UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO TECNOLÓGICO
DEPARTAMENTO DE INFORMÁTICA E ESTATÍSTICA
BACHARELADO EM SISTEMAS DE INFORMAÇÃO

PUBLICAÇÃO DE DADOS ABERTOS GOVERNAMENTAIS NO FORMATO *LINKED DATA*

GESIEL DA SILVA

GREICI BARETTA FRANZEN

Florianópolis

2013

UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO TECNOLÓGICO
DEPARTAMENTO DE INFORMÁTICA E ESTATÍSTICA
BACHARELADO EM SISTEMAS DE INFORMAÇÃO

GESIEL DA SILVA

GREICI BARETTA FRANZEN

PUBLICAÇÃO DE DADOS ABERTOS GOVERNAMENTAIS NO FORMATO *LINKED DATA*

Trabalho de conclusão de curso submetido à banca examinadora do Curso de Graduação em Sistemas de Informação da Universidade Federal de Santa Catarina como requisito parcial para a obtenção do titulo de Bacharel em Sistemas de Informação.
Orientador: Prof. Dr. Jose Leomar Todesco
Coorientador: Prof. Dr. Gauthier

Florianópolis

2013

GESIEL DA SILVA

GREICI BARETTA FRANZEN

PUBLICAÇÃO DE DADOS ABERTOS GOVERNAMENTAIS NO FORMATO *LINKED DATA*

Este trabalho de conclusão de curso foi julgado adequado para obtenção do Titulo de Bacharel em Sistemas de Informação e aprovado em sua forma final pelo Curso de Graduação em Sistemas de Informação.

Florianópolis, julho de 2013.

_____

Prof. Leandro Jose Komosinski, Dr.
Coordenador do Curso

Banca examinadora

_____

Prof., Dr. José Leomar Todesco,
Orientador
Universidade Federal de Santa Catarina

_____

Prof., Dr. Fernando Alvaro Ostuni Gauthier,
Coorientador
Universidade Federal de Santa Catarina

_____

Renato Deggau,
EPAGRI

AGRADECIMENTOS

Neste momento alcançamos o final de mais uma fase de nossas vidas.
Batalhamos, estudamos, mas tudo que conquistamos até agora,
não o fizemos sozinhos.
Assim, agradecemos primeiramente a nossos companheiros,
Mayara e Lúcio, e a nossas famílias, pelo apoio, encorajamento e
compreensão nos momentos de ausência.
Agradecemos aos orientadores e membros da banca que nos auxiliaram no
desenvolvimento, seja compartilhando conhecimentos, seja corrigindo desvios.
Agradecemos aos amigos conquistados nesses mais de 4 anos de universidade,
em especial a Eduarda Zanette Ferreira, Mateus Patrício Mello e Matheus Rosado
Vill, companheiros desde os almoços de comemoração até as provas de
Linguagens Formais e Compiladores.
De forma geral, agradecemos também a todos que contribuíram de alguma maneira
para a concretização deste Trabalho de Conclusão de Curso.

RESUMO

This paper describes the experiment conducted with the statistical data of agricultural production of the state of Santa Catarina, supplied by the public EPAGRI in a historical series from 1970 to 2011, which were published on the principles of Linked Data. Therefore, the data were processed and stored according to a dimensional model and represented by three complementary ontologies. Then were published in RDF triples according to Data Cube Vocabulary and published on a server triplets. Finally, were consumed by a web application open source, which aims to demonstrate, through graphs, data in RDF, in the dimensional model. The experiment allowed to expand the boundaries of publishing open government data, enabling consolidation, and knowledge discovery.

***Resumo**. Este artigo descreve o experimento realizado com os dados estatísticos de produção agropecuária do estado de Santa Catarina, fornecidos pela empresa pública EPAGRI, em uma série histórica de 1970 a 2011, que foram publicados nos princípios do Linked Data. Para tanto, os dados foram tratados e armazenados seguindo um modelo dimensional e representados por três ontologias complementares. Em seguida, foram publicados em triplas RDF segundo o Data Cube Vocabulary e publicados em um servidor de triplas. Por fim, foram consumidos por uma aplicação web de código aberto, que visa demonstrar, através de gráficos, dados em RDF, no modelo dimensional. O experimento permitiu expandir as fronteiras de publicação de dados abertos governamentais, possibilitando a consolidação e a descoberta de conhecimentos.*

LISTA DE FIGURAS

ABREVIATURAS E SIGLAS

ACARESC - Associação de Crédito e Assistência Rural de Santa Catarina

ACARPESC - Associação de Crédito e Assistência Pesqueira de Santa Catarina

ASP - *Active Server Pages*

CEPA - Centro de Socioeconomia e Planejamento Agrícola

DAML - *DARPA Agent Markup Language*

DSD - *qb:DataStructureDefinition*, elemento do *Data Cube Vocabulary*

E/R - Modelo Entidade-Relacionamento

EMPASC - Empresa Catarinense de Pesquisa Agropecuária S.A.

EPAGRI – Empresa de Pesquisa Agropecuária e Extensão Rural de Santa Catarina.

ETL - *Extract, Transform, Load*

HTML - *HyperText Markup Language*

HTTP - *HyperText Transfer Protocol*

IAB - *Interactive Advertising Bureau*

IASC - Instituto de Apicultura de Santa Catarina

OIL - *Ontology Interchange Language*

OLAP - *On-line Analytical Processing*

OWL - *Ontology Web Language*

PHP - *PHP*: Hypertext Preprocessor

PIB – Produto Interno Bruto

RDF - *Resource Description Framework*

REST - *Representational State Transfer*

SHOE - *Simple HTML Ontology Extension*

SIAGRO - Sistema de Informações Agropecuárias

SKOS - *Simple Knowledge Organization System*

SOAP - *Simple Object Access Protocol*

SPARQL - *SPARQL Protocol And RDF Query Language*

SQL – *Structured Query Language*

URI - *Uniform Resource Identifier*

VSP - *Virtual Services Platform*

W3C - *World Wide Web Consorcium*

WEB - *World Wide Web*

XML - *Extensible Markup Language*

SUMÁRIO

1. INTRODUÇÃO

1.1. Apresentação do Problema

O grande alcance da internet e sua disseminação já são públicos e notórios, como bem expressa a Figura 1, que demonstra a quantidade de dados produzidos online em somente um minuto, no mundo todo.

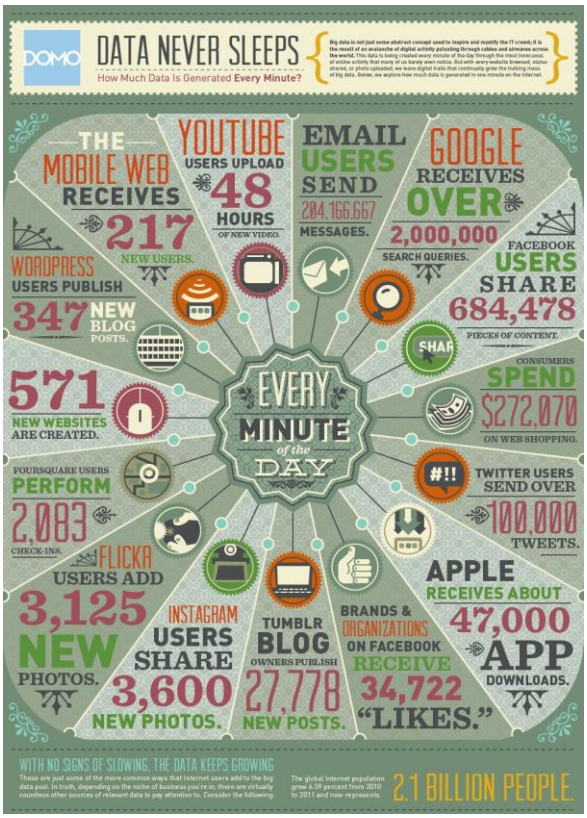

Figura 1: Quantidade de dados produzidos em um minuto (JAMES, 2012)

A nível de Brasil, em setembro de 2012, o número de pessoas com acesso à internet em casa atingiu a marca de 67,8 milhões, se considerarmos o acesso em todos os locais, como trabalho, *lan houses*, o número aumenta para 83,4 milhões de pessoas(IBGE, 2012). Outra pesquisa, realizada pelo IAB – *Interactive Advertising Bureau*, constatou que 70% dos brasileiros entre 15 e 55 anos, se possuíssem 15 minutos de tempo livre, gastariam preferencialmente na internet, mídia está considerada a mais importante dentre 80% dos entrevistados (IAB, 2013). Estes números já são altos e seguem uma tendência de crescimento constante, sem previsão de declínio.

Todas estas pessoas geram e consomem dados a cada acesso. Tais dados estão, em sua maioria, não formatados, não padronizados e pouco se pode inferir

deles. O grande desafio está em trabalha-los, tanto no sentido de conseguir manipular a enorme quantidade, quanto no de retirar informações com significado.

Seguindo a tendência, também as entidades públicas aumentam sua presença na internet, publicando seus dados abertos.

Diante deste cenário, a evolução natural está ligada a possibilidade de que os computadores interpretem os dados e a compartilhem com outros computadores, o que é chamado de web semântica (CARDOSO, 2011).

No contexto deste trabalho de conclusão de curso, são aplicados os conceitos de *Linked Data* e do *Data Cube* Vocabulary aos dados de produção agropecuária do Estado de Santa Catarina, mantidos pela empresa pública EPAGRI, através do DATACEPA. Nesta base estão contidos indicadores de produção como área colhida, área destinada à colheita, área plantada, quantidade produzida agrícola e pecuária, quantidade abatida e efetivo[1].

Em sua forma original, estes dados não são utilizados para nenhum tipo de análise e tampouco estão à disposição do público para consultas e processamento. Também não se apresentam em um formato que facilite seu entendimento e compartilhamento.

1.2. Objetivos

1.2.1. Objetivo Geral

Este trabalho de conclusão de curso procura alcançar o objetivo de publicar dados estatísticos levantados de acordo com a base de dados do sistema DATACEPA. Este sistema é utilizado pelo órgão público EPAGRI, Empresa de Pesquisa Agropecuária e Extensão Rural de Santa Catarina.

1.2.2. Objetivos Específicos

- Desenvolver um cubo de dados estatísticos, levando em consideração informações relevantes armazenadas na base de dados do DATACEPA.
- Publicar dados estatísticos multidimensionais, em formato *Linked Data* utilizando o vocabulário RDF *Data Cube*.

[1] Entrevista concedida pelo Sr. Renato Deggau, na sede da EPAGRI em Florianópolis, no dia 26/03/2013, conforme roteiro no apêndice A.

- Estudar e aplicar conhecimentos atuais de *web* semântica, *Linked Data*, dados abertos e dados abertos governamentais, levando sempre em consideração as boas práticas da *web* e do *Linked Data*.
- Preparar adequadamente ambientes de desenvolvimento que permitam a publicação de dados estatísticos no formato *Linked Data.*
- Acessar os dados estatísticos por meio de uma aplicação *web*.

1.3. Justificativa

O Brasil apresenta grande capacidade de produção nos setores de agropecuária e pecuária, sendo considerado por muitos como o celeiro do mundo, já que o país tem grande participação nas exportações internacionais de produtos do setor.

Diante do fato apresentado é natural que seja gerada uma enorme quantidade de dados com informações detalhadas do setor, porém essas informações não estão ainda disponíveis de forma satisfatória para o consumo da população.

Portanto, a web semântica e o *Linked Data*, tecnologias atuais e de ampla aceitação por diversos governos, apresentam-se como boa alternativa na publicação desses dados de tal forma que facilite a interpretação dos mesmos, tanto por máquinas quanto, posteriormente, por humanos.

1.4. Organização dos Capítulos

Este trabalho está organizado em cinco capítulos. O primeiro apresenta o problema e o contexto em que está inserido, a justificativa para a escolha do tema e os objetivos perseguidos.

O segundo traz a fundamentação teórica dos conceitos abordados, como a web semântica, *Data Cube Vocabulary, Data Warehouse, Linked Data* e Ontologias.

O terceiro apresenta o desenvolvimento, ou seja, todos os passos seguidos para atingir os objetivos definidos inicialmente.

No quarto encontram-se as conclusões e os resultados obtidos a partir da execução das atividades.

Finalmente, no quinto capítulo estão as oportunidades de trabalhos futuros, identificadas ao longo do desenvolvimento.

2. FUNDAMENTAÇÃO TEÓRICA

2.1. Produção Agrícola em Santa Catarina

Já faz parte do conhecimento popular a noção de que o Brasil tem uma "vocação agrícola", em virtude da vasta extensão territorial e dos fatores climáticos e de solo extremamente favoráveis ao cultivo de alimentos. O país é um dos líderes mundiais na produção e exportação de inúmeros produtos agrícolas (Ministério da Agricultura, 2012).

Em Santa Catarina não poderia ser diferente. Este estado apresenta um dos maiores índices de produtividade por área e encontra-se entre os seis principais estados produtores de alimentos. Tal sucesso pode ser explicado pelo alto nível de inovação do agricultor predominantemente familiar e ao emprego de tecnologia de ponta. Estatisticamente, 12,8 % do PIB estadual tem origem na produção agrícola, ocupando o quinto lugar entre estados exportadores (CEPA/SC, 2012).

O ano 2000 foi do setor primário em Santa Catarina, destacando-se principalmente pela produção de cebola, maçã e carnes suínas. O estado ainda foi o segundo produtor de carne de frango, alho, fumo e mel de abelha e o terceiro de arroz e banana. A agroindústria constitui-se na principal atividade desenvolvida no estado, contribuindo com 20% do PIB e com 50% do total de exportações (CEPA/SC, 2012).

Outro ano muito favorável na produção agrícola foi o de 2010, quando foram registrados recordes na produção de milho, soja e arroz, superando as expectativas dos produtores (EPAGRI, 2011).

2.1.1. A imigração e a agricultura familiar

A colonização do estado de Santa Catarina, em especial o oeste catarinense (ALVES, 2006), seguiu um modelo minifundiário, onde famílias de imigrantes recebiam pequenos pedaços de terra a fim de garantir sua subsistência. Neste sistema, as principais culturas eram de produtos básicos como milho, arroz, feijão, mandioca, esta herança dos indígenas, e de animais de serviço e consumo, como gado, suínos e aves. Há pouco registro de excedentes comercializáveis, principalmente no início da ocupação do estado, no século XVII (ZOLDAN, 2004).

O planalto, em sua ocupação, possuía características extrativistas, com destaque para a madeira e a erva-mate. A produção de gado, alavancada pela passagem dos tropeiros no século XVIII, também foi destaque durante a colonização. No litoral, é predominante a pesca e o cultivo da mandioca, para produção de farinha (ZOLDAN, 2004).

A imigração de Europeus foi essencial para marcar a forma de cultivo predominantemente familiar e de policultura com pequenos excedentes, dada a abundância de terras e a distância dos grandes centros de comercialização (ZOLDAN, 2004).

Assim, até hoje percebe-se no estado de Santa Catarina que a produção em pequenas propriedades e com variedades de cultura é a forma predominante de cultivo, garantindo fortes laços com os mercados regionais locais (SANTOS, 2011).

2.1.2. EPAGRI

A Empresa de Pesquisa Agropecuária e Extensão Rural de Santa Catarina, foi criada em 1991 a partir da união de quatro órgãos estaduais relacionados ao setor agrícola, sendo eles a Empresa Catarinense de Pesquisa Agropecuária S.A. (Empasc), a Associação de Crédito e Assistência Rural de Santa Catarina (Acaresc), a Associação de Crédito e Assistência Pesqueira de Santa Catarina (Acarpesc) e o Instituto de Apicultura de Santa Catarina (Iasc). Em 2005 também passou a fazer parte da instituição o Instituto de Planejamento e Economia Agrícola de Santa Catarina (Instituto Cepa/SC) (EPAGRI, [2013?]).

Tem por missão a frase: “Conhecimento, tecnologia e extensão para o desenvolvimento sustentável do meio rural, em benefício da sociedade”. Seus objetivos fins são: “Promover a preservação, recuperação, conservação e utilização sustentável dos recursos naturais. Buscar a competitividade da agricultura catarinense frente a mercados globalizados, adequando os produtos às exigências dos consumidores. Promover a melhoria da qualidade de vida do meio rural e pesqueiro” (EPAGRI, [2013?]).

A empresa atua em todo o Estado por meio das gerências regionais que são compostas por escritórios municipais, unidades de pesquisa, com seus campos experimentais, e centros de treinamento. No nível político-estratégico, há a sede administrativa, integrada pelos órgãos deliberativos e de fiscalização, a diretoria executiva, as gerências estaduais e as assessorias (EPAGRI, [2013?]).

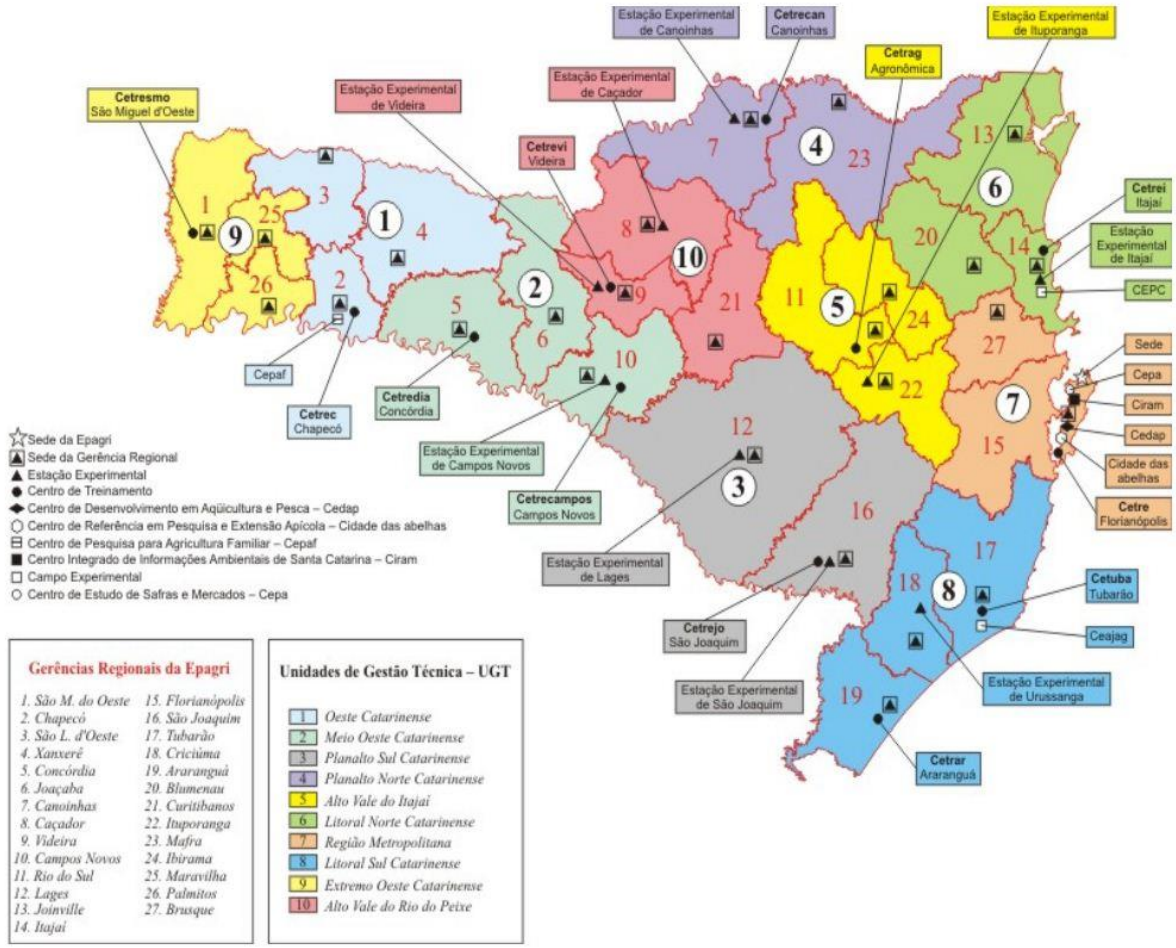

Figura 2: Atuação da Epagri no Estado de Santa Catarina (EPAGRI, [2013?])

2.1.3. CEPA

O Centro de Socioeconomia e Planejamento Agrícola foi criado em 2005, substituindo Instituto de Planejamento e Economia Agrícola de Santa Catarina (Instituto Cepa/SC). Este teve origem em 1982, a partir da Comissão Estadual de Planejamento Agrícola, criada em 1975. Em toda sua existência, esteve atuando nas áreas de informação e planejamento agrícola catarinense (CEPA, [2013?]).

Sua missão é: “Buscar o desenvolvimento sustentável de Santa Catarina, por meio da elaboração de pesquisas socioeconômicas, estudos, projetos e disseminação de informações nas áreas econômica, tecnológica, científica e organizacional”. Tem como objetivos: “Realizar o monitoramento e análise da produção do mercado agrícola e das políticas públicas, atuar no desenvolvimento local e regional, desenvolver estudos e pesquisas sobre o espaço rural, gerar e disseminar informações e prestar serviços para os governos do Estado, da União e

municipais, iniciativa privada, organizações de produtores e universidades" (CEPA, [2013?]).

2.1.4. SIAGRO e DATACEPA

O SIAGRO (Sistema de Informações Agropecuárias) foi desenvolvido em parceria com a EPAGRI/CEPA, para o governo do Maranhão, tendo como objetivo geral: "Desenvolver, implantar e manter um sistema de captação, armazenamento, recuperação e disseminação de dados, relativo ao setor rural e segmentos afins, compatíveis com as necessidades do perfil de seus usuários" (DEGGAU, 2007).

Tendo em vista que a parceria com a empresa desenvolvedora do software foi interrompida, o sistema não foi completamente finalizado. Posteriormente, foi adotado pela EPAGRI, tendo seu nome modificado para DATACEPA[2].

Este sistema serve como base para a disponibilização na internet dos dados coletados referente aos preços diários dos produtos, enquanto os indicadores de produção anual das cultivares no estado de Santa Catarina não são utilizados[3].

2.1.5. Resgate do Problema

Como explanado até agora, a empresa pública EPAGRI provém uma enorme quantidade de indicadores de produção e preços agropecuários, que até então não são utilizados para tomada de decisão internamente na empresa, nem são disponibilizados adequadamente para consultas do público, dificultando qualquer tipo de análise.

Além disso, a forma na qual as informações de preços diários são publicadas hoje em dia não atingem todo seu potencial, pois o sistema não fornece meios para combinação irrestrita das informações, já que sua forma de consulta não é adequada. Outro aspecto relevante diz respeito ao formato dos dados, que se apresentam brutos e sem semântica, o que impede o seu pleno uso e combinação com outras fontes de dados, no intuito de obter a máxima expressividade.

[2] Entrevista concedida pelo Sr. Renato Deggau, na sede da EPAGRI em Florianópolis, no dia 26/03/2013, conforme roteiro no apêndice B.
[3] Entrevista concedida pelo Sr. Renato Deggau, na sede da EPAGRI em Florianópolis, no dia 26/03/2013, conforme roteiro no apêndice C.

2.2. *Linked Data*

Tendo como base tecnológica do presente trabalho a publicação de dados no formato *Linked Data*, faz-se necessária a introdução a termos técnicos presentes neste contexto de estudo, conforme detalhado a seguir.

