CAST (avg(e1999.nu_nota_global_redacao) AS NUMERIC), 2) mediaObjetiva,
  substring(e1999.codmunic_insc from 1 for 1) regiao,
  '1999' as ano
 from enem1999_conceitos_ies e1999
 where  (e1999.nu_nota_objetiva <> 0 or e1999.nu_nota_global_redacao <> 0)
  and substring(e1999.codmunic_insc from 1 for 1) is not null
 group by substring(e1999.codmunic_insc from 1 for 1), ano)

 union

 (select
  round(CAST (avg(e2000.nu_nota_objetiva) AS NUMERIC), 2)  mediaRedacao,
  round(CAST (avg(e2000.nu_nota_global_redacao) AS NUMERIC), 2) mediaObjetiva,
  substring(e2000.codmunic_insc from 1 for 1) regiao,
  '2000' as ano
 from enem2000_conceitos_ies e2000
 where  (e2000.nu_nota_objetiva <> 0 or e2000.nu_nota_global_redacao <> 0)
  and substring(e2000.codmunic_insc from 1 for 1) is not null
 group by substring(e2000.codmunic_insc from 1 for 1), ano)

 union

 (select
  round(CAST (avg(e2001.nu_nota_objetiva) AS NUMERIC), 2)  mediaRedacao,
  round(CAST (avg(e2001.nu_nota_global_redacao) AS NUMERIC), 2) mediaObjetiva,
  substring(e2001.codmunic_insc from 1 for 1) regiao,
  '2001' as ano
 from enem2001_conceitos_ies e2001
 where  (e2001.nu_nota_objetiva <> 0 or e2001.nu_nota_global_redacao <> 0)
  and substring(e2001.codmunic_insc from 1 for 1) is not null
```

```
group by substring(e2001.codmunic_insc from 1 for 1), ano)

union

(select
 round(CAST (avg(e2002.nu_nota_objetiva) AS NUMERIC), 2)  mediaRedacao,
 round(CAST (avg(e2002.nu_nota_global_redacao) AS NUMERIC), 2) mediaObjetiva,
 substring(e2002.codmunic_insc from 1 for 1) regiao,
 '2002' as ano
from enem2002_conceitos_ies e2002
where  (e2002.nu_nota_objetiva <> 0 or e2002.nu_nota_global_redacao <> 0)
 and substring(e2002.codmunic_insc from 1 for 1) is not null
group by substring(e2002.codmunic_insc from 1 for 1), ano)

union

(select
 round(CAST (avg(e2003.nu_nota_objetiva) AS NUMERIC), 2)  mediaRedacao,
 round(CAST (avg(e2003.nu_nota_global_redacao) AS NUMERIC), 2) mediaObjetiva,
 substring(e2003.codmunic_insc from 1 for 1) regiao,
 '2003' as ano
from enem2003_conceitos_ies e2003
where  (e2003.nu_nota_objetiva <> 0 or e2003.nu_nota_global_redacao <> 0)
 and substring(e2003.codmunic_insc from 1 for 1) is not null
group by substring(e2003.codmunic_insc from 1 for 1), ano)

union

(select
 round(CAST (avg(e2004.nu_nota_objetiva) AS NUMERIC), 2)  mediaRedacao,
 round(CAST (avg(e2004.nu_nota_global_redacao) AS NUMERIC), 2) mediaObjetiva,
 substring(e2004.codmunic_insc from 1 for 1) regiao,
 '2004' as ano
from enem2004_conceitos_ies e2004
where  (e2004.nu_nota_objetiva <> 0 or e2004.nu_nota_global_redacao <> 0)
 and substring(e2004.codmunic_insc from 1 for 1) is not null
group by substring(e2004.codmunic_insc from 1 for 1), ano)

union

(select
 round(CAST (avg(e2005.nu_nota_objetiva) AS NUMERIC), 2)  mediaRedacao,
 round(CAST (avg(e2005.nu_nota_global_redacao) AS NUMERIC), 2) mediaObjetiva,
 substring(e2005.codmunic_insc from 1 for 1) regiao,
 '2005' as ano
from enem2005_conceitos_ies e2005
where  (e2005.nu_nota_objetiva <> 0 or e2005.nu_nota_global_redacao <> 0)
 and substring(e2005.codmunic_insc from 1 for 1) is not null
group by substring(e2005.codmunic_insc from 1 for 1), ano)

union

(select
 round(CAST (avg(e2006.nu_nota_objetiva) AS NUMERIC), 2)  mediaRedacao,
 round(CAST (avg(e2006.nu_nota_global_redacao) AS NUMERIC), 2) mediaObjetiva,
 substring(e2006.codmunic_insc from 1 for 1) regiao,
 '2006' as ano
from enem2006_conceitos_ies e2006
where  (e2006.nu_nota_objetiva <> 0 or e2006.nu_nota_global_redacao <> 0)
```