2.2.1. Conhecimento Aberto

O conceito de conhecimento aberto surgiu em agosto de 2005 quando um grupo de autores, entre eles Peter Suber, Cory Doctorow, Tim Hubbard, Peter Murray-Rust, Jo Walsh e Prodromos Tsiavos, escreveram o primeiro rascunho com a definição de conhecimento (dados) aberto. Em outubro do mesmo ano, uma segunda versão foi publicada na web, porém, somente em julho de 2006, quando o projeto foi movido para seu atual website, http://opendefinition.org/, que a versão 1.0 foi lançada. Desde então pequenas alterações são realizadas ao projeto de acordo com o *feedback* do público (OPEN KNOWLEDGE FOUNDATION, [2011?]).

A definição em questão não diz respeito apenas a dados, mas a conhecimento. Conhecimento esse que faz referência a conteúdos como música, filmes e livros, dados de qualquer área de conhecimento e informações governamentais e administrativas. De acordo com a definição, foi determinado que para ser considerado aberto, algumas prerrogativas devem ser contempladas, são elas (OPEN KNOWLEDGE FOUNDATION, [2012?]b):

1. Acesso: Obras disponíveis na integra e se houver algum custo será relacionado apenas a forma de reprodução do conteúdo.
2. Redistribuição: A licença não deve restringir a venda ou doação da obra e nem garantir *royalty* ou taxas correlatas.
3. Reutilização: A licença não deve proibir modificações e trabalhos derivados da obra.
4. Ausência de restrições tecnológicas: A apresentação da obra deve ser de tal forma que não exista limitações tecnológicas para execução das tarefas já citadas.
5. Atribuição: Pode ser exigido na redistribuição e reutilização, por meio de licença, a atribuição dos colaboradores e criadores da obra.
6. Integridade: A licença pode exigir que trabalhos derivados devem ser distribuídos com número de versão e título diferentes da obra original.

7. Sem discriminação contra pessoas ou grupos: A obra não deve conter discriminação a qualquer pessoa ou grupo de pessoas.
8. Sem discriminação contra campos de trabalho: A utilização da obra não deve ser restringida a campos específicos de atuação.
9. Distribuição da licença: Os direitos associados a obra devem ser aplicadas a todos que tiverem acesso a mesma.
10. Licença não deve ser específica de uma coletânea: Os direitos associados a obra não devem depender de uma coletânea.
11. Licença não deve restringir a distribuição de outras obras: A licença não pode limitar os direitos associados a outras obras distribuídas na mesma coletânea.

Resumindo, “dados abertos são dados que podem ser livremente usados, reutilizados e redistribuídos por qualquer pessoa – sujeitos, no máximo, à exigência de atribuição da fonte e compartilhamento pelas mesmas regras” (OPEN KNOWLEDGE FOUNDATION, [2012?]b).

O conceito apresentado anteriormente é importante pois garante a interoperabilidade entre conjunto de dados. Desse modo, centros comunitários de dados podem ser livremente conectados entre si para gerar maior valor agregado e possibilitar a criação de novos e melhores produtos e serviços (OPEN KNOWLEDGE FOUNDATION, [2012?]a).

Deve ser ressaltado que o foco na publicação de dados abertos está nos dados não-pessoais, ou seja, que não contêm informação sobre pessoas específicas. Assim sendo, as instituições governamentais tornaram-se grandes candidatas para a prática de publicação dos dados abertos, tanto pela quantidade e centralidade dos dados que coleta, quanto pelos dados serem públicos conforme legislações nacionais (LEI nº 12.527, 2011). Diante de tal cenário surgiu o termo “Dados Abertos Governamentais” (OPEN KNOWLEDGE FOUNDATION, [2012?]a).

A invenção da internet reacendeu a chama de que o acesso as informações governamentais é um diferencial para as sociedades, ainda mais nos tempos contemporâneos, onde a geração atual cresceu junto com a WEB e tende a exercer sua cidadania nesse espaço colaborativo cibernético. A disponibilização de informações governamentais através da WEB possibilita novas oportunidades para acessar, comentar e exigir melhorias no governo (SALM JÚNIOR, 2012).

Como um complemento à definição proposta pelo *Open Knowledge Foundation*, em 2007 um grupo de trabalho realizado na Califórnia definiu oito princípios para servirem como parâmetro de avaliação para os dados governamentais abertos. Os princípios são apresentados a seguir (OPEN GOVERNMENT WORKING GROUP, 2007):

1. O dado deve ser completo, garantindo que todos os dados públicos sejam publicados.
2. O dado deve ser primário, ou seja, publicado da mesma maneira como foi coletado.
3. O dado deve ser atual. Para tal, os dados coletados devem ser publicados o mais rápido possível, garantindo assim maior valor agregado.
4. O dado deve ser acessível, publicado para o maior público e propósito possível.
5. O dado deve ser processável por máquina, ou seja, deve ser publicado de tal forma que computadores sejam capazes de processa-lo através de rotinas automatizadas.
6. O acesso ao dado deve ser não discriminatório, garantindo que qualquer pessoa tenha acesso ao mesmo.
7. O formato do dado deve ser não proprietário, publicado de tal forma que nenhuma entidade possua controle exclusivo.
8. O dado não deve estar vinculado a alguma licença. Objetivando impossibilitar a regulação de direitos autorais, marcas, patentes e correlatos.

De maneira mais simples e minimalista, David Eaves, ativista dos dados governamentais abertos, contribuiu com três leis que julgava serem essenciais para que dados, não somente governamentais, fossem admitidos como abertos. São as leis a seguir, que garantem que os dados possam ser encontrados, utilizados e compartilhados (EAVES, 2009):

1. O dado não existe se não pode ser indexado ou encontrado na WEB;
2. O dado não pode ser reaproveitado quando não estiver publicado em formato compreensível por máquina; e
3. O dado não é útil quando algum dispositivo legal impede sua replicação.

Os dados abertos governamentais vêm sendo uma das principais ferramentas do governo aberto. Governo este cujos princípios são baseados em transparência,

participação e colaboração. A transparência para promover o *accountability*, ou seja, uma prestação de contas do governo com seus cidadãos, e para garantir que o povo saiba das ações governamentais. Participação no que diz respeito a aumentar a eficácia da atuação governamental e para levar em consideração as opiniões e experiências coletivas na tomada de decisões e elaboração de novas políticas públicas. Já a colaboração deve garantir que o povo esteja ativamente envolvido nas obras governamentais (OBAMA, 2009).

No Brasil, a iniciativa de divulgação de dados abertos é implementada pelo Portal Brasileiro dos Dados Abertos, no sítio dados.gov.br. O portal é uma ferramenta que visa disponibilizar dados e informações públicas para utilização de todos. Foi desenvolvido a partir de um compromisso assumido pelo governo que foi formalizado no Plano de Ação de Governo Aberto. Complementando a abertura dos dados, em 18 de novembro de 2011 foi sancionada a Lei de Acesso a Informação Pública (Lei 12.527/2011), que preconiza que a informação solicitada pelo cidadão deve seguir as três leis dos dados abertos, já citadas nesta seção (MINISTÉRIO DO PLANEJAMENTO ORÇAMENTO E GESTÃO, [2013?]).

2.2.2. Web Semântica

A *world wide web*, ou simplesmente web, foi idealizada pra ser a web de documentos, basicamente, um conjunto de páginas interconectadas através de links. Desde seu surgimento, o compartilhamento de conhecimento e o acesso a publicações foram amplamente favorecidos (BIZER; HEATH; BERNERS-LEE, 2009). Apesar de todos os benefícios o foco ainda se direciona em conteúdo especifico para a interpretação por pessoas, não permitindo o acesso e a interpretação automática de seu conteúdo por computadores.

Hoje os computadores são capazes de interpretar a estrutura de uma página web distinguindo, por exemplo, o cabeçalho de um *web site* do seu corpo ou rodapé. Porém, os navegadores ainda não conseguem entender os significados contidos nos documentos da web, como por exemplo, a quem pertence determinado documento ou para onde apontam seus *links* (BERNERS-LEE; HENDLER; LASSILA, 2001).

A web semântica, apresentada em 2001 por Tim Berners Lee, James Hendler e Ora Lassila, tem o intuito de adicionar sentido a todos os componentes da página, preenchendo a lacuna existente e permitindo que não somente humanos utilizem a

informação, mas também máquinas. A partir desta ideia, os dados disponibilizados estariam conectados, ligados uns aos outros em diversas bases de dados. Surge, assim, a web de dados.

Essa abordagem pode permitir que agentes, programas capazes de varrer o conteúdo da web e combina-lo com outras diversas fontes de informações publicadas na internet, sejam capazes de automatizar processos como marcar consultas médicas automaticamente de acordo com o prontuário e agenda do paciente, descobrir onde estão os amigos do usuário em determinado momento ou até mesmo diminuir o volume de dispositivos sonoros ao atender um telefone (BERNERS-LEE; HENDLER; LASSILA, 2001).

Para alcançar este objetivo de inserir semântica na Web atual, tendo como base a estrutura em camadas já instalada e sem precisar modifica-la, o que demandaria um esforço de difícil aceitação por todos, foi proposta a adição de novas camadas sobre as já existentes, adicionando gradativamente a semântica. Desta forma a implementação pode ser feita gradativamente e com uso imediato (BREITMAN, 2010).

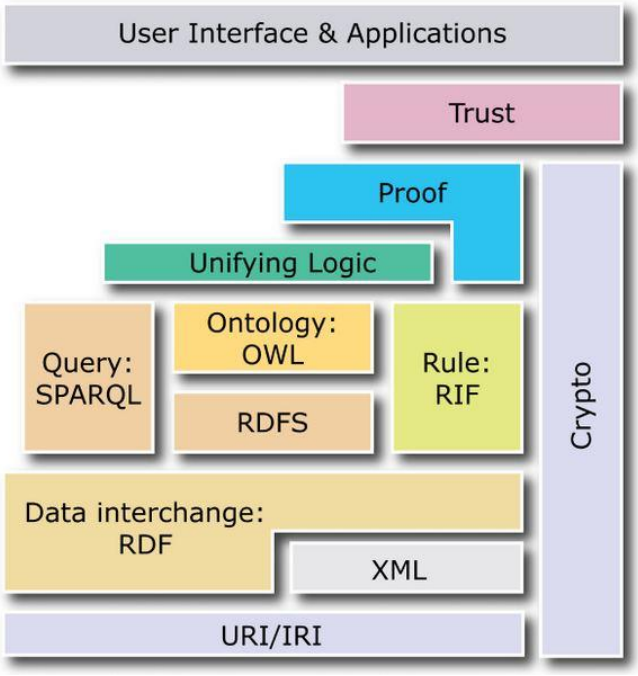

Figura 3: modelo da web semântica em camadas[4]

As camadas na base da pirâmide, URI/IRI, já se encontram em funcionamento adequado na web atual de documentos, uma vez que os portais

[4] WORLD WIDE WEB CONSORTIUM (W3C). Disponível em: <http://www.w3.org/2007/03/layerCake.png>. Acesso em: 30 abr. 2013.

publicados na web hoje são identificados por URIs únicas (BREITMAN, 2010). Em breve essas URIs tendem a se tornarem internacionalizadas, por isso o novo nome IRI.

A camada XML é responsável pela linguagem utilizada na web semântica, sendo que o RDF, assunto que será tratado em seções futuras deste trabalho, é uma evolução desta linguagem. Entre os pontos positivos proporcionados pelo XML estão o suporte para criação conexões entre documentos e recursos de rede, o fato de possibilitar a separação do conteúdo da estrutura do documento e por garantir interoperabilidade entre sistemas (BREITMAN, 2010).

O ponto chave da web semântica é a representação do conhecimento, de forma a possibilitar que computadores façam mais do que somente mostrar informações na tela, mas também possam interpretar e retirar da massa de dados a informação desejada por quem os programou. Pretende-se alcançar estas características e objetivos através de tecnologias já existentes, como XML, RDF, SPARQL, Ontologias para descrever o conhecimento, todos unidos em torno do conceito de *Linked Data* (BERNERS; HENDLER; LASSILA, 2001). Estes termos serão discutidos no decorrer do presente trabalho.

2.2.3. Ontologias

Como mencionado anteriormente, a web semântica propõe a conexão de bases de dados distintas. É nesse contexto que se encaixam as ontologias, proporcionando uma maneira de softwares descobrirem que termos diferentes possuem o mesmo significado semântico e assim cruzando informações entre bases de dados (BERNERS; HENDLER; LASSILA, 2001).

O termo ontologia tem origem nas palavras gregas “*onto*”, ser, e “*logos*”, palavra (ALMEIDA, 2003). Assim sendo, na filosofia é uma disciplina com foco no estudo do ser, da natureza existencial das coisas (BERNERS; HENDLER; LASSILA, 2001). A área de inteligência artificial também deu significado para o termo em questão, não se preocupando, porém, com a existência ou não de um conceito no mundo físico, mas se este conceito é capaz de proporcionar raciocínio útil em um determinado domínio (BORST, 1997). Porém, na área de tecnologia da informação, o termo tem um significado especial (ALMEIDA, 2003).

A definição de ontologia adotada por este trabalho é de “uma especificação formal de uma conceitualização compartilhada” (BORST, 1997), onde

conceitualização é definida como “uma coleção de objetos, conceitos e outras entidades que se assume existirem em um domínio e os relacionamentos entre eles”, ou seja, uma abstração simplificada do domínio que é necessário ser representado (GENESERETH; NILSSON, 2003).

Neste sentido, de especificação formal, há várias classificações para as ontologias (BREITMAN, 2010).

Segundo seu espectro semântico, são dispostas em um ranking que as coloca entre as mais leves (*lightweight*) e as mais pesadas (*heavyweight*). Assim, a classificação inicia com ontologias leves, como catálogos informais de termos, e vai gradativamente aumentando a complexidade de acordo com o conteúdo e a estrutura interna, passando por glossários, tesauros, hierarquias tipo-de informais e formais entre outras, até chegar a ontologias que exprimem restrições em lógica de primeira ordem. O que varia é o grau de formalismo demonstrado e a expressividade de cada representação (BREITMAN, 2010).

A fim de evitar dúvidas é importante ressaltar as diferenças entre alguns conceitos utilizados neste contexto classificatório (BREITMAN, 2010):

- **Taxonomias X Ontologias**: A primeira serve para classificar informação em uma hierarquia simples, com relacionamentos tipo-de, classicamente utilizada para estabelecer classificação ordenada de animais e plantas. Sua única forma de relacionamento é de pai-filho e é esta característica que a difere da Ontologia. Nesta última é possível inserir relacionamentos, características e propriedades como parte-de, localização, causa-efeito, etc.
- **Tesauros X Ontologias:** Tesauro trata-se de uma taxonomia adicionada de um conjunto de relacionamentos semânticos, porém em número limitado e bem definido, não sendo possível estender os conceitos para novas possibilidades. Em sendo necessário aumentar a pluralidade dos relacionamentos, deve-se utilizar uma ontologia.

Outra classificação é segundo a sua generalidade. Aqui, parte-se de ontologias de nível superior, que descrevem conceitos genéricos como espaço, tempo; passando por ontologias de domínio e de tarefas e por último, há as ontologias de aplicação. Cada uma delas especializa um termo mais específico da anterior, diminuindo o espectro de abrangência, porém especificando de forma mais detalhada os termos.

Finalmente, classificam-se ontologias conforme o tipo de informação que representam. São elas:

- Ontologias para representação do conhecimento;
- Ontologias gerais e de uso comum: expressam conhecimento de senso comum;
- De topo ou de nível superior: expressam conceitos gerais;
- De domínio: por exemplo, domínio médico, farmacêutico, jurídico;
- De tarefas;
- De domínio-tarefa: onde é modelada uma tarefa em um determinado domínio que não pode ser reaproveitado em outros;
- De método;
- De aplicação.

Com o conceito compreendido, a próxima questão refere-se à como representar uma ontologia no ambiente web. Diversas linguagens já foram criadas par tal fim, atualmente destacando-se o RDF e RDFSchema, a SHOE, Oil, DAML, DAML + Oil e a OWL (BREITMAN, 2010), sendo que esta última será a linguagem utilizada para representar a ontologia desenvolvida no presente trabalho.

2.2.4. OWL

Criada pelo Consórcio W3C, a partir de uma revisão da linguagem DAML + Oil, tem o objetivo de construir ontologias, explicitar fatos sobre um determinado domínio e racionalizar sobre ontologias e fatos (BREITMAN, 2010). Foi criada para uso em aplicações que precisam processar informações, ao invés de somente apresenta-la e sua estrutura apresenta mais facilidades para declarar termos do que suas antecessoras (MCGUINNESS; HARMELEN, 2004).

Está dividida em três sublinguagens (MCGUINNESS; HARMELEN, 2004):

1. OWL Lite: oferece uma estrutura simplificada de hierarquia e relacionamentos. É a que possui uma menos complexidade formal;
2. OWL DL: assim chamada por sua correspondência com a lógica de descrição. É mais complexa que a Lite, mas ainda garante integridade computacional.

3. OWL Full: garante a máxima expressividade, porém sem garantir a integridade computacional, ou seja, é pouco provável que um software consiga suportar toda a sua complexidade.

Seus elementos básicos são (BREITMAN, 2010):

1. *Namespaces*: São declarações típicas do início de um documento, ajudando a interpretar os termos presentes, sem ambiguidades.

```
<rdf:RDF
    xml:base="http://www.w3.org/2002/03owlt/description-logic/inconsistent001"
    xmlns:owl="http://www.w3.org/2002/07/owl#"
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#" xmlns:xsd="http://www.w3.org/2001/XMLSchema#">
```

Figura 4: exemplo de *namespaces* (CARROLL; ROO, 2004)

2. Cabeçalhos: sentenças que registram comentários acerca da versão do documento ou outros comentários explicativos. Demarcada com a *tag* owl: Ontology;

```
<rdf:RDF
    xml:base="http://www.w3.org/2002/03owlt/description-logic/inconsistent001"
    xmlns:owl="http://www.w3.org/2002/07/owl#"
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#" xmlns:xsd="http://www.w3.org/2001/XMLSchema#">
    <owl:Ontology rdf:about=""/>
```

Figura 5: exemplo de cabeçalho (CARROLL; ROO, 2004)

3. Classes: representa um conjunto de indivíduos, que possuem características em comum, capazes de diferenciá-los dos demais;

```
    <owl:Class rdf:about="http://oiled.man.example.net/test#b">
        <rdfs:subClassOf>
            <owl:Class>
                <owl:complementOf>
                    <owl:Class rdf:about="http://oiled.man.example.net/test#c"/>
                </owl:complementOf>
            </owl:Class>
        </rdfs:subClassOf>
    </owl:Class>
```

Figura 6: exemplo de classe (CARROLL; ROO, 2004)

4. Indivíduos: são as instâncias da classe e relacionam-se com outros indivíduos e classes através de propriedades;
5. Propriedades: descrevem os relacionamentos entre indivíduos ou classes e podem se referir a todos os indivíduos ou a somente um. Existem dois tipos: as propriedades de objeto, que ligam classes; e as propriedades de dado, que ligam indivíduos a literais.

```
<owl:ObjectProperty rdf:ID="hasHead">
   <owl:equivalentProperty>
      <owl:ObjectProperty rdf:ID="hasLeader"/>
   </owl:equivalentProperty>
</owl:ObjectProperty>
```

Figura 7: exemplo de propriedade de objeto (CARROLL; ROO, 2004)

```
<owl:DatatypeProperty rdf:ID="convertedAbsoluteValue">
  <rdfs:domain rdf:resource="http://www.w3.org/2001/XMLSchema#integer"/>
  <rdfs:range rdf:resource=
      "http://www.w3.org/2001/XMLSchema#nonNegativeInteger" />
</owl:DatatypeProperty>
```

Figura 8: exemplo de propriedade de dado (CARROLL; ROO, 2004)

2.2.5. Linked Data

O termo *Linked Data* surge no contexto da Web Semântica, para assumir a função de conectar os dados espalhados em bases pela web, tratando os dados de tal forma que possibilita a execução de consultas como em um banco de dados comum (W3C, [2013?]a).

Neste sentido, quanto mais estruturado e padronizado for o dado, maior seu valor, mais fácil será para criar ferramentas que visem reutilizá-lo, mais sofisticado será o processamento (HEATH, BIZER, 2011). Vários padrões foram criados com esta finalidade, sendo *Linked Data* um deles.

Berners-Lee define o termo (em tradução livre) como sendo “um conjunto de boas práticas para publicar e conectar dados estruturados, sendo a base da evolução da web” (BERNERS; HENDLER; LASSILA, 2001).

A fim de que os dados realmente façam parte de um banco global e os objetivos da web de dados sejam atingidos, quatro regras foram estabelecidas (BERNERS-LEE, 2006):

1. Usar URIs como nome para as coisas descritas;
2. Usar HTTP URIs para que as pessoas possam consultar esses nomes;
3. Quando consultadas as URIs, informações úteis devem ser retornadas, utilizando uma forma padronizada (RDF, SPARQL);
4. Incluir *links* para outras URIs, para que mais coisas possam ser descobertas.

O uso padronizado da mesma tecnologia surge como um ponto importante para garantir o fácil e irrestrito acesso aos dados, o que permite que estes sejam (BIZER, CYGANIAK, HEATH, 2008):

1. Mais facilmente acessados por mecanismos de busca;
2. Acessados usando um navegador genérico;
3. Ligados entre fontes diferentes.

Similarmente a navegar entre páginas através de links HTML, *Linked Data* permite a navegação entre diferentes fontes de dados, através de links RDF. Isso permite ao usuário começar com uma fonte de dados e mover-se para uma web de dados potencialmente sem fim (HEATH, BIZER, 2011).

2.2.5.1. URIs

Quando pensamos em publicar coisas na internet, precisamos inicialmente nos preocupar em identificar tais coisas, suas propriedades e relacionamentos. As coisas representadas são chamadas de recursos e a fim de identificar de forma simples um recurso representado no mundo digital, foi criada a URI (*Uniform Resource Identifier)*, que se trata de uma cadeia de caracteres única para cada representação. O termo é definido por (BERNERS-LEE, 1998):

1. Identificador, usado no sentido de que qualquer objeto que tenha a função de ser referência a algo é um identificador;
2. Uniforme, onde recursos de diferentes fontes e consumidos por diferentes mecanismos são representados da mesma forma;
3. Recursos, que podem ser qualquer coisa que possua uma identidade e possa ser representado.

No contexto de *Linked Data* é padrão utilizar HTTP URIs, pois além de proporcionar uma forma não centralizada de gerenciar os identificadores, também permite que o recurso seja acessado na web por um navegador comum (HEATH, BIZER, 2011).

2.2.5.2. RDF

O formato RDF (*Resource Description Framework*) é um modelo para representar recursos na Web. Neste modelo, a descrição do recurso é feita por um simples, porém eficiente, conjunto de triplas. Cada tripla é composta por três

elementos: sujeito, predicado e objeto. O sujeito e o objeto são representados por nós e o predicado por um arco de ligação entre os dois (KLYNE; CARROLL, 2004).

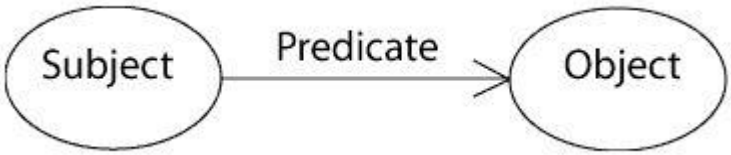

Figura 9: Estrutura da tripla RDF (KLYNE; CARROLL, 2004)

O sujeito é identificado por uma URI que representa o recurso em foco. O objeto pode ser um valor, uma data, uma palavra ou até mesmo uma outra URI, apontando para outra base de dados, esta última forma sendo a de maior valor para a web semântica. O predicado, que também é uma URI, representa a ligação entre os dois termos, tipificando a relação e conectando os dois recursos (HEATH, BIZER, 2011).

Este modelo é ideal para uso na web de dados, pois tem a capacidade de conectar bases de dados de forma simples, fazendo com que diferentes bases sejam naturalmente entrelaçadas. Com ele, os clientes podem descobrir informações adicionais sobre a área de interesse simplesmente seguindo as URIs e partindo de qualquer uma delas. As URIs, por sua vez, proveem uma forma única de representar o recurso (HEATH, BIZER, 2011).

2.2.5.3. Publicando *Linked Data*

O primeiro passo para publicar dados no formato *Linked Data* é transformá-los. Existem diversas ferramentas para executar essa tarefa, cujo uso depende principalmente do formato original dos dados.

Independentemente da ferramenta utilizada, alguns passos devem ser seguidos (BIZER, C., HEATH, T., BERNERS-LEE,2009):

1. Atribuir URIs aos recursos descritos e prover um retorno no formato RDF quando essa URI for consultada;
2. Definir ligações RDF para outras bases de dados na web;
3. Prover metadados sobre as publicações.

No tocante à escolha da URI pode acontecer de duas bases se referirem ao mesmo recurso com URIs diferentes e isso não é necessariamente um problema. Esta escolha reflete diferentes pontos de vista sobre o mesmo assunto, já que em

uma pesquisa pelo termo, duas abordagens podem ser exibidas (HEATH; BIZER, 2011).

A interconexão entre diferentes bases é o fundamento da web de dados, é o que transforma ilhas de dados em um espaço global. Tecnicamente, um link externo é uma tripla RDF onde o sujeito é a URI que identifica uma base e o objeto aponta para a identificação da outra base (BIZER; CYGANIAK; HEATH, 2008).

Existem três tipos de links RDF (HEATH; BIZER, 2011):

1. Links de relacionamento: indicam coisas relacionadas à descrita, como por exemplo, local de residência, de trabalho;
2. Links de identificação: apontam para descrições alternativas do mesmo recurso;
3. Links de vocabulário: indicam o vocabulário usado para representar os dados.

Os metadados acompanhantes devem fornecer informações sobre quem criou, quando e a forma de criação dos dados, para que os clientes possam escolher entre fontes de autores diferentes. Também, podem acompanhar os dados, informações técnicas sobre a base ou formas de interligação com outras bases (HEATH; BIZER, 2011).

Após a conversão, é necessário disponibilizar estes dados na web. Para isso existem formatos de serialização, que tratam de transformar as triplas em um fluxo de dados. O mais comum é o formato RDF/XML, que é baseada nos padrões XML. Outro formato conhecido é o RDFa, o qual insere as triplas em documentos HTML (BIZER; HEATH; BERNERS-LEE, 2009).

Para disponibilizar as triplas na web, existem diversas ferramentas que servem tanto para armazenar quanto para prover uma interface Linked Data para visualizar dados de fontes não-RDF. Estas ferramentas ocultam detalhes técnicos e garantem que os dados publicados obedecem aos princípios básicos de Linked Data (BIZER; HEATH; BERNERS-LEE, 2009).

2.2.5.1. SPARQL

Assim como bases relacionais utilizam SQL para efetuar consultas, a web semântica também necessita de um padrão para o mesmo fim, função desempenhada pela linguagem de consulta SPARQL (W3C, [2013?]b). Sendo o

RDF a base para o armazenamento dos dados em *Linked Data*, a sintaxe e semântica da consulta SPARQL são definidas por ele (W3C, 2013).

A maioria das consultas contém por padrão um conjunto de triplas, chamado de padrão básico de grafo, onde cada sujeito, predicado ou objeto pode ser uma variável, que pode ou não ser selecionada para exibição. Sua estrutura é semelhante à SQL (W3C, 2013).

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
SELECT  ?name
FROM    <http://example.org/foaf/aliceFoaf>
WHERE   { ?x foaf:name ?name }
```

Figura 10: Exemplo de consulta básica SPARQL (W3C, 2013)

Além de *SELECT*, que identifica as variáveis que aparecem nos resultados, *FROM*, que define o grafo para consulta, e *WHERE*, que provê o padrão básico de grafo a ser combinado com os dados, existem outras cláusulas complementares. Algumas delas são (W3C, 2013):

- *Construct:* gera grafos com base em triplas RDF resultantes do padrão básico de grafo;

```
PREFIX foaf:   <http://xmlns.com/foaf/0.1/>
PREFIX org:    <http://example.com/ns#>

CONSTRUCT { ?x foaf:name ?name }
WHERE  { ?x org:employeeName ?name }
```

Figura 11: Exemplo de uso de *Construct* (W3C, 2013)

- *Filter:* restringe resultados para os quais a expressão de filtro é avaliada como verdadeira;

```
PREFIX  dc:  <http://purl.org/dc/elements/1.1/>
SELECT  ?title
WHERE   { ?x dc:title ?title
          FILTER regex(?title, "^SPARQL")
        }
```

Figura 12: Exemplo de uso de *Filter* (W3C, 2013)

- *Optional:* Usado para que a consulta retorne resultados, mesmo que a condição descrita após a palavra chave não seja satisfeita;

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
SELECT ?name ?mbox
WHERE  { ?x foaf:name  ?name .
         OPTIONAL { ?x  foaf:mbox  ?mbox }
       }
```

Figura 13: Exemplo de uso de *Optional* (W3C, 2013)

- *Union:* um meio de combinação de padrão básico de grafo;

```
PREFIX dc10:  <http://purl.org/dc/elements/1.0/>
PREFIX dc11:  <http://purl.org/dc/elements/1.1/>

SELECT ?title
WHERE  { { ?book dc10:title  ?title } UNION { ?book dc11:title  ?title } }
```

Figura 14: Exemplo de uso de *Union* (W3C, 2013)

- *Ask:* testa se o padrão básico tem solução, retornando um boolean;

```
PREFIX foaf:    <http://xmlns.com/foaf/0.1/>
ASK  { ?x foaf:name  "Alice" ;
          foaf:mbox  <mailto:alice@work.example> }
```

Figura 15: Exemplo de uso de *Ask* (W3C, 2013)

- *Describe:* retorna o RDF contendo dados sobre o recurso resultado da consulta; entre outros.

```
PREFIX foaf:   <http://xmlns.com/foaf/0.1/>
DESCRIBE ?x
WHERE    { ?x foaf:mbox <mailto:alice@org> }
```

Figura 16: Exemplo de uso de *Describe* (W3C, 2013)

Além das cláusulas descritas, outras são utilizadas com o mesmo uso de seu equivalente em SQL, como o *ORDER BY, LIMIT* e *DISTINCT* (W3C, 2008).

2.3. Publicação de Séries Estatísticas

2.3.1. Data Warehouse

Um *Data Warehouse* pode ser definido como um repositório de dados que podem ter sido coletados de diversas fontes e cuja função é prover apoio à tomada de decisão (KIMBALL, 1998). Está organizado por assuntos, ou seja, possui estruturas menores especializadas: os *Data Marts*; é integrado; não-volátil, o que significa dizer que depois de carregados, os dados não serão mais alterados; e variável no tempo (INMON, 1997).

A base de um *Data Mart* é o modelo dimensional. Ele contém as mesmas informações que foram retiradas do modelo Entidade-Relacionamento (E/R), ou seja, da base de dados transacional, alvo do processo de *Data Warehousing*, porém armazena os dados com o intuito de favorecer sua consulta e entendimento. É composto por fatos e dimensões (KIMBALL et al., 1998).

O fato é a tabela que contém as medidas, ou seja os valores propriamente ditos. Possui duas ou mais chaves estrangeiras, funcionando como chaves primárias, que apontam para as dimensões. Estas, por sua vez, contém a descrição dos itens relacionados no fato. Sua chave primária é referenciada no fato. Em virtude desta característica clássica de um fato e várias dimensões, é comumente chamado de modelo-estrela (KIMBALL et al., 1998).

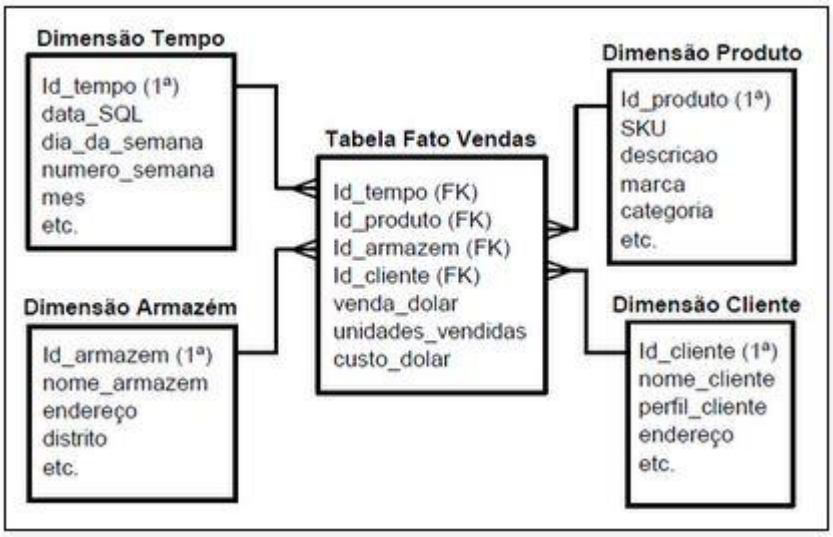

Figura 17: exemplo de modelagem dimensional (WAGNER, 2012)

A combinação de fato(s) e dimensões, que frequentemente se restringem a um processo de negócio, ou grupo de processos análogos, constitui um *Data Mart*. A união dos vários *Data Marts* compõe um *Data Warehouse* (KIMBALL et al., 1998).

A transição dos dados de um modelo E/R para o modelo dimensional de um *Data Mart* é chamada de processo de estagiamento dos dados. Ela inclui basicamente as fases de (KIMBALL et al., 1998):

1. Extração: ler os dados do banco de dados de origem e copiar o que for necessário para posterior trabalho;
2. Transformação: trabalhar os dados de forma a adequá-los de acordo com o modelo multidimensional. Nesta fase é comum que haja a limpeza dos dados, ou seja, tratar formatos de valores, excluir valores não interessantes, combinar fontes, criar as chaves primárias das dimensões e vinculá-las ao fato e formar agregações para melhorar a performance das consultas;

3. Carga e indexação: envolve a passagem dos dados transformados da área de estagiamento para o *Data Mart*, bem como sua indexação para facilitar consultas; e
4. Checagem de qualidade: verificação da qualidade dos dados, das fórmulas aplicadas, das séries temporais.

2.3.2. Data Cube Vocabulary

Ao analisar dados estatísticos, percebe-se que há um conjunto de valores observados organizados em torno de um grupo de dimensões, juntamente com os metadados associados. A publicação deste formato de dados se adapta perfeitamente ao modelo de cubo, provido pelo *Data Cube Vocabulary*, que utiliza RDF para representar a informação multidimensional e segue os princípios de *Linked Data (W3C, 2013)*.

O modelo de cubo apresentado pelo *Data Cube Vocabulary* é composto por *(W3C, 2013)*:

- Dimensões: são a descrição que identifica a observação;
- Medidas: representam o que está sendo observado;
- Atributos: qualificam e interpretam o valor observado, como por exemplo, unidades de medida, fatores de escala, entre outros.

Sobre o cubo já pronto é possível definir *slices*, que agrupam um subconjunto de observações, mantendo fixos os valores de algumas dimensões, podendo então ter todas as observações em um só indicador. Cada *slice*, assim, pode representar uma série temporal dos valores, o que é comum quando se trabalha com dados estatísticos *(W3C, 2013)*.

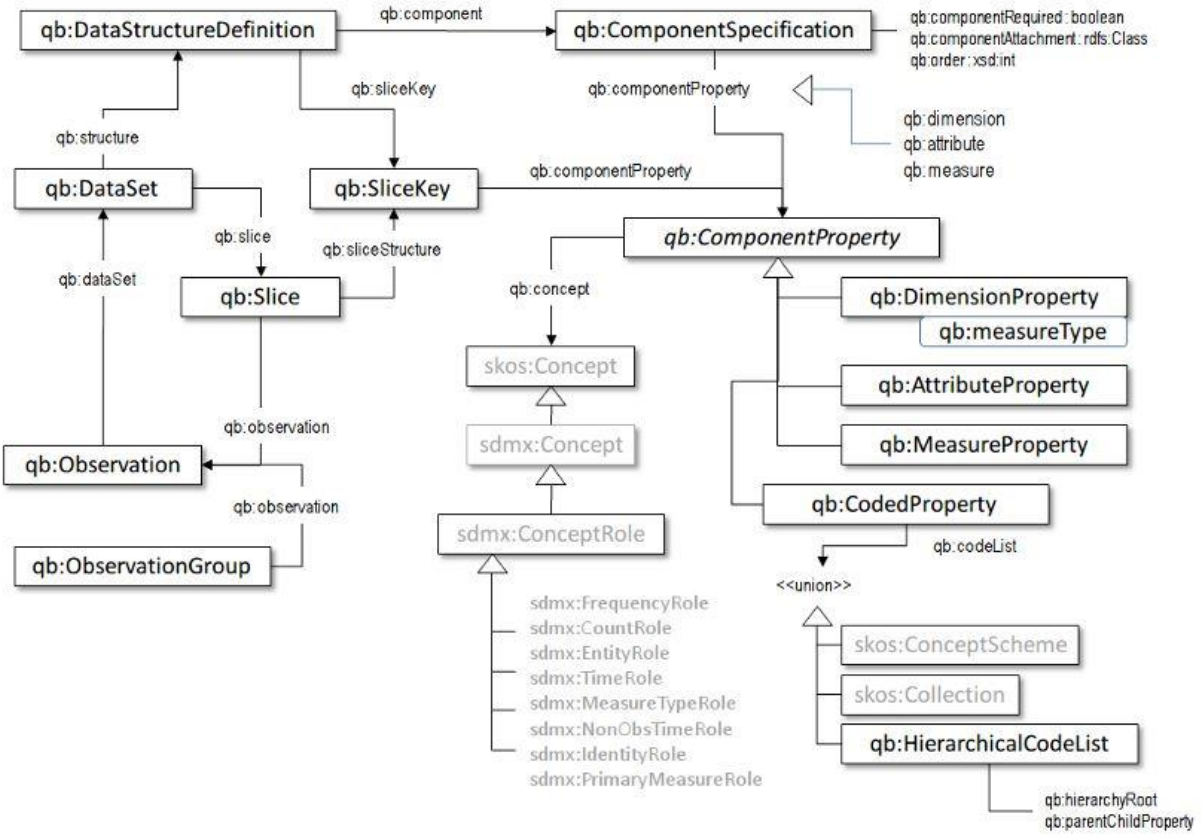

Figura 18: Estrutura básica do *Data Cube Vocabulary (W3C, 2013)*

A Figura 18 ilustra como estão organizados os conceitos a serem aplicados aos dados. O elemento inicial é o *qb:DataStructureDefinition* (DSD), que define a estrutura de organização dos dados, permitindo que uma série de publicações possam seguí-la, garantindo coerência em todas as séries nas quais é aplicada *(W3C, 2013)*.

Os componentes de dimensão, atributo e medida são representados por propriedades RDF, sendo cada um uma instância de *qb:ComponentProperty*, o qual representa um *qb:ComponentSpecification.* A propriedade do componente encapsula o conceito representado, a natureza do componente e o código usado para representar o valor. Complementarmente e visando o reuso, o conceito representado pode referenciar um outro conceito já existente, neste caso, sugere-se a utilização do vocabulário SKOS, encontrado no endereço http://www.w3.org/TR/2009/NOTE-skos-primer-20090818/. O código pode ser representado por uma *qb:CodeList*, que, como sugestão, pode ser vinculada a uma outra estrutura do vocabulário SKOS, a *skos:ConceptScheme , skos:Collection* ou *qb:HierarchicalCodeList (W3C, 2013).*

```
eg:dsd-le a qb:DataStructureDefinition;
    # The dimensions
    qb:component [qb:dimension eg:refArea;         qb:order 1];
    qb:component [qb:dimension eg:refPeriod;       qb:order 2];
    qb:component [qb:dimension sdmx-dimension:sex; qb:order 3];
    # The measure(s)
    qb:component [qb:measure eg:lifeExpectancy];
    # The attributes
    qb:component [qb:attribute sdmx-attribute:unitMeasure;
                  qb:componentRequired "true"^^xsd:boolean;
                  qb:componentAttachment qb:DataSet;] .
```

Figura 19: Exemplo de DSD de uma base de dados (W3C, 2013)

```
eg:refPeriod  a rdf:Property, qb:DimensionProperty;
    rdfs:label "reference period"@en;
    rdfs:subPropertyOf sdmx-dimension:refPeriod;
    rdfs:range interval:Interval;
    qb:concept sdmx-concept:refPeriod .
```

Figura 20: Exemplo de DSD de Dimensão (W3C, 2013)

```
eg:lifeExpectancy  a rdf:Property, qb:MeasureProperty;
    rdfs:label "life expectancy"@en;
    rdfs:subPropertyOf sdmx-measure:obsValue;
    rdfs:range xsd:decimal .
```

Figura 21: Exemplo de DSD de Medida (W3C, 2013)

Um *DataSet* é um conjunto de dados estatísticos, cuja estrutura é definida pela *qb:DataStructureDefinition*. Cada observação é representada com o tipo *qb:Observation*. Já as *slices* são representadas com *qb:Slice* (W3C, 2013).

Existem casos em que os dados observados possuem várias medidas. Nestas condições há duas maneiras de representa-las no *Data Cube Vocabulary* (W3C, 2013):

- Adicionando em cada grupo de dimensões todas as medidas observadas, utilizando o tipo *qb:measure.* Permite que vários valores sejam anexados a uma observação individual. Aplica-se quando os valores compartilham os mesmos atributos, já que não será possível diferenciá-los, uma vez que são declarados na mesma instância.
- Criando uma nova dimensão com o tipo *qb:measureType*, que indica qual medida está sendo apresentada. Assim, cada observação possui seus atributos individuais. Neste formato haverá muita duplicação de informação, porém esta é aceitável para garantir a correta disponibilização dos dados.

Este vocabulário foi estruturado de forma generalista para adequar-se a vários tipos de conjunto de dados, tais como dados de pesquisas, planilhas e cubos de dados OLAP (W3C, 2013).

2.4. Ferramentas

A fim de facilitar a publicação dos dados da EPAGRI no formato *Linked Data* neste trabalho, foram adotados algumas ferramentas. Nessa seção será apresentada uma breve descrição das plataformas D2RQ, que conta com a linguagem declarativa D2RQ, o servidor RDF D2R Server, bem como o banco de dados e também servidor RDF Virtuoso Open-Link Open-Source Edition, o Pentaho Data Integration, as ferramentas para desenvolvimento de ontologias Protégé e OntoKEM e a aplicação *web* OntoWiki, juntamente com sua extensão CubeViz.

2.4.1. Linguagem declarativa D2RQ

Antes da difusão da *web* semântica, ou *web* de dados, e até os dias atuais é comum que todos os dados sejam armazenados em um banco de dados relacional. Por tanto, é necessário certo esforço para publicar todo esse dado na *web* como *Linked Data*. Com tal propósito surgiu a plataforma D2RQ. Uma linguagem declarativa de mapeamento entre esquemas de bases relacionais de sistemas legados e ontologias RDF, descartando assim a necessidade de replicação dos dados em bases RDF (BIZER; SEABORNE, [2004?].

Sobre essa plataforma é possível realizar consultas em bases não RDF utilizando SPARQL, além de acessar os dados no formato *Linked Data* através da *web*. Também é disponibilizada a funcionalidade de criar *dumps* RDF para posterior carregamento em bases de triplas RDF e também acesso aos dados através da API Apache Jena[5], um *framework* para criação de aplicações da *web* semântica[6].

O mapeamento D2RQ especifica como as entidades abstraídas no banco de dados relacional e suas respectivas propriedades devem ser consideradas quando representadas através de triplas RDF. Esse trabalho é realizado basicamente por dois objetos, o *ClassMap*, que representa as entidades ou grupos de entidades, e o

[5] ***D2RQ**: Accessing Relational Databases as Virtual RDF Graphs*. Disponível em: <http://d2rq.org/>. Acesso em: 13 fev. 2013

[6] APACHE SOFTWARE FOUNDATION. **Apache Jena.** Disponível em: <http://jena.apache.org/>. Acesso em: 13 fev. 2013.

PropertyBridges, que especificam as descrições das entidades. O conteúdo de cada propriedade pode ser preenchido diretamente com os dados contidos em tabelas ou ser oriundo de um padrão ou tradução previamente definido. O D2RQ suporta também mapeamentos condicionais a nível de *ClassMap* e *PropertyBridges* (BIZER, CYGANIAK, [2006?]).

2.4.2. D2R Server

O D2R Server é uma ferramenta para publicação das bases de dados relacionais na web semântica por meio do mapeamento D2RQ[7]. A ferramenta é responsável por garantir a navegabilidade pelos dados não RDF, seja por navegadores RDF ou HTML, além de permitir consultas SPARQL (BIZER, CYGANIAK, [2006?]).

O servidor recebe todas as requisições da web e transforma em consultas SQL através do mapeamento D2RQ, essa estratégia garante a navegabilidade sobre os dados em um tempo de resposta aceitável (BIZER, CYGANIAK, [2006?]).

Caso o usuário necessite de uma implantação rápida, o D2R Server dispõe da funcionalidade de geração do mapeamento automaticamente a partir de um esquema relacional, gerando um vocabulário RDF para cada banco de dados. Os nomes de tabelas e colunas são utilizados como nomes de classes e propriedades respectivamente. Esse mapeamento pode ser customizado posteriormente de acordo com as necessidades do usuário (BIZER, CYGANIAK, [2006?].

2.4.3. Pentaho Data Integration

Esta aplicação, também conhecida por Kettle, tem como objetivo fornecer um ambiente gráfico e intuitivo para o processo de Extração, Tratamento e Carga (ou ETL, em sua sigla em inglês) de dados. Com uma interface baseada em arrastar e soltar apresenta-se como uma boa alternativa para o processo de ETL de *Data Warehouse.* Possui duas versões, a Community Edition, que é um opensource, e a versão Enterprise Edition versão paga que contempla suporte técnico, atualizações e recursos corporativos. Encontra-se atualmente em sua versão 4.4.0[8].

---

[7] ***D2R Server****: Accessing databases with SPARQL and as Linked Data*. Disponível em: <http://d2rq.org/d2r-server>. Acesso em: 13 fev. 2013.

[8] PENTAHO. Pentaho Data Integration: Kettle. Disponível em: <http://kettle.pentaho.com/>. Acesso em: 30 abr. 2013.