```
  and substring(e2006.codmunic_insc from 1 for 1) is not null
 group by substring(e2006.codmunic_insc from 1 for 1), ano)

 union

 (select
  round(CAST (avg(e2007.nu_nota_objetiva) AS NUMERIC), 2)  mediaRedacao,
  round(CAST (avg(e2007.nu_nota_global_redacao) AS NUMERIC), 2) mediaObjetiva,
  substring(trim(to_char(cod_municipio_insc, '9990999')) from 1 for 1) regiao,
  '2007' as ano
 from enem2007_conceitos_ies e2007
 where  (e2007.nu_nota_objetiva <> 0 or e2007.nu_nota_global_redacao <> 0)
  and cod_municipio_insc is not null
 group by substring(trim(to_char(cod_municipio_insc, '9990999')) from 1 for 1),
     ano)

 union

 (select
  round(CAST (avg(e2008.nu_nt_objetiva) AS NUMERIC), 2)  mediaRedacao,
  round(CAST (avg(e2008.nu_nota_redacao) AS NUMERIC), 2) mediaObjetiva,
   substring(trim(to_char(cod_munic_insc, '9990999')) from 1 for 1) regiao,
  '2008' as ano
 from enem2008_conceitos_ies e2008
 where  (e2008.nu_nt_objetiva <> 0 or e2008.nu_nota_redacao <> 0)
  and e2008.cod_munic_insc is not null
 group by  substring(trim(to_char(e2008.cod_munic_insc, '9990999')) from 1 for 1),
      ano)
)x
group by x.regiao, x.ano
order by x.regiao asc, x.ano  desc;
```

Consulta C.4: Quantidade de alunos que tiraram zero na redação agrupados por ano e região para as tabelas do ENEM que compreendem os anos de 1998-2008

```
select
    x.qtdRedacaoZeros,
    case
        when x.regiao = '1' then 'NORTE'
        when x.regiao = '2' then 'NORDESTE'
        when x.regiao = '3' then 'SUDESTE'
        when x.regiao = '4' then 'SUL'
        when x.regiao = '5' then 'CENTRO-OESTE'
    end regiao,
    x.ano
from(
    (select
        count(e1998.nu_nota_global_redacao)  qtdRedacaoZeros,
        substring(e1998.codmunic_insc from 1 for 1) regiao,
        '1998' as ano
    from enem1998_conceitos_ies e1998
    where  (e1998.nu_nota_objetiva <> 0
   or e1998.nu_nota_global_redacao = 0)
        and substring(e1998.codmunic_insc from 1 for 1) is not null
    group by substring(e1998.codmunic_insc from 1 for 1), ano)

    union
```

```
 (select
     count(e1999.nu_nota_global_redacao)  qtdRedacaoZeros,
     substring(e1999.codmunic_insc from 1 for 1) regiao,
     '1999' as ano
 from enem1999_conceitos_ies e1999
 where  (e1999.nu_nota_objetiva <> 0
or e1999.nu_nota_global_redacao = 0)
     and substring(e1999.codmunic_insc from 1 for 1) is not null
 group by substring(e1999.codmunic_insc from 1 for 1), ano)

 union

 (select
     count(e2000.nu_nota_global_redacao)  qtdRedacaoZeros,
     substring(e2000.codmunic_insc from 1 for 1) regiao,
     '2000' as ano
 from enem2000_conceitos_ies e2000
 where  (e2000.nu_nota_objetiva <> 0
or e2000.nu_nota_global_redacao = 0)
     and substring(e2000.codmunic_insc from 1 for 1) is not null
 group by substring(e2000.codmunic_insc from 1 for 1), ano)

 union

 (select
     count(e2001.nu_nota_global_redacao)  qtdRedacaoZeros,
     substring(e2001.codmunic_insc from 1 for 1) regiao,
     '2001' as ano
 from enem2001_conceitos_ies e2001
 where  (e2001.nu_nota_objetiva <> 0
or e2001.nu_nota_global_redacao = 0)
     and substring(e2001.codmunic_insc from 1 for 1) is not null
 group by substring(e2001.codmunic_insc from 1 for 1), ano)