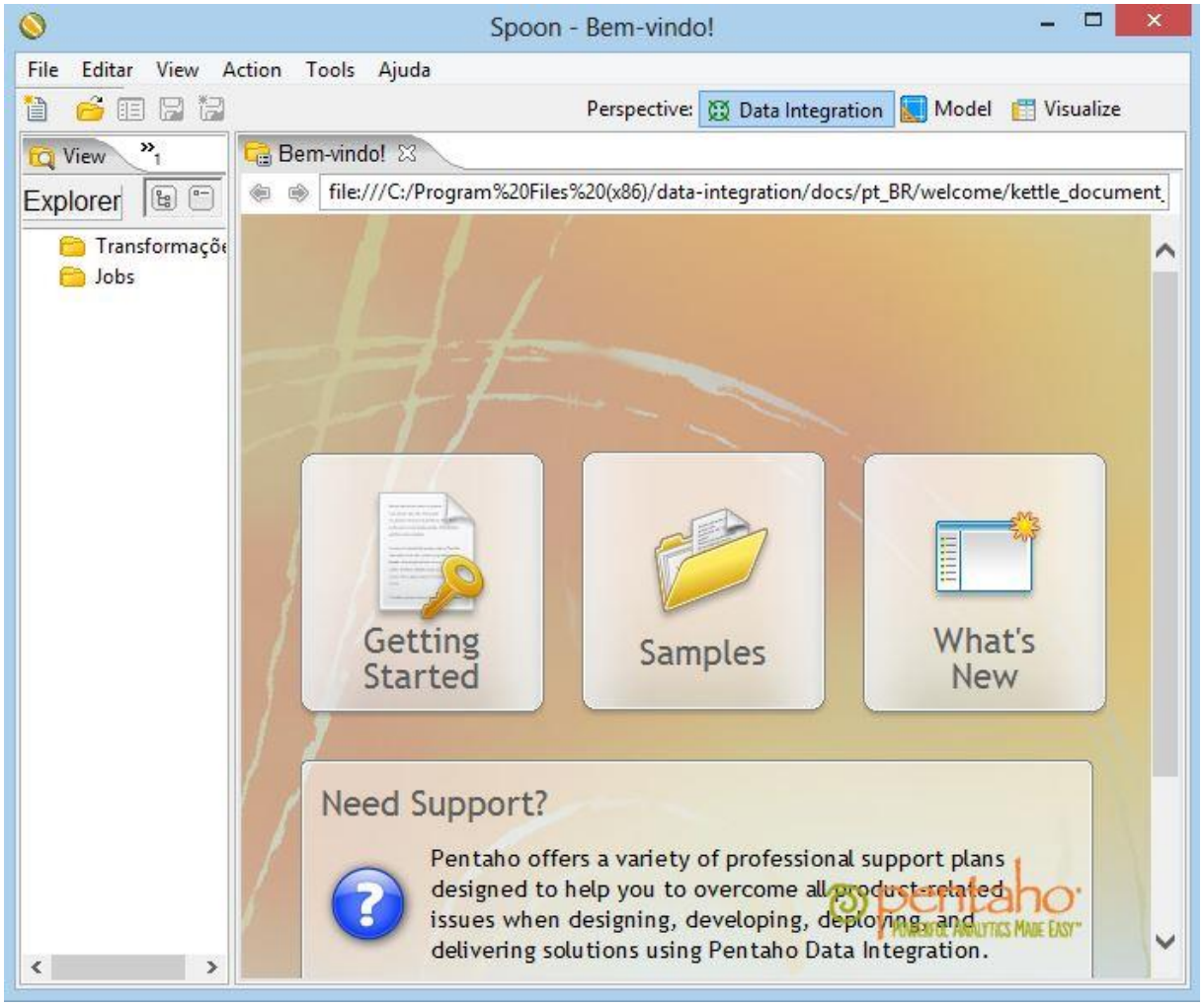

Figura 22: tela inicial do programa

Desenvolvida em 2002, por Matt Casters, foi posteriormente integrada à suíte Pentaho de *Business Inteligence*. O Pentaho *Data Integration* é constituído por várias aplicações, sendo o *Spoon* a responsável por modelar graficamente a entrada, tratamento e saída dos dados (PEIXOTO, 2011).

O Kettle é composto por *Transformations*, que é onde o tráfego de dados é modelado, possuindo inúmeras combinações de formas de entrada, de saída e de transformação dos dados. Uma ou mais *transformations* podem ser agrupadas em *Jobs* e estes também em outros *Jobs*, assim como manipular arquivos, e-mails entre outros. Cada componente de uma *transformation* ou *job* é chamado de *step*, que podem ocorrer em número indeterminado e são conectados por um *hop*, ou linha, que determina a direção do fluxo de dados (PEIXOTO, 2011).

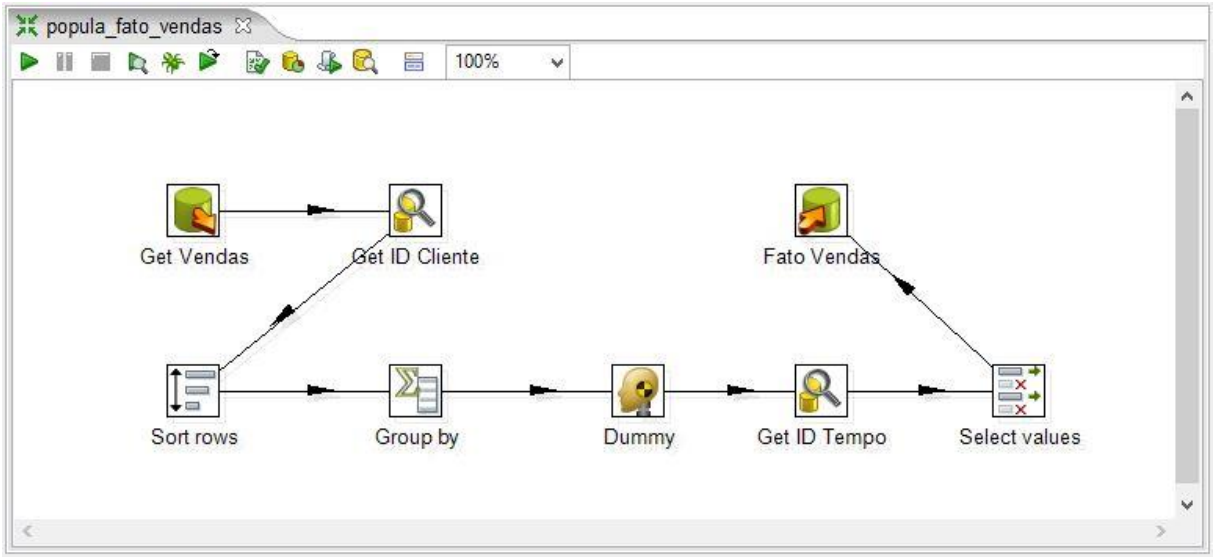

Figura 23: exemplo de transformation (PEIXOTO, 2011)

2.4.4. OntoKEM

Aplicação desenvolvida pelo Laboratório de Engenharia de Conhecimento do Programa de Engenharia e Gestão de Conhecimento da UFSC. A ferramenta tem contribuído e facilitado projetos de pesquisa e desenvolvimento e atividades de ensino de graduação e pós-graduação[9].

Trata-se de uma ferramenta para construção e documentação de ontologias que combina basicamente três metodologias. A metodologia 101, que propõem sete passos iterativos na construção da mesma, sendo eles: determinar o escopo, considerar reuso, listar termos, definir classes, definir propriedades, definir restrições e definir instâncias. Também a metodologia On-to-Knowledge, que sugere a elaboração de questões de competência como maneira de determinar mais facilmente o escopo da ontologia. E por fim a metodologia METHONTOLOGY, que formaliza uma série de artefatos para documentação do processo de construção de uma ontologia[10].

[9] LABORATÓRIO DE ENGENHARIA DO CONHECIMENTO EGC-UFSC. **O que é a OntoKEM:** Histórico. Disponível em: <http://ontokem.egc.ufsc.br/>. Acesso em: 26 maio 2013.
[10] LABORATÓRIO DE ENGENHARIA DO CONHECIMENTO EGC-UFSC. **O que é a OntoKEM:** O que é OntoKEM. Disponível em: <http://ontokem.egc.ufsc.br/>. Acesso em: 26 maio 2013.

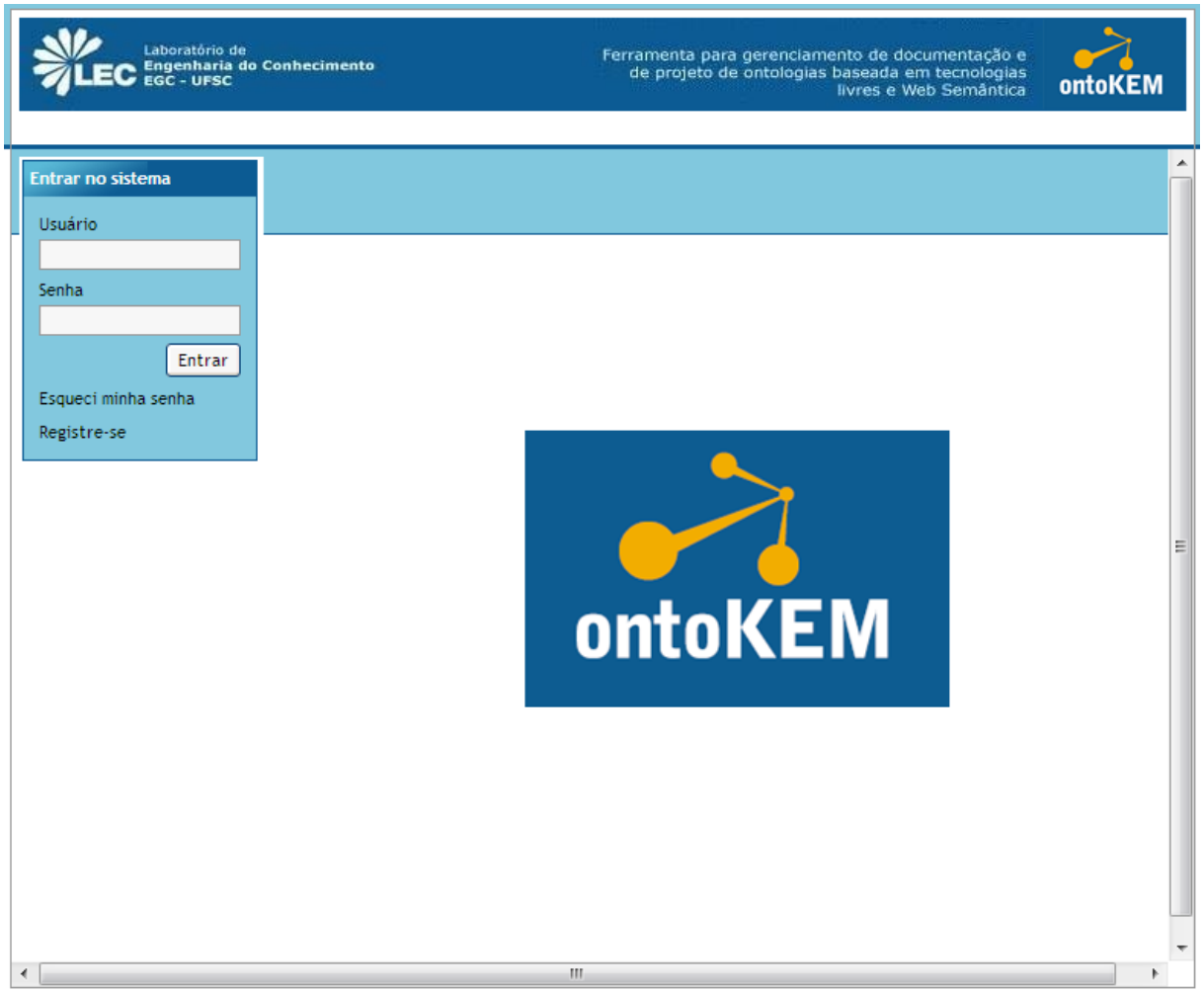

Figura 24: Tela de acesso ao OntoKEM

2.4.5. Protégé

O Protégé é uma plataforma de distribuição gratuita e de código aberto que possui uma série de funcionalidades para a construção de ontologias de representação de conhecimento e de modelos de negócio. A ferramenta implementa uma gama de estruturas e ações que possibilitam além da construção, também a visualização e manipulação de ontologias em várias formatos de representação.[11]

[11] STANFORD CENTER FOR BIOMEDICAL INFORMATICS RESEARCH. ***What is protégé?*** Disponível em: <http://protege.stanford.edu/overview/>. Acesso em: 26 maio 2013.

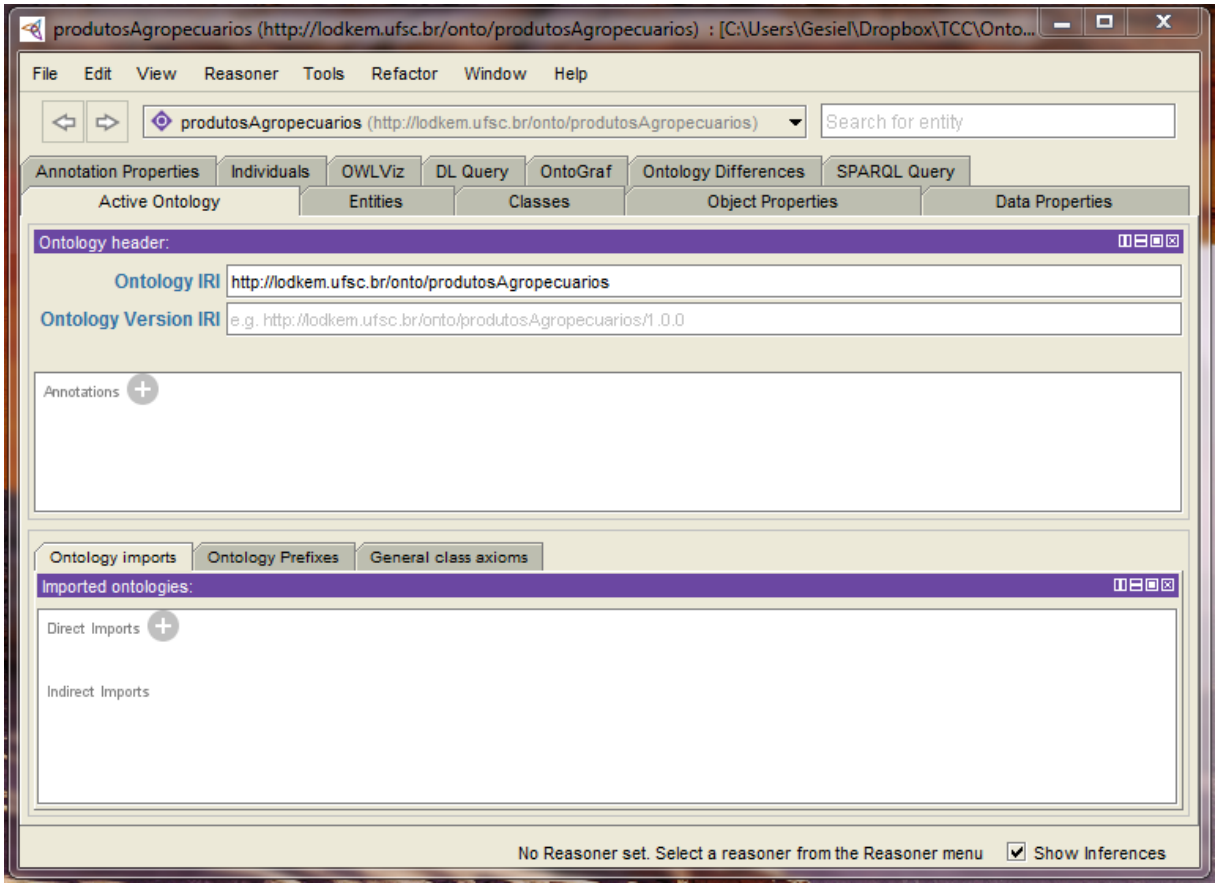

Figura 25: Tela do Protégé

2.4.5.1. Virtuoso Open-Link Open-Source Edition

A aplicação Virtuoso Open-Link é na verdade um banco de dados objeto-relacional de alto desempenho. Como uma plataforma de banco de dados, ela fornece suporte à transações, um compilador SQL e um processador de *stored-procedure* com a possibilidade de utilização de Java e .NET.[12]

Além disso, o Virtuoso possui também um servidor de aplicações web embutido, capaz de prover páginas dinâmicas escritas em linguagens como VSP, PHP, ASP e .NET. Esse servidor ainda é capaz de fornecer acesso através de requisições SOAP e REST à procedimentos armazenados na ferramenta.[13]

O que faz este trabalho tirar proveito dessa ferramenta, porém, é a capacidade do Virtuoso de suportar consultas SPARQL em uma base de dados RDF mantida pela própria plataforma. Através de uma aplicação web, chamada

[12] OPEN-LINK SOFTWARE. ***Virtuoso OpenSource Edition Introduction:****What is Virtuoso?.* Disponível em: <http://virtuoso.openlinksw.com/dataspace/doc/dav/wiki/Main/VOSIntro>. Acesso em: 26 maio 2013.
[13] OPEN-LINK SOFTWARE. **Virtuoso OpenSource Edition Introduction:**What is Virtuoso? Disponível em: <http://virtuoso.openlinksw.com/dataspace/doc/dav/wiki/Main/VOSIntro>. Acesso em: 26 maio 2013.

conductor, é possível a carga de arquivos RDF ao banco de dados de triplas do Virtuoso e também a execução de consultas SPARQL em cima desses dados.[14]

Figura 26: Página inicial do conductor, aplicação web fornecida pelo Virtuoso

2.4.5.2. OntoWiki e CubeViz

Trata-se de uma aplicação de código aberto que funciona também como um framework facilmente extensível para aplicações da Web Semântica. Um de seus principais propósito é o gerenciamento de bases de conhecimento, onde conhecimento é na verdade uma coleção de dados estruturados para leitura de máquinas, no formato RDF[15].

O CubeViz, por sua vez, é uma extensão ao OntoWiki que cria uma interface que proporciona consultas avançadas para dados estatísticos representados através do *Data Cube Vocabulary*. Baseado nos componentes do vocabulário RDF, o CubeViz disponibiliza uma série de opções que podem ser selecionadas pelos usuários para geração de gráficos com os indicadores que foram publicados[16].

[14] OPEN-LINK SOFTWARE. **Virtuoso OpenSource Edition Introduction:**What is Virtuoso? Disponível em: <http://virtuoso.openlinksw.com/dataspace/doc/dav/wiki/Main/VOSIntro>. Acesso em: 26 maio 2013.
[15] AGILE KNOWLEDGE ENGINEERING AND SEMANTIC WEB. ***What is OntoWiki?*** *What can it do for you?* Disponível em: <https://github.com/AKSW/OntoWiki/wiki>. Acesso em: 02 jun. 2013
[16] AGILE KNOWLEDGE ENGINEERING AND SEMANTIC WEB. ***CubeViz:*** *The RDF DataCube Browser*. Disponível em: <http://aksw.org/Projects/CubeViz.html>. Acesso em: 02 jun. 2013.

3. PROPOSTA

3.1. Procedimentos metodológicos

- Definição da base de dados alvo;
- Entrevista com o responsável pela base de dados na empresa;
- Revisão da literatura;
- Definição de escopo;
- Construção do modelo estrela;
- Carga e transformação dos dados para o novo banco de dados;
- Criação de uma ontologia para representar os dados;
- Publicação dos dados no formato *Linked Data*;
- Publicação dos dados com o *Data Cube Vocabulary* e interligando com a ontologia;
- Acessar os dados por meio de uma aplicação web.
- Avaliação do resultado obtido;

3.2. Definição de escopo

Ao analisar os dados e a estrutura do DATACEPA, percebe-se que este é um sistema generalista e que nele pode ser inserida e controlada qualquer informação que dependa de indicadores e qualificadores, como é o caso dos dados sobre indicadores econômicos, cadastro de produtores e outros que existem na base alvo.

Para fins deste trabalho de conclusão de curso e atendendo às expectativas da EPAGRI, foi levada em consideração, em um primeiro momento, a publicação de informações relacionadas à produção e preço agrícola dos produtos contidos na base de dados. Porém, a amostra da base para a qual foi liberado acesso ainda não continha dados de preço, resultando, assim, na restrição da área modelada, que tornou-se somente a referente aos dados de produção, tendo como grão o município, produto agropecuário e o ano para todas as medidas.

Com relação à produção agrícola foram consideradas as medidas de área plantada, área destinada à colheita, área efetivamente colhida e quantidade produzida. Para a produção pecuária, considerou-se os dados de quantidade

produzida, quantidade abatida e produção efetiva. Para garantir uma análise precisa, foi também inserida uma unidade de medida para cada um destes indicadores.

Assim sendo, com estas informações devidamente publicadas na web de dados, os produtores do mundo inteiro terão acesso a esses indicadores e serão capazes de desenvolver agentes automatizados a fim de realizarem todo o tipo de análise em cima dos dados de produção das cidades de Santa Catarina, podendo inclusive obter comparativos com seus próprios indicadores.

3.3. Construção do modelo estrela

Com base no escopo, foram analisadas várias opções de modelo, principalmente no relacionado a quantos e quais seriam os fatos. Discutiu-se sobre a possibilidade de existir vários fatos, de acordo com o tipo de dado analisado, ou seja, um fato para área plantada, outro para área produzida, e para as demais medidas apresentadas na seção 3.2. Outra opção seria a de fazer fatos de acordo com o produto analisado: alho, milho, feijão, bovino, etc.

A fim de otimizar as consultas, obter um melhor desempenho da aplicação final, uma melhor adaptação ao modelo do *Data Cube Vocabulary* e até facilitar o entendimento dos futuros usuários, optou-se pela utilização de dois fatos: um para produção agrícola e outro para produção pecuária. Decisão esta apoiada pelo fato de que as unidades de medida são diferentes entre os dos tipos de produção e também visando uma melhor organização das informações. Em complemento aos fatos há as dimensões tempo, localidade, produto e unidade de medida.

A figura 27 ilustra, de forma básica, o modelo dimensional criado. O modelo completo pode ser conferido no Apêndice A. No Apêndice B encontra-se o Script de criação do banco de dados relacional, que armazena o modelo dimensional, em MySQL.

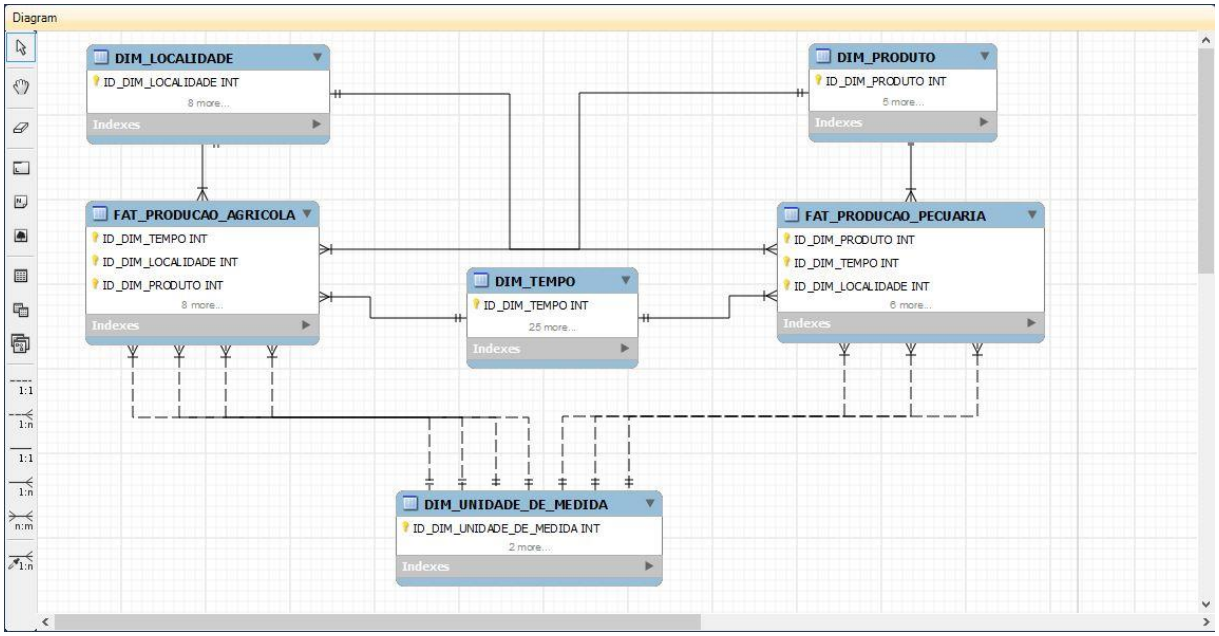

Figura 27: Modelagem dimensional resumida

As dimensões tempo e localidade são responsáveis por garantir a granularidade das medidas em ano e município, mas, apesar dos dados de produção serem anuais, decidiu-se que a dimensão tempo teria granularidade diária, possibilitando assim uma possível evolução do modelo para armazenar outro tipo de indicador.

Para dimensão produto, foi pensado em manter o mesmo nome utilizado na base de dados original do DATACEPA, tema, para facilitar a manutenção, mas com o decorrer do projeto percebeu-se que o impacto de mudar o nome para produto seria mínimo e tornaria a compreensão do modelo estrela mais natural. Esta dimensão possui a descrição dos dados sobre o escopo do projeto, ou seja, categoriza e detalha a informação dos fatos no que diz respeito ao produto analisado, seja ele agrícola ou pecuário.

Por último surgiu a necessidade da dimensão unidade de medida. Ela foi adicionada ao modelo, pois todos os indicadores poderiam possuir unidades de medida distintas, de acordo com o produto a que estivesse relacionado.

3.4. Transformação e carga dos dados para o novo banco de dados

Tendo em mãos o modelo estrela completo e aprovado, iniciou-se a fase de extração, tratamento e carga dos dados da base do DATACEPA para a nova base modelada, utilizando a ferramenta Pentaho Data Integration. Nesta fase, o primeiro passo foi elaborar as consultas em SQL para popular as tabelas correspondentes às

dimensões localidade, produto e unidade de medida e os fatos produção agrícola e produção pecuária. Uma visão parcial pode ser constatada nas figuras a seguir, e os comandos na íntegra estão nos apêndices C a G.

cargaDimLocalidade.sql

```
1  SELECT  DISTINCT
2          l.codlocal,
3          l.nomelocal,
4          l.uf,
5          l.codibge,
6          gerEpagri.nomeregiao as regiaoEpagri,
7          microRegEst.nomeregiao as microRegEstadual,
8          regiaoNacional.nomeregiao as regiaoNacional
```

Figura 28: Carga da dimensão Localidade

cargaDimProduto.sql

```
1  SELECT DISTINCT
2         t.codtema,
3         t.nometema,
4         g.codgrupo,
5         g.nomegrupo
```

Figura 29: Carga da Dimensão Produto

cargaDimUnidadeDeMedida.sql

```
1  SELECT nomeUnidadeDeMedida,
2         codUnidadeDeMedida
3
```

Figura 30: Carga da Dimensão UnidadeDeMedida

cargaFatoAgricola.sql

```sql
SELECT  DISTINCT
        tema.codgrupo,
        tema.codsubgrupo,
        tema.codtema,
        tema.codlocal,
        TO_CHAR(tema.datainformacao, 'yyyy') AS anoRel,
        TO_CHAR(tema.datainformacao, 'mm') AS mesRel,
        TO_CHAR(tema.datainformacao, 'dd') AS diaRel,
        qtdadeProduzida.codcaracterizacao AS caracterizacaoQP,
        qtdadeProduzida.valorNumerico AS qtdadeProduzida,
        qtdadeProduzida.decimaisexibicao AS decimQP,
        qtdadeProduzida.codunidadedemedida AS unidMedidaQP,
        areaPlantada.codcaracterizacao AS caracterizacaoAP,
        areaPlantada.valorNumerico AS areaPlantada,
        areaPlantada.decimaisexibicao AS decimAP,
        areaPlantada.codunidadedemedida AS unidMedidaAP,
        areaDestinada.codcaracterizacao AS caracterizacaoAD,
        areaDestinada.valorNumerico AS areaDestinadaAcolheita,
        areaDestinada.decimaisexibicao AS decimAD,
        areaDestinada.codunidadedemedida AS unidMedidaAD,
        areaColhida.codcaracterizacao AS caracterizacaoAC,
        areaColhida.valorNumerico AS areaColhida,
        areaColhida.decimaisexibicao AS decimAC,
        areaColhida.codunidadedemedida AS unidMedidaAC
```

Figura 31: Carga do fato ProduçãoAgrícola

cargaFatoPecuaria.sql

```sql
SELECT  DISTINCT
        tema.codgrupo,
        tema.codsubgrupo,
        tema.codtema,
        tema.codlocal,
        TO_CHAR(tema.datainformacao, 'yyyy') AS anoRel,
        TO_CHAR(tema.datainformacao, 'mm') AS mesRel,
        TO_CHAR(tema.datainformacao, 'dd') AS diaRel,
        qtdadeProduzida.codcaracterizacao,
        qtdadeProduzida.valorNumerico AS qtdadeProduzida,
        qtdadeProduzida.decimaisExibicao AS decimQP,
        qtdadeProduzida.codUnidadeDeMedida AS unidMedidaQP,
        efetivo.codcaracterizacao,
        efetivo.valorNumerico AS efetivo,
        efetivo.decimaisExibicao AS decimEF,
        efetivo.codUnidadeDeMedida AS unidMedidaEF,
        qtdadeAbatida.codcaracterizacao,
        qtdadeAbatida.valorNumerico AS qtdadeAbatida,
        qtdadeAbatida.decimaisExibicao AS decimQA,
        qtdadeAbatida.codUnidadeDeMedida AS unidMedidaQA
```

Figura 32: Carga do fato ProduçãoPecuária

A tabela que possui os dados da dimensão tempo foi populada diretamente no banco de dados de destino, através de uma procedure MySQL, criada especificamente para este fim, cuja estrutura básica consta na Figura 33. Seu conteúdo completo pode ser analisado no apêndice H.

populaDimensaoTempo.sql

```
1  DELIMITER //
2  CREATE PROCEDURE dm_siagro.populaDimensaoTempo(dataInicial DATE,dataFinal DATE)
3  BEGIN
4      DECLARE dataAtual DATE;
5      DECLARE ano SMALLINT;
6      DECLARE mes SMALLINT;
7      DECLARE dia SMALLINT;
8      DECLARE diaSemana SMALLINT;
9      DECLARE fimSemana CHAR(1);
10     DECLARE nomeFeriado VARCHAR(20);
11     DECLARE preFeriado CHAR(1);
12     DECLARE feriado CHAR(1);
13     DECLARE posFeriado CHAR(1);
14     DECLARE nomeMes VARCHAR(20);
15     DECLARE nomesMesAbreviado CHAR(3);
16     DECLARE diaUtil CHAR(1);
17     DECLARE nomeDiaSemana VARCHAR(20);
18     DECLARE nomeDiaSemanaAbreviado CHAR(3);
19     DECLARE bimestre SMALLINT;
20     DECLARE trimestre SMALLINT;
21     DECLARE numeroSemanaMes SMALLINT;
22     DECLARE estacaoAno VARCHAR(10);
23     DECLARE anoChar CHAR(4);
24
```

Figura 33: Procedure de criação da dimensão tempo

Em seguida, utilizando como base as consultas no banco relacional e as funcionalidades do programa Petaho Data Integration, criou-se a estrutura de *steps* para tratar e carregar os dados no Data Mart. Para cada um dos fatos ou dimensões utilizou-se uma transformação e, ao final, um *job* para uni-los em sequência.

Nas três transformações que proveem os dados para as dimensões, os *steps* basicamente inserem a chave primária artificial de cada tabela. Para as tabelas Localidade e Produto, existem informações adicionais, necessárias para adequar a tabela à Ontologia, que não puderam ser inseridas automaticamente. São elas a informação de praça para Localidade, bem como a categoria para Produto, inseridas manualmente no banco de dados.

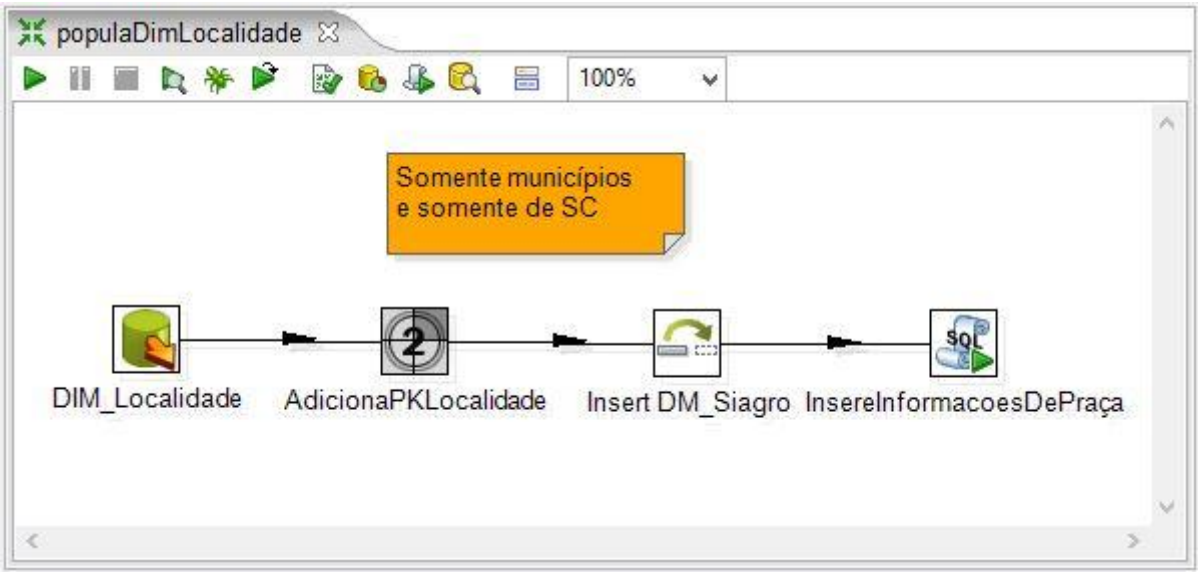

Figura 34: Estrutura de *steps* para popular a dimensão Localidade

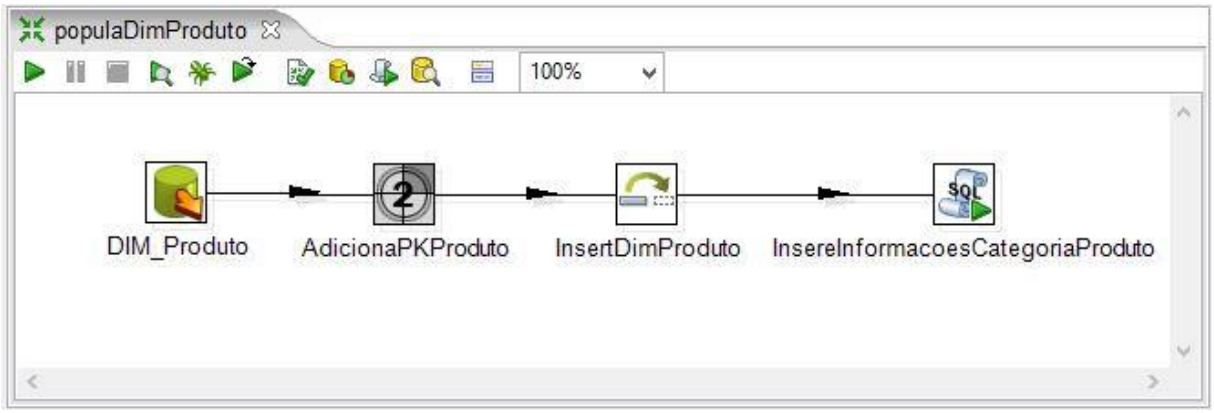

Figura 35: Estrutura de *steps* para popular a dimensão Produto

Figura 36: Estrutura de *steps* para popular a dimensão UnidadeDeMedida

Já para popular os fatos, diversas ações foram tomadas, como mapear a chave estrangeira correspondente às dimensões em cada registro, ajustar os valores

conforme a unidade decimal definida no banco de dados relacional e adequação de valores para atender às exigências do modelo dimensional.

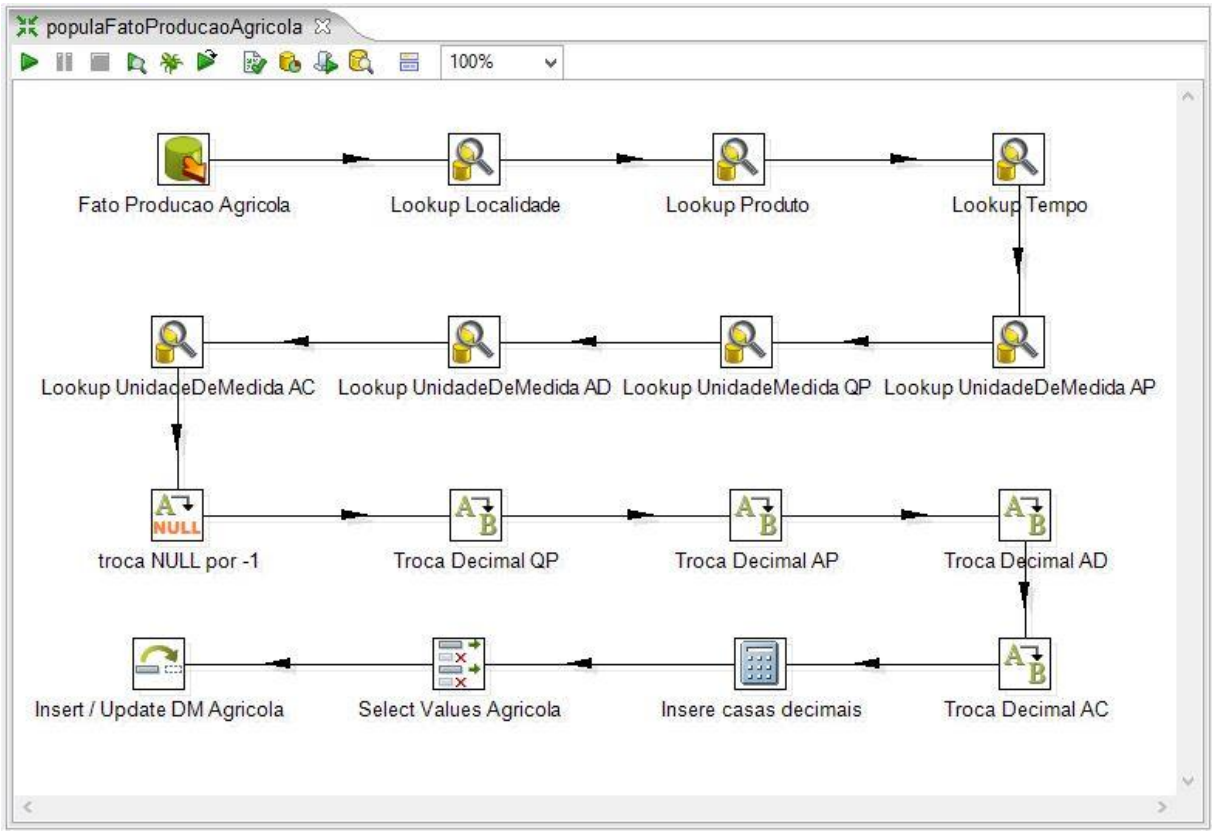

Figura 37: Estrutura de *steps* para popular o fato ProduçãoAgrícola

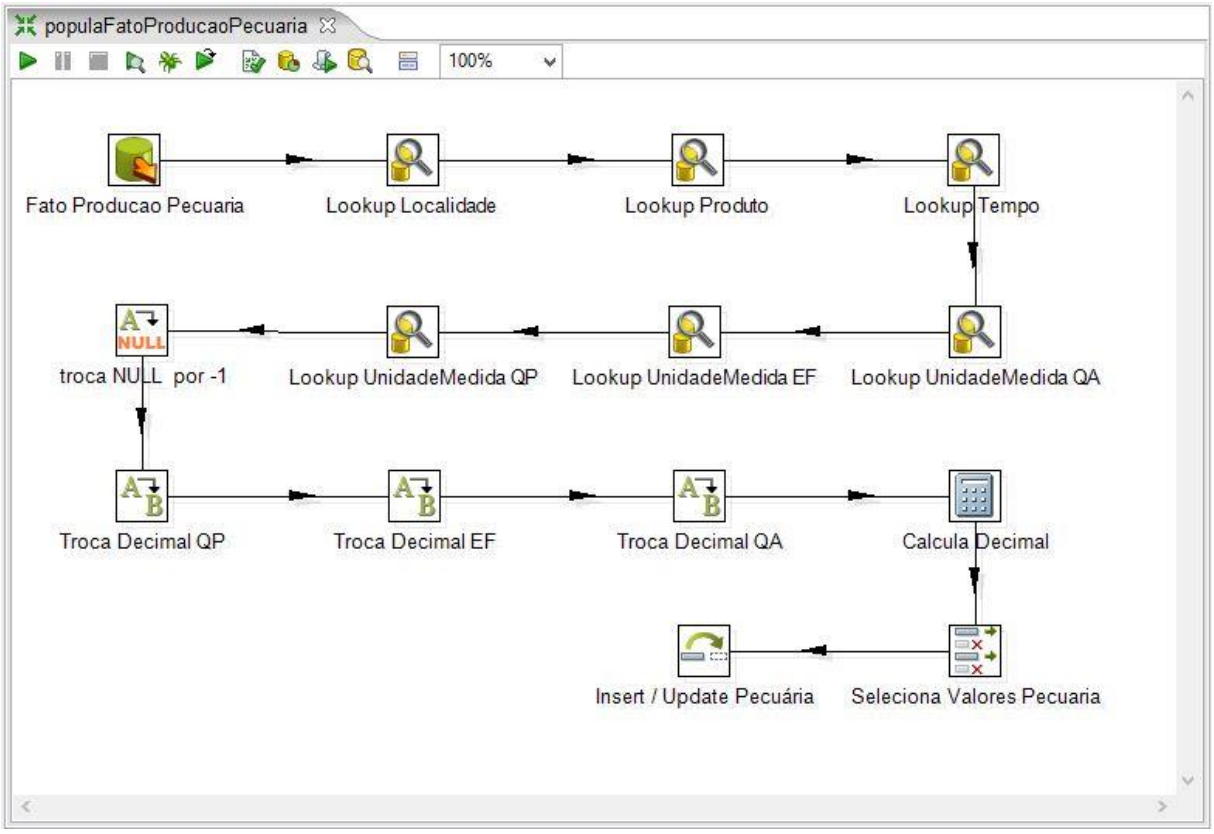

Figura 38: Estrutura de *steps* para popular o fato ProduçãoPecuária

O *job* descrito na Figura 39 tem por objetivo facilitar a correta execução sequencial das transformações da carga inicial dos dados.

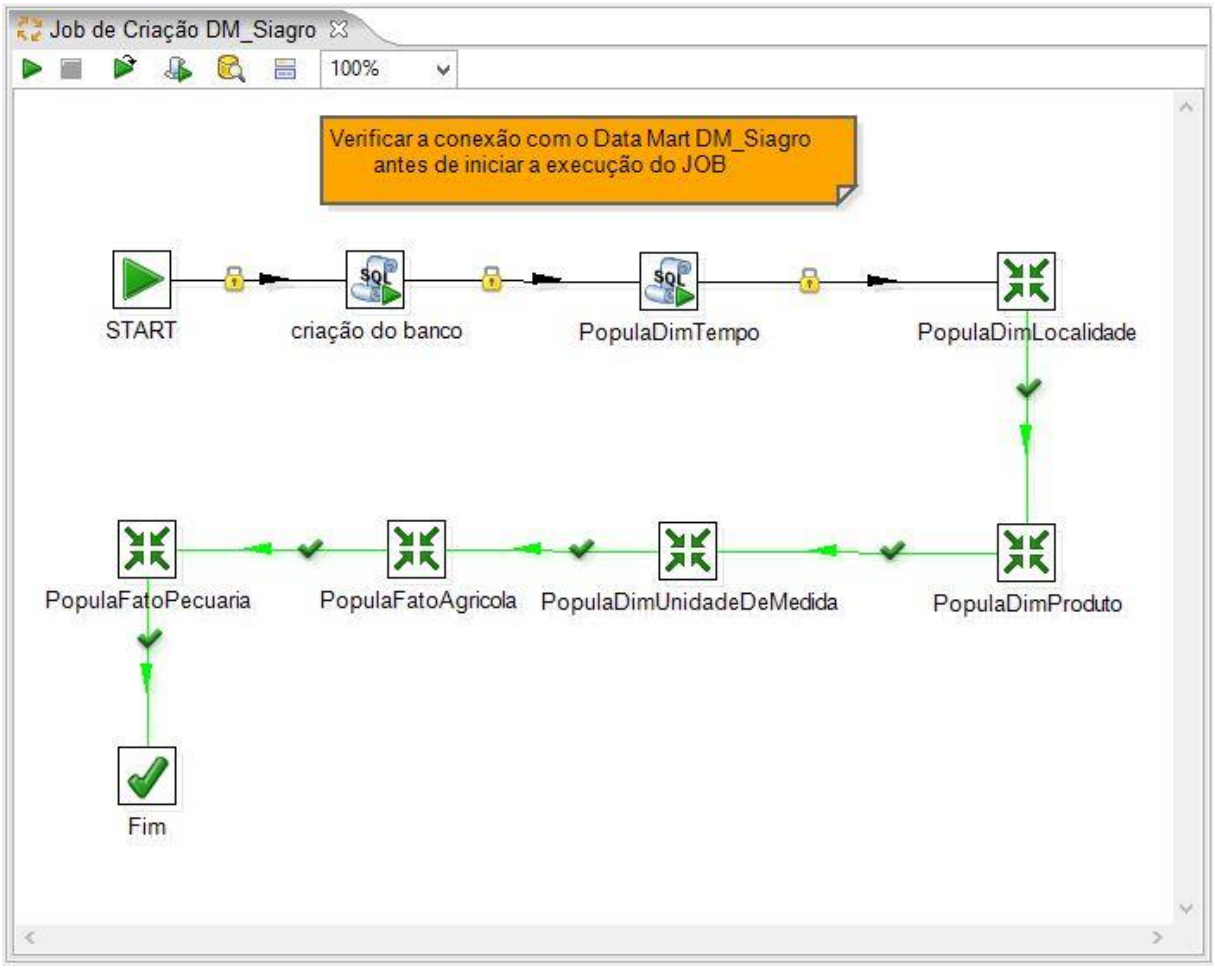

Figura 39: *Job* de carga e transformação dos dados do *Data Mart* DM_Siagro

3.5. Ontologia

Com o modelo estrela construído e devidamente carregado, faz-se necessária a criação de uma ontologia capaz de representar os dados de produção agropecuária cedidos pela EPAGRI.

Como explicitado na seção 2.2.3, essa ontologia será responsável por garantir uma comunicação formal entre o *data set* construído e publicado neste trabalho de conclusão de curso com os demais já disponíveis na web, além de possibilitar que aplicativos sejam capazes de realizar leituras automatizadas dos dados.