 union

 (select
     count(e2002.nu_nota_global_redacao)  qtdRedacaoZeros,
     substring(e2002.codmunic_insc from 1 for 1) regiao,
     '2002' as ano
 from enem2002_conceitos_ies e2002
 where  (e2002.nu_nota_objetiva <> 0 or
e2002.nu_nota_global_redacao = 0)
     and substring(e2002.codmunic_insc from 1 for 1) is not null
 group by substring(e2002.codmunic_insc from 1 for 1), ano)

 union

 (select
     count(e2003.nu_nota_global_redacao)  qtdRedacaoZeros,
     substring(e2003.codmunic_insc from 1 for 1) regiao,
     '2003' as ano
 from enem2003_conceitos_ies e2003
 where  (e2003.nu_nota_objetiva <> 0
or e2003.nu_nota_global_redacao = 0)
     and substring(e2003.codmunic_insc from 1 for 1) is not null
 group by substring(e2003.codmunic_insc from 1 for 1), ano)

 union
```

```
 (select
     count(e2004.nu_nota_global_redacao)  qtdRedacaoZeros,
     substring(e2004.codmunic_insc from 1 for 1) regiao,
     '2004' as ano
 from enem2004_conceitos_ies e2004
 where  (e2004.nu_nota_objetiva <> 0
or e2004.nu_nota_global_redacao = 0)
     and substring(e2004.codmunic_insc from 1 for 1) is not null
 group by substring(e2004.codmunic_insc from 1 for 1), ano)

 union

 (select
     count(e2005.nu_nota_global_redacao)  qtdRedacaoZeros,
     substring(e2005.codmunic_insc from 1 for 1) regiao,
     '2005' as ano
 from enem2005_conceitos_ies e2005
 where  (e2005.nu_nota_objetiva <> 0
or e2005.nu_nota_global_redacao = 0)
     and substring(e2005.codmunic_insc from 1 for 1) is not null
 group by substring(e2005.codmunic_insc from 1 for 1), ano)

 union

 (select
     count(e2006.nu_nota_global_redacao)  qtdRedacaoZeros,
     substring(e2006.codmunic_insc from 1 for 1) regiao,
     '2006' as ano
 from enem2006_conceitos_ies e2006
 where  (e2006.nu_nota_objetiva <> 0
or e2006.nu_nota_global_redacao = 0)
     and substring(e2006.codmunic_insc from 1 for 1) is not null
 group by substring(e2006.codmunic_insc from 1 for 1), ano)

 union

 (select
     count(e2007.nu_nota_global_redacao)  qtdRedacaoZeros,
     substring(trim(to_char(cod_municipio_insc,
'9990999')) from 1 for 1) regiao,
     '2007' as ano
 from enem2007_conceitos_ies e2007
 where  (e2007.nu_nota_objetiva <> 0
or e2007.nu_nota_global_redacao = 0)
     and substring(trim(to_char(cod_municipio_insc,
'9990999')) from 1 for 1) is not null
 group by substring(trim(to_char(cod_municipio_insc,
'9990999')) from 1 for 1), ano)

 union

 (select
     count(e2008.nu_nota_redacao)  qtdRedacaoZeros,
     substring(trim(to_char(cod_munic_insc,
'9990999')) from 1 for 1) regiao,
     '2008' as ano
 from enem2008_conceitos_ies e2008
 where  (e2008.nu_nt_objetiva <> 0
```

```
   or e2008.nu_nota_redacao = 0)
        and substring(trim(to_char(cod_munic_insc,
  '9990999')) from 1 for 1) is not null
    group by substring(trim(to_char(cod_munic_insc,
  '9990999')) from 1 for 1), ano)