3.5.1. Metodologia de criação da ontologia

O desenvolvimento da ontologia para este trabalho foi realizado com base no estudo de Natalia F. Noy e Deborah L. McGuinness, que propuseram uma metodologia de criação de ontologias (NOY; MCGUINNESS, 2001). Assim, como determina o primeiro passo dessa metodologia, a primeira preocupação foi em definir o domínio e escopo da ontologia.

Levando em consideração o principal objetivo deste trabalho, publicar os dados estatísticos de produção agropecuária no formato *Linked Data* utilizando o *Data Cube Vocabulary,* e que as observações criadas pelo *Data Cube* apontariam para instâncias de municípios, produtos agropecuários e da unidade temporal ano, faz-se necessária a definição de uma ontologia que permita a criação de instâncias desses conceitos. Porém, já existem ontologias para alguns dos conceitos apresentados anteriormente, como por exemplo municípios e ano.

Chegou-se à conclusão de que era necessária uma ontologia capaz de categorizar os produtos agropecuários de acordo com aspectos como agrícola, pecuário e o tipo do produto, uma vez que essa ontologia seria utilizada na publicação de dados estatísticos e isso seria fundamental para agrupar os produtos em consultas analíticas. Enquanto que, para representação dos municípios e da unidade temporal ano, seria criada uma extensão à ontologia Geopolítica do Brasil e utilizada a ontologia SDMX[17], respectivamente.

Ainda nessa mesma etapa, Noy e McGuinness propõem o levantamento de questões de competências, ou seja, questões que a ontologia deveria ser capaz de responder no intuito de facilitar a definição do escopo. Esse procedimento não foi realizado, pois o escopo deste trabalho era bem definido e específico, além de que todos os dados necessários para que o objetivo final fosse atingido já se encontravam no *Data Mart* desenvolvido em passos anteriores.

Após determinado o que exatamente a ontologia deveria representar e quais ontologias poderiam ser reutilizadas, chegou o momento de enumerar todos os termos relevantes que poderiam fazer parte da ontologia em desenvolvimento. Esse procedimento foi realizado com o auxílio da ferramenta ontoKEM, desenvolvida pelo Laboratório de Engenharia de Conhecimento da UFSC, como mostra a Figura 40.

[17] ***SDMX****: Statistical Data and Metadata eXchange*. Disponível em: <http://sdmx.org/>. Acesso em: 19 maio 2013.

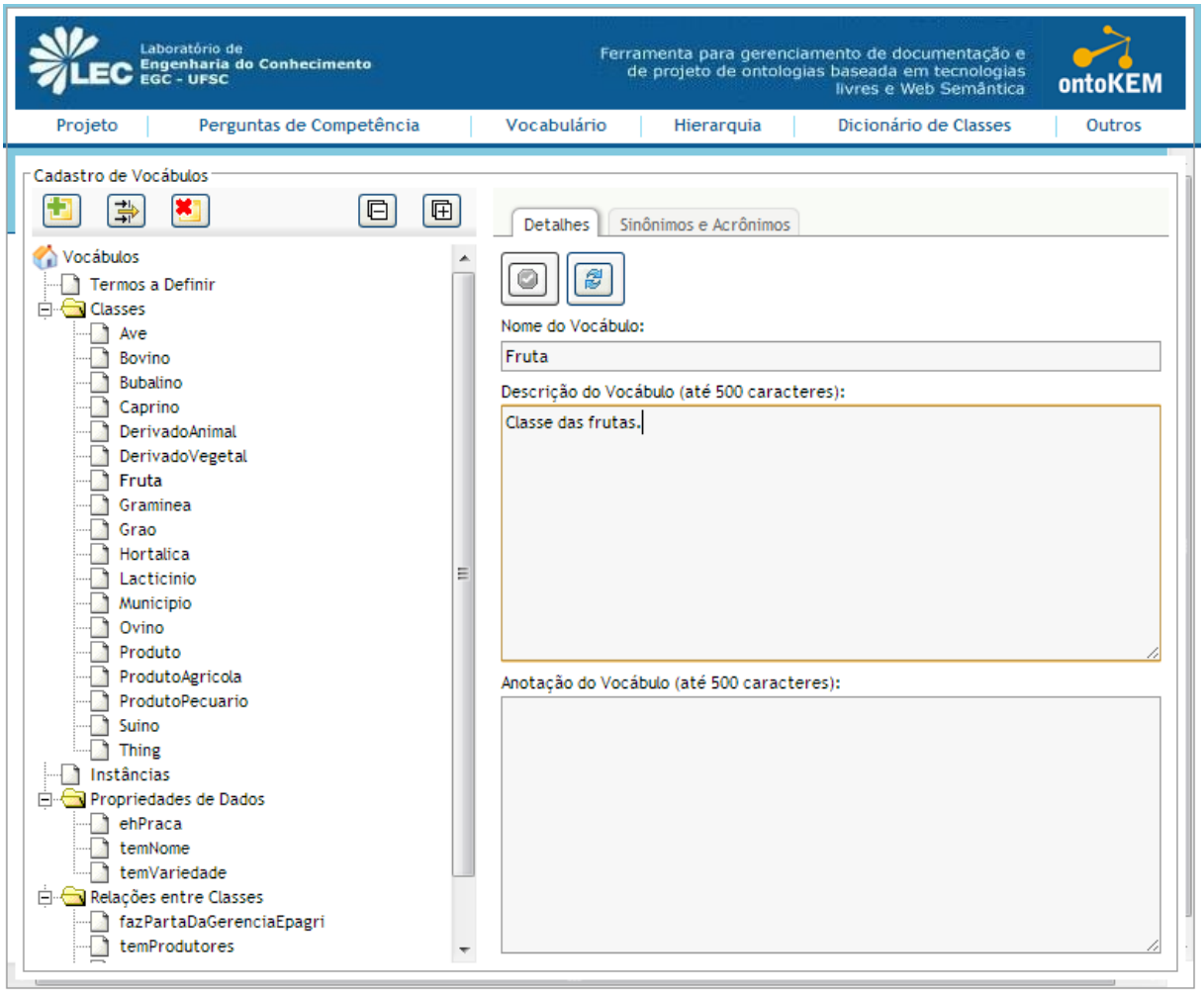

Figura 40 - Enumeração de termos no sistema ontoKEM

Como nessa etapa já era de conhecimento dos desenvolvedores deste trabalho todas as propriedades que foram definidas anteriormente na criação do modelo estrela, além de definir os termos, eles foram classificados em classes, propriedades e relacionamento entre classes, descartando, quando necessário, as propriedades já incluídas na ontologia GeopoliticaBR. Levou-se em consideração também que esses termos fariam parte das URIs identificadoras de recursos na web e, como esse identificador ainda não é internacionalizado nos dias atuais, os acentos e caracteres especiais foram desconsiderados. Diante desses fatos, os termos definidos foram:

- Classes
    - Ave
    - Bovino
    - Bubalino
    - Caprino

  - DerivadoAnimal
  - DerivadoVegetal
  - Fruta
  - Graminea
  - Grao
  - Hortalica
  - Lacticinio
  - Municipio
  - Ovino
  - Produto
  - ProdutoAgricola
  - ProdutoPecuario
  - Suino
- Propriedades de dados
  - ehPraca
  - temNome
  - temVariedade
- Relações entre classes ou propriedades de classes
  - fazParteDaGerenciaEpagri
  - temProdutores
  - temProdutos

Com a conclusão desta etapa, a metodologia adotada prevê a definição da hierarquia entre as classes classificadas no passo anterior. Esse processo também foi auxiliado pelo ontoKEM, uma vez que a ferramenta provê funcionalidades que facilitam o processo. O resultado desta etapa se encontra na Figura 41.

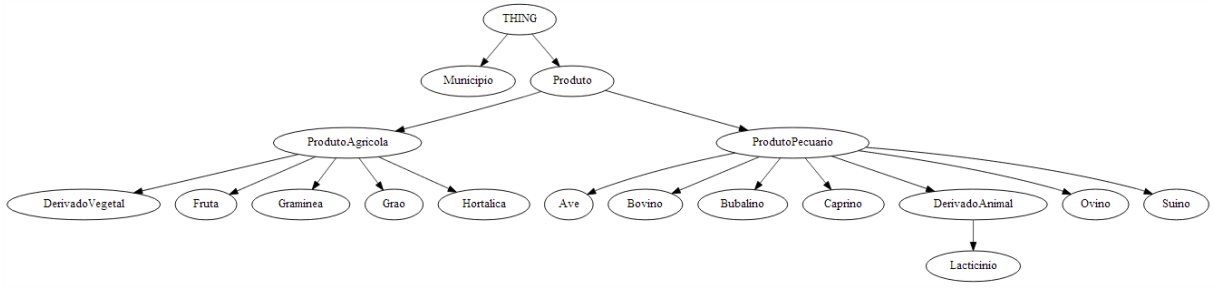

Figura 41: Hierarquia entre classes

Iniciou-se então o último processo auxiliado pelo OntoKEM, que se trata da definição do dicionário de classes. Nesse procedimento é definido a qual classe

cada propriedade pertence, ou seja, o *domain* da propriedade, e também para qual classe ou tipo de dado a propriedade vai apontar, obtendo-se assim o *range* da propriedade.

Dentre as propriedades de dados foi definido que temNome e temVariedade são propriedades da classe Produto e possuem como *range* o tipo *string*. As propriedades fazParteDaGerenciaEpagri e temProdutos possuem como *domain* a classe Municipio. Como fazParteDaGerenciaEpagri representa a gerência da EPAGRI reponsável por determinado município, determinou-se que seu *range* será a instância da classe Município, onde se encontra sediada a gerência em questão. Por sua vez, temProduto é a propriedade que identifica os produtos produzidos no município, definindo-se assim como seu *range* a classe Produto. Por fim, resta a propriedade temProdutores, que representa exatamente o inverso da propriedade temProduto.

Já com as classes, propriedades, relacionamentos e hierarquias definidos pelo ontoKEM, iniciou-se a criação das instâncias de cada classe. Para essa tarefa, o software Protégé surgiu como a melhor opção e, portanto, através de uma funcionalidade do ontoKEM de exportar a ontologia que estava em desenvolvimento até então, o processo continuou pela ferramenta Protégé. A ontologia definida até este momento pode ser conferida no apêndice I.

Pelo fato de a quantidade de produtos carregados no modelo estrela ser bem reduzida, optou-se pela criação manual dessas instâncias e nesse processo foi percebido que a ontologia criada até então não estava tão expressiva quanto era necessário. Houve a necessidade de criar mais de uma instância dos produtos feijão e batata, sendo essencial, por sua vez, agrupar essas instâncias em consultas posteriores através de uma super classe em comum. Portanto, já utilizando o Protégé, foram adicionadas novas classes e propriedades à ontologia, garantindo assim maior expressividade na representação de produtos agropecuários. Os novos termos definidos são:

- Classes
    - Fumo
    - Banana
    - Laranja
    - Maca
    - Uva

    - CanaDeAcucar
    - Arroz
    - Feijao
    - Milho
    - Soja
    - Trigo
    - Alho
    - Batata
    - Cebola
    - Mandioca
    - Tomate
    - Codorna
    - Galinha
    - GaloFrangoPinto
    - VacaOrdenhada
    - MelDeAbelha
    - Ovos
    - Leite
    - GerenciaEpagri
- Propriedades de dados
    - temSafra
    - ehGerenciaEpagriDe

Com a definição dessas novas classes e propriedade, foi necessária também a atualização da hierarquia entre as classes e do próprio dicionário de classes, tudo sendo realizado pelo próprio Protégé. A nova propriedade temSafra possui como *domain* a classe ProdutoAgricola, uma vez que faz sentido somente para essa categoria de produtos, e como *range* o tipo *string*. Além disso, a propriedade fazParteDaGerenciaEpagri teve seu *range* atualizada para a nova classe GerenciaEpagria, enquanto também foi criada uma nova propriedade inversa a esta, chamada ehGerenciaEpagriDe, com o *domain* GerenciaEpagri e o *range* Municipio. A hierarquia atualizada e final é representada pela Figura 42.

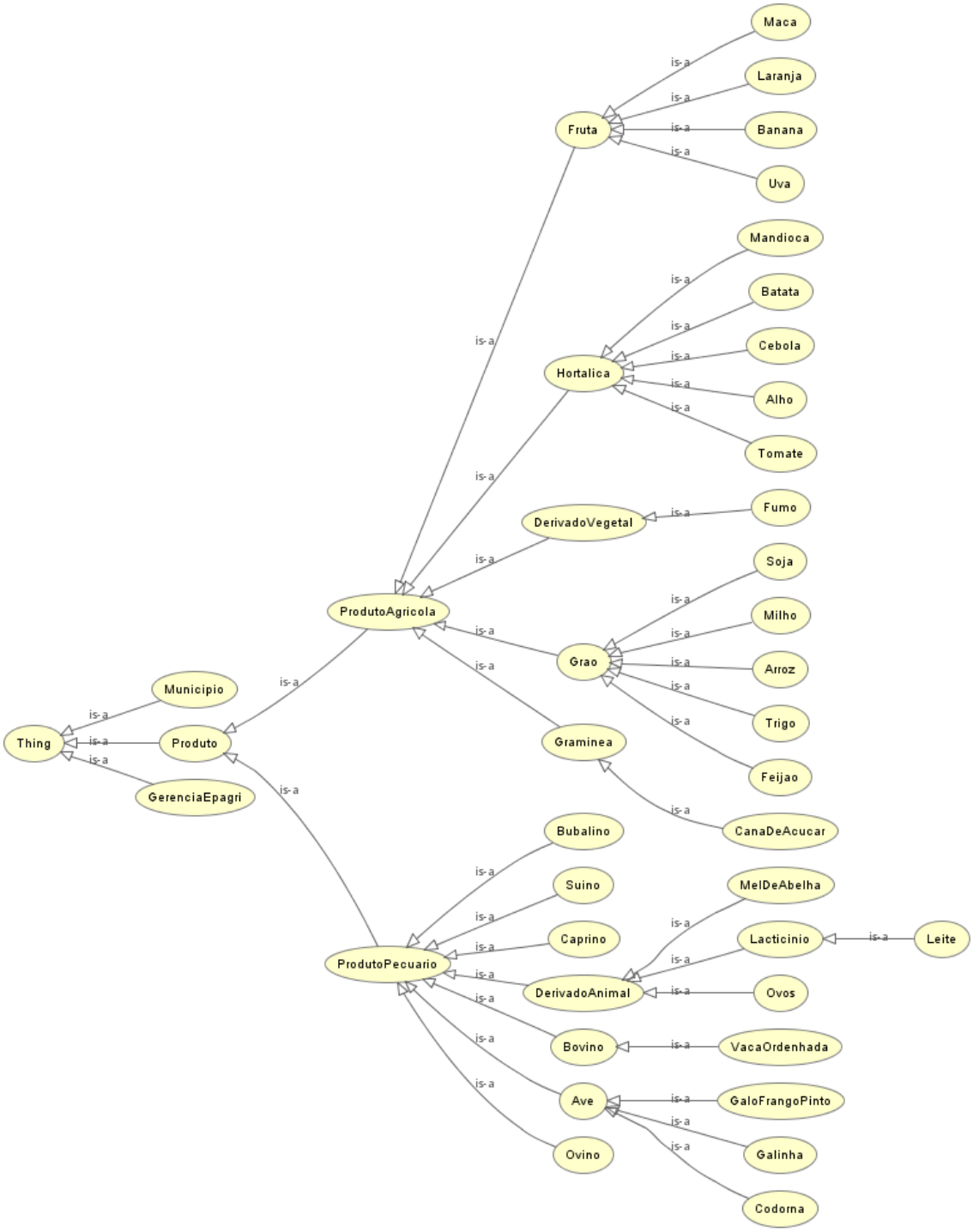

Figura 42: Hierarquia de classes final

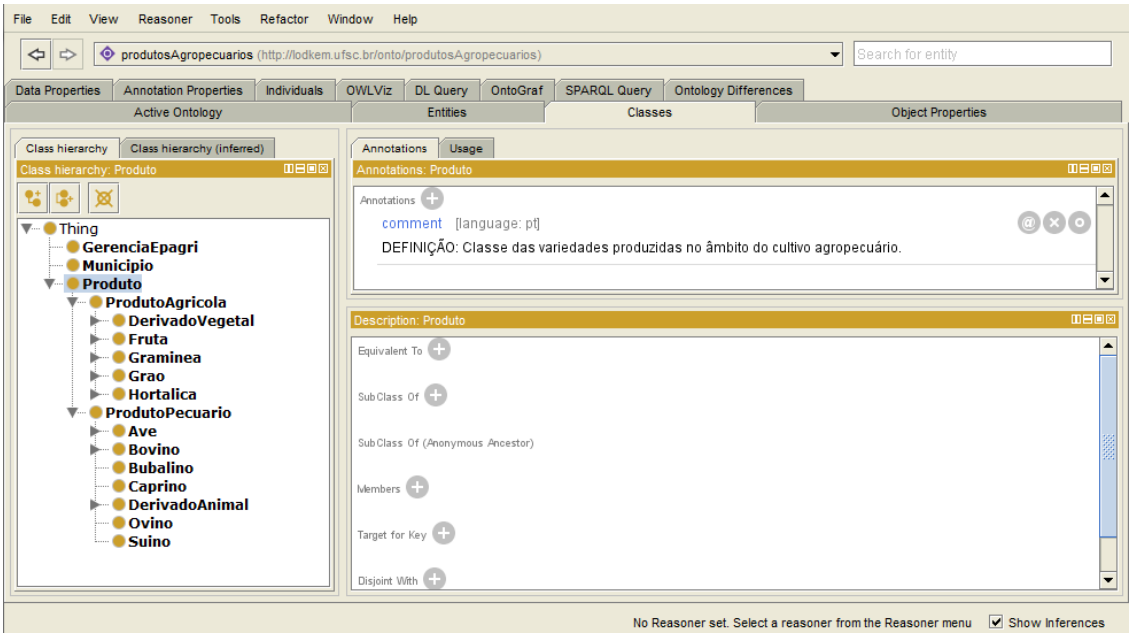

Figura 43 - Aprimoramento da ontologia no Protégé

Por fim, foram criadas as instâncias de cada produto contido no *Data Mart* carregado anteriormente na seção 3.4 e as instância da nova classe GerenciaEpagri. O mesmo procedimento, porém, não era trivial de ser aplicado para criação das instâncias de município, levando em consideração o número de instâncias que deveriam ser criadas uma a uma em um processo manual. Para tanto, a ferramenta D2RQ foi utilizada como uma alternativa de automatizar esse procedimento.

O primeiro passo para criação das instâncias de município foi tirar proveito da funcionalidade *generate-mapping*, já citada anteriormente na seção 2.2.6.2, que gera automaticamente o mapeamento entre uma base de dados relacional e a ontologia a ser utilizada. Essa funcionalidade tem como resultado um arquivo de mapeamento, que será utilizado posteriormente na geração de um *dump* RDF que carrega todas as instâncias de município. A execução desse comando pode ser visualizada na Figura 44 e o arquivo resultante na Figura 45.

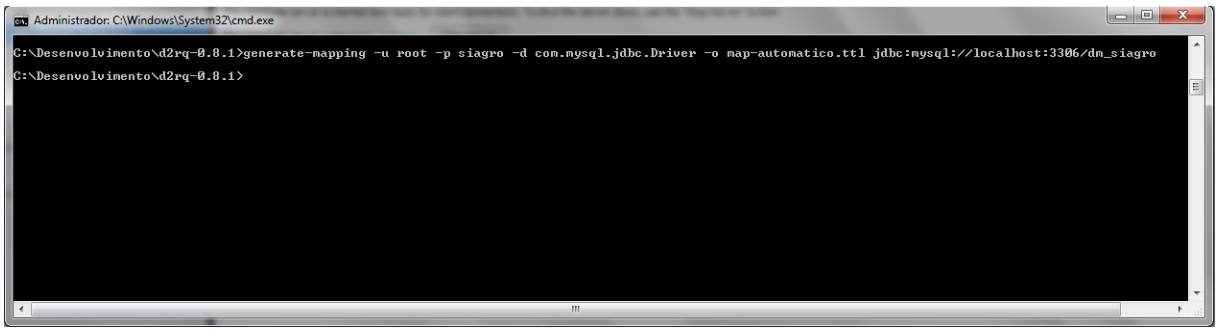

Figura 44 - Execução do comando generate-mapping

map-automatico.ttl

```
map:database a d2rq:Database;
    d2rq:jdbcDriver "com.mysql.jdbc.Driver";
    d2rq:jdbcDSN "jdbc:mysql://localhost:3306/dm_siagro";
    d2rq:username "root";
    d2rq:password "siagro";
    jdbc:autoReconnect "true";
    jdbc:zeroDateTimeBehavior "convertToNull";
    .

# Table dim_localidade
map:dim_localidade a d2rq:ClassMap;
    d2rq:dataStorage map:database;
    d2rq:uriPattern "dim_localidade/@@dim_localidade.ID_DIM_LOCALIDADE@@";
    d2rq:class vocab:dim_localidade;
    d2rq:classDefinitionLabel "dim_localidade";
    .
map:dim_localidade__label a d2rq:PropertyBridge;
    d2rq:belongsToClassMap map:dim_localidade;
    d2rq:property rdfs:label;
    d2rq:pattern "dim_localidade #@@dim_localidade.ID_DIM_LOCALIDADE@@";
    .
map:dim_localidade_ID_DIM_LOCALIDADE a d2rq:PropertyBridge;
    d2rq:belongsToClassMap map:dim_localidade;
    d2rq:property vocab:dim_localidade_ID_DIM_LOCALIDADE;
    d2rq:propertyDefinitionLabel "dim_localidade ID_DIM_LOCALIDADE";
    d2rq:column "dim_localidade.ID_DIM_LOCALIDADE";
    d2rq:datatype xsd:integer;
    .
map:dim_localidade_CODIGO_IBGE a d2rq:PropertyBridge;
    d2rq:belongsToClassMap map:dim_localidade;
    d2rq:property vocab:dim_localidade_CODIGO_IBGE;
    d2rq:propertyDefinitionLabel "dim_localidade CODIGO_IBGE";
    d2rq:column "dim_localidade.CODIGO_IBGE";
    .
```

Figura 45 - Arquivo resultante da execução do generate-mapping

A geração automática do arquivo de mapeamento é ideal como um primeiro passo no processo de automatização de criação de instâncias, no entanto, a execução desse comando mapeia todas as tabelas e propriedades contidas no banco de dados para classes, propriedades e relacionamento entre classes.