)x
order by ano, regiao
```

Apêndice D

Ontologias utilizadas no Babel

XML D.1: XML Ontologia Babel

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/2/22-rdf-syntax-ns#"
    xmlns:rdfs="http://www.w3.org/2/1/rdf-schema#"
          xmlns:owl="http://www.w3.org/22/7/owl#"
          xmlns:dc="http://purl.org/dc/elements/1.1/">

 <!-- OWL Header -->
 <owl:Ontology rdf:about="http://webpide.ledes.net/ontology">
  <dc:title>Ontology for enem database</dc:title>
  <dc:description>Ontology for enem database</dc:description>
 </owl:Ontology>

 <!-- OWL Class Definition FOR YEARS-->
 <owl:Class rdf:about=http://webpide.ledes.net/ontology/Years">
  <rdfs:label>Years</rdfs:label>
  <rdfs:comment>Contains all the years Enem</rdfs:comment>
  <rdfs:isDefinedBy>http://webpide.ledes.net/ontology</rdfs:isDefinedBy>
 </owl:Class>

 <!-- OWL ObjectProperty Definition FOR YEAR-->
 <owl:ObjectProperty rdf:about="http://webpide.ledes.net/ontology/Years#Year">
  <rdfs:label>Year</rdfs:label>
  <rdfs:comment>Each year of application Enem</rdfs:comment>
  <rdfs:domain>http://webpide.ledes.net/ontology/Years</rdfs:domain>
  <rdfs:isDefinedBy>http://webpide.ledes.net/ontology</rdfs:isDefinedBy>
 </owl:Class>

 <!-- OWL ObjectProperty Definition FOR Type-->
 <owl:ObjectProperty rdf:about="http://webpide.ledes.net/ontology/Years#Type">
  <rdfs:label>Type</rdfs:label>
  <rdfs:comment>Tipo de dados contido: Dimensao ou Measure</rdfs:comment>
  <rdfs:domain>http://webpide.ledes.net/ontology/Years</rdfs:domain>
  <rdfs:isDefinedBy>http://webpide.ledes.net/ontology</rdfs:isDefinedBy>
 </owl:Class>

 <!-- OWL Class Definition FOR GENDERS-->
 <owl:Class rdf:about=http://webpide.ledes.net/ontology/Genders">
  <rdfs:label>Genders</rdfs:label>
  <rdfs:comment>Contains all the genders Enem</rdfs:comment>
  <rdfs:isDefinedBy>http://webpide.ledes.net/ontology</rdfs:isDefinedBy>
```

```
</owl:Class>

<!-- OWL ObjectProperty Definition FOR GENDER-->
<owl:ObjectProperty rdf:about="http://webpide.ledes.net/ontology/Genders#Gender">
 <rdfs:label>Gender</rdfs:label>
 <rdfs:comment>Each gender of application Enem</rdfs:comment>
 <rdfs:domain>http://webpide.ledes.net/ontology/Genders</rdfs:domain>
 <rdfs:isDefinedBy>http://webpide.ledes.net/ontology</rdfs:isDefinedBy>
</owl:Class>

<!-- OWL Class Definition FOR AVERAGE WRITINGS-->
<owl:Class rdf:about=http://webpide.ledes.net/ontology/AverageWritings">
 <rdfs:label>AverageWritings</rdfs:label>
 <rdfs:comment>Contains all the average writings of Enem</rdfs:comment>
 <rdfs:isDefinedBy>http://webpide.ledes.net/ontology</rdfs:isDefinedBy>
</owl:Class>

<!-- OWL ObjectProperty Definition FOR AVERAGE WRITING-->
<owl:ObjectProperty rdf:about="http://webpide.ledes.net/ontology/AverageWritings#
    AverageWriting">
 <rdfs:label>AverageWriting</rdfs:label>
 <rdfs:comment>Each average writing of application Enem</rdfs:comment>
 <rdfs:domain>http://webpide.ledes.net/ontology/AverageWritings</rdfs:domain>
 <rdfs:isDefinedBy>http://webpide.ledes.net/ontology</rdfs:isDefinedBy>
</owl:Class>

<!-- OWL ObjectProperty Definition FOR Type-->
<owl:ObjectProperty rdf:about="http://webpide.ledes.net/ontology/AverageWritings#
    Type">
 <rdfs:label>Type</rdfs:label>
 <rdfs:comment>Tipo de dados contido: Dimensao ou Measure</rdfs:comment>
 <rdfs:domain>http://webpide.ledes.net/ontology/AverageWritings</rdfs:domain>
 <rdfs:isDefinedBy>http://webpide.ledes.net/ontology</rdfs:isDefinedBy>
</owl:Class>

<!-- OWL Class Definition FOR AVERAGE Objectives-->
<owl:Class rdf:about=http://webpide.ledes.net/ontology/AverageObjectives">
 <rdfs:label>AverageObjectives</rdfs:label>
 <rdfs:comment>Contains all the average objectives of Enem</rdfs:comment>
 <rdfs:isDefinedBy>http://webpide.ledes.net/ontology</rdfs:isDefinedBy>
</owl:Class>

<!-- OWL ObjectProperty Definition FOR AVERAGE Objective-->
<owl:ObjectProperty rdf:about="http://webpide.ledes.net/ontology/
    AverageObjectives#AverageObjective">
 <rdfs:label>AverageObjective</rdfs:label>
 <rdfs:comment>Each average objective of application Enem</rdfs:comment>
 <rdfs:domain>http://webpide.ledes.net/ontology/AverageObjectives</rdfs:domain>
 <rdfs:isDefinedBy>http://webpide.ledes.net/ontology</rdfs:isDefinedBy>
</owl:Class>

<!-- OWL ObjectProperty Definition FOR Type-->
<owl:ObjectProperty rdf:about="http://webpide.ledes.net/ontology/
    AverageObjectives#Type">
 <rdfs:label>Type</rdfs:label>
 <rdfs:comment>Tipo de dados contido: Dimensao ou Measure</rdfs:comment>
 <rdfs:domain>http://webpide.ledes.net/ontology/AverageObjectives</rdfs:domain>
 <rdfs:isDefinedBy>http://webpide.ledes.net/ontology</rdfs:isDefinedBy>
</owl:Class>
```

```
</rdf:RDF>
```