Levando em conta que o intuito da utilização da geração automática do mapeamento nesse trabalho foi apenas para gerar as instâncias da classe Município, o arquivo resultado da execução do comando *generate-mapping* teve de ser alterado para esse propósito. O ajuste foi realizado manualmente, removendo o mapeamento das demais tabelas que não a de município e das propriedades irrelevantes. Além disso, foram ajustados também a URI identificadora da ontologia criada até então e os relacionamentos entre a classe Município e suas propriedades fazParteDaGerenciaEpagri e ehPraca.

mapeamento-municipio.ttl

```
@prefix map: <#> .
@prefix db: <> .
@prefix vocab: <http://lodkem.ufsc.br/onto/produtosAgropecuarios#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix d2rq: <http://www.wiwiss.fu-berlin.de/suhl/bizer/D2RQ/0.1#> .
@prefix jdbc: <http://d2rq.org/terms/jdbc/> .
```

Figura 46: Adição da URL identificadora da ontologia ao prefixo gerado pelo D2RQ

mapeamento-municipio_disc.ttl

```
# Table dim_localidade
map:dim_localidade a d2rq:ClassMap;
    d2rq:dataStorage map:database;
1   d2rq:uriPattern "http://lodkem.ufsc.br/resource/municipio@@dim_localidade.CODIGO_IBGE@@";
    d2rq:class producao:Municipio;
    .

map:dim_localidade_gerencia_reg_epagri a d2rq:PropertyBridge;
    d2rq:belongsToClassMap map:dim_localidade;
2   d2rq:property producao:fazPartaDaGerenciaEpagri;
    d2rq:join "dim_localidade.GERENCIA_REG_EPAGRI => gerencia_epagri.CIDADE";
    d2rq:alias "dim_localidade AS gerencia_epagri";
    d2rq:uriSqlExpression "CONCAT('http://lodkem.ufsc.br/resource/gerenciaEpagri', gerencia_epagri.codigo_ibge)";
    .

map:dim_localidade_PRACA a d2rq:PropertyBridge;
    d2rq:belongsToClassMap map:dim_localidade;
3   d2rq:property producao:ehPraca;
    d2rq:column "dim_localidade.PRACA";
    d2rq:translateWith map:Praca;
    d2rq:datatype xsd:boolean;
    .

map:Praca a d2rq:TranslationTable;
4       d2rq:translation [ d2rq:databaseValue "0"; d2rq:rdfValue false; ];
        d2rq:translation [ d2rq:databaseValue "1"; d2rq:rdfValue true; ];
        .
```

Figura 47: Alterações realizadas ao arquivo de mapeamento

Como demonstra a Figura 46, foi necessário adicionar ao prefixo *vocab*, gerado automaticamente pelo comando *generate-mapping*, a URL base para identificação da ontologia que descreve os produtos agropecuários, dessa maneira a classe Município e as demais propriedades definidas na ontologia puderam ser referenciadas por este arquivo.

A Figura 47 aponta todas as alterações que foram efetuadas ao arquivo de mapeamento. O mapeamento marcado com o número 1 faz a conexão entre a tabela *dim_localidade* do modelo estrela com a classe Município da ontologia. O número 2 é responsável pelo mapeamento da propriedade fazParteDaGerenciaEpagri, indicando como deve ocorrer o *join* na camada do banco relacional para que a propriedade possa então ser preenchida corretamente. Os mapeamentos indicados pelos números 3 e 4 indicam como a propriedade ehPraca deve ser persistida, indicando que o valor 0 em um registro do banco de dados deve ser traduzido para o valor *false* no RDF e o valor 1 para *true*.

Com o arquivo de mapeamento devidamente ajustado basta agora gerar o dump RDF, um arquivo com todas as instâncias da classe Município que será publicado, juntamente com o arquivo de ontologia e instâncias de produtos confeccionados no Protégé, no servidor RDF Virtuoso Open Link. Para geração desse arquivo, o D2RQ provém da funcionalidade *dump-rdf*, um comando que recebe como entrada o arquivo de mapeamento e produz como resultado o arquivo

dump RDF com todas as instâncias da classe Município. A execução do mesmo segue na Figura 48.

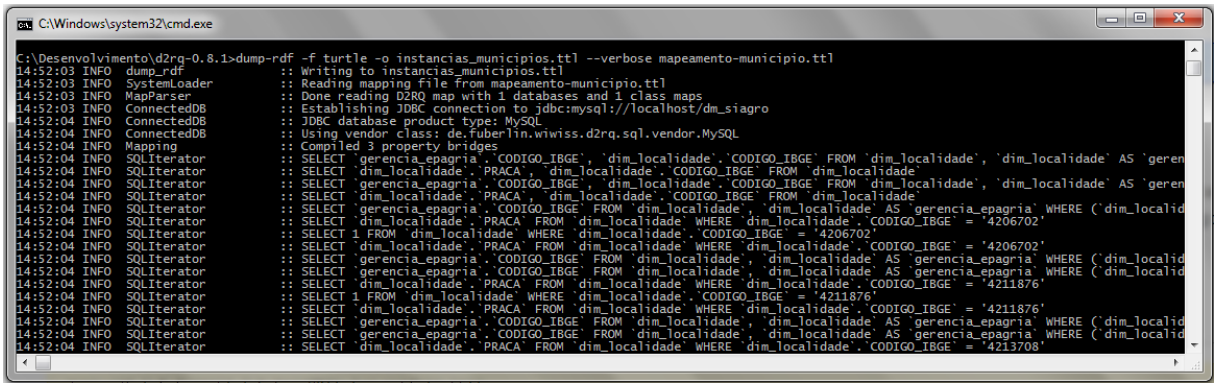

Figura 48: Execução do comando *dump-rdf* para gerar arquivo com as instâncias de Município

Tendo todas as instâncias para as classes da ontologia criadas neste trabalho, o passo consecutivo foi a geração dos arquivos RDF com os dados estatísticos utilizando o *Data Cube Vocabulary*, para tal, foi utilizado um aplicativo desenvolvido pelo laboratório de Engenharia do Conhecimento da UFSC. Trata-se de um aplicativo web, responsável por ler os dados em um arquivo CSV, onde a primeira linha deve conter as URIs que identificam as classes que representam as dimensões e medidas, enquanto as demais linhas devem conter as URIs das instâncias dessas classes e o valor das medidas.

Para gerar o arquivo CSV que servirá como entrada do aplicativo em questão, foram desenvolvidas uma série de consultas SQL que, com algumas vantagens disponibilizadas pelo MySQL, geraram arquivos CSV com o resultado já devidamente formatado com as devidas URLs para cada instância de classe. A seguir a Figura 49 exemplifica uma dessas consultas.

SQL File 1 | consultaParaGerarCSV*

```
-- consulta de Quantidade Produzida
SELECT
    concat('http://lodkem.ufsc.br/resource/municipio', l.codigo_ibge),
    concat('http://lodkem.ufsc.br/resource/produto', p.id_dim_produto),
    concat('http://reference.data.gov.uk/id/year/', tp.ano),
    'http://lodkem.ufsc.br/onto/produtosAgropecuarios#quantidadeProduzida',
    pa.quantidade_produzida,
    um.UNIDADE_DE_MEDIDA
FROM
    fat_producao_agricola pa
    INNER JOIN dim_localidade l ON pa.id_dim_localidade = l.id_dim_localidade
    INNER JOIN dim_produto p ON pa.id_dim_produto = p.id_dim_produto
    INNER JOIN dim_tempo tp ON pa.id_dim_tempo = tp.id_dim_tempo
    INNER JOIN dim_unidadedemedida um ON pa.ID_DIM_UNIDADE_DE_MEDIDA_QP = um.ID_DIM_UNIDADE_DE_MEDIDA
WHERE
    pa.quantidade_produzida IS NOT NULL AND pa.quantidade_produzida >= 0
ORDER BY
    1,2,3
LIMIT
    0, 8000
INTO OUTFILE
    'C:/temp/agrQuantidadeProduzida1.csv'
CHARACTER SET
    latin1
FIELDS TERMINATED BY
    ';'
LINES TERMINATED BY
    '\n';
```

Figura 49: Exemplo de consulta gerando arquivo CSV

Os resultados dessas consultas foram então utilizados na aplicação citada anteriormente para gerar o RDF com as observações do *Data Cube Vocabulary*, porém, com uma pequena edição, que adicionou as classes da ontologia na primeira linha dos arquivos CSVs. A aplicação pode ser visualizada na Figura 50.

Figura 50: Aplicativo para geração do RDF com *Data Cube Vocabulary*

Apesar do resultado gerado por esta aplicação já criar automaticamente uma estrutura do *Data Cube Vocabulary*, foi optado pela utilização somente das observações geradas pela aplicação, sendo necessário então a criação dos componentes *DataStructureDefinition* e *DataSet* manualmente, uma vez que era essencial a utilização de componentes bastante descritivos e específicos.

As limitações do Data Cube Vocabulary para representação de múltiplas medidas fez com que fosse criado um componente DataStructureDefinition e um DataSet para cada indicador, evitando assim possíveis problemas de representação dos dados. Essa decisão foi apoiada pela falta de compatibilidade do CubeViz, extensão da aplicação web exemplo que será utilizada neste trabalho na seção 3.7, com definições que possuem mais de uma medida.

ontologia_datacube.ttl

```
95  datacube:dsdQuantidadeProduzida a qb:DataStructureDefinition ;
96      rdfs:label "Indicador quantidade produzida."^^<http://www.w3.org/2001/XMLSchema#string> ;
97      qb:component  <http://lodkem.ufsc.br/onto/produtosAgropecuarios/datacube/cs/Municipio>,
98                    <http://lodkem.ufsc.br/onto/produtosAgropecuarios/datacube/cs/Produto>,
99                    <http://lodkem.ufsc.br/onto/produtosAgropecuarios/datacube/cs/Ano>,
100                   <http://lodkem.ufsc.br/onto/produtosAgropecuarios/datacube/cs/unidadeMedida>,
101                   <http://lodkem.ufsc.br/onto/produtosAgropecuarios/datacube/cs/quantidadeProduzida> .
102
```

Figura 51: Exemplo de um componente DataStructureDefinition

ontologia_datacube.ttl

```
143 #
144 # Data Set
145 #
146 datacube:dsQuantidadeProduzida a qb:DataSet ;
147     rdfs:label "Quantidade produzida"^^<http://www.w3.org/2001/XMLSchema#string> ;
148     rdfs:comment "Representa a coleção de observações anuais de quantidade produzida dos municípios de Santa Catarina." ;
149     qb:structure datacube:dsdQuantidadeProduzida .
```

Figura 52: Exemplo de um componente DataSet

O desenvolvimento dos componentes DataStructureDefinition e DataSet somado com o resultado do processamento da aplicação desenvolvida pelo LEC sobre os dados de entrada, juntamente com os demais arquivos com as instâncias das classes Município e Produto, pôde então ser publicado no Virtuoso Open Link, assunto este tratado com maiores detalhes na seção seguinte.

agrQuantidadeProduzida1.rdf

```
17  <qb:Observation rdf:about="http://lodkem.ufsc.br/onto/produtosAgropecuarios/datacube/quantidadeProduzidaObsA1">
18     <qb:dataSet rdf:resource="http://lodkem.ufsc.br/onto/produtosAgropecuarios/datacube/dsQuantidadeProduzida"/>
19     <ns1:Municipio rdf:resource="http://lodkem.ufsc.br/resource/municipio4200051"/>
20     <ns1:Produto rdf:resource="http://lodkem.ufsc.br/resource/produto1001"/>
21     <ns2:CalendarYear rdf:resource="http://reference.data.gov.uk/id/year/1990"/>
22     <ns3:unitMeasure>t</ns3:unitMeasure>
23     <ns1:quantidadeProduzida>360</ns1:quantidadeProduzida>
24  </qb:Observation>
25
26  <qb:Observation rdf:about="http://lodkem.ufsc.br/onto/produtosAgropecuarios/datacube/quantidadeProduzidaObsA2">
27     <qb:dataSet rdf:resource="http://lodkem.ufsc.br/onto/produtosAgropecuarios/datacube/dsQuantidadeProduzida"/>
28     <ns1:Municipio rdf:resource="http://lodkem.ufsc.br/resource/municipio4200051"/>
29     <ns1:Produto rdf:resource="http://lodkem.ufsc.br/resource/produto1001"/>
30     <ns2:CalendarYear rdf:resource="http://reference.data.gov.uk/id/year/1991"/>
31     <ns3:unitMeasure>t</ns3:unitMeasure>
32     <ns1:quantidadeProduzida>24</ns1:quantidadeProduzida>
33  </qb:Observation>
```

Figura 53: Arquivo gerado pela aplicação desenvolvida pelo LEC com as observações do *Data Cube Vocabulary*

3.6. Publicando os dados no formato *Linked Data*

Após realizada a carga dos dados no *Data Mart*, desenvolvimento da ontologia para representação do modelo na web e geração de todas as instâncias de

classes, o próximo passo é publicar efetivamente os dados no formato *Linked Data*. Para tal, fez-se necessário a utilização de um servidor RDF, nesse caso a versão *open source* do Virtuoso Open Link, software escolhido em função de trabalhos anteriores já o terem utilizado com sucesso e também por ser utilizado no servidor LodKEM, onde os dados serão publicados efetivamente para consulta do público ao final desse projeto.

O Virtuoso, além de servidor de triplas RDF, oferece uma gama de funcionalidades, como banco de dados, gerenciamento de replicações, servidor de aplicações e entre outras, o objeto principal deste trabalho, o servidor de triplas que possibilita o *Linked Data*. Através de uma interface web, chamada *Conductor*, disponibilizada pelo Virtuoso, é possível realizar duas operações básicas para se conseguir efetivamente o *Linked Data*: subir arquivos RDF e efetuar consultas SPARQL.

Figura 54: conductor, interface web disponibilizada pelo Virtuoso

Subir arquivos RDF para o servidor é uma tarefa simples, basta acessar o menu *Linked Data* e posteriormente acessar a opção *Quad Store Upload*. Nessa tela, é possível selecionar o arquivo RDF desejado e informar a URI do grafo a qual o arquivo selecionado pertence. Essa URI torna mais simples a manutenção dos arquivos carregados no servidor, uma vez que é possível excluir todos os RDFs vinculadas a determinado grafo, tarefa bastante relevante quando há problemas em informações já publicadas.

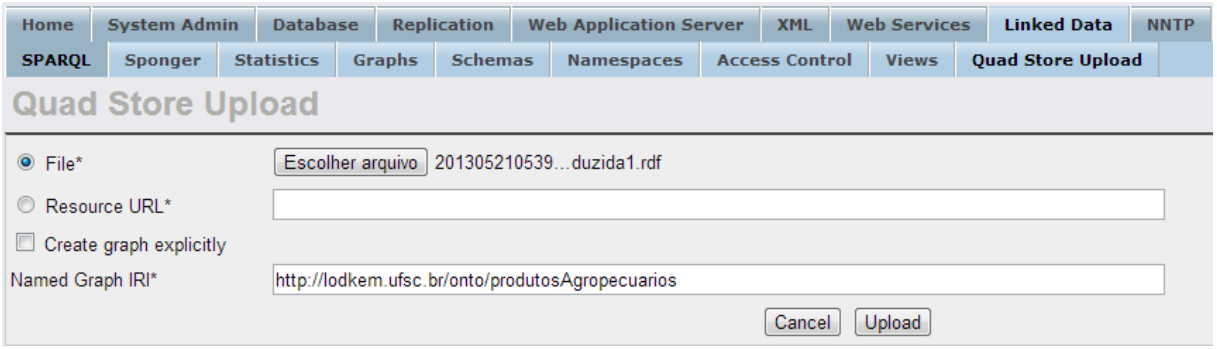

Figura 55: Opção que permite subir arquivos RDF ao servidor Virtuoso

Com os arquivos RDFs devidamente carregados no servidor torna-se possível realizar consultas através do *endpoint* SPARQL. Para exemplificar essa funcionalidade foi preparada uma consulta SPARQL que recupera as URLs que identificam instâncias de municípios, a respectiva gerencia da EPAGRI responsável por este município, uma URL de produto que seja instância da classe Batata, a safra desse produto e a medida quantidadeProduzida, filtrando pelo município de Água Doce, cujo código do IBGE é 4200408 e pelo ano de 2008. Essa consulta e seu resultado se encontram na Figura 54.

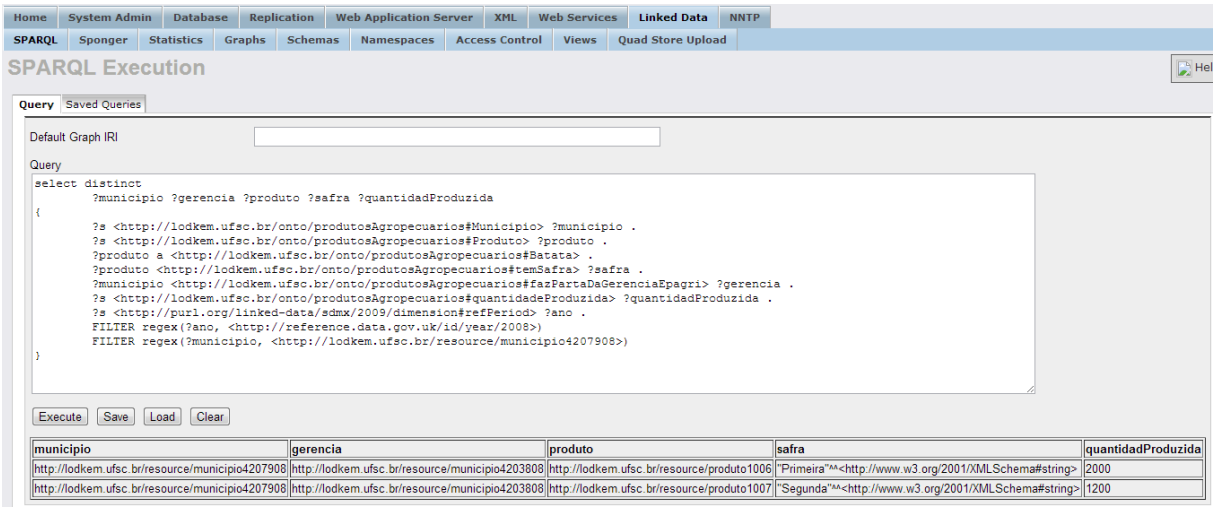

Figura 56: Execução de uma consulta SPARQL na interface do Virtuoso

Assim sendo, qualquer agente na web agora pode realizar consultas automatizadas nesse *endpoint* SPARQL, realizando qualquer tipo de análise em cima dos indicadores publicados com auxílio da ontologia criada neste trabalho e de todas as ferramentas apresentadas até aqui.

3.7. Acessando os dados através do OntoWiki e CubeViz

Também é objetivo deste trabalho acessar os dados publicados no formato *Linked Data* através de uma aplicação *web*. Graças a isso surgiu a oportunidade de utilização da aplicação OntoWiki, um gerenciador de bases de conhecimento

baseadas em ontologias criadas com a linguagem OWL, junto com uma extensão denominada de CubeViz, responsável por ler dados estatísticos estruturados com o *Data Cube Vocabulary*.

Essa aplicação é adequada, pois já possui um módulo que proporciona comunicação direta da aplicação com a base de triplas RDF mantida pelo Virtuoso, tornando-se desnecessária a utilização de outro banco de dados. Assim sendo, já foi alcançado um dos pré-requisitos para utilização do OntoWiki, uma vez que já existia um banco de dados em funcionamento, no caso deste trabalho o Virtuoso. Além disso, foi necessária a instalação e configuração do servidor de aplicações web Apache.

Com a ferramenta adequadamente instalada e utilizável foi possível carregar os arquivos RDF, criados até então, através da interface disponibilizada pelo OntoWiki, criando-se assim uma base de conhecimento. Esse processo foi de fácil execução e é exemplificado através da Figura 57.

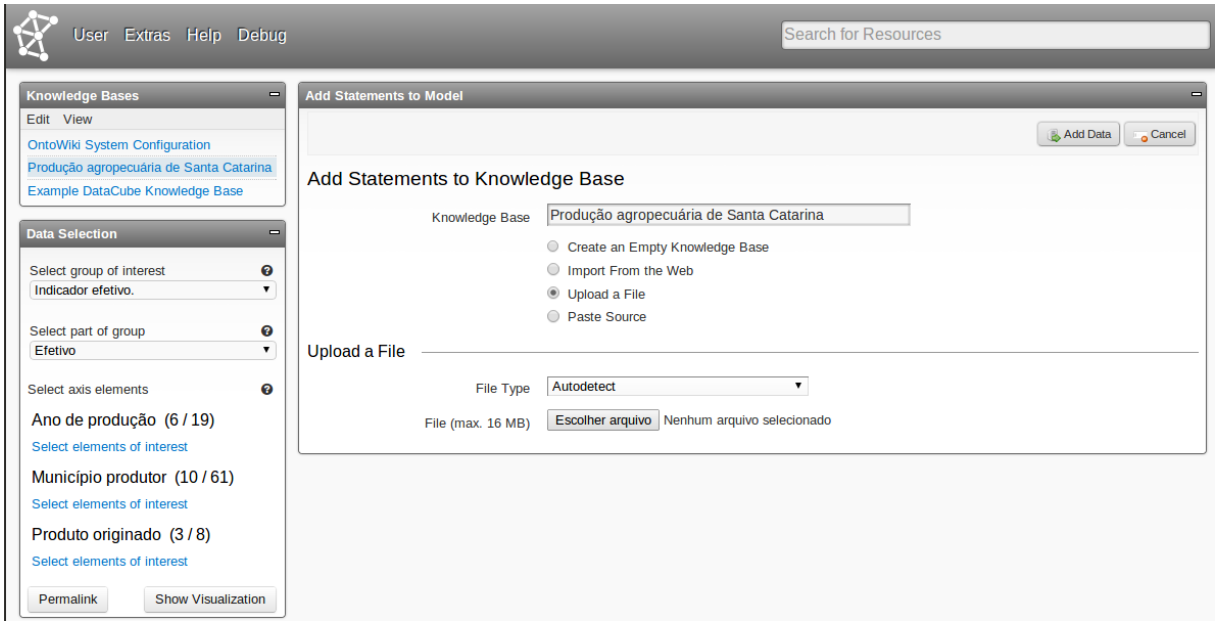

Figura 57: Interface do OntoWiki para criação de base de conhecimento ou adição de dados a bases já existentes

Já com todos os dados carregados na aplicação, a extensão CubeViz possibilitou a visualização dos indicadores em formato de gráficos em barra, coluna, linha e demais formas de representações gráficas. Criando a condição inclusive de realizar filtros pelas dimensões ano, município e produto, definidas através do *Data Cube Vocabulary*.