Apêndice E

Configuração template XML do Babel

XML E.1: XML Template Babel

```
<babel-config>
 <dataStore>
  <descriptor>
   <param name="DRIVER_CLASS" value="org.postgresql.Driver"/>
   <param name="URL_CONNECTION" value="jdbc:postgresql://localhost:5432/postgres"
       />
   <param name="PASSWORD" value="postgres"/>
   <param name="USER_NAME" value="postgres"/>
  </descriptor>
  <collection id="region">
   <param name="SQL" value="select id_dim_regiao as id_localidade,
           municipio || ' - ' || descr_municipio mun,
           regiao || ' - ' || descr_regiao reg,
           uf || ' - ' || descr_uf uf
         from dw.dim_regiao"/>
  </collection>
  <collection id="year">
   <param name="SQL" value="select id_dim_ano as id_year, ano as year from dw.
       dim_ano"/>
  </collection>

  <collection id="gender">
   <param name="SQL" value="select id_dim_sexo as id_gender, descr_sexo as gender
       from dw.dim_sexo "/>
  </collection>

  <collection id="writing">
   <param name="SQL" value=" select  distinct (round(CAST (media_redacao AS
       NUMERIC), 2)) as average_writing from dw.fato_enem " />
  </collection>

  <collection id="objective">
   <param name="SQL" value=" select  distinct (round(CAST (media_objetiva AS
       NUMERIC), 2)) as average_objective from dw.fato_enem  " />
  </collection>
```

```
 </dataStore>
 <template>

  <rdf:RDF xmlns:cc="http://creativecommons.org/ns#"
   xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
   xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
   xmlns:owl="http://www.w3.org/2002/07/owl#"
   xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
   xmlns:lgdo="http://linkedgeodata.org/ontology/"
   xmlns:webpide="http://webpide.ledes.net/ontology/">


   <lgdo:Place rdf:about="http://localhost:8890/enem/place/{region.id_localidade}"
       >
    <lgdo:City>{region.mun}</lgdo:City>
    <lgdo:State>{region.uf}</lgdo:State>
    <lgdo:Region>{region.reg}</lgdo:Region>
    <lgdo:Country>Brasil</lgdo:Country>
    <webpide:Type>Dimension</webpide:Type>
   </lgdo:Place>

   <webpide:Years rdf:about="http://localhost:8890/enem/year/{year.id_year}">
    <webpide:Year>{year.year}</webpide:Year>
    <webpide:Type>Dimension</webpide:Type>
   </webpide:Years>

   <webpide:Genders rdf:about="http://localhost:8890/enem/gender/{gender.id_gender
       }">
    <webpide:Gender>{gender.gender}</webpide:Gender>
    <webpide:Type>Dimension</webpide:Type>
   </webpide:Genders>

   <webpide:AverageWritings rdf:about="http://localhost:8890/enem/average_writing
       /{writing.average_writing}">
    <webpide:AverageWriting>{writing.average_writing}</webpide:AverageWriting>
    <webpide:Type>Measure</webpide:Type>
   </webpide:AverageWritings>

   <webpide:AverageObjectives rdf:about="http://localhost:8890/enem/
       average_objective/{objective.average_objective}">
    <webpide:AverageObjective>{objective.average_objective}</webpide:
        AverageObjective>
    <webpide:Type>Measure</webpide:Type>
   </webpide:AverageObjectives>

  </rdf:RDF>

 </template>
</babel-config>
```