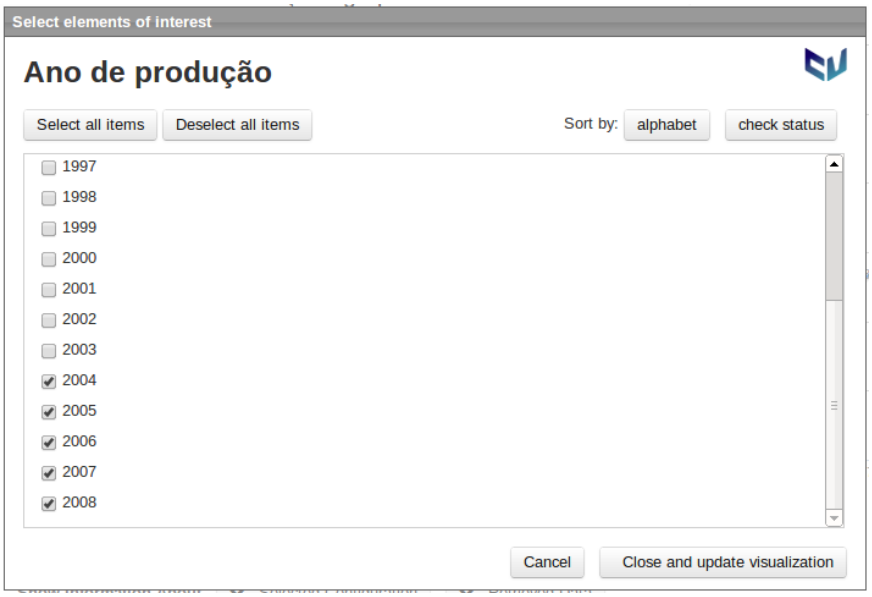

Figura 58: Filtro pela dimensão ano

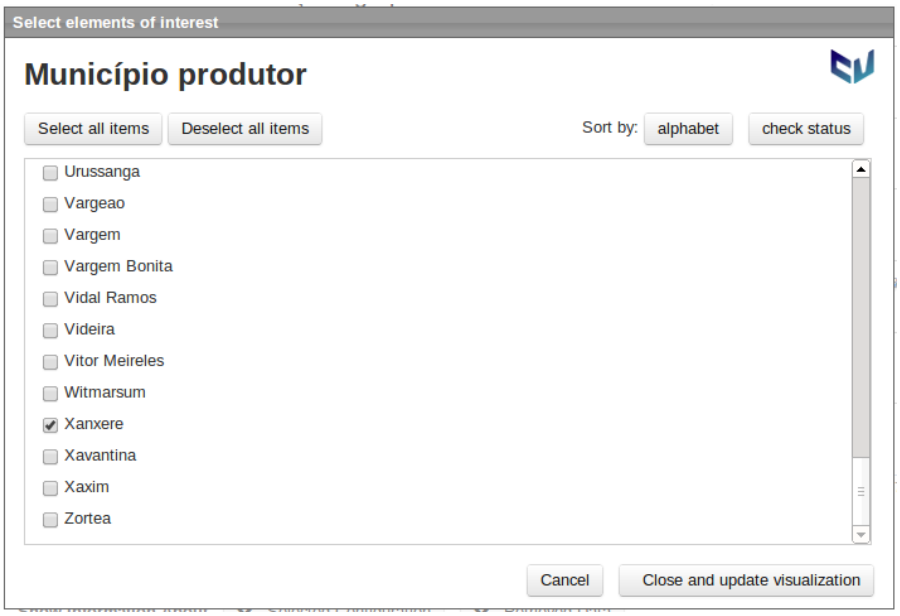

Figura 59: Filtro pela dimensão município

Através do OntoWiki e da extensão CubeViz foi possível então exemplificar o funcionamento de uma aplicação web capaz de, internamente e de forma transparente para o usuário, realizar consultas SPARQL e proporcionar análise dos dados publicados com o formato Linked Data.

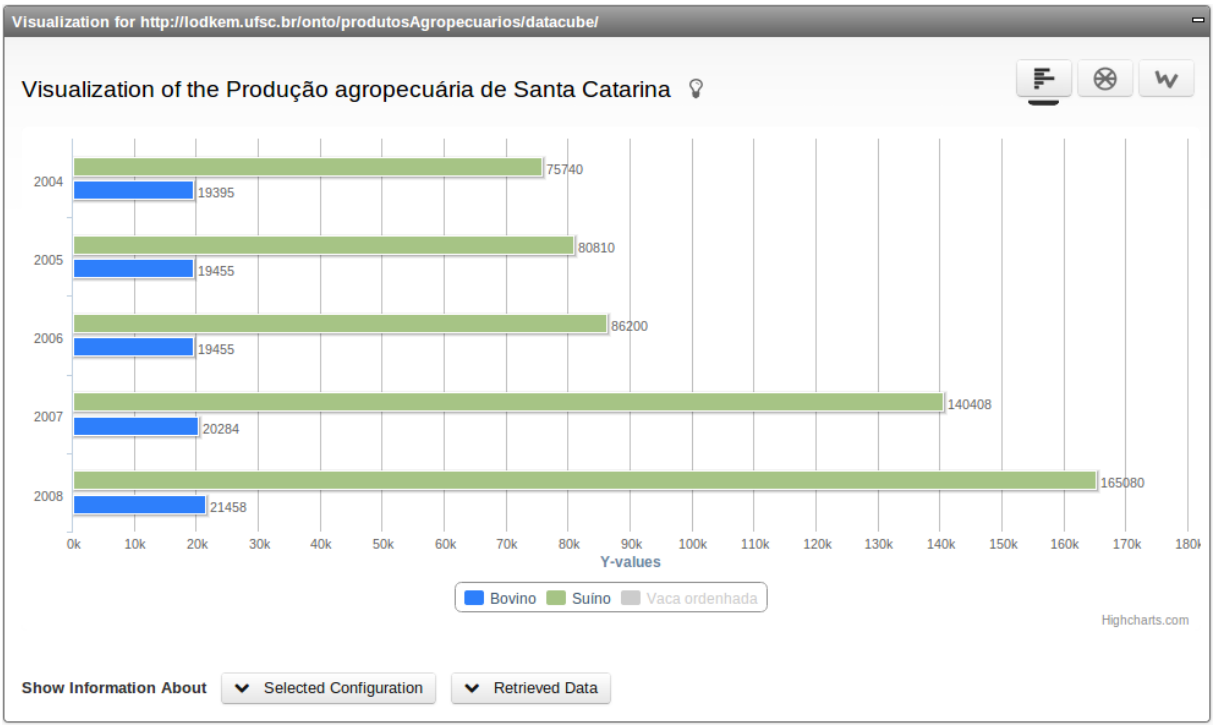

Figura 60: Exemplo do gráfico de produção de bovinos e suínos no município de Xanxerê

4. CONCLUSÕES

Considerando os objetivos definidos na seção 1.2, este trabalho pode ser considerado um sucesso. Todo o esforço exercido para que os objetivos pudessem ser alcançados com êxito possibilitou aos alunos integrantes deste projeto um enorme ganho em conhecimento, tanto nos aspectos teóricos relacionados a publicação dos dados interligados, quanto na utilização de ferramentas até então desconhecidas pelos desenvolvedores.

A definição do tema do trabalho não apresentou grandes dificuldades, sendo que a base de dados abertos foi-nos oferecida como uma oportunidade, já que a EPAGRI não possui uma forma adequada de divulgação destes dados.

O primeiro grande desafio a ser superado, foi o levantamento de fontes relevantes para a elaboração da fundamentação teórica, uma vez que o assunto *Linked Data* unido com os conceitos do *Data-Cube Vocabulary* é relativamente novo e todo o conteúdo presente na web, ou em artigos publicados pelo mundo, encontram-se em língua estrangeira.

Outro fator que exigiu demasiada atenção dos integrantes foi a base de dados fornecida pela EPAGRI, uma vez que a modelagem relacional da mesma não representava com clareza o domínio de negócio a ser publicado na web através da execução do trabalho. Esse fato dificultou a elaboração do modelo estrela e o processo de extração, transformação e carga dos dados, que por diversas vezes precisou ser revisto e refeito.

A etapa de criação da ontologia para representação dos dados a serem publicados não apresentou grande dificuldade para ser efetuada, porém, foi complicado identificar o quão necessário deveria ser a expressividade desta ontologia, surgindo nesta etapa, então, diversas situações de incerteza. Por outro lado, o desenvolvimento da ontologia proporcionou contato direto com ferramentas deveras importantes e completas dessa área, desenvolvendo assim a capacidade em elaboração de ontologias de ambos os participantes deste trabalho de conclusão de curso.

O momento em que mais ocorreram dificuldades foi na publicação efetiva dos dados utilizando o vocabulário para dados estatísticos. A principal dificuldade foi em encontrar alguma ferramenta que automatizasse o processo de leitura de uma estrutura OLAP e de transformação dessa estrutura em triplas RDF. Quando uma

ferramenta candidata surgia, não possuía documentação adequada que possibilitasse sua utilização. Diante de tal cenário foi optado pela exportação dos dados em arquivos separados por vírgula para que pudessem ser processados pela ferramenta desenvolvida pelo Laboratório de Engenharia de Conhecimento da UFSC.

Para atingir o objetivo de consultar os dados através de uma aplicação web, optou-se por utilizar aplicações já disponíveis no mercado ao invés de desenvolvimento de algo mais especifico. Essa escolha cumpriu com o objetivo, mas não se mostrou a mais adequada, uma vez que as análises que poderiam ser executadas com ajuda do OntoWiki e CubeViz não cumpriram com as expectativas dos integrantes deste trabalho. A aplicação se mostrou com problemas de desempenho devido à grande quantidade de observações carregadas no Virtuoso e também não possuía compatibilidade adequada com coleções de dados de múltiplas medidas, mostrando que ainda há muito para evoluir nesse aspecto.

Mesmo deparando-se com diversos contratempos, foi possível supera-los e alcançar o objetivo principal deste trabalho, publicar os dados de produção agropecuária mantidos pela empresa pública EPAGRI no formato *Linked Data* e com auxílio do *Data-Cube Vocabulary*, possibilitando, daqui pra frente, a leitura e análise desses dados por meio de agentes automatizados na web.

5. TRABALHOS FUTUROS

Mesmo com os objetivos cumpridos adequadamente, surgiram no decorrer do projeto diversas oportunidades de desenvolvimento e até mesmo melhorias no processo percorrido.

Como principal sugestão fica o desenvolvimento de uma ferramenta que seja capaz de ler um banco de dados organizado com um modelo estrela exportando todos os dados diretamente em triplas RDF no formato *Linked Data* e utilizando o padrão proposto pelo W3C, o *Data-Cube Vocabulary*. Essa funcionalidade pode ser desenvolvida até mesmo como forma de um *plugin* das ferramentas de criação de ontologias já consolidadas no mercado.

Outro ponto que pode ser levado como trabalho futuro e o aperfeiçoamento da ontologia dos produtos agropecuários criada neste trabalho de conclusão de curso, uma vez que ela foi desenvolvida visando os dados que estavam disponíveis na base de dados da EPAGRI e não efetivamente o domínio de negócio.

E por fim, propomos o desenvolvimento de uma aplicação web que realize de maneira bem específica uma série de análises estatísticas em cima dos dados publicados por este trabalho e também que suporte o formato de múltiplas medidas do *Data Cube Vocabulary*.

6. BIBLIOGRAFIA

AGILE KNOWLEDGE ENGINEERING AND SEMANTIC WEB. ***CubeViz:*** *The RDF DataCube Browser*. Disponível em: <http://aksw.org/Projects/CubeViz.html>. Acesso em: 02 jun. 2013.

AGILE KNOWLEDGE ENGINEERING AND SEMANTIC WEB. ***What is OntoWiki?*** *What can it do for you*? Disponível em: <https://github.com/AKSW/OntoWiki/wiki>. Acesso em: 02 jun. 2013

ALMEIDA, Mauricio B.; BAX, Marcello P.. **Uma visão geral sobre ontologias:** pesquisa sobre definições, tipos, aplicações, métodos de avaliação e de construção, 2003. Disponível em: <http://www.scielo.br/pdf/ci/v32n3/19019>. Acesso em: 13 fev. 2013.

ALVES, Pedro Assumpção; MATTEI, Lauro Francisco. **Migrações no Oeste Catarinense:** História e Elementos Explicativos, 2006. Disponível em: <http://www.abep.nepo.unicamp.br/encontro2006/docspdf/ABEP2006_598.pdf>. Acesso em: 21 nov. 2012.

APACHE SOFTWARE FOUNDATION. **Apache Jena.** Disponível em: <http://jena.apache.org/>. Acesso em: 13 fev. 2013.

BERNERS. T.B.; HENDLER; J., LASSILA, O. ***The Semantic Web:*** *A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities*. Scientific American, maio 2001. Disponível em: <http://www.scientificamerican.com/2001/050lissueberners-lee.html>. Acesso em: 30 jan. 2013.

BERNERS-LEE, T. ***Linked Data: Design Issues***. 27 jul. 2006. Disponível em: <http://www.w3.org/DesignIssues/LinkedData.html>. Acesso em: 03 fev. 2013.

BERNERS-LEE, Tim et al. ***Uniform Resource Identifiers (URI):*** *Generic Syntax*, 1998**.** Disponível em: <http://www.ietf.org/rfc/rfc2396.txt>. Acesso em: 12 fev. 2013.

BIZER, C., HEATH, T., BERNERS-LEE, T. ***Linked Data:*** *The Story So* Far. *International Journal on Semantic Web and Information Systems*, *5*(3), 1–22, 2009. IEEE. Disponível em: http://tomheath.com/papers/bizer-heath-berners-lee-ijswis-linked-data.pdf>. Acesso em: 02 fev. 2013.

BIZER, Christian; CYGANIAK, Richard; HEATH, Tom. ***How to Publish Linked Data on the Web***, 2008**.** Disponível em: <http://wifo5-03.informatik.uni-mannheim.de/bizer/pub/LinkedDataTutorial/>. Acesso em: 07 fev. 2013.

BIZER, Christian; CYGANIAK, Richard. **D2R Server:** *Publishing Relational Databases on the Semantic Web*, [2006?]. Disponível em: <http://richard.cyganiak.de/2008/papers/d2r-server-iswc2006.pdf>. Acesso em: 13 fev. 2013.

BIZER, Christian; SEABORNE, Andy. **D2RQ:** *Treating Non-RDF Databases as Virtual RDF Graphs*, [2004?] . Disponível em: <https://files.ifi.uzh.ch/ddis/iswc_archive/iswc/ab/2004/iswc2004.semanticweb.org/posters/PID-SMCVRKBT-1089637165.pdf>. Acesso em: 13 fev. 2013.

BORST, Willem Nico. ***Construction of Engineering Ontologies for Knowledge Sharing and Reuse.*** Enschede, The Netherlands: Dutch Graduate School For Information And Knowledge Systems, 1997. Disponível em: <http://doc.utwente.nl/17864/1/t0000004.pdf>. Acesso em: 13 fev. 2013.

BREITMAN, Karin. **Web Semântica:** a internet do futuro. Rio de Janeiro: LTC, 2010

CARDOSO, Ismael. **Futuro da web está na interpretação de dados pelos PCs**, 2011. Disponível em: <http://tecnologia.terra.com.br/futuro-da-web-esta-na-interpretacao-de-dados-pelos-pcs,6d0890365e1ea310VgnCLD200000bbcceb0aRCRD.html>. Acesso em: 21 maio 2013.

CARROLL, Jeremy J.; ROO, Jos De. ***OWL:*** *Web Ontology Language: Test Cases*, 2004. Disponível em: <http://www.w3.org/TR/2004/REC-owl-test-20040210/>. Acesso em: 30 abr. 2013.

CEPA. **Cepa:** Institucional, [2013?]. Disponível em: < http://cepa.epagri.sc.gov.br//>. Acesso em: 13 fev. 2013.

CEPA/SC. **Santa Catarina:** Características e Potenciais. Disponível em: <http://cepa.epagri.sc.gov.br/aspectos/menu_sc.htm>. Acesso em: 22 nov. 2012.

D2R Server*: Accessing databases with SPARQL and as Linked Data*. Disponível em: <http://d2rq.org/d2r-server>. Acesso em: 13 fev. 2013.

D2RQ: *Accessing Relational Databases as Virtual RDF Graphs*. Disponível em: <http://d2rq.org/>. Acesso em: 13 fev. 2013

DEGGAU, Renato et al. **SIAGRO:** Sistema de Informações Agropecuárias do Maranhão. Florianópolis: Estado de Santa Catarina, 2007.

EAVES, David. ***The Three Laws of Open Government Data***, 2009**.** Disponível em: <http://eaves.ca/2009/09/30/three-law-of-open-government-data/>. Acesso em: 21 maio 2013.

EPAGRI. **Epagri:** Institucional, [2013?]. Disponível em: <http://www.epagri.sc.gov.br/>. Acesso em: 13 fev. 2013.

EPAGRI. **Notícias Gerais:** Situação da safra 2009/10 e desempenho da produção animal de 2009, 2011. Disponível em: <http://www.epagri.sc.gov.br/>. Acesso em: 22 nov. 2012.

GENESERETH, Michael R.; NILSSON, L. Logical foundation of AI. San Francisco, Los Altos, Califórnia: Morgan Kaufman, 1987 apud ALMEIDA, Mauricio B.; BAX, Marcello P.. **Uma visão geral sobre ontologias:** pesquisa sobre definições, tipos, aplicações, métodos de avaliação e de construção, 2003. Disponível em: <http://www.scielo.br/pdf/ci/v32n3/19019>. Acesso em: 13 fev. 2013.

HEATH, Tom; BIZER, Christian. **Linked *Data:*** *Evolving the Web into a Global Data Space*. Nova York: Morgan & Claypool, 2011. Disponível em: <http://linkeddatabook.com/editions/1.0/>. Acesso em: 05 fev. 2013.

IAB. **Brasil Conectado 2**: Segunda pesquisa Brasil Conectado, 2013. Disponível em: <http://iabbrasil.net/portal/brasilconectado2/>. Acesso em: 21 maio 2013.

IBGE. **Acesso à internet em domicílios continua a crescer no Brasil**, 2012. Disponível em: <http://www.ibope.com.br/pt-br/noticias/Paginas/Acesso-a-internet-em-domicilios-continua-a-crescer-no-Brasil.aspx>. Acesso em: 21 maio 2013.

INMON, William H.. **Como Construir o Data warehouse.** 2. ed. Rio de Janeiro: Campus, 1997. Tradução Ana Maria Netto Guz.

JAMES, Josh. ***How Much Data is Created Every Minute*?**, 2012. Disponível em: <http://www.domo.com/blog/2012/06/how-much-data-is-created-every-minute/?dkw=socf3>. Acesso em: 21 maio 2013.

KIMBALL, Ralph et al. ***The Data Warehouse Lifecycle Toolkit:*** *Experts Methods for designing, developing and deploying Data Warehouses*. Nova York: John Wiley &sons, 1998.

KIMBALL, Ralph. ***Data warehouse Toolkit:*** *The Complete Guide to Dimensional Modeling*. 2. ed. Nova York: John Wiley &sons, 1998.

KLYNE, Graham; CARROLL, Jeremy J.. ***Resource Description Framework (RDF):*** *Concepts and Abstract Syntax. W3C Recommendation* 10 fev. 2004. Disponível em: <http://www.w3.org/TR/rdf-concepts/>. Acesso em: 07 fev. 2013.

LABORATÓRIO DE ENGENHARIA DO CONHECIMENTO EGC-UFSC. **O que é a OntoKEM:** Histórico. Disponível em: <http://ontokem.egc.ufsc.br/>. Acesso em: 26 maio 2013.

LABORATÓRIO DE ENGENHARIA DO CONHECIMENTO EGC-UFSC. **O que é a OntoKEM:** O que é OntoKEM. Disponível em: <http://ontokem.egc.ufsc.br/>. Acesso em: 26 maio 2013.

LEI nº 12.527, de 18 de Novembro de 2011. Disponível em: <http://www.lexml.gov.br/urn/urn:lex:br:federal:lei:2011-11-18;12527>. Acesso em: 28 jan. 2013.

MCGUINNESS, Deborah L.; HARMELEN, Frank Van. ***OWL:*** *Web Ontology Language: Overview*, 2004. Disponível em: <http://www.w3.org/TR/owl-features/>. Acesso em: 30 abr. 2013.

Ministério da Agricultura. **Vegetal:** Estatísticas. Disponível em: <http://www.agricultura.gov.br/vegetal/estatisticas>. Acesso em: 21 nov. 2012.

MINISTÉRIO DO PLANEJAMENTO ORÇAMENTO E GESTÃO. **Sobre o dados.gov.br**, [2013?]. Disponível em: <http://dados.gov.br/sobre/>. Acesso em: 21 maio 2013.

NOY, Natalya F.; MCGUINNESS, Deborah L.. ***Ontology Development 101:*** *A Guide to Creating Your First Ontology.* Stanford: Stanford University, 2001. Disponível em: <http://www.ksl.stanford.edu/people/dlm/papers/ontology-tutorial-noy-mcguinness.pdf>. Acesso em: 19 maio 2013.

OBAMA, B. ***Transparency and Open Government****. Presidential Memorandum, Federal Register.* v. 74, n. 15, p. 26, 2009.

OPEN GOVERNMENT WORKING GROUP. **8 *Principles of Open Government Data***, 2007**.** Disponível em: <http://www.opengovdata.org/home/8principles>. Acesso em: 21 maio 2013.

OPEN KNOWLEDGE FOUNDATION. ***History: History of the Open Definition***, [2011?]. Disponível em: <http://opendefinition.org/history/>. Acesso em: 02 dez. 2012.

OPEN KNOWLEDGE FOUNDATION. ***Open Data Handbook*****,** [2012?]a Disponível em: <http://opendatahandbook.org/pt_BR/what-is-open-data/index.html>. Acesso em: 08 dez. 2012.

OPEN KNOWLEDGE FOUNDATION. ***Open Definition***, [2012?]b. Disponível em: <http://opendefinition.org/okd/>. Acesso em: 02 dez. 2012.

OPEN-LINK SOFTWARE. ***Virtuoso OpenSource Edition Introduction:****What is Virtuoso*?. Disponível em: <http://virtuoso.openlinksw.com/dataspace/doc/dav/wiki/Main/VOSIntro>. Acesso em: 26 maio 2013.

PEIXOTO, Luciano. ***Kettle Beginners:*** Kettle ou Pentaho Data Integration?, 2011. Disponível em: <http://kettlebeginners.blogspot.com.br/2011/08/kettle-ou-pentaho-data-integration.html>. Acesso em: 30 abr. 2013.

PENTAHO. *Pentaho Data Integration*: Kettle. Disponível em: <http://kettle.pentaho.com/>. Acesso em: 30 abr. 2013.

SALM JúNIOR, José Francisco. **Padrão de projeto de ontologias para inclusão de referências do novo serviço público em plataformas de governo aberto.** 2012. 303 f. Tese (Doutorado) - UFSC, Florianópolis, 2012.

SANTOS, Alex Alves Dos et al. **Síntese Anual da Agricultura de Santa Catarina:** 2009-2010, 2011. Disponível em: <http://cepa.epagri.sc.gov.br/Publicacoes/Sintese_2010/sintese%202010_inteira.pdf>. Acesso em: 21 nov. 2012.

SDMX: *Statistical Data and Metadata eXchange.* Disponível em: <http://sdmx.org/>. Acesso em: 19 maio 2013.

STANFORD CENTER FOR BIOMEDICAL INFORMATICS RESEARCH. ***What is protégé?*** Disponível em: <http://protege.stanford.edu/overview/>. Acesso em: 26 maio 2013.

WAGNER, Cláudio. ***Data Warehouse* (DW)**, 2012**.** Disponível em: <http://cacau-indicou.blogspot.com.br/2012/02/data-warehouse-dw.html>. Acesso em: 04 maio 2013.

WORLD WIDE WEB CONSORTIUM (W3C). ***Linked Data*:** *What is Linked Data*, [2013?]a**.** Disponível em: <http://www.w3c.br/Padroes/WebSemantica >. Acesso em: 30 jan. 2013.

WORLD WIDE WEB CONSORTIUM (W3C). ***Query****: What is query used for?*, [2013?]b*.* Disponível em: <http://www.w3.org//standards/semanticweb/query>. Acesso em: 25 maio 2013.

WORLD WIDE WEB CONSORTIUM (W3C). ***SPARQL:*** *Query language for RDF*, 2008*.* Disponível em: <http://www.w3.org/TR/rdf-sparql-query/#sparqlDefinition>. Acesso em: 25 maio 2013.

WORLD WIDE WEB CONSORTIUM (W3C). ***The RDF Data Cube Vocabulary***, 2013. Disponível em: <http://www.w3.org/TR/vocab-data-cube/>. Acesso em: 25 maio 2013.

ZOLDAN, Paulo; CAPPELINI, Carlos. **Museu do agricultor de Santa Catarina:** estudo para implantação. Florianópolis. Instituto Cepa/SC/Fepa, 2004.. Disponível em: <http://cepa.epagri.sc.gov.br/Publicacoes/museu.pdf>. Acesso em: 21 nov. 2012